THE TUNING OF DNA MUTABILITY VIA CODON CONTEXT AND USAGE BIAS: IDENTIFYING PREDISPOSITIONS TO NONNEUTRAL EVOLUTION WITHIN HUMAN GENES

APPROVED BY SUPERVISORY COMMITTEE

Harold R. Garner, Ph.D

Nick Grishin, Ph.D

Philip J. Thomas, Ph.D

Andrew R. Zinn, M.D., Ph.D.

DEDICATION

This body of work is completed in honor of Dolores Mitschke Patton,

and all other strong, intelligent women.

THE TUNING OF DNA MUTABILITY VIA CODON CONTEXT AND USAGE BIAS: IDENTIFYING PREDISPOSITIONS TO NONNEUTRAL EVOLUTION WITHIN HUMAN GENES

by

MONICA MARIE HORVATH, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

December, 2004

THE TUNING OF DNA MUTABILITY VIA CODON CONTEXT AND USAGE BIAS: IDENTIFYING PREDISPOSITIONS TO NONNEUTRAL EVOLUTION WITHIN HUMAN GENES

Publication No. _____

Monica Marie Horvath, B.S.

The University of Texas Southwestern Medical Center at Dallas, 2004

Supervising Professor: Harold R. Garner, Ph.D

Nonrandom human point mutation trends have been identified across numerous SNP databases to show that CpG dinucleotides in particular display hierarchal mutabilities depending upon the surrounding DNA sequence microenvironment. This information can be harnessed to create a scoring system to contrast the relative mutability of gene sequences, which as a result highlights a gene's rigidity or malleability towards point mutation throughout its evolution. Nonsynonymous mutation probabilities for human genes are calculated and contrasted using four mutation models derived from distinct sources: Disease-causing variants, single nucleotide polymorphisms, intronic mutations, and interspecific substitutions from aligned orthologs. The most mutable human genes are those that mediate reaction to environmental stimuli, including those involved in immune response, pathogen response, and olfaction. As expected, genes using context inclining low point mutation are those involved in essential processes such as cell proliferation and DNA repair. Coupled with observations from studies indicating these classes have experienced positive selection in humans, such results imply that codon usage may shape the size and diversity of the mutation pool on which selection acts. A preinclination towards either radical or safe mutation can be encoded by a gene through using a set of codons with innate tendencies that enhance variation in the required evolutionary direction. The importance of such 'internal forces' in shaping genome evolution signals a need for adjustment of a key principle underlying Neo-Darwinism, which holds that natural selection is the single driving force of genome evolution because underlying point mutation is thought to be random and ubiquitous.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

Horvath, MM and HR Garner (2004). "Understanding human gene evolution with genome-wide mutation spectra." *Submitted.*

Horvath, MM, Fondon, JW, and HR Garner. (2003). "Low hanging fruit: a subset of human cSNPs is both highly non-uniform and predictable." Gene **312**: 197-206.

Flood, EM, Tang, F. Horvath, MM, Pertsemlidis, A and HR Garner. (2002). "SNPCEQer: detecting SNPs in sequences generated by the Beckman CEQ2000 DNA Analysis System." Biotechniques **33**(4): 814, 816, 818-20 passim.

Horvath, MM and NV Grishin (2001). "The C-terminal domain of HPII catalase is a member of the type I glutamine amidotransferase superfamily." Proteins **42**(2): 230-6.

CONFERENCE PRESENTATIONS

Horvath, MM, Fondon, JW, and HR Garner. *Point mutation trends across cSNP databases: Similarities and Predictions.* Poster presentation, Oct 8-11 2003 6th Computational Genomics – Hyatt Regency Boston, MA.

Horvath, MM and HR Garner. *SNP prediction as a strategy to detect rare alleles implicated in complex disease.* Poster presentation, Nov 20-23 2003 6th International Meeting on SNPs and Complex Genome Analysis– Westfield Marriott, Washington DC.

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF DEFINITIONS AND ABBREVIATIONS

CDCV – Common disease common variant hypothesis. This holds that disease which are common within a population of individuals is caused by equally common genome lesions.

CGAP – Cancer Genome Anatomy Project. This project, sponsored by the National Cancer Institute, has the goal of discovering genetic point mutations relevant to cancer genes for mouse, rat, and human models.

cSNP – Coding region Single Nucleotide Polymorphism.

dbSNP – database of SNPs and other genome lesions held by the National Center for Biotechnology Institute

DCM – Dilated cardiomyopathy

HGMD – Human Genome Mutation Database.

iSNP – Intronic Single Nucleotide Polymorphism.

jSNP – Japanese SNP database

Multi-equivalent risk model – This disease model holds that common disease is not caused by common mutations, but by a large body of interacting rare mutations that as a whole sum to a frequency that is close to the frequency of the genetic component of a common disease.

Mutation spectrum – The distribution of point mutation events according to some method of mutation classification. In this work, point mutations are classified according to sequence context motifs meaning that the mutation spectrum here is a description of all sequence contexts coupled with the frequency of point mutations occurring in those motifs.

SNP – Single Nucleotide Polymorphism. These are germ-line point mutations that occur with a frequency >1% in a population. However, this term is used sloppily by the genetics community to refer to any point mutation seen within a population, no matter how global or specific, so that in this study the term 'SNP' is synonymous with 'point mutation'.

TSC – The SNP Consortium. This is a collaborative effort between both academic and industrial leaders to develop a SNP map of the human genome.

CHAPTER ONE
INTRODUCTION

1-1 Inundation of mutational data in the post-genomic era

Since the completion of a draft human genome sequence, post-genomic science has had the information to empower whole organism-driven research that complements current technique-driven and molecule-driven methods. For example, the interaction of many proteins may be studied on a tissue level to elucidate complex problems such as cell-cell signaling and its relation to disease. However, the term "post-genomic" biology is overused, often referring to "pre-genomic" methods simply performed on a massive scale. True post-genomic approaches represent a new way of thinking about science where the best start for a new experiment might be a computational approach. Such is the case for many projects attempting to catalogue all forms DNA variation, a goal that would afford biologists a detailed view of specific genetic differences between human populations, which potentially revolutionizes common practices of modern medicine. Large web-based databases exist for a wide range of experimental data that, when analyzed, may provide invaluable knowledge that can increase the chance of in-house experimental success. All too often, however, such repositories are examined in a somewhat short-sighted manner where the mutations are reviewed only in terms of their projected medical use as agents causative of a phenotype or as points of reference in the genome. Often it is overlooked that each discovered mutation represents an event causing one individual to differ from another, and the persistence of that mutation can illustrate what selective benefit (if any) it affords carriers within the human population.

As a whole, the spectrum of human mutation, which is the distribution of mutation throughout the genome and their associated frequencies, tells an interweaved tale of mutation rates, population dynamics, and instances of natural selection. Researchers then should be plowing through the large public databases not only to look for medical relevance, but to develop hypotheses that allow insight to be gained into our evolutionary history as well as our projected molecular future. It is my thesis that this large volume of existing mutation data can be used to develop a human mutation spectrum that in turn can be applied to individual genes in order to obtain a measure of their inherent mutational load. That is, existing data in mutation databases can be used as a statistical prior to identify genes most likely to change in functionally relevant ways over the course of many generations. This goal requires sound statistical methods to empirically analyze the available datasets as well as comparative analysis of the mutational tendencies of other species in order to put the results in an evolutionary context. Although this idea could certainly underpin a long tenure of laboratory research, for my graduate study I have chosen to focus on single DNA base changes (point mutation) that occur in gene protein-coding regions because at this point in time, proteins are the cellular machines best understood in terms of their impact on the function of biological systems. Success of this project would then naturally engender similar studies within other areas of the genome.

1-2 Creation of a mutation spectrum

Traditionally, mutability of a gene product is estimated by examining the distribution of point mutations, a mutation spectrum, discovered upon genotyping a pool of individuals

(Rogozin, Kondrashov et al. 2001). A comprehensive point mutation spectrum of a gene coding region allows one to identify which areas of a DNA sequence are tolerant to change in the natural human population. A practical issue arises, however, when one intends to contrast differing mutational spectra of large groups of genes that have been grouped by their ontology, such as the biological process they take part in (e.g. transcription) or their molecular function (e.g. works as a polymerase). This is because a prohibitive amount of mutation detection experiments, such as those often done over the course of DNA sequencing, would be required in order to sample a gene set in a large enough population of individuals to achieve statistical relevancy. A second option is to mine large point mutation databases such as dbSNP (Sherry, Ward et al. 2001) for point variants (SNPs, single nucleotide polymorphisms) that map to a particular class of genes. However, many studies of differing genotyping depths compose this database which would cause mutation saturation in a gene to reflect only its level of experimental scrutiny by researchers. Additionally, although there are several million known reference SNPs in this database, few of them have been sampled across enough populations to build a rigorous experimental mutation spectrum.

Given these problems, Shapiro and coworkers devised a novel, predictive solution where an experimentally-derived spectrum of mutations known for immunoglobulin V genes was statistically massaged to accurately predict mutations in paralogs. The immunoglobin V gene sequence was broken down into its component di- and trinucleotides, and experimentally discovered somatic point mutations were categorized for that gene according to the di- or trinucleotide class in which it occurred. This of course creates a hierarchy of mutability preferences for di- and trinucleotide sequences that can then be applied to

unstudied immunoglobin V genes. As a result, the use of an empirical mutation spectrum from one gene allowed the mutational load of a second gene region to be predicted without interrogating it experimentally (Shapiro, Aviszus et al. 1999; Shapiro, Aviszus et al. 2002). Given this proof-of-principle experiment showing the power of DNA sequence context to predict as-of-yet unseen mutation, the intent of my research has been to expand upon this concept by applying it universally to all major human gene classes using as training data the large body of point mutations in publicly available databases. My first goal then is to develop metrics describing point mutability from these datasets.

Since there are so many point mutation datasets of varying size and scope from which to develop metrics to describe gene mutability, each training dataset therefore defines a distinct mutation spectrum. Each spectrum, when applied to a gene sequence, will reflect a different perspective of mutability. For example, if a collection of genes were sequenced in cancer cell lines, it is expected that the resulting spectrum of mutations may look very different than a spectrum obtained from genotyping healthy tissue. This is because cancerous systems are known to be deficient in a variety of mutation repair processes. Additionally, the spectrum of mutations seen in those same genes could differ dramatically depending on the sample size used. As sample size increases, the number of observed, rare mutations will increase considerably. More sequencing depth would be needed to detect mutational events that are quite likely but whose frequencies are attenuated by selection. Although individual genes have their own characteristics such as protein structure and function that dictate the impact of a mutation, the effects of different types of missense substitutions can be summarized to obtain a ‘world view’ of protein impact for a large set of genes.

I define a mutation spectrum in this study, *M*, as the distribution of observed point mutations across a DNA sequence. My primary objectives in studying gene mutation spectra are to define hypermutable regions and contrast different genes' mutation propensities. Ideally if enough data were available for each gene in the human genome, one would be able to contrast mutation spectra and correlate findings to the underlying sequence context directly from gene to gene. Unfortunately, few (if any) genes satisfy this criteria, so a classification system must be defined to categorize point mutation by local sequence context thereby engendering discovery of mutation trends. In doing this, I have chosen both a method of categorizing variants as well as datasets of mutations. Since I am interested in a mutation's sequence context, a variant is categorized according to its wild-type and mutant trinucleotide—its trinucleotide mutation class (TMC), such as a GGG→GAG TMC for a particular G/A mutation. Since there are 64 different trinucleotide sequences that can be built from the bases A, T, G, and C, there are then $3 \cdot 3 \cdot 4^3 = 576$ possible TMCs that can be populated by a database of point mutations. Additionally, as single DNA base can belong to three different trinucleotides depending on which 'frame' is used when analyzing a DNA sequence. Therefore, a given mutation spectrum is composed of three 576-member mutability tables, one for each frame. Approaching the problem in this way avoids needing to make a mutation spectrum based on pentanucleotide classification.

In total, TMC frequencies are calculated for a wide variety of different point mutation datasets to subsequently develop numerous versions of *M* that each individually summarize observed variants in human gene regions. Each of these spectra differ from the next due to biases inherent in the training dataset. This is because two principle components of *M* shape

the variant pool observed within a given dataset: i) *m*, the inherent mutability of a DNA sequence and ii) *s*, natural selection forces, so that

$$M \propto m + s.$$

Each different database from which mutability trends are derived is a product of a distinct experimental study. Mutation detection studies can only detail what variants were observed within a specific population, but by comparing numerous mutation spectra derived from complementary sources one may ascertain which genes have the greatest propensity to mutate and generate variants prior to selection pressure. This is rationalized as follows. Imagine there exists mutation spectrum *a* where

$$M_a \propto m_a + s_a$$

and there also exists mutation spectrum *b* where

$$M_b \propto m_b + s_b.$$

Since *m* represents the inherent mutability of those DNA sequences that compose a mutation spectrum as a result of cellular processes and nucleotide physiochemistry, I estimate that *m* will be roughly constant in all gene coding regions. Therefore,

$$m_a = m_b = m$$

In comparing and contrasting two mutation spectra, one wants to note their differences because it is those changes that indicate the selective differences inherent in the underlying training data. That is, since

$$m_a = m_b \qquad \text{then}$$

$$M_a - M_b \propto s_a - s_b$$

This formalism represents how mutation spectra will be contrasted in this study.

1-3 Point mutation in gene coding regions is expected to be nonrandom

An organism's gene mutation is balanced throughout its evolution to both ensure diversity for survival and protect essential functions. Most Neo-Darwinists hold that new mutations occur randomly across a genome and then selection acts upon those variants to drive them to fixation or purification (Lenski and Mittler 1993). However, evidence of strong adaptive evolution has been recently reported in numerous comparative genomics studies of gene coding regions, that is, the rate of positive selection is far higher than can be permitted if the neutral theory of evolution holds without amendment (Fay, Wyckoff et al. 2002). This is because the observation of a particular mutation in a population is shaped by two entangled forces: i) the natural mutational tendencies acting upon the genome that may predispose a spontaneous event to occur and ii) selective forces on this new variant. It is my thesis that the effects of these two forces acting on different point mutation datasets can be analyzed and summarized to describe a gene's propensity to mutate and change the sequence of the encoded protein. My first goal is to therefore take calculated human mutation spectra and show that the inherent mutational events within a gene, measured by the calculation of mutationally warm- and coolspots, are undeniably nonrandom. Since by definition a nonrandom capacity allows one to be predictive, that is, know the odds of observing mutations in a population, I can ask whether the trends are strong enough to predict point mutation for testing in an experimental setting. Such results would suggest that contrary to the view of many neo-Darwinists, natural selection is not the sole creative force that drives the processes underlying the evolution of genes. The rate and direction of mutation can

prime genomes for certain evolutionary fates as well. This makes sense from a fitness standpoint, for it is easy to imagine that if a species has to constantly rise to the same challenges, such as immune response or pathogen defense, then it will be more fit if genes meeting those challenges are predisposed to a particular level of mutation.

If the test for prediction accuracy as determined by experiments is not successful, however, that does not indicate that the trends discovered are not informative. It is not hard to imagine that a series of small yet nonrandom tendencies towards mutation can culminate over the course of many generations to enhance the chance of a biased set of amino acid mutations. The totality of such changes may not therefore be visible within a single population. Therefore the predictive metrics can be applied to a gene to describe not what variants are expected to be seen currently in the human population, but how a gene will evolve over time. Is such a gene more prone to undergo radical mutation and evolve new functions, or does it "stack the deck" using clever codon usage to ensure only conservative mutation? Is this gene by way of its sequence context prone to any mutation at all? Such questions will allow me to gain insight into the future of human genes in a relative manner.

1-4 The mutagenic potential of differing gene classes can lend insight into gene evolution

With a mutation spectrum in hand that describes the chances of different categories of mutation to occur, it can be applied mathematically to a gene to describe its inherent mutability. If this is done for all human genes, they then can all be compared and contrasted by their mutability values. In doing this, I expect to see that essential genes have a low

mutational load while those part of environmental responses, such as those involved in pathogen defense, would be more permissive of radical mutation due to their need to harbor variety. Given that the definition of mutability is derived from DNA sequence context, any proclivities discovered reflect the influence of codon bias.

If these nonrandom inclinations towards point mutation indeed exist, it then follows that such proclivities effectively can possibly either enhance or diminish the presence of risky amino acid substitutions *before the forces of selection ever come into play*. Whether that pool of alleles is more loaded with risky or safe changes to the protein product is a function of codon usage, that is, the trinucleotide synonyms within a gene that code different amino acids. Consequently, each gene would therefore have its own distinctive nonrandom pool of alleles by virtue of the mutabilities of the composite codons, even before selective processes begin. It is this pool that is then available to the forces of natural selection to shape a gene's evolutionary future—not a random soup envisioned by Neo-Darwinists. Furthermore, it is not hard to imagine that a large number of slightly nonrandom mutational events, which are nonrandom by nature of the codon choices within a gene sequence, can alter the gene pool as to increase the chances of positive selection on the amino acid product. In other words, the group of amino acid variants seen within a population is somewhat predetermined by the coding DNA sequence. Therefore, the search for nonrandomness requires the identification of DNA sequence elements that predispose point mutations to occur. A preinclination towards either radical or safe mutation can be encoded by a gene through using a set of codons with innate mutational tendencies that direct mutation towards

the required direction. This process could be the mechanism encouraging positive selection that has not yet been found within genomics.

CHAPTER TWO
BACKGROUND

2-1 Point mutations and SNPs in the post-genomic era

Of all of the resources that have been born out of the human genome project, few have gained as much press and expectation of revolutionary medicine as SNPs, single nucleotide polymorphisms. SNPs are germline point mutations that occur at a frequency of >1% in the global human population, although there is poor adherence to this definition within the SNP community. Often an ethnically or disease-stratified population (<100 individuals) is genotyped and any point variation discovered within that small group is described as a SNP (Brookes 1999). With this variance in the SNP definition, the probability of discovering a polymorphic allele is dependent on the luck of amassing the correct population stratification. For example, the frequencies of SNPs discovered in the *BRCA1* gene from a group of several hundred individuals diagnosed with advanced breast cancer might inaccurately portray the global variation of the gene. However, such mutations discovered with allele frequencies of >1% in that focused group of people will be championed as a SNP and form the meat of many disease association studies. Therefore, a consistent definition of SNPs simply does not exist so I use the terms SNP, variant, and point mutation interchangeably.

SNPs have a wide variety of uses as tools to understand what makes one individual differ from another. Given such interest, a whole industry of SNP genotyping has emerged. Although a standard protocol of capillary DNA sequencing can certainly type SNPs, if only one base in a genome is being queried this results in a large waste of experimental resources. As a result, numerous exotic methods for SNP detection have been developed including

genotyping by mass spectroscopy (Tang, Fu et al. 1999) and pyrosequencing (Ahmadian, Lundeberg et al. 2000), amongst others. Also, a wide range of bioinformatics tools has arisen to aid the researcher in compiling and analyzing point mutations. The largest repository of DNA point mutations is undoubtedly dbSNP, which currently has over ten million unique SNPs detailed (Sherry, Ward et al. 2001). dbSNP is searchable as an Entrez database and is a tightly linked to the available bioinformatics resources held at NCBI (Geer and Sayers 2003).

Once a SNP is found, the next challenge becomes to understand what it may suggest about population structure and its associated impact on function. Two of the most widely used cSNP analysis tools, PolyPhen and SIFT, are not available directly through the NCBI portal. Polyphen uses a hierarchy of prediction rules to determine whether a SNP is likely to show a phenotypic effect (Sunyaev, Hanke et al. 1999). These rules include functional annotation from the SWALL database (Hermjakob, Lang et al. 1998), mappings on to protein structure, and conservation with a multiple alignment of homologs. The main competitor to Polyphen is SIFT, a sequence homology-based tool from the Henikoff lab that predicts whether a SNP will cause a phenotype using an alignment of sequences belonging to a protein family (Ng and Henikoff 2002). SIFT does not use any external functional or structural information, but its SNP scoring matrices perform better than PolyPhen when that information is not available.

2-1-1 Disease association studies and SNP maps

A major challenge in medicine is understanding what genetic factors predispose individuals to disease. Most of the fanfare currently surrounding DNA variation discovery is caused by studies that strive to correlate disease susceptibility to SNP occurrence out of the expectation that this will lead to an era of personalized medicine. When SNP databases experienced their initial burst in growth, many researchers predicted that the most common of these events would underlie much of complex disease. The hope was that if a high frequency allele SNP map of the human genome was created (approximately one SNP marker per 1000 bases), then researchers would simply only need to genotype those markers in a group of both affected and unaffected individuals and look for a statistical difference in SNP frequencies between the cases and controls (Sachidanandam, Weissman et al. 2001). This methodology, known as the common disease, common variant (CDCV) hypothesis, would relate SNPs to a disease state in one of two ways: Direct association, where the discovered allele has a deleterious effect on fitness, or indirect association, where the variant is in linkage disequilibrium with the actual allele that confers a phenotype. The latter method relies on the hypothesis that each linked allele must have arisen concomitantly in a particular individual at some time in the past causing the profile of linked polymorphisms in the altered region to be inherited along with the disease-causing allele. Therefore with a marker SNP in hand, one can zero-in on the specific genomic area causing a phenotype as well as develop clinical assays so that a patient can know if he carries a disease-predictive marker. The fact that a common allele may be the root of a common disease seems somewhat antithetical to the principles of natural selection. That is, if deleterious alleles were so frequent in the

population, it may be expected that we should have died out as a species. But most common diseases hit individuals later in life—after reproduction—meaning that it is quite likely that they are nearly invisible to selective forces.

Unfortunately, SNP association has not been very successful in unraveling the threads weaving complex disease, and many problems inherent to the philosophy of such studies have been identified. Researchers must know that both the SNPs they are investigating as well as the population queried are relevant to the disease state. A crippling problem with the CDCV model as studied by marker SNPs is that phenotypic frequency does not necessarily estimate the genetic risk if the common disease in question has a large environmental component. Polymorphisms that map to a disease can be very common, but contribute actually only a small risk to someone's chance in displaying a complex disorder. For example, cardiac disease, the leading cause of death in the US, has been estimated to have a maximum heritability of 34% in whites and 53% in blacks, and a correlation of cardiac disease incidence has been found between spouses (Katzmarzyk, Perusse et al. 2000). Smoking, obesity, and physical inactivity are solely environmental factors that are known to play a considerable role in disease risk even in absence of a genetic component. Therefore it does not necessarily follow that a common disease can be wholly described by comparably common alleles. A second issue is that so-called common diseases are often a semantic lumping together of many disorders displaying similar phenotypes, as is the case for long QT syndrome, cardiomyopathy, and atherosclerosis, all of which are tersely described as cardiac disease but can be heavily influenced by mutations in separate genes. Again, this problem reiterates the importance of choosing a disease cohort quite carefully, as well as measuring a

variety of phenotypes. Even if a set of important SNPs is found in disease cases, pinning down the pertinent phenotype is non-trivial.

Finally, the existing experimental support for the CDCV hypothesis comes solely from detection methods having poor allele frequency resolution (Wright, Caraothers et al. 1999), that is, they are only competent to discover common variants. The ease of detection of these SNPs therefore belies this body of supporting data, for only a small, biased subset of all disease alleles is available. In most studies, only a handful of individuals (~24) are screened due to the time and expense associated with DNA sequencing. As a result, often variants having frequencies of even 1-5% are missed. Additionally, recent SNP-discovery projects turn up very low numbers of nonsynonymous point mutations (~ 1/3000 bases). Thus the most sought after variants, those that change the protein product, are far and few between which is unfortunate given that they are of the class most easily described in terms of disease risk (Cargill, Altschuler et al. 1999; Halushka, Fan et al. 1999; Ohnishi, Tanaka et al. 2000). In fact, the rates of nonsynonymous and silent SNPs detected are very close in two studies (Cargill, Altschuler et al. 1999; Ohnishi, Tanaka et al. 2000), suggesting that the missense SNPs discovered are under the same level of selection. This difficulty is compounded by the fact that the sequencing error rate is often higher than the allele frequency causing many false positives (Cargill, Altschuler et al. 1999).

Given the problem with CDCV-based association studies, a competing model of allelic diversity underlying disease susceptibility, the multi-equivalent risk model, has been recognized. It assumes that for any disease there is a large pool of risk alleles, each of which individually have a low population frequency. Therefore the cumulative frequency of the

risk alleles may be considerable, but the exact frequency of any one allele is low. This view complements natural selection theory in that point mutations having a marked effect on phenotype, such as nonconservative mutations in the coding regions of genes, would be expected to have low population frequencies. Therefore, the problem with the CDCV hypothesis is that it engenders mutation discovery studies that are biased against rare variants, which means those alleles mostly likely to be functionally important will be missed. Leading members of the "SNP-o-typing" community that had been working under the CDCV hypothesis, such as Aravinda Chakravarti at Case Western Reserve University and Eric S. Lander, director of the genome center at the Whitehead Institute for Biomedical Research, are beginning to realize that the standard small population sizes used for disease allele discovery may miss nearly all SNPs that alter the protein product given that the allele frequencies are expected to be quite low. Chakravarti now admits, "The right way to go is to take a set of candidate genes and assess them directly in as many patients as possible" (Hagmann 1999). Figure 2.1 demonstrates mathematically the reduction in scope caused by genotyping only small sample sizes. One has only a 64% chance of discovering an allele of frequency of 1% using a population of 50 individuals. This of course means that for all alleles of even lower frequency a geneticist will on average miss them more often than discovering them in that 50-person population.

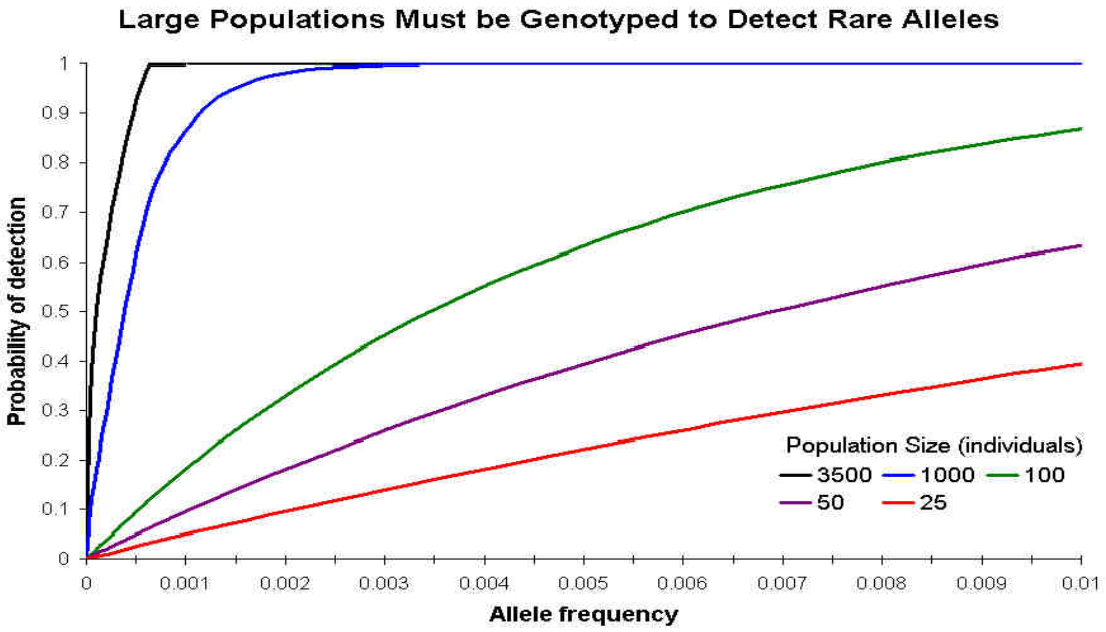


Figure 2.1. The probability of detection is calculated as $P=1-(1-X)^{2Y}$ where X is the allele frequency and Y is the population size. Rare alleles (frequency <1%) are unlikely to be discovered in populations smaller than 50 individuals. A population of 3500 is sufficient (97% chance) to detect alleles having frequencies as low as 0.0005.

The multi-equivalent risk and CDCV models represent two extremes, and the correct model is probably some combination of the two. To maximize chances of success in disease mapping, it is critical that experiments are competent to detect subtle genetic effects under a variety of genetic models. Current variation discovery projects, most notably the SNP Consortium, fail to satisfy this requirement because only a small and often unstratified population is screened rendering it impossible to discover the rare variants existing under the multi-equivalent risk model. Thus, the multi-equivalent risk model has been systematically ignored in nearly all disease allele discovery studies. There is an overwhelming preference

for the CDCV hypothesis in the SNP genotyping community because it supports the status quo of low resolution genotyping.

Since it is therefore highly unlikely that the CDCV hypothesis can be the only model of the association between alleles and disease, there is obviously a need for high-throughput, post-genomic technologies so that both common and rare alleles may be resolved in a panel of several thousand individuals. This is a task difficult to perform with a traditional round of DNA sequencing due to time and cost considerations.

2-1-2 Haplotype mapping

Another popular way to use SNPs is to search for blocks of SNPs that seem to segregate as groups and then correlate these blocks, or haplotypes, to multigenic traits. Such SNPs are said to be within linkage disequilibrium of each other, that is, they are statistically found together at rates greater than expected given recombination. Perfect linkage disequilibrium is defined as the minor allele frequencies of the alleles being perfectly correlated, that is, they would have a correlation coefficient (r^2) of 1. Such measurements allow the identification of haplotypes and subsequent designation of "haplotype-tagging" SNPs when exploring the genome variation in a population. Not only would genotyping for all known closely linked SNPs waste resources, but it would not in theory provide any new information about the genetic structure of a population. The demarcation of haplotype tagging SNPs therefore can allow one to surmise the profile of linked variation upon experimental interrogation of only one DNA base.

In examining how SNPs cosegregate, common SNPs are typically very old while new SNPs are often rare given that they have not had time to achieve fixation within a population or be selected against. When a rare SNP initially appears, it becomes associated with the background of common SNPs already on the chromosome. Therefore, a rare variant nearby a chromosome will be in linkage disequilibrium with one of these frequent SNPs, and consequently segregate with it. If such a rare variant actually causes a disease, it would be enriched in a group of affected individuals and logically, the nearby common SNPs would be enriched as well.

This reasoning underpins many studies that look for marker/disease associations using only haploytpe tagging SNPs. The HapMap Consortium (http://www.hapmap.org) is an initiative across six countries to develop a public haplotype map of the human genome so that association studies could presumably be completed efficiently and quickly (TheHapMapConsortium 2003). The goal of the HapMap team is to take 270 individuals of wide-ranging ethnicity and genotype them for approximately 1 million SNPs spaced at 5kb intervals. From this, haplotypes can be inferred and the best set of haplotype-tagging SNPs would then be set aside for future use. The Achilles heal of the haplotype method is that if only common haplotypes are mapped, which is a function of the initial population screened for SNPs, the identified tagging SNPs may not segregate with rare, disease-causing SNPs at a rate great enough to be detected as a statistical difference between a disease cohort and control population. That is, the minor allele frequencies of the tagging SNP and the rare, disease-causing SNP depends on the r^2 value, such that the power to detect the rare SNP indirectly in N samples is the same as the power to detect it directly in Nr^2 samples.

Therefore, although haplotype methods may decrease the number of SNPs required to elucidate patterns of alleles in linkage disequilibrium, the population size of genotyping studies is still a factor that must be weighed carefully. Additionally, there has been much debate over how far linkage disequilibrium stretches across chromosomes, and it seems that this magic number depends strongly upon what region of the genome is being examined, for regions of recombination hotspots have been identified (Visser, Shimokawa et al. 2004; Yamada, Mizuno et al. 2004; Yeadon, Bowring et al. 2004).

.

2-1-3 The emergence of pharmacogenetics

The burgeoning interest in disease association and SNP markers has engendered a new field: Pharmacogenetics. The hope is that SNP markers can be genotyped for any patient so that scenarios ranging from understanding one's potential for an allergic reaction to predicting the likelihood of cancer onset become possible. One challenge of pharmacogenetics is that as medicine is increasingly specialized to meet the needs of individual genomes, the financial returns for testing and marketing treatments for these small groups begins to diminish.

One company that has enjoyed unprecedented success in correlating SNP markers and complex disease is Decode Genetics located in Iceland. The key to their success lies in the Icelandic people, who are highly isolated, interrelated, and have kept genealogical information describing those relations for over 1100 years. Matched with both genetic and phenotypic data on over half the adult population of Iceland, Decode Genetics then can literally trace the inheritance of disease-related SNP markers throughout the population.

Decode Genetics is currently in the process of clinical trials for a drug developed using this information that is purported to help prevent heart attacks. They also have similar projects in place to identify and treat genetic susceptibility to common disorders such as atherosclerosis, asthma, stroke, schizophrenia, type 2 diabetes, and obesity (Halapi and Hakonarson 2004; Helgadottir, Manolescu et al. 2004; Stefansson, Steinthorsdottir et al. 2004). Given such advances, Decode Genetics is currently the most successful drug company in the field of pharmacogenetics.

2-2 Mutational mechanisms and mutation spectra

Understanding the relationship between genomic lesions and the resultant phenotypic effects has always depended heavily on the analysis of mutational mechanisms. Point mutations fall within two classes: Transitions, which are purine to purine or pyrimidine to pyrimidine substitutions, and transversions, which are purine to pyrimidine (or vice versa) mutations. The latter class changes the number of hydrogen bonds between the DNA bases at the substitution site in the DNA double helix, meaning those mutations can greatly affect DNA annealing temperature. Transversions are much less frequent than transitions, for they occur at roughly half the mutation rate. Point mutations can arise either in the germline, meaning they will be transmitted to an individual's progeny, or in somatic cells, meaning they will only affect the individual in which they occur as they will not be transmitted genetically. Somatic mutations are often those that underlie common cancers such as skin cancer. Given the type of datasets analyzed in this study, all mutations are expected to be

inherited and thus any derived mutation trends is indicative of those pertinent to germline point mutation.

2-2-1 Factors influencing point mutation occurrence

The average point mutation rate in mammals has been estimated to be between 10^{-8} to 10^{-9} events per base pair per year (Kumar and Subramanian 2002). Most models of nucleotide substitution assume that neighboring DNA sites evolve independently. Although this certainly makes certain statistical tests and phylogenetic tree building easier, this assumption is in error. It has been demonstrated numerous times that both germline and somatic nucleotide substitution is context dependent, where a base's propensity to mutate is strongly affected by the nucleotides flanking either side (Krawczak, Ball et al. 1998; Horvath, Fondon et al. 2003; Lunter and Hein 2004). In general, point mutations are caused in mammalian genes whenever a blatant error occurs in processing the DNA, such as random misincorporation of a nucleotide, or when certain types of DNA microenvironments make the DNA sequence more labile. Methylation, polymerase arrest sites, recombination, and inequities in the DNA repair mechanisms due to different efficiencies of DNA motifs to interact with these cellular processes will contribute to escalating mutation rates as well as make some sequences more mutable than others. For example, it has been shown that errors introduced by DNA polymerase η during synthesis of the nontranscribed DNA strand in immunoglobulin variable region genes contributes to the occurrence of mutational hotspots at AT base pairs (Rogozin, Pavlov et al. 2001). The most well-known sequence motif prone to point mutation is the CpG dinucleotide. In mammals, methylated cytosine when flanked by a

guanine will spontaneously deaminate to thymine at a rate an order of magnitude greater than all other point mutation (Cooper and Krawczak 1993). Methylation of DNA is a mechanism by which the mammalian cell can keep a gene inactive. Methylated CpG dinucleotides are also preferential targets of UV radiation-induced mutagenesis (You, Li et al. 1999). Although individual sequence environments are strong enough to effect a point mutation, most substitutions can be explained by multiple mutagenesis mechanisms that in an additive fashion inclines mutation (Todorova and Danieli 1997).

2-2-2 Traditional mutation spectra

The effects of each of these individual mutation processes on the body of mutations observed within a population can be determined using a mutation spectrum. Such spectra consist merely of the distribution of mutation frequencies that occur along a DNA sequence. In absence of an external source of phenotypic selection or inducement on the part of the experimenter, a gene's mutation spectrum is representative of the result of the normal mutational processes that can occur within a cell. Mutation spectra are usually employed to compare differences between populations that have been subjected to differing levels of chemical stimuli. A statistically significant increase of mutations at a certain site in a DNA sequence is known as a mutational hotspot.

2-3 Calculation of selection acting on gene sequences

2-3-1 The neo-Darwinian synthesis and neutral evolution

The neo-Darwinian synthesis represents a marriage of Charles Darwin's theory of evolution via natural selection with a rediscovery of Gregor Mendel's work in genetics. The gene is seen as the unit of evolutionary change within an organism, where inheritable, random mutations arise and their allele frequencies randomly drift through populations. Natural selection may acts upon these alleles to either drive their frequencies to fixation or suppress them to zero because they lower fitness. The presence of sequence variation is the fodder for the evolution of a species and explains how the microevolutionary process of selection, which acts on the allele level, changes allele frequencies to lead to macroevolutionary events such as adaptation and speciation. Because variation is so abundant, natural selection is seen as the driver of evolution.

The neutral theory of evolution holds that the vast majority of point mutations in a population are neutral or nearly neutral, that is, they do not affect species fitness for better or for worse (Kimura 1987). This is because it is believed that deleterious variation is rapidly removed by selection which of course leaves the neutral alleles behind. Kimura recognized that the random drift of these alleles may have a profound influence on species evolution. Through random genetic drift, these mutations may disappear completely or become fixed within a population at a rate equal to their rate of occurrence. These alleles then drift randomly in frequency throughout populations due to mating choices and environmental disasters. Eventually such alleles may become fixed due to drastic changes in a population, such as isolation or bottlenecks, that eventually causes speciation. The neutral theory

embraces selection, but holds that it is small compared to random genetic drift. The neutral theory typically serves as a null hypothesis for how evolution unfolds when a group of genes are being studied. According to the neutral theory most alleles that arise in a population have no impact on gene function, although it is possible that a small number will be selected against (negative selection) or selected for (positive selection).

2-3-2 Discovering positive selection using point mutations

Positive selection, also termed adaptive evolution, is the process where an allele increases the fitness of carrier organisms in a population. Typically, instances of adaptive evolution can be calculated directly from multiple sequence alignments where orthologs from two closely related organisms are compared to glean interspecific point mutations. An organism representing a common ancestor is often also aligned so that it can be used as an outgroup to determine the ancestral allele. If the two compared species are sufficiently related, one can assume that only one point mutation has occurred per site since divergence from their common ancestor. This assumption is often employed in chimpanzee-human comparisons because the rate of point mutation is on the order of $\sim 10^{-8}$ events per year, and the chimpanzee-human divergence was approximately 5 million years ago. When comparing mice and humans, however, maximum likelihood methods must be employed to correct point mutations statistics.

The interspecific point mutations can be divided into two groups: Nonsynonymous (amino acid changing) and silent (non-amino acid changing) substitutions. The ratio of the frequencies per nonsynonymous or silent site (the 'Ka/Ks' ratio) can be used to determine

whether adaptive evolution has occurred in a gene sequence. If Ka/Ks >> 1, it is often inferred that adaptive evolution, also termed as positive selection, has occurred.

Often the Ka/Ks ratio calculation, derived from interspecific substitutions in a multiple alignment, is augmented with a calculation of the ratio of nonsynonymous to synonymous polymorphism seen within a gene. The neutral theory predicts that the relative frequency of nonsynonymous and silent substitutions should be the same both within a species (polymorphism) and between related species (interspecific substitutions). But when a difference is observed, this is indicative of a change in selective pressure, and new statistical methods have been developed to evaluate such occurrences (McDonald and Kreitman 1991). These refinements to the methods of determining positive selection have therefore only become widely available since the explosion of SNP information in public databases.

The major problem in using adaptive evolution measurements such as Ka/Ks to infer mutationally hot or cold regions of the genome, as has been done recently (Chuang and Li 2004), is that the point mutations culled are only those that have been accepted by evolution. Even in genome areas where there has been positive selection for variation, there may still be some portion of the mutation spectrum that had been selected against and therefore will not appear in whatever metric is calculated to designate a genome region as mutationally 'hot'. The fact that this information is not included therefore means that any conclusions made about genome mutability are suspect.

Instances of adaptive evolution have been identified in many systems. Venom gland phospholipases have a rate of nonsynonymous substitution that exceeds the rate of silent substitution (Ogawa, Nakashima et al. 1996). Genes that are involved in pathogen defense or

immune response also often show signs of adaptive evolution. Even genes involved in sperm-egg interactions have been under positive selection (Galindo, Vacquier et al. 2003). Adaptive evolution allows these recurring challenges to be met by organisms as well as engender the development of new genes and gene families. Given that the evidence of widespread positive selection within mammals has been growing, it has even been suggested that the universal genetic code itself may represent an adaptation to maximize the efficiency of adaptive evolution (Freeland 2002). This means that genes might have a preferential codon usage pattern that defines the spectrum of possible amino acid changes.

Therefore, the analysis of point mutations has much to offer the field of evolutionary theory in understanding both where organisms are coming from and where they are heading on a genome level. Although the presence and frequency of a SNP is due to entangled forces such as selection, mutation, migration, recombination, and genetic drift, if any of these components can be identified by carefully subdividing large SNP datasets, calculation of mutational preferences and inferences of gene mutability can be made.

CHAPTER THREE
CONSTRUCTION OF HUMAN POINT MUTATION SPECTRA

3-1 Collation of point mutation information from public datasets

A large variety of point mutation databases is needed in order to deduce genome-wide coding region point mutation trends. This is because each separate resource exhibits a profile of SNPs dependent on the resource's typical genotyping experiment, which is a function of population stratification—cohort size, cohort scope, and degree of sample pooling. In other words, ascertainment bias inherent in the experiments generating a SNP database naturally will determine the spectrum of mutations that can be seen. This body of observed protein-encoding DNA variants results from the constant balancing act between a mutation's intrinsic rate and the typical impact on the encoded protein. For example, figure 2.1 shows that if a population of only 25 individuals is genotyped in a study, then that study is sufficient to detect rare alleles with frequencies of <1% only ~35% of the time. This is significant because the mutations that cause a greater impact on the genome are those that will be found at low frequencies. Each distinct mutation database therefore engenders its own distinct mutation spectrum, M. Given this, mutation spectra composed throughout this study are identified by their training dataset. For example, a mutation spectrum calculated from dbSNP mutations would be referred to as M_{dbSNP}. Details of the wide variety of point mutation databases used throughout this dissertation project are summarized in table 3.1.

Table 3.1: Point mutation datasets analyzed in this study

Database	Build or release date	# point mutations (#nonsynonymous/ #synonymous)	# genes (mutations/gene)
HGMD	5/24/2001	15,118 (15,118/0)	964 15.7
	10/17/2002	17,170 (17,170/0)	1,157 14.8
dbSNP intron (iSNPs)	RefSeq genome contigs build 34.3	1,506,740 (n/a)	17,723 85.0
dbSNP cSNPs	RefSeq release 1	42,237 (24,496/17,741)	15,508 2.7
Interspecific point mutations, chimpanzee and human	*Science* 302(5652): 1960-3. 2003.	14,073 (4,915/9,158)	7,645 1.8
Cancer genome anatomy project, mouse (CGAP)	4/17/2002	4,232 (1,567/2,665)	1,108 3.8
SNP Consortium	release 10	6,016 (3,926/2,090)	3,663 1.6
jSNP	5/12/2003	4,595 (1,924/2,671)	3,022 1.5

* All datasets represent coding region point mutation except for the dbSNP iSNPs

3-1-1 Human Genome Mutation Database

The human gene mutation database (HGMD) is a highly specialized resource detailing nonsynonymous, disease-causing point mutations manually culled from published studies (Krawczak, Ball et al. 2000) and maintained by the Institute of Medical Genetics in Cardiff. The goal of the HGMD has been to collate the wide variety of published gene

lesions and provide the information in a centralized, publicly available setting. This database, which includes information on point mutations, deletions, insertions, duplications, and complex DNA rearrangements, is maintained by a combination of computerized scans of over 250 scholarly journals as well as manual curation. Over the course of this study, the HGMD was referenced to glean annotated, nonsynonymous mutations that were judged to be causative of a clinical phenotype. The HGMD contains far too few synonymous mutations (<20) to collect them for the purpose of mutational trend analysis. This dearth of silent mutation data is most likely due to bias in the curation process, for such variants are thought to be neutral due to their lack of impact on the protein product. This may be a fallacious assumption when applied so broadly and indiscriminately; however, for silent mutations can most certainly affect mRNA structure as well as alter crucial sequence motifs involved in exon splicing. It is expected that as the number of large-scale phenotype association studies grow (and more importantly, mature), the number of silent mutations judged to be phenotype-inducing will grow in concert.

The most often-used HGMD build in this study was downloaded on 10/17/2002 where 15,118 total gene coding region point mutations were retrieved from the HGMD website (http://www.hgmd.org). These mutations spanned 964 genes, which is indicative of a SNP density of 15.7 mutations per gene. The high density of SNPs within the HGMD reflects the experimental design and goals of the studies that compose the database. Disease relevant mutations are most often found within high-throughput surveys of gene sequences or through the use of a highly specialized cohort. All analyses of the HGMD were repeated with updated releases as appropriate throughout this study.

3-1-2 dbSNP

The largest public SNP database available from which to derive point mutation trends is NCBI's dbSNP project (Smigielski, Sirotkin et al. 2000), which catalogs polymorphisms from any genotyping experiment. This database was first created near the beginning of my graduate career, making this ensuing study of point mutation quite timely. The goal of dbSNP is to be a central depository for any polymorphism discovered in any organism during the course of any type of genotyping study. Most of the point mutation alleles found in human populations, however, have been discovered during reduced representation shotgun sequencing of less than ten individuals (Altshuler, Pollara et al. 2000) or via detection of bacterial clone overlap as a by-product of the human genome project (Mullikin, Hunt et al. 2000). This means that these SNPs are by nature of the discovery methods quite common in humans, and any derived trends would represent those reflecting common point mutation only. Therefore, dbSNP is expected to be heavily biased towards neutral variants, that is, those that confer no phenotype. Mutation rates based on this resource would therefore be underestimated due to the lack of high impact alleles.

Throughout this work two types of dbSNP SNP collections were maintained: One derived from introns and a second from gene coding regions (cSNPs). This was done so that point mutation trends could be calculated from collections that were both free from selection (intronic SNPs, iSNPs) and under the influence of selection on the protein product (cSNPs). The iSNPs become very important to interpreting the coding region spectra, which is discussed at length in chapters 4 and 6. Intronic dbSNP mutations were obtained by parsing

all Genbank-style human RefSeq (NCBI build 34.3) genomic contigs for '/variation' tags. The coordinates of the gleaned iSNPs were matched to the annotated coordinates of non-pseudogenic, protein coding genes in order to identify intronic mutations. In cases where multiple intron/exon arrangements were possible, the arrangement giving the longest protein coding sequence was used. Intronic SNPs were further filtered by dbSNP reference number (rs#) to retain only those mapping unambiguously to one genomic region. This process found 1,586,740 variants across 17,723 gene sequences.

To collect only reviewed dbSNP cSNPs, NCBI's LocusLink (Pruitt and Maglott 2001) (build 10/13/2002) was queried for human, protein coding, non-pseudogene loci. The corresponding RefSeq (Pruitt and Maglott 2001) cDNA GenBank accession numbers were collated. Annotated cSNPs were obtained by parsing all "/variation" tags in GenBank records and sorted by reference ID (rs#) to obtain 42,237 nonredundant mutations across 15,538 protein-encoding genes (2.7 cSNPs/ gene). The low SNP density in this dataset as compared to coding region point mutations found in the HGMD reflects the low-throughput nature of the dataset. For both different dbSNP datasets, the major allele was assumed to be the nucleotide in the reference sequence. Multi-allelic SNPs were included as independent mutation events that happen to occur at identical nucleotide positions.

3-1-3 The SNP Consortium

The SNP Consortium (TSC) resource (Sachidanandam, Weissman et al. 2001) details human point variants identified by sequencing DNA from an ethnically diverse group of 20 individuals for construction of a high-density genomic SNP map. It is accessible in the form

of a full MySQL dump at http://snp.cshl.org. Unlike the case of dbSNP, TSC SNPs are all obtained from only a single experiment and are therefore expected to display less mosaicity in any resultant mutation trend analysis. A local copy of TSC release 10 was downloaded (Sachidanandam, Weissman et al. 2001) and 1,148,402 human SNPs with at least 50 bases each of 5' and 3' flanking sequence were extracted. Gapped BLAST (Altschul, Madden et al. 1997) was used to map the variants to protein coding RefSeq cDNA sequences (RefSeq, build 3/15/2002). A mapped cSNP was defined as any match whose alignment had no gaps and at least 95% sequence identity over 60 nucleotides or more (6,016 cSNPs total). These cSNPs were spread across 3,663 genes giving a density of 1.6 cSNPs/gene. Although these stringent parameters likely eliminated genuine cSNPs, they afforded high confidence in the annotation. As with the other datasets, the cDNA reference sequence was assumed to contain the major allele.

3-1-4 jSNP

jSNP is a repository of SNPs identified in the Japanese population as derived from a panel of 24 unrelated (yet ethnically Japanese) individuals (Hirakawa, Tanaka et al. 2002). The goal of this project has been to identify common variants that can later be used to draw ties to common diseases. Given that this isolated population is more inbred, jSNP stands a greater chase at identifying common disease-relevant SNPs. This dataset is similar in scope to the SNP Consortium resource, however, only one ethnic group is genotyped while the SNP Consortium chooses an ethnically random cohort. Therefore, cSNP trends derived from jSNP are representative of a mutation spectrum derived from a more inbred and isolated

population than any other human cohort analyzed in this study. 4,595 cSNPs were gleaned by downloading the dataset from the jSNP website (http://snp.ims.u-tokyo.ac.jp/). These cSNPs spanned 3,022 genes giving a SNP density of 1.5 SNPs/gene. The ancestral allele was determined by referencing the point of mutation in the listed RefSeq GenBank record.

3-1-5 Cancer Genome Anatomy Project

NCI's Cancer Genome Anatomy Project (CGAP) has created the Genomic Annotation Initiative (GAI) (Riggins and Strausberg 2001) to locate human and mouse germline point variants in cancer-relevant genes. ESTs are generated in massive amounts from healthy mice and aligned to flag high quality discrepancies as SNPs (Buetow, Edmonson et al. 1999). Considering that as few as 5 individuals' ESTs compose this alignment, the database is heavily biased towards high frequency alleles as in the case of the human dbSNP collection. This resource was probed in order to retrieve a cSNP dataset complementary to dbSNP but yet originating from a distinct mammalian system. This would later facilitate comparison between the two species in order to ask whether mammals that diverged from a common ancestor a very long time ago share point mutation trends. Unique *Mus musculus* cSNPs (4,232 across 1018 genes, 4.2 cSNPs/gene) were obtained from the CGAP-GAI web site (http://lpgws.nci.nih.gov:82/perl/snp2ref, 4/17/2002). The snp2ref tool maps their discovered alleles onto RefSeq cDNAs. Only non-pseudogene mutations with a SNP confidence score, the chance that the given nucleotide is correctly called as calculated from sequence quality values developed by CGAP researchers, of greater than or equal to

0.99 were considered (Clifford, Edmonson et al. 2000). The authors of the CGAP-GAI project also assume that the major allele is that in the associated RefSeq cDNA.

3-1-6 Interspecific substitutions between primates

All datasets discussed thus far are composed of polymorphisms seen within a single species. However, it is of great interest to examine the profile of point substitution *between* species given that this spectrum is merely the summation of a large body of intraspecific mutation that eventually gives rise to speciation. Since nearly all datasets described thus far are human, an interspecific substitution dataset that reflects human speciation is useful. Therefore, human gene sequences should be aligned to a homolog from another species so that DNA base differences can be resolved. In doing this, the first problem is to choose an organism for interspecific comparison. This is a function of how 'far back' in the evolutionary tree one wants climb in order to find a common ancestor with another organism and to a lesser, practical extent, the availability of gene sequences. It is clear that one would want to only chose an organism where it could be safely assumed that in nearly all cases only one point mutation event occurred per DNA base when comparing aligned homologs. Otherwise, any derived trends would not accurately reflect point mutation trends. Additionally, once two species are chosen for comparison it is critical that an ancestral homolog be aligned so that the direction of any interspecific substitutions can be ascertained. Fortuitously, Clark and coworkers published a large number of high-quality cDNA multiple sequence alignments of *Homo Sapiens*, *Mus musculus*, and *Pan troglodytes* orthologs in early December 2003 as part of the chimpanzee genome project (Clark, Glanowski et al. 2003).

Since the estimated time of divergence between humans and chimpanzees is 5 million years ago while the rate of point mutation is on the order of 10^{-8}, it can be reasonably assumed that any point substitution seen within a multiple alignment represents one mutational event (Bazykin, Kondrashov et al. 2004).

In order to glean interspecific substitutions, any ungapped coding region position within the alignments that differed in nucleotide identity between the chimpanzee and human cDNA sequences was treated as an interspecific point mutation, with the aligned mouse allele defining the ancestral state so that the direction of the mutation could be estimated. In the case that the mouse allele did not match either the aligned chimp or human base then that mutation was discarded. 4,915 DNA base positions causing a missense or nonsense change in the aligned amino acid were gleaned from 7,645 cDNA alignments.

3-2 Elucidation of point mutation spectra from diverse point mutation databases

3-2-1 The definition of and meaning underlying of a point mutation spectrum

As discussed in chapter 1, I have chosen to define a mutation spectrum, M, as the distribution of observed point mutations across a DNA sequence. M has two components that shape the variant pool observed within a given dataset: i) m, the inherent mutability of a DNA sequence and ii) s, natural selection forces, so that $M \propto m + s$. Each spectrum contains point mutations that are categorized according to their three-nucleotide sequence context before and after the mutation, such as a GGG→GAG for a particular G/A mutation. Such a grouping is termed a trinucleotide mutation class (TMC). There are 576 TMCs possible given that there are four DNA bases. Additionally, since a single base within a DNA

sequence can belong to three distinct trinucleotides, each mutation spectrum is composed of three mutability tables (e.g. one for each frame), each one describing the observed frequencies of the 576 TMCs. Table 3.2 lists the mutation spectra calculated in this study, which are derived from the datasets listed in table 3.1.

The principal spectrum examined by which all others are compared, M_{intron}, is derived from assessing the observed frequency of SNPs that have been mapped to RefSeq (Pruitt and Maglott 2001) intron sequences across all possible TMCs. These frequencies are weighted by the nucleotide composition of those same intron sequences so that a log odds score can be calculated. Since intronic DNA is not subjected to the same selective forces exerted upon coding region sequences, I estimate that s_{intron} is nearly zero and therefore M_{intron} reduces to the actual mutability of the TMCs, or m_{intron}. This is the only mutation spectrum in the entire dataset that can be treated in this manner. Such an approach has been used in other studies where it was necessary to avoid selective forces complicating mutation observation in coding regions (Majewski and Ott 2003). Although there are functional sequence motifs within introns indeed subject to selection pressure, the SNPs in these motifs, as well as the motifs themselves, are expected to be quite diluted across the whole dataset.

Since m represents the inherent mutability of those DNA sequences that compose a mutation spectrum as a result of cellular processes and nucleotide physiochemistry, I assume that m will be roughly constant in all gene coding regions. Therefore, if I calculate an additional mutation spectrum (M_{coding}) derived directly from mutations occurring in gene coding regions where s_{coding} is certainly nonzero, I can compare it to M_{intron} and make

inferences about the nature of s_{coding} from qualitative differences. That is, if $M_{intron} \propto m_{intron}$ (since $s_{intron} \approx 0$) and

$$M_{coding} \propto m_{coding} + s_{coding}, \text{ then}$$

$$M_{coding} - M_{intron} \propto s_{coding}.$$

The spectra listed in table 3.2 calculated from datasets of cSNPs (table 3.1) were chosen for comparison to M_{intron}.

Table 3.2: Mutation spectra derived in this study

Database	Spectrum name	Mutation scoring tables
dbSNP intron	M_{intron}	M_{intron}_TMC
dbSNP cDNAs	M_{dbSNP}	M_{dbSNP}_TMC$_{\text{coding}}$_nonsynonymous M_{dbSNP}_TMC$_{+1}$_nonsynonymous M_{dbSNP}_TMC$_{+2}$_nonsynonymous M_{dbSNP}_TMC$_{\text{coding}}$_synonymous M_{dbSNP}_TMC$_{+1}$_synonymous M_{dbSNP}_TMC$_{+2}$_synonymous M_{dbSNP}_TMC$_{\text{coding}}$_all M_{dbSNP}_TMC$_{+1}$_all M_{dbSNP}_TMC$_{+2}$_all
Interspecific point mutations, primates	$M_{\text{interspecific}}$	$M_{\text{interspecific}}$_TMC$_{\text{coding}}$_nonsynonymous $M_{\text{interspecific}}$_TMC$_{+1}$_nonsynonymous $M_{\text{interspecific}}$_TMC$_{+2}$_nonsynonymous $M_{\text{interspecific}}$_TMC$_{\text{coding}}$_synonymous $M_{\text{interspecific}}$_TMC$_{+1}$_synonymous $M_{\text{interspecific}}$_TMC$_{+2}$_synonymous $M_{\text{interspecific}}$_TMC$_{\text{coding}}$_all $M_{\text{interspecific}}$_TMC$_{+1}$_all $M_{\text{interspecific}}$_TMC$_{+2}$_all
CGAP, mouse	M_{CGAP}	M_{CGAP}_TMC$_{\text{coding}}$_nonsynonymous M_{CGAP}_TMC$_{+1}$_nonsynonymous M_{CGAP}_TMC$_{+2}$_nonsynonymous M_{CGAP}_TMC$_{\text{coding}}$_synonymous M_{CGAP}_TMC$_{+1}$_synonymous M_{CGAP}_TMC$_{+2}$_synonymous M_{CGAP}_TMC$_{\text{coding}}$_all M_{CGAP}_TMC$_{+1}$_all M_{CGAP}_TMC$_{+2}$_all

Database	Spectrum name	Mutation scoring tables
TSC	M_{TSC}	M_{TSC}_TMC$_{coding}$_nonsynonymous M_{TSC}_TMC$_{+1}$_nonsynonymous M_{TSC}_TMC$_{+2}$_nonsynonymous M_{TSC}_TMC$_{coding}$_synonymous M_{TSC}_TMC$_{+1}$_synonymous M_{TSC}_TMC$_{+2}$_synonymous M_{TSC}_TMC$_{coding}$_all M_{TSC}_TMC$_{+1}$_all M_{TSC}_TMC$_{+2}$_all
HGMD	M_{HGMD}	M_{HGMD}_TMC$_{coding}$_nonsynonymous M_{HGMD}_TMC$_{+1}$_nonsynonymous M_{HGMD}_TMC$_{+2}$_nonsynonymous M_{HGMD}_TMC$_{coding}$_synonymous M_{HGMD}_TMC$_{+1}$_synonymous M_{HGMD}_TMC$_{+2}$_synonymous M_{HGMD}_TMC$_{coding}$_all M_{HGMD}_TMC$_{+1}$_all M_{HGMD}_TMC$_{+2}$_all
jSNP	M_{jSNP}	M_{jSNP}_TMC$_{coding}$_nonsynonymous M_{jSNP}_TMC$_{+1}$_nonsynonymous M_{jSNP}_TMC$_{+2}$_nonsynonymous M_{jSNP}_TMC$_{coding}$_synonymous M_{jSNP}_TMC$_{+1}$_synonymous M_{jSNP}_TMC$_{+2}$_synonymous M_{jSNP}_TMC$_{coding}$_all M_{jSNP}_TMC$_{+1}$_all M_{jSNP}_TMC$_{+2}$_all

3-2-2 Determination of point mutation spectra for gene intron regions

When a trinucleotide undergoes a single point mutation, such as a GGG→GAG transition, this defines a trinucleotide mutation class (TMC). Since the mechanisms invoking a NNX→NNY mutation may be distinct from those causing a NNY→NNX variant, TMCs have directionality and assume both a wild-type and mutant allele. Overall, the TMC matrix has 576 nonzero values because there are $3 \cdot 3 \cdot 4^3$ total ways that the set of 64 all possible trinucleotides can point mutate. I tallied how many intronic SNPs (iSNPs) belonged to each possible TMC so that the most (or least) mutable sequences could be highlighted and placed into a table that would describe their probabilities. Each intronic SNP was counted and categorized three times, once for each possible trinucleotide context, which is analogous to considering mutations in different coding frames and reflects mutation occurrence in the context of its four surrounding nucleotides. An example of this process is illustrated in figure 3.1a. The frequency of TMCs in the intronic SNP dataset was therefore related by the equation:

$$f(TMC) = \frac{n_{TMC}}{n_{total}}$$

where n_{TMC} is the number of SNPs of a specific TMC and n_{total} is total number of SNPs in the dataset (1,586,740 iSNPs). The TMC frequencies are weighted against the usage of the wild-type trinucleotide calculated directly from the intron sequences as the fraction of all trinucleotides in the introns that are a specific trinucleotide. Since a single base in a DNA sequence can belong to three distinct trinucleotide stretches, the total number of trinucleotides was three times the combined length of all intron sequences. The logarithm

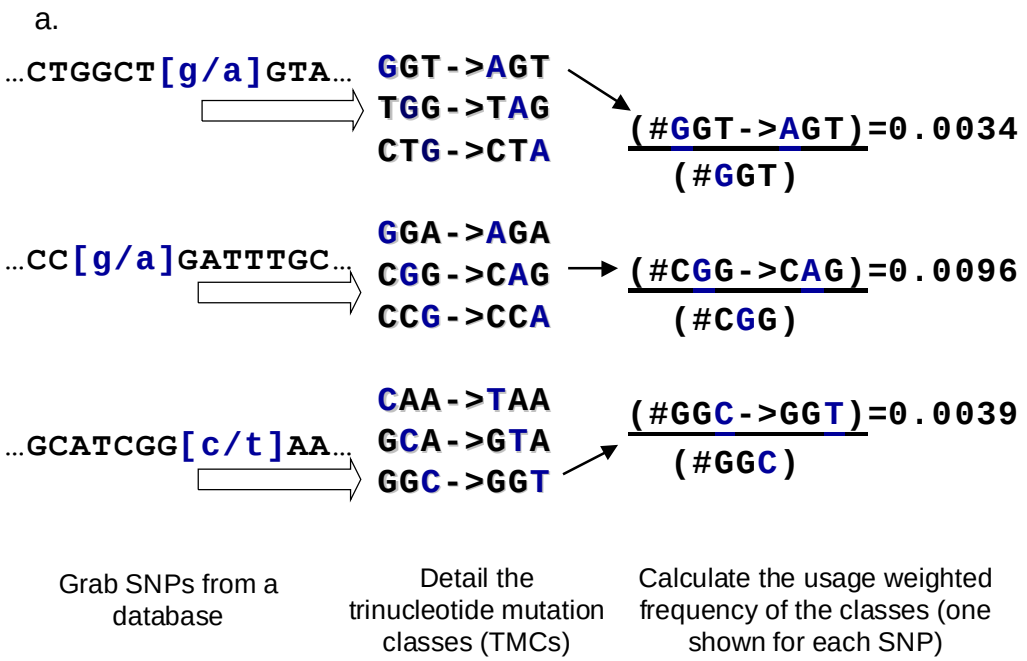
was then taken of the quotient of these two values to obtain a log likelihood score for a specific TMC. For example, the log likelihood score for the TMC $NaN \rightarrow NbN$ was calculated as:

$$score(NaN \rightarrow NbN) = \log\left(\frac{f(NaN \rightarrow NbN)}{usage(NaN)}\right)$$

where *f(* $NaN \rightarrow NbN$ *)* is the frequency of the TMC $NaN \rightarrow NbN$ in the database examined, and *usage(NaN)* is the frequency of that trinucleotide in all introns examined. A TMC with a positive score represents a mutation class more likely to occur than expected based on the underlying trinucleotide usage. One with a negative score is less likely than random to occur based on the underlying trinucleotide composition.

Often throughout this study there were situations where it was more useful to represent trinucleotide mutation class scoring tables as a set of usage-weighted frequencies (figure 3.1b), as opposed to log odds ratios. In this case, the logarithm was not taken of the usage-weighted frequency shown above. Instead, the entire body of usage-weighted frequencies was normalized to 1. One advantage of this version of TMC scores is that one can more easily visualize the magnitude of difference between the most- and least- likely mutation classes. The disadvantage is that one cannot quickly determine which of the TMCs are least likely than random to occur, for all the frequencies have positive numbers. Just like a log odds score, a TMC frequency value is considered to be an estimate of the probability of such a mutation being observed during a genotyping experiment of a cohort having similar population structure as the source database. To avoid confusion, the log likelihood representation is referred to throughout this document as the “log odds ratio score” while the

usage-weighted frequency version is referred to as the "weighted frequency score". The reader is reminded, however, that both versions of calculating mutability observation are entirely equivalent.



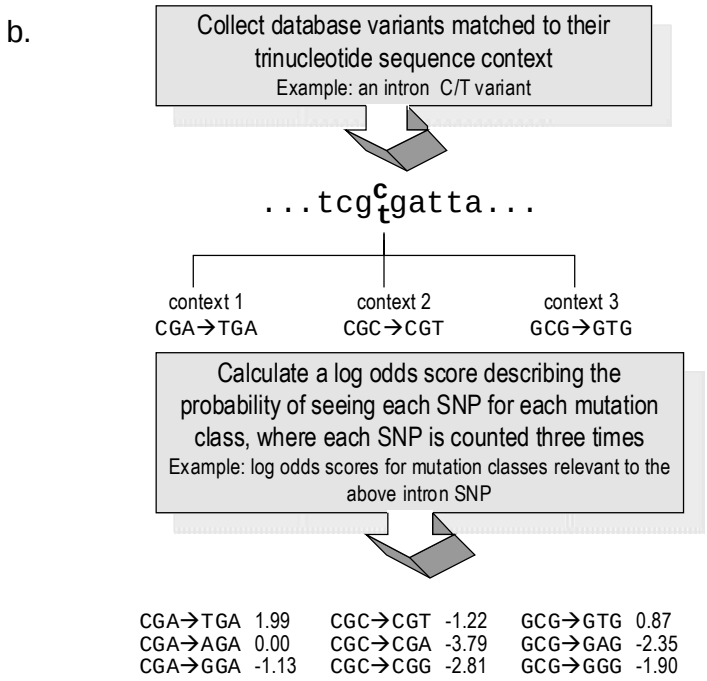


Figure 3.1: Calculation of trinucleotide mutation class (TMC) mutability tables using iSNPs collated from dbSNP. a) Demonstration of calculation of TMC scoring table in terms of usage-weighted frequencies. b) Demonstration of calculation of TMC scoring table in terms of log odds ratios.

3-2-3 Determination of point mutation spectra for gene coding regions

The calculation of log scores for various gene coding region spectra derived from datasets in table 3.1 was identical to that done for M_{intron} except that since cSNPs are utilized, they must be categorized according to specific frames. Consequently, a total of three scoring tables were required per spectrum: One for the actual coding frame (the TMC_{coding} table) plus two representing the noncoding trinucleotide frames due to a +1 or +2 base frameshift (tables TMC_{+1} and TMC_{+2}, respectively). The latter noncoding tables allowed the mutation data to be categorized in the context of its contiguous codons and the frequencies represent mutation propensity influenced by the 5' and 3' codons. Since the HGMD lacks any silent point mutations and M_{intron}, derived from dbSNP intronic SNPs, cannot be categorized by coding frames, in some cases three distinct mutation spectra (each of course having three scoring tables, one for each trinucleotide context) were calculated: One for nonsynonymous SNPs, silent SNPs, and all SNPs. This facilitates proper comparison of the datasets throughout the study. Table 3.2 lists all mutation spectra and the corresponding scoring tables used in this study.

In generating a given mutation spectrum, each point mutation was evaluated to determine its TMC_{coding}, TMC_{+1}, and TMC_{+2} classes after frameshifts. Coding and noncoding frame trinucleotide usages were calculated directly from the cDNAs in each dataset. An exception to this was the human HGMD datasets, where noncoding frame trinucleotide usages were calculated from 104,170 non-pseudogene human cDNAs retrieved from UniGene (Hs.seq.uniq 8/20/2002) because the total number of HGMD genes were few in number. The codon usage was taken from the codon usage database (Nakamura, Gojobori

et al. 2000). Although the process of constructing these tables for cSNPs is nearly identical to doing so for iSNPs, figure 3.2 reiterates the method of deriving log odds ratio scores using the HGMD dataset in order to avoid any confusion.

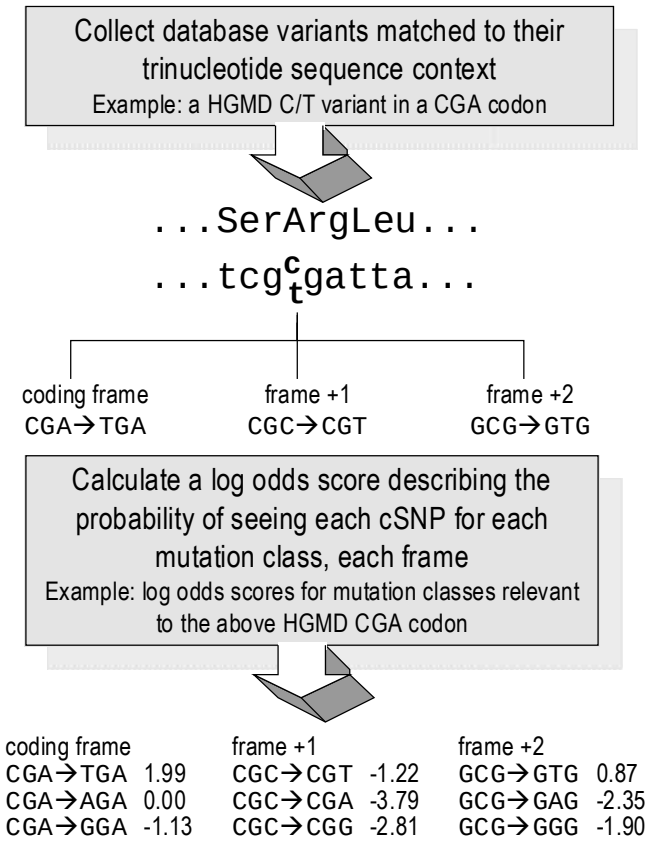


Figure 3.2. Calculation of trinucleotide mutation class (TMC) mutability tables using cSNPs collated from the HGMD.

A frame-based method of classifying cSNPs is critical to understanding coding region point mutation trends. Selection operates on each position in a codon differently depending on the impact of the amino acid replacement. For any given codon in a functional gene, it is impossible to know the intrinsic mutation rate at each position because many variants are quickly selected against and consequently never achieve an appreciable frequency in a population. As a result, the perceived quantity of neutral mutation, such as most silent variants, is disproportionately high compared to the number of nonsynonymous mutations found (Cargill, Altschuler et al. 1999). A genotyping experiment that probes a large number of individuals has greater power to detect deleterious alleles that exist at low frequencies. Therefore, representing a database by three different trinucleotide mutation class (TMC) distribution, one for each frame, not only contrasts the relative mutation tendencies of codons, but allows quantitative comparison of the efficiency of different genotyping experiments using statistical methods.

CHAPTER FOUR
ANALYSIS OF HUMAN POINT MUTATION TRENDS

4-1 Mutation classes defined by trinucleotide sequence context effectively characterizes different SNP datasets

From evolution's viewpoint two inseparable properties shape observed trinucleotide mutation class (TMC) frequencies: The new allele's impact on the encoded protein and the rate of point mutation due to cellular mechanisms. On a superficial level this statement seems to conflate the distinct phenomena of mutation rate, selection, and phenotypic consequence. But this problem cannot be avoided when examining a mutation spectrum individually. Selection acts to suppress a fitness-decreasing mutation by penalizing its allele frequency (perhaps even to zero). Attempts to estimate basal gene mutation rates in real populations via genotyping suffer from an inability to disentangle the intertwined effects of phenotype impact and selection from the underlying mutation rate, especially when a variant may have a role in a complex phenotype (Cooper and Krawczak 1990). Instead of improperly equating a derived TMC frequency to a mutation rate when examining cSNPs under selection, it is best described as the projected incidence of that point variant class in an experimental cohort of similar size and breadth as the training SNP dataset (e.g. the SNP dataset from which the TMCs were derived). If a TMC frequency is quite high, such a mutation type is likely to be found in another population of similar stratification because the typical impact of that mutation on the numerous genes within each dataset will be the same.

The calculation of a table of TMC frequencies in their log odds ratio form is analogous to the philosophy behind creating BLOSUM substitution matrices from sets of protein alignments. To develop the BLOSUM matrix, a "population" of protein homologs is

aligned and the usage-weighted frequency of each amino acid at each position in the alignment estimates the probability of one amino acid substituting another in a homolog (Wilbur 1985; Henikoff and Henikoff 1993). The BLOSUM protein "population" varies depending on the sequence identity limit for any two proteins in the alignment, such as 62% for BLOSUM62. BLOSUM scores indicate that on average Ile to Leu will occur more often in a multiple alignment of homologs than a charge altering mutation such as Asp to Leu. It is understood that this assumption will not be valid for all proteins at all positions; however, BLOSUM scoring matrices are used extensively in studies as the first attempt to quantify whether a mutation preserves or disrupts the function of a protein. Likewise, a matrix of 576 TMC frequencies derived from categorizing a database's cSNPs performs the same function by stating the average likelihood of finding a new allele due to the intertwined effects of mutation rate and selection. Thus the TMC frequency matrix enables one to classify mutations, compare gene mutational load, and contrast SNP databases populated by different genotyping methods.

4-2 Most point mutation can be described by only a handful of DNA sequence contexts

The frequency of TMC observation exhibits a highly nonuniform distribution in all databases examined. According to classic mutation spectra analysis, if all cSNP events are equally likely to be observed, the null model states that the TMCs should follow a multinomial distribution with respect to the number of variants in each class (Adams and Skopek 1987). For all cases, the observed distribution of TMCs is nonuniform having a set of classes that are far more or far less likely to be found in a given database relative to the

null model. The null model is the distribution that would be seen if all nucleotide sites in gene coding regions were equally likely to be mutated. Figure 4.1 a-d Shows the distribution of point mutations across the TMCs (weighted frequency score method, see chapter 3) for the coding frame of four cSNP mutation spectra derived in this study, and is representative of the trends seen in the databases. Figure 4.2 shows a similar distribution of log odds scores (see chapter 3) but only for the M_{intron} mutation spectrum built upon dbSNP intronic mutations.

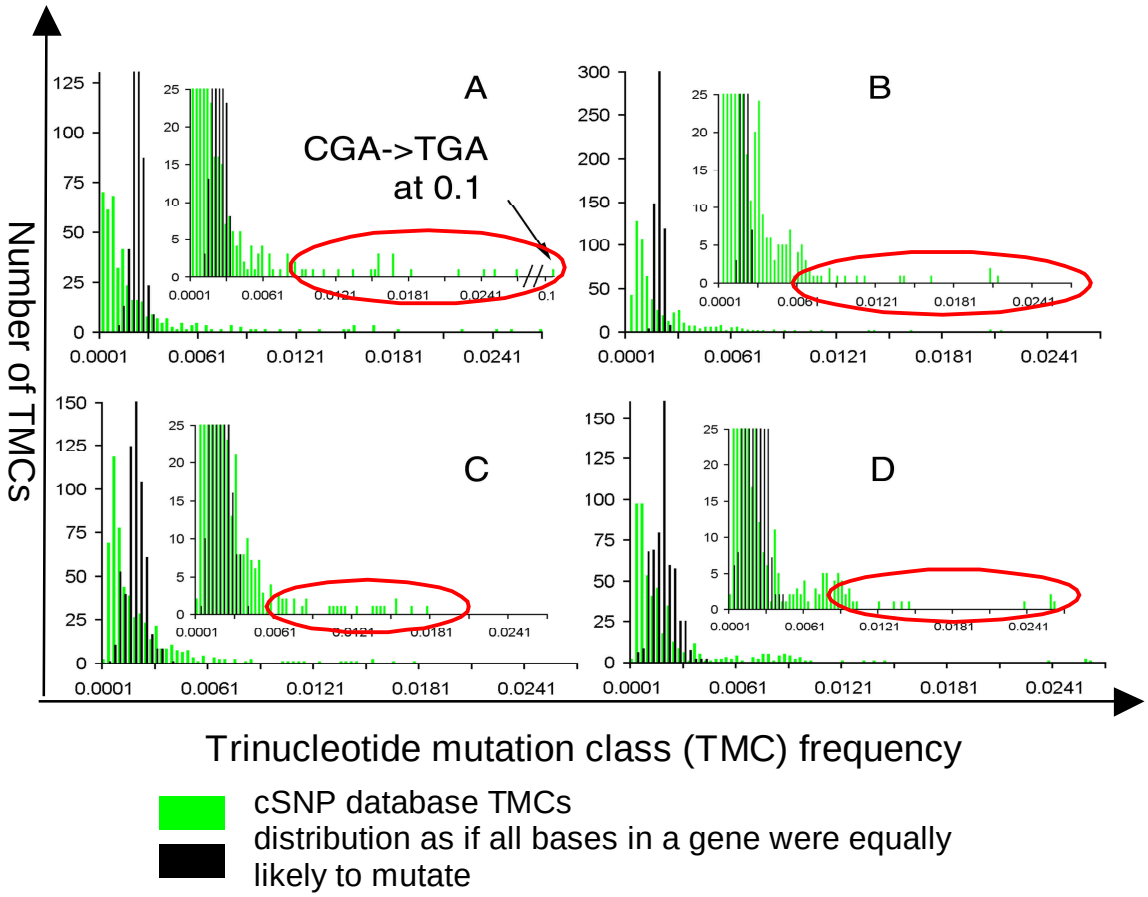


Figure 4.1 Point mutation is a highly nonrandom process: A large fraction of human gene coding region mutation can be described by a handful of sequence context motifs. Shown in green are the human trinucleotide mutation class (TMC_{coding}) distributions (weighted frequency scores method, see chapter 3) for each analyzed dataset with the extremely likely mutation classes encircled in red. A) M_{HGMD} (nonsynonymous-only) mutations B) M_{dbSNP} (all mutations) C) *Mus musculus* M_{CGAP} (all mutations) and D) M_{TSC} data (all mutations). The x-axis is TMC frequency while the y-axis is the number of TMCs that have frequency x. The black histograms represent what would be expected if the TMCs for each dataset were equally likely to be populated under the null model for mutation spectra, that is, if each type of mutation was equally likely to occur. Insets display the graphs with a reduced y-axis maximum so that the high frequency bars are visible. CGA→TGA is actually off-axis in A (break shown) and represents 10% of all causative mutation in the HGMD.

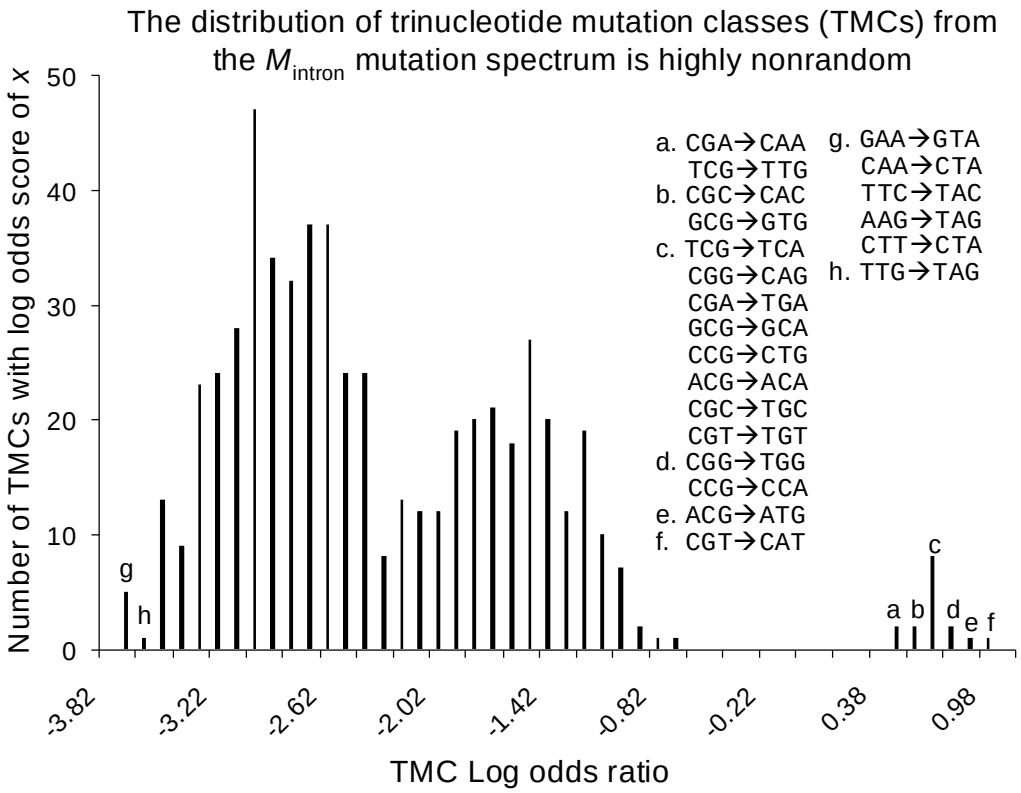


Figure 4.2: Intronic point mutation is also highly nonrandom when it is examined according to its distribution across all possible trinucleotide mutation classes (576 possible). In contrast to figure 4.1, log odds scores are shown meaning that the mutation class is less likely than random to occur (e.g. if mutations were scattered evenly across the intron sequences) when the value is negative. The mutation class is more likely than random when the value is positive. The x-axis is the TMC log odds ratio score while the y-axis is the number of TMCs that have score x. Mutation classes with positive log odds scores are marked as bars a-f while representative highly unlikely mutation classes are groups g and h.

The fact that mutation is nonrandom in itself is not surprising, but what is intriguing is the extremely high departure from uniformity in all datasets. For example, only the top 5% (29/576) of TMCs account for 27.4% of the observed coding region variants in dbSNP while the bottom 5% account for only 0.02% of dbSNP variants. By contrast, the expected values for the top and bottom 5% taken by sampling all 576 possible TMCs with equal probability 42,237 times (the number of variants in dbSNP) are 6.3% and 3.9% respectively. Fig. 4.1 graphs the observed and expected distributions which show the dramatic difference in TMC dispersion for four mutation spectra. The other databases have similar statistics where the top 5% of TMCs describe 28.8% of TSC variants, 34.2% of CGAP variants, and 40.8% (22/438 nonsynonymous classes) of disease-causing HGMD mutations. The null model expects that these values should be only 8.5%, 9.4%, and 5.1% respectively. This convergent result from four independent datasets shows that a considerable fraction of point mutation in functional genes can be described by only a handful of TMCs. Such sequence contexts have high propensity for observable cSNPs regardless of gene, population stratification in the genotyping experiment, or specialized nature of the database, such as the HGMD, where only disease-causing mutations are detailed. Additionally, such trends hold true in the mouse model system (CGAP data). The fact that the mouse CGAP data has TMC characteristics akin to human databases demonstrates that these results are mirrored in highly-related mammals. Although CpG dinucleotides are hotspots even in bacteria, there is the shared trend between mouse and human mutation data showing that certain CpG-containing TMCs are more prevalent than others. Consequently, relative gene mutational load may be

estimated by examining a gene's coding sequence to register how many of these "hotspot" TMCs are possible.

Figure 4.2 shows the distribution of dbSNP SNPs in intronic regions according to the TMCs (M_{intron}). In this depiction of the data, the comparison to what is expected under the null model (random mutation) is inherent in the log odds score shown. The distribution of SNPs is expected to reflect only the differential mutation propensities of the mutation classes and are unaffected by selection at the level of amino acids. The first important piece of information to glean from this spectrum is that 33.4% of all intronic point mutations are due to only sixteen mutation classes—all of which involve transitions at a CpG dinucleotide. This suggests that 1/3 of all point mutation can be very easily predicted and represents the 'low hanging fruit' of naturally occurring point mutation. Additionally, it is clear that not all CpG dinucleotides have equal mutation potential, for all have significantly different mutabilities. Because each mutation is counted three times in order to categorize the intronic SNPs in all three frames, each TMC represents SNPs categorized according to four nucleotides—two on either side of the site of mutation. This is significant because so many studies of mutation rates uses a dinucleotide model, and these results suggest that conclusions based on such work may need to be reevaluated.

4-3 SNP datasets differ primarily in their distribution of point mutations that change the sequence of the encoded protein.

In analyzing point mutation spectra differences, the task was broken down to considering silent and nonsynonymous mutations separately. This is because silent mutations are more often expected to be selectively neutral than nonsynonymous mutations, for the former do not change the amino acid sequence of the encoded protein.

4-3-1 Analysis of nonsynonymous variants

In examining the differences between mutation spectra, the most focus was given to spectra that represented a distinct form of genotyping bias: M_{intron} (iSNPs used to gauge mutation tendencies in the absence of selection), M_{dbSNP} (low-throughput genotyping of human gene sequences), M_{HGMD} (disease-causing mutations only), and $M_{\text{interspecific}}$ (interspecific substitutions found between primates). Spectra based on TSC and jSNP data are not discussed rigorously with respect to M_{intron} because they are both cSNP datasets derived from databases composed from low-throughput genotyping. This type of spectrum is best represented by mining the dbSNP cSNP database (M_{dbSNP}), for it is the largest in terms of actual SNP number.

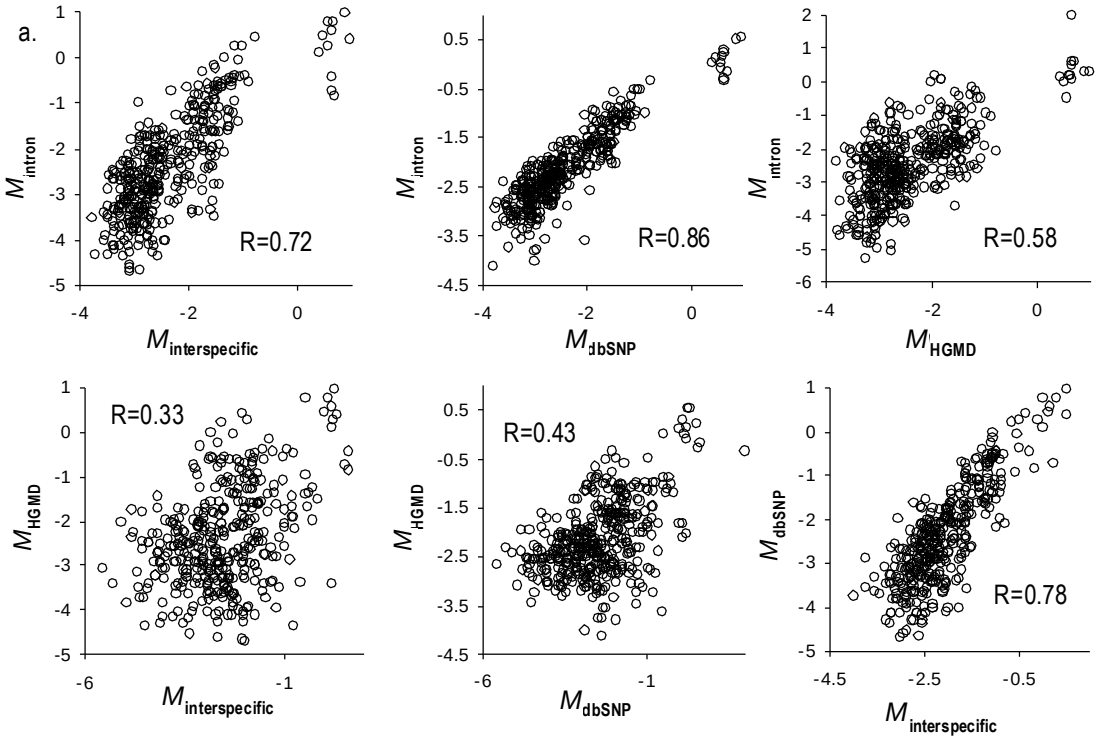
As discussed in chapter 3, a mutation spectrum 'M' is shaped by two major components: i) m, the inherent mutability of a DNA sequence and ii) s, natural selection forces, so that $M \propto m + s$. Therefore, each nonsynonymous point mutation spectrum, given the ascertainment bias inherent in the point mutations on which it is trained, imparts a different view of s and therefore a different view of what types of variants remain in a gene

pool after different levels of natural selection. Firstly, M_{dbSNP} is calculated from cSNPs (coding region SNPs) within dbSNP and will be biased towards conservative mutations with respect to the protein product given that much of the data was generated from low-throughput genotyping. M_{TSC} and M_{jSNP} are similarly calculated. At the opposite extreme, an HGMD-based spectrum (M_{HGMD}) would exhibit an excess of radical mutations given that the training dataset contains mostly Mendelian disease mutations. The final spectrum to undergo rigorous comparison to M_{intron}, $M_{\mathrm{interspecific}}$, is trained on single base differences seen between orthologous human and chimpanzee cDNAs aligned to a mouse ortholog treated as the ancestral sequence. With respect to the protein products, this spectrum represents an intermediate between M_{dbSNP} and M_{HGMD} because few of the substitutions are expected to dramatically change the function of the encoded proteins; however, enough differences are present to effect speciation.

Additional spectra calculated from datasets of nonsynonymous mutations are compared to M_{intron} in this manner. Each spectrum, given the ascertainment bias inherent in the point mutations on which it is trained, imparts a different view of s and therefore a different view of what types of variants remain in a gene pool after different levels of natural selection. Firstly, M_{dbSNP} is calculated from cSNPs (coding region SNPs) within dbSNP and will be biased towards conservative mutations with respect to the protein product given that much of the data was generated from low-throughput genotyping. At the opposite extreme, an HGMD-based spectrum is calculated (M_{HGMD}) which exhibits an excess of radical mutations given that the training dataset contains mostly Mendelian disease mutations. $M_{\mathrm{interspecific}}$ is trained on single base differences found between orthologous human and

chimpanzee cDNAs aligned to a mouse ortholog, which is taken as the ancestral sequence. With respect to the protein products, this spectrum represents an intermediate between M_{dbSNP} and M_{HGMD} because few of the substitutions are expected to dramatically change the function of the encoded proteins; however, enough differences are present to effect speciation.

All four of these spectra correlate positively to each other as shown in Figure 4.3a, but it is clear that the magnitude of that correlation depends on the nature of the training data underlying each spectrum, that is, the nature of s in each case. Note, CpG-centric effects dominate these patterns within figure 4.3, yet the specific context of the most mutable CpG's differs for each dataset (figure 4.3b). It is known that the mutation rate of transitions at CpG dinucleotides are 17-18-fold greater (Lunter and Hein 2004) than other point mutations due to the tendency of methylated cytosine to rapidly deaminate to thymine.



b.

M_{HGMD}	M_{intron}	$M_{interspecific}$	M_{dbSNP}
1.99 CGA→TGA	0.98 CGT→CAT	0.95 ACG→ATG	0.54 CGT→CAT
0.63 CGG→TGG	0.87 ACG→ATG	0.77 GCG→GTG	0.53 ACG→ATG
0.60 CGC→TGC	0.69 CGG→TGG	0.77 CCG→CTG	0.30 CCG→CTG
0.53 CGT→TGT	0.64 CCG→CTG	0.58 CGG→CAG	0.24 CGT→TGT
0.32 CGT→CAT	0.63 CGC→TGC	0.47 TCG→TTG	0.17 CGG→CAG
0.28 ACG→ATG	0.63 CGT→TGT	0.42 GTA→ATA	0.11 TCG→TTG
0.21 CGC→CAC	0.63 CGA→TGA	0.39 CGT→CAT	0.02 CGA→CAA
0.19 CGG→CAG	0.62 CGG→CAG	0.25 GTC→ATC	0.01 GCG→GTG
0.19 TGG→TAG	0.56 CGC→CAC	0.25 CGC→CAC	
0.16 CGA→CAA	0.56 GCG→GTG	0.23 ATA→GTA	
0.11 CAG→TAG	0.47 TCG→TTG	0.11 CGA→CAA	
0.11 TGG→TGA	0.44 CGA→CAA		
0.08 CCG→CTG			

Figure 4.3. Differences between point mutation trends seen in global mutation spectra reflect the hand of natural selection. A) Shown are scatter plots of all pair wise Spearman correlations between trinucleotide mutation class (TMC) log odds scores for the four mutation spectra determined in this study. All Spearman correlation coefficients have p-values <0.0001. B) Shown are all TMCs (coding frame only) for each spectrum that have positive log odds scores. These DNA sequence contexts are those most prone to mutation according to each metric and are later scrutinized in this study to determine which gene ontology (GO) categories are hypermutable in the human genome.

The attenuated correlations between the TMC rankings in figure 4.3 are reflected by the Spearman R-values and illustrate the hand of selection operating at different strengths within each of the training datasets. The poorest correlations occur between M_{HGMD} and M_{dbSNP} or $M_{\mathrm{interspecific}}$ because the former is biased towards radical mutation while the latter two spectra are dominated by conservative variants. The M_{intron} spectrum captures all variants regardless of selection and therefore correlates to any of the other datasets better than they correlate to each other. Clearly, this indicates that for the relationship where $M \propto m + s$, m (mutability) is the dominant component shaping the observed spectrum M. That is, as the human genome currently stands, the setup of point mutation game pieces as influenced by nucleotide usage is a stronger force than the hand of natural selection which moves them in determining the outcome of the evolutionary game that populations undertake in their struggle to survive.

Table 4.1 details the top ten nonsynonymous TMCs from each dataset. Note that frequent TMCs of one SNP database often rank highly in any one of the other databases. In fact, only 18 TMCs describe the top ten most observable events in all cSNP resources. BLOSUM62 values are given to provide a rough estimate of the typical impact magnitude of the amino acid substitution. Nearly all highly ranked TMCs involve CpG dinucleotides, but three prominent classes (GGT→AGT (G→S), GTC→ATC (V→I), and GTA→ATA (V→I)) involve G→A transitions at non-CpG sites. Upon manual investigation of individual mutations in these classes, it was found that a substantial portion of this variation is due to a 5’ cytosine that would create a codon-spanning CpG site, and the CpG would be present if a table was drawn similar to 4.1 showing the other two noncoding frames. The CGAP mouse

data is dominated by conservative TMCs and an excess of silent mutations which is a consequence of the small, unstratified population examined by CGAP's EST-alignment process. Such a low-throughput method will have power to detect only extremely common and therefore mostly benign alleles. The data from chimp-mouse-human multiple sequence alignments ($M_{\text{interspecific}}$) are alleles that have been tolerated between similarly-functioning orthologs, and therefore mutation class rank correlates strongly to increasing BLOSUM62 score, an average measure of mutation impact on the protein product. This data illustrates how mutation rate and impact are coupled to shape the observable allele frequencies in a genotyping study. The identities of the most observable TMCs in any SNP database are reshuffled according to the population stratification of the studies that generated it.

This analysis illustrates that codon categorization of SNPs powerfully contrasts databases. Examining the TMC distribution of a dataset and comparing it to the matrices presented in this study can potentially illuminate both its net neutral character (e.g. is the dataset more dbSNP-like or HGMD-like) and identify systematic genotyping errors masquerading as unusual, high-frequency TMC classes.

Table 4.1: cSNP datasets differ primarily in the dispersion of nonconservative variants

trinucleotide mutation class (coding frame)	BLOSUM62 score	Rank of TMC frequency[a] for listed nonsynonymous mutations spectrum					
		dbSNP	jSNP	interspecific	CGAP	TSC	HGMD
CGA→TGA (R→X)	Stop[b]	12	195	16	17	6	1
CGG→TGG (R→W)	-3	10	14	44	45	11	2
CGC→TGC (R→C)	-3	11	9	20	11	9	3
CGT→TGT (R→C)	-3	4	5	39	30	4	4
CGT→CAT (R→H)	0	1	4	7	6	2	5
ACG→ATG (T→M)	-3	2	1	1	2	1	6
CGG→CAG (R→Q)	+1	5	3	4	8	10	7
CGC→CAC (R→H)	0	9	6	9	20	13	8
CGA→CAA (R→Q)	+1	7	7	11	9	7	9
TGG→TAG (W→X)	Stop[b]	158	270	314	153	57	10
CCG→CTG (P→L)	-3	3	2	3	1	3	11
TCG→TTG (S→L)	-2	6	11	5	70	5	14
GCG→GTG (A→V)	0	8	8	2	76	8	21
GGT→AGT (G→S)	0	21	21	48	10	18	22
TCT→CCT (S→P)	-1	67	96	69	7	155	101
GTC→ATC (V→I)	+3	15	12	8	4	21	127
GTA→ATA (V→I)	+3	13	10	6	3	14	154
GTT→GCT (V→A)	0	16	117	75	5	54	189

a. High TMC rank represents a frequently observed mutation class in a database.
b. Stop mutations do not have BLOSUM62 values but highly impact the encoded protein.

4-3-2 Analysis of synonymous variants

Table 4.2 shows that synonymous TMC matrices from all databases correlate very strongly, as expected if silent substitutions were mostly neutral. Nearly all confidence intervals shown in table 4.2 overlap. This underscores that it is then the nonsynonymous TMC frequencies that create the differences between databases, that is, the frequency differentials are due to the impact of the amino acid substitution. The most mutation-prone codons contain CpG dinucleotides, which are known to be hypermutable via deamination of methylated cytosine to yield thymine (Cooper and Krawczak 1990). The most observed

synonymous TMCs are those involving such a mutation: TC**G**→TC**A** (Ser), CC**G**→CC**A** (Pro), GC**G**→GC**A** (Ala), and AC**G**→AC**A** (Thr). These four classes alone represent 16.5% of all silent cSNPs observed in a mouse or human genome (if silent mutation occurs randomly, this value would be 2.9%). Again, it is clear that these groups of mutation represent predictable 'low-hanging fruit' of the naturally-occurring synonymous point mutation spectrum in genes. These trends are strong enough that I expect that if one were to indicate ten bases in any gene expected to undergo frequent silent mutation, most mutations would be found in a random cohort of individuals.

Table 4.2: Spearman correlation coefficients between human synonymous mutation spectra ($p<0.0001$)

M_{dbSNP}	M_{jSNP}	M_{TSC}	**Mutation spectrum**
R=0.71 (CI =0.65-0.76)	R=0.66 (CI =0.58-0.72)	R=0.65 (CI =0.58-0.71)	$M_{interspecific}$
	R=0.85 (CI =0.81-0.88)	R=0.86 (CI =0.83-0.89)	M_{dbSNP}
		R=0.81 (CI =0.76-0.85)	M_{jSNP}

CHAPTER FIVE

PREDICTION OF HUMAN GENE CODING REGION POINT MUTATION

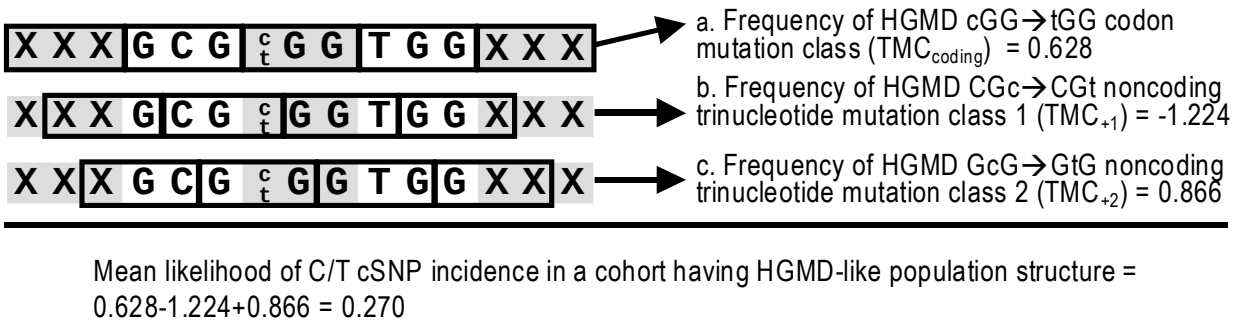
5-1 Elucidation of a method to utilize calculated point mutation trends for *de novo* prediction of gene point mutations

A highly desirable application of calculated whole-genome mutation spectra would be to use that information in a predictive manner. Hotspot prediction not only gives geneticists valuable targets for disease association/causation studies, but also could shed light onto how mutation propensity is distributed across the genome which may foretell gene evolution. Because of these reasons, it is critical that the predictive capacity of the spectra be tested.

Computational testing of predicted mutational spectra can be performed in two ways: A) taking a human gene sequence, making predictions, and searching for the predictions in available public point mutation databases and B) predicting a whole spectrum for a set of genes and comparing that to the entire point mutation makeup of whole databases.

A large range of mutation spectra has been calculated to facilitate this study meaning that for any given coding region sequence, a variety of point mutations (nonsynonymous, silent, or both) can be predicted as if they were found under a range of differing experimental conditions (e.g. in a low-throughput sequencing survey as for M_{dbSNP}). Once a mutation spectrum model (M) and the class of mutations to predict are decided (nonsynonymous or silent), the cDNA is examined to list each possible point mutation and their associated likelihood is assigned from the mutability scoring tables. Figure 5.1 illustrates this process for a GCG-**C**GG-TGG portion of a coding sequence where a C->T mutation occurs in the

CGG codon and is scored using the M_{HGMD} scoring tables, which do not contain any values for silent point mutations. The trinucleotide mutation class of the point variant was evaluated for each frame (one coding and two noncoding), and then assigned a log odds score by referencing the appropriate frame-specific scoring table. The three log likelihood scores are summed to obtain the total likelihood of observing the C->T variant given its DNA sequence context. Once every possible nonsynonymous variant was scored for an entire coding sequence, the entire body of point mutations was ranked by descending log likelihood scores. This process was employed to derive M_{intron}–, M_{dbSNP}–, M_{HGMD}–, and $M_{\mathrm{interspecific}}$–like mutation likelihoods for sets of genes. Note that as an exception, the M_{intron} spectrum only has one scoring table because the SNPs from the training dataset do not occur in coding sequences. Therefore, when this spectrum is applied to a gene, each of the three TMCs corresponding to a particular mutation was summed from the single scoring table.



Mean likelihood of C/T cSNP incidence in a cohort having HGMD-like population structure = 0.628-1.224+0.866 = 0.270

Fig. 5.1. Demonstration of the sequence context-dependent cSNP prediction method. A hypothetical gene sequence undergoes HGMD-like mutation. To predict a C/T cSNP (lower case) based on point mutation trends calculated from the HGMD database, the putative variant is first classified by each of the three possible trinucleotide contexts. Outlined in black are the boundaries of each trinucleotide for the a) protein coding frame and noncoding frames upon a b) +1 and c) +2 nucleotide frameshift of the gene sequence. For reference, the true protein coding frame is represented by the shaded boxes in all three sequence representations. The log odds score of the cSNP's codon mutation class is referenced in the HGMD TMC$_{coding}$ mutability table for (a) and the log odds score of the trinucleotide housing the cSNP according to the other two trinucleotide mutation classes (TMCs) (b,c) are referenced in the TMC$_{+1}$ and TMC$_{+2}$ tables. X's represent the remainder of the hypothetical sequence. The summed log odds score is the likelihood of observing this specific C/T cSNP in a population similarly stratified as those which compose the HGMD.

5-2 Computational validation of predicted point mutations in disease-relevant genes

The predominant question resulting from the TMC distribution analysis (chapter 4) is how well such spectra estimate the aggregate mutational trends of a single gene. This can be computationally assessed by taking a gene coding sequence, making a list of all missense, nonsense, and silent point mutations possible at each position in a transcript, and assigning a TMC frequency value to each mutation. This value acts as a probability estimate for

observing that mutation in a population stratification similar to the one that built the applied mutability table. If the mutation probability values were ranked in descending order, one could examine a top fraction and see if any of these predicted mutations have actually been discovered. The major challenge in using this historical approach to validate *de novo* mutation prediction is that many mutations per gene are required; therefore, the best dataset for this analysis is the HGMD. This is because SNP density is quite low within most point mutation databases, and the HGMD possess the highest SNP saturation per gene.

For a set of seven unrelated genes representing a broad spectrum of allele frequencies and inheritance modes, the HGMD TMC distribution (10/17/2002 build, table 3.1) is recalculated with those genes' mutations excluded from the training set. Predictions are referenced in the HGMD to determine the method's effectiveness in calling disease-causing mutation. To determine if mined cSNP trends encoded in trinucleotide mutation class frequencies have predictive power, gene sequences were selected and trinucleotide mutation class scores derived from the HGMD database (M_{HGMD}) were employed to estimate of the likelihood of that event at the wild-type codon. For a whole coding sequence, the entire body of possible nonsynonymous cSNPs are ranked in descending order by score and a slice of the top predictions for the gene (0.25%, 0.5%, 2.5%, and 5%) are referenced in the HGMD to determine the rate at which they are experimentally observed. Accuracy (the percentage of predictions detailed in the HGMD) and completeness (the percentage of all HGMD-detailed point mutations predicted by our methods) statistics are reported in table 5.1 for seven genes representing a range of inheritance modes. This validation technique suffers from the fact

that for no gene are all of the disease-relevant alleles known; therefore, the perceived accuracy will be a lower estimate only.

The accuracies in table 5.1, however, only have meaning in contrast to the accuracy rate of predicting mutations randomly. If a gene is saturated with mutations (i.e. nearly all existing germline mutations have been discovered), as would be the case for historically well-studied genes such a hemoglobin-β, random prediction accuracy would also be high therefore indicating that the empirical method presented here does not lend a mutation discovery improvement. To simulate prediction by this random, null model, all log likelihood scores for the body of HGMD nonsynonymous mutation predictions were scrambled and randomly assigned back to each prediction. These re-assigned mutations were ranked by descending probability and accuracy was computed at the thresholds in table 5.1. This procedure was repeated for 10,000 cycles to acquire an average prediction accuracy for the null model.

Table 5.1: A portion of disease-causing cSNPs can be accurately predicted[a]

GENE (# possible nonsyn mutations)	**Top 0.25%**	**Top 0.5%**	**Top 2.5%**	**Top 5%**	**Ratio[e]**
	Percent accuracy for cSNP prediction[b] ***Percent accuracy using the null (random) model[c]*** **Percent completeness[d]**				
F9[f] (3,275)	100% (8/8) *15.7* ±12.9% 1.6% (8/515)	93.8% (15/16) *15.9* ±9.1% 2.9% (15/515)	65.4% (53/81) *16.0* ±4.0% 14.2% (73/515)	61.6% (101/164) *16.2* ±2.8% 19.6% (101/515)	**6.3**
CFTR[g] (10,391)	38.5% (23/26) *5.4* ±4.4% 4.1% (23/565)	69.2% (36/52) *5.4* ±*3.1%* 6.4% (36/565)	40.0% (104/260) *5.4* ±2.2% 18.4% (104/565)	28.7% (149/519) *5.4* ±*1.8%* 26.4% (149/565)	**16.3**
GJB1[h] (1,942)	100.0% (5/5) *4.3* ±*9.1%* 5.9% (5/85)	60.0% (6/10) *4.3* ±*6.4%* 7.1% (6/85)	37.5% (18/48) *4.3* ±*2.9%* 21.2% (18/85)	24.7% (24/97) *4.3* ±2.0% 28.2% (24/85)	**23.3**
HMBS[i] (2,471)	100% (6/6) *3.3* ±*7.2%* 7.4% (6/81)	75.0% (12/16) *4.9* ±*5.3%* 7.6% (12/157)	45.6% (37/81) *4.9* ±*2.3%* 23.6% (37/157)	36.0% (58/161) *4.9* ±*1.7%* 36.9% (58/157)	**30.5**
PAX6[j] (2,925)	85.7% (6/7) *1.4* ±*4.5%* 14.6% (6/41)	57.1% (8/14) *1.4* ±*3.0%* 19.5% (8/41)	20.5% (15/73) *1.4* ±*1.4%* 36.6% (10/73)	11.0% (16/146) *1.4* ±*0.9%* 39.0% (16/41)	**61.7**
SERPINA1[k] (2,927)	71.4% (5/7) *0.9* ±*3.6%* 18.5% (5/27)	46.7% (7/15) *1.0* ±*2.5%* 25.9% (7/27)	13.7% (10/73) *0.9* ±*1.1%* 37.0% (10/27)	7.5% (11/146) *0.9* ±*0.8%* 40.7% (11/27)	**77.0**
XPA[l] (1,953)	60.0% (3/5) *0.3* ±*2.6%* 50.0% (3/6)	30.0% (3/6) *0.3* ±*2.2%* 50.0% (3/6)	6.1% (3/49) *0.3* ±*0.8%* 50.0% (3/6)	3.7% (3/97) *0.3* ±*0.6%* 50.0% (3/6)	**181.0**

a. The frequencies of HGMD TMCs were used to blindly predict nonsynonymous (nonsyn) coding region point mutations for each gene.

b. Percentage of predicted point mutations that have been experimentally observed according to the HGMD. (# correct predictions / # predictions made)*100%

c. Calculated as in (b) but using the null model, which predicts mutations randomly

d. Percentage of HGMD-detailed point mutations that were predicted for each gene using the M_{HGMD} point mutation spectrum. (# correct predictions / # known mutations)*100%

e. Ratio of (% accuracy of cSNP prediction)/(% accuracy of null model) for the top 0.25% mutation prediction threshold level.

f. Factor 9, g. cystic fibrosis transmembrane receptor, h. connexin 32, i. hydroxymethylbilane synthase, j. paired box homeotic 6, k. alpha-1-antitrypsin, l. xeroderma pigmentosum, complementation group A

Table 5.1 indicates all seven genes have an easily-predicted volume of mutational space despite the wide spectrum of allele inheritance modes represented, which illustrates the generality of point mutation trends. When considering only the top quarter percentile of ranked nonsynonymous substitutions to be potential disease-causing alleles, 56/64 predictions (87.5%) in this select fraction are already known to cause disease. Depending on a gene's cSNP saturation, this method is between 6.3 to 181-fold more accurate than predicting causative cSNPs using the null model of mutation, which assumes that all cSNPs are equally likely to occur and therefore predicts mutations randomly. There is an obvious correlation between prediction accuracy and the number of clinically identified alleles; therefore, I believe that many of the false positive mutations predicted in table 5.1 (i.e. they are not detailed in the HGMD) exist and cause disease but either have not yet been discovered or are annotated in a separate database. Although a tedious task given the nonuniformity of both central and locus-specific mutation databases (Claustres, Horaitis et al. 2002), matching predictions against any other known polymorphisms would increase accuracy to even higher rates than reported here. For example, manual searching of the *CFTR* Mutation Database (Bobadilla, Macek et al. 2002) reveals that three unconfirmed predictions at the 0.5% level in table 5.1 do exist: 31R→H, 170R→H, and 1453R→Q (HGMD base numbering used), but their relation to disease is uncertain. But in terms of predicting real mutations in CFTR, the accuracy rate increases to 39/52 predictions or 75.0%.

Thousands of new disease-relevant point mutations are added to the HGMD every few months. Often many of these new entries are recognized from the unverified prediction list of the older dataset. For each gene *AR, RPE65, DMD, CHM, MSH2, KEL, MCOLN1,*

and *ATP7A* two to three out of the top five most likely but uncataloged mutations (18/40 predictions) were described in the HGMD as causative of disease over the next 5 months (e.g. *KEL* R128→X (Lee, Russo et al. 2001), *RPE65* R91→W (Morimura, Fishman et al. 1998), and *ATP7A* R980→X (Gu, Kodama et al. 2001)). If the same test was performed another six months from now with a current list of unverified predictions, it stands to reason that a similar fraction will have been annotated in the HGMD. Based on these results, for a newly discovered, poorly characterized gene in the human genome one can immediately predict a handful of point variants that are both likely to exist and possibly even cause disease.

With candidate disease alleles in hand, a researcher can employ high-throughput genotyping technologies, such as oligonucleotide microarrays (Bell, Chaturvedi et al. 2002), mass spectroscopy (Buetow, Edmonson et al. 2001), or Pyrosequencing (Ahmadian, Lundeberg et al. 2000), where the identity of a single user-defined DNA base is queried (instead of a whole amplicon). These methods can determine thousands of genotypes a day if they have the foreknowledge of the exact base that is expected to be multiallelic. Targeted bases could be acquired from the prediction methods presented here to boost success of association studies, especially when the number of candidate genes is quite large. Once a few mutations are found researchers may then elect to sequence the entire gene. The advantage in this protocol is that candidate genes are first screened at the most likely places of variation to decrease the number of amplicons undergoing costly and lengthy resequencing. The TMC method therefore has the power to tell a researcher today what disease-causing alleles will be found tomorrow by guiding genotyping experiments.

The computational prediction validation underscores that since the mutation propensity of genes is based upon sequence context, the cumulative mutation likelihood for genes or gene fragments implies how codon bias is modulated in those objects so that the chance of and type of point mutation would be skewed towards that required by selection pressures, either towards a more stable or more hypermutable sequence. Using these methods a *de novo* mutation spectra estimated from empirical data can be used to approximate the relative mutability of genes.

5-3 Experimental search for predicted point mutations expected within candidate genes for dilated cardiomyopathy

A standard disease gene-association study was designed to look for point mutations that may be associated with dilated cardiomyopathy. The goal was to determine how many predicted point mutations could be found if only the exon predicted to be most mutable within a candidate gene underwent DNA sequencing. This portion of the dissertation study was done within the first full year of working with the Garner lab where I was given freedom to direct the study myself. Looking back on the project, there are many places where the hand of a novice graduate student are apparent in the design of the search strategy. Many decisions were simply quite naive. However, this time was invaluable to my development into an independent scientist because not only did I learn new laboratory techniques, but I gained the experience necessary to appreciate the difficulties underlying genotyping studies. Such insight is often lacking from bioinformaticists who spend most of their time glued to a monitor and somewhat just assume that experiments will ‘work.’ Although the data

generated did not provide quantitative insight into point mutation prediction efficacy, it was employed to test a new piece of software generated by other Garner lab members. SNPCEQer is a tool similar to PolyPhred (Nickerson, Tobe et al. 1997) that calls point mutations from sequence traces, but it is optimized for the Beckman CEQ chemistry whereas PolyPhred was benchmarked on ABI systems. It was determined that in certain scenarios SNPCEQer performs better than PolyPhred on Beckman traces (Flood, Tang et al. 2002).

5-3-1 Design of search strategy

Genomic DNA from a cohort of 128 patients who had either familial or idiopathic dilated cardiomyopathy were obtained from our collaborator, Dr. Ralph Shohet. All of these individuals died of heart failure. In lieu of a panel of normal individuals, 64 cancer cell line samples (all from different individuals and different cancer types) were obtained from Ryan Weil in collaboration with Dr. John Minna. Cancer cell lines are expected to be hypermutable given that cancer is often defined by both excessive cell growth and the inactivation of mutation repair mechanisms. At this point in time, a completely unaffected control population was not selected because I was searching for variants that I knew would be rare. Six candidate genes were selected based upon expert opinion from Dr. Shohet and Dr. Edward Rame, as well as literature research. For each candidate gene, LocusLink ids were found and the corresponding RefSeq mRNA was pulled from GenBank (Pruitt and Maglott 2001). At the time of this study, only M_{HGMD} was available to generate predictions. The amplicon with the highest summed mutability log odds score was chosen for subsequent PCR amplification and DNA sequencing. Once variants were found, they were to be

analyzed with respect to the protein product for projected impact. It was my intent to sequence normal individuals (e.g. without any disorders that were known) once a series of point mutations was found at an appreciable frequency within the dilated cardiomyopathy population, or at least a population subtype. For those genes, it was also intended to sequence the rest of the exons was well as intron/exon boundaries in case any closely linked variants existed.

5-3-2 Candidate gene selection

Dilated cardiomyopathy (DCM) is a complex disorder characterized by left ventricle enlargement with concomitant weakening of the heart muscle resulting in inefficient blood pumping and often heart failure. It is estimated that 1 out of 2500 people are afflicted, and it is a chronic disease yet without a cure (Ku, Feiger et al. 2003). Most cases of dilated cardiomyopathy are idiopathic, although genetic and infectious components have been identified. Six candidate genes were chosen based on literature searches and consultation with Dr. Ralph Shohet here at UT Southwestern (table 5.2). Overall the most mutable amplicon (typically one exon per gene) was determined for these six genes by scoring each coding region DNA sequence using the M_{HGMD} nonsynonymous mutability tables. At this point in time, M_{HGMD} was the most robust mutation spectrum available from which to make point mutation predictions.

Table 5.2: Candidate genes for dilated cardiomyopathy examined in this study

Gene (symbol)	GenBank accession, base ranges amplified
myocyte-enriched calcineurin-interacting protein (MCIP1)	XM_048637, 109773-110435
bradykinin B2 receptor (BDKRB2)	S45489, 342-941
endothelin receptor type A (EDNRA)	D11145, 79-525
cAMP responsive element binding protein 1 (CREB1)	AC009298, 133373-133867
runt-related transcription factor 2 (RUNX2)	NM_004348, 1-371
beta-1-adrenergic receptor (ADRB1)	AL355543, 36117-36750
beta-2-adrenergic receptor (ADRB2)	Y00106, 1287-1868

MCIP1 (myocyte-enriched calcineurin-interacting protein 1) is known to use a negative feedback circuit to inhibit calcineurin, a calcium/calmodulin-regulated protein phosphatase expressed in cardiac and skeletal cycles during periods of environmental stress (Yang, Rothermel et al. 2000). *MCIP1* protein has been shown to inhibit DCM in mice when its expression is increased, suggesting that any mutation that lowers *MCIP1* dosage through either down regulation or alteration of protein function would enhance the likelihood of DCM (Rothermel, McKinsey et al. 2001).

The bradykinin beta 2 transmembrane receptor (*BDKRB2*) binds the bradykinin peptide and associates with G proteins that stimulate a phosphatidylinositol-calcium second

messenger system . Knockout mice develop dilated cardiomyopathy as they age as evidenced by hypertension, left ventricle remodeling, and heart function impairment. Mice heterozygous for the knockout develop the symptoms more slowly, suggesting a direct role between *BDKRB2* function and structural preservation of the heart (Emanueli, Maestri et al. 1999).

Endothelin receptor type A (*EDNRA*) is a G-protein coupled receptor that mediates vasoconstrictor actions, and was a target specifically requested by Dr. Ralph Shohet. The endothelin system is known to play a role in the pathophysiology of idiopathic dilated cardiomyopathy, and a silent polymorphism within EDNRA has been associated with decreased survival rates of individuals with dilated cardiomyopathy-like cardiac phenotypes (Herrmann, Schmidt-Petersen et al. 2001).

CREB1 (cAMP responsive element binding protein 1) has been implicated as a member of a transcriptional pathway that regulates cardiac myocyte function, which makes it a good candidate for dilated cardiomyopathy because myocyte proliferation is associated with DCM (Beltrami, Di Loreto et al. 1997). Additionally, a dominant-negative *CREB1* mutation (Ser133Ala) expressed in transgenic mice resulted in the development of DCM that closely resembled the human form (Fentzke, Korcarz et al. 1998).

RUNX2 (runt-related transcription factor 2) is a favorite of Dr. John Fondon and has been studied within the Harold Garner lab for its relationship to craniofacial development of dogs. It is known to be highly expressed in many tissues, including the heart, and increased *RUNX2* expression has been found within heart valves of individuals with calcific aortic stenosis, which is the third most common cardiovascular disease in the United States

(Rajamannan, Subramaniam et al. 2003). Given that *RUNX2* has this unclear, yet real, association with such a common cardiac disorder and was of interest to other members of the Harold Garner lab, this gene was added to my candidate gene list for sequencing.

Beta-1- and beta-2-adrenergic receptors (*ADRB1* and *ADRB2*, respectively) have both been extensively studied with respect to various aspects of cardiac disease and were chosen by request of Dr. Ralph Shohet. Interestingly, *CREB1* mRNA levels are down-regulated following chronic beta-adrenergic stimulation in rats, and this sort of stimulation is thought to be a major component of progressive human heart failure (Fentzke, Korcarz et al. 1998).

5-3-3 Experimental methods

Gene regions of interest were PCR amplified from genomic DNA samples (obtained from Dr. Ralph Shohet) derived from individuals diagnosed with various forms of dilated cardiomyopathy. Primers were designed with Primer3 (Rozen and Skaletsky 2000). Table 5.3 details PCR primers, sequencing primers, and PCR conditions. PCR reactions were set up using the Epicentre FailSafe™ PCR system (Epicentre, WI, USA) with 0.625 U enzyme, approximately 100ng template DNA, and 1 μM each primer (Operon, CA, USA). Thermal cycling was conducted in a MJ Research PTC100 thermocycler (MJ Research, MA, USA). Reactions were subjected to 96°C for 3 minutes; 40 cycles of 96°C for 45 seconds, annealing temperature for 45 seconds, 72°C for 2.5 minutes; and 72°C for 4 minutes. PCR products were purified using Qiagen's PCR Purification Kit (Qiagen, CA, USA). PCR product concentration was estimated from electrophoresis of sample against GeneChoice Ladder II (PGC Scientifics, MD, USA). Cycle sequencing reactions (both forward and reverse reads)

employed Beckman Coulter CEQ DTCS kits (Beckman-Coulter, CA, USA). 1.25μM sequencing primer, ~100 ng template DNA, and 50 fmol PCR product were used. Sequencing reactions were subjected to 96°C, 1 minute; 41 cycles of 96°C for 40 seconds, annealing temperature for 20 seconds, 61°C for 4 minutes; and 61°C for 4 minutes. Sequencing was conducted on Beckman CEQ™2000 sequencers using the Long Fast Read method: Injection at 2.0 kV for 10 seconds, separation at 50°C, 4.2 kv for 75 minutes.

The sequence traces were assembled, called, and visualized using Phred/Phrap/Consed (Ewing and Green 1998; Ewing, Hillier et al. 1998; Gordon, Abajian et al. 1998) and SNPs were called with the associated PolyPhred package using default parameters. Chromatograms containing putative SNPs were visually examined to confirm the SNP. A SNP was confirmed when it appeared clearly in both directions and had appropriate height for a diploid sample.

Table 5.3: Candidate gene PCR Conditions

Gene	PCR primers (forward/reverse)	PCR reaction, annealing temp., premix used	Sequencing primer	Sequencing annealing temp.
EDNRA	AATTTGCCTCAAGATGGAAACC TGGTTACTTCCTACCTTAAATACATTG	Epicentre, 60.0 Premix D	CTCAAGATGGAAACCCTTTGC	60.0
BDKRB2	AGAGATCTACCTGGGGAACCTG AGGAAGGTGCTGATCTGGAAG	Epicentre, 61.0 Premix K	CGCAGCAGACCTGATCCTG	61.0
RUNX2	ATGCGTATTCCTGTAGATCCGAG ACGGGCAGGGTCTTGTTGC	Epicentre, 64.0 Premix J	ATGCGTATTCCTGTAGATCCGAG	61.0
CREB1	GCCAAAGAGTGTCTGGGATG GGTACTTCAGTCCTACAGTTTCTCC	Epicentre, 59.5 Premix G	GGTACTTCAGTCCTACAGTTTCTCC	59.5
ADRB1	GAGCCCCGGTAACCTGTCG CCGGTTGGTGACGAAGTC	Epicentre, 60.5 Premix J	CCGGTTGGTGACGAAGTC	60.5
ADRB2	CCTTCTTGCCCATTCAGATG TGCCGTTGCTGGAGTAGC	Epicentre, 59.3 Premix G	TGCCGTTGCTGGAGTAGC	59.3
MCIP1	ACAGTCCCAAATGTCCTTGTG CTGGACTCCCAGAATGTTTG	Epicentre, 59.5 Premix L	ACAGTCCCAAATGTCCTTGTG	59.5

5-3-4 Genotyping results

The SNPs are reported in Table 5.4. No genuine SNPs were found in gene *CREB1*. 19/21 variants had not been detailed in dbSNP before this study. Since five of these were found solely in cancer cell lines, it is possible that these are somatic and do not occur naturally in the population at the calculated frequency. None of the 21 discovered variants listed in table 5.4 showed departure from Hardy Weinberg equilibrium within the total population or any cohort subgroups. Most variants were too rare within the population to be able to associate them to a specific group of individuals, and none of the other alleles appeared to segregate according to any phenotype.

In Table 5.4, each mutation can be described by three separate trinucleotide mutation classes (TMCs), one for each frame. As a group this describes the point mutation relative to its nucleotide context within a 5-base DNA segment. Interestingly, 5/21 (23.8 %) of the point mutations occur with at least one frame having a log odds score greater than zero according to the M_{intron} mutation spectrum. Additionally, nearly all mutations occur within the M_{intron} mutability table at a high rank for at least one of the TMCs. For example, for *BDKRB2* mutation GGG-->GGA the rank of the trinucleotide class GGG-->GGA is 93. However, when examining the other two mutation classes possible for that mutation given the noncoding frame nucleotide contexts (not shown in Table 5), one noncoding frame TMC has a rank of 47. Most of the mutations found can be rationalized by having a rank that maps to the top 10% most mutable contexts as judged by comparison to the M_{intron} mutation spectrum. In fact, if one takes the most mutable (of three possible) TMC for each discovered point mutation, the median rank is 59. Naively, one may expect that the ranks will correlate

with the minor allele frequencies of each mutation given that rank expresses the mutation likelihood of that SNP. However, population effects, not to mention selection, greatly impact the frequency of the alleles. Most common mutations in the human population are common because of population dynamics such as selective sweeps or bottlenecks, not because the mutation rate is high. A bottlenecking of a population due to circumstances such as inbreeding can elevate even neutral or deleterious mutations to a rate much higher than would be expected based on their intrinsic mutation likelihoods. The most common discovered mutation in my study, *ADRB1* AGC-->GGC, has minor allele frequency of 15% yet an average rank of 128 across all three TMCs.

Table 5.4: SNPs found in dilated cardiomyopathy candidate genes

Gene	GenBank gi# (or accession), base position	Trinucleotide mutation (amino acid change)	Genotypes	Allele freq	Rank in M_{intron} for each TMC	Phenotype
RUNX2	10863884 215	GCG-->GCA (Ala-->Ala)	102 GG 6 GA 0 AA 74 undet.	p = 0.9722 q = 0.0277	6, 12, 45	4 DCM 2 cancer
ETAR	NM_001957 838	CTG-->CTA (Leu-->Leu)	180 GG 0 GA 2 AA 10 undet.	p = 0.9890 q= 0.04109	123, 196, 77	cancer
BDKRB2	4557358, 378	GGG-->GGA (Gly-->Gly)	184 GG 1 GA 1 AA 6 undet.	p = 0.9919 q = 0.0080	93, 133, 47	cancer
	4557358 383	ACG-->ATG (Thr-->Met)	184 CC 1 CT 0 TT 6 undet.	p = 0.9972 q = 0.0027	2, 9, 28	DCM
	4557358 565	ACG-->ACA (Thr-->Thr)	174 GG 1 GA 0 AA 17 undet.	p = 0.9971 q = 0.0028	8, 13, 66	DCM
	4557358 568	CTG-->CTA (Leu-->Leu)	184 GG 1 GA 0 AA 6 undet.	p = 0.9972 q = 0.0027	123, 151, 66	DCM
	4557358 626	ACC-->AGC (Thr-->Ser)	184 CC 1 GC 0 GG 6 undet.	p = 0.9972 q = 0.0027	342, 354, 253	cancer
	4557358 792	ACG-->ACA* (Thr-->Thr)	155 GG 20 GA 4 AA 13 undet.	p = 0.9217 q = 0.0782	8, 12, 47	cancer
MCIP1	14780300 696	gCg-->gTg (intronic)	183 CC 2 CT 0 TT 6 undet.	p = 0.9945 q = 0.0054	14, 10, 61	DCM
	14780300 685	gCg-->gTg (intronic)	167 CC 10 CT 2 TT 13 undet.	p = 0.9608 q = 0.0391	14, 35, 10	DCM
	14780300 649	CCG-->CCA (Pro-->Pro)	177 GG 6 GA 2 AA 6 undet.	p = 0.9729 q = 0.0270	3, 12, 47	7 DCM 1 cancer

Table 5.4 continued... SNPs found in dilated cardiomyopathy candidate genes, continued

Gene	GenBank gi# or accession, base position	Trinucleotide mutation (amino acid change)	Genotypes	Allele freq	Rank in M_{intron} (out of 576)	Phenotype
MCIP1	14780300 645	AGG-->AAG (Arg-->Lys)	182 GG 1 GA 0 AA 9 undet.	p = 0.9972 q = 0.0027	37, 45, 84	DCM
ADRB2	NM_000024 1109	GTG-->GCG (Val-->Ala)	175 TT 7 CT 0 CC 10 undet.	p = 0.9807 q = 0.0192	99, 140, 170	6 DCM 1 cancer
	NM_000024 1075	TGG-->CGG (Trp-->Arg)	181 TT 1 CT 0 CC 10 undet.	p = 0.9972 q = 0.0027	81, 113, 169	DCM
	NM_000024 1058	GGC-->GGT (Gly-->Gly)	181 CC 1 CT 0 TT 10 undet.	p = 0.9972 q = 0.0027	55, 168, 131	DCM
ADRB1	NM_000684 231	AGC-->GGC* (Ser-->Gly)	134 AA 38 GA 8 GG 12 undet.	p = 0.8500 q = 0.1500	161, 179, 44	37 DCM 9 cancer
	NM_000684 626	TGC-->TGT (Cys-->Cys)	170 CC 5 CT 0 TT 17 undet.	p = 0.9857 q = 0.0142	61, 168, 131	3 DCM 2 cancer
	NM_000684 312	AAT-->GAT (Asn-->Asp)	179 TT 1 CT 0 CC 12 undet.	p = 0.9972 q = 0.0027	222, 150, 54	DCM
	NM_000684 293	GTG-->GTA (Val-->Val)	178 GG 2 AG 0 AA 12 undet.	p = 0.9944 q = 0.0055	122, 151, 66	DCM
	NM_000684 323	GTG-->GTA (Val-->Val)	179 GG 0 GA 2 AA 11 undet.	p = 0.9889 q = 0.0110	122, 196, 77	cancer
	NM_000684 490	ACC-->GCC (Thr-->Ala)	180 AA 1 GA 0 GG 12 undet.	p = 0.9972 q = 0.0027	78, 132, 163	DCM

* = SNPs discovered that had been previously listed in dbSNP

5-3-4 Search strategy redesign: A post-mortem

At the closure of my graduate tenure, various ways to have improved this study are now apparent. Firstly, the choice of population panel was flawed in that too many forms of dilated cardiomyopathy (familial, idiopathic, ischemic, or undetermined subtypes) were sequenced. Only a specific subgroup should have been examined given that the causes underlying dilated cardiomyopathy are multifarious. Another area for improvement is the choice of mutability spectrum used to predict point mutations. When this study began, only M_{HGMD} was available to score DNA bases' mutational load. Currently I have M_{intron} at my disposal which is not biased against conservative mutations. In fact, this spectrum is expected to not be biased by the expected selective impact of the predicted mutation. Expanding on this, given that there are a large variety of candidate genes for dilated cardiomyopathy, if the study were to be redone I would want to develop two gene lists: One where the genes have quite large summary mutational load values with respect to the rest of the human genome and a second where the summary mutational load values are small. The comparison of discovered mutations within each of the gene lists would help determine whether choosing resequencing amplicons by applying global mutation spectra increases the efficiency of SNP discovery. The use of cancer cell lines as a control population is not ideal because discovered mutations could quite easily be somatic, meaning additional sequencing of undifferentiated cells would be needed. Additionally, it would be quite beneficial to add a few chimpanzees to the panel so that the ancestral allele could be determined with certainty. In this study the ancestral allele was taken as the base in the mRNA reference sequence from

RefSeq. However, this base is often the most frequent allele within the global human population and does not necessarily reflect the ancestral state given that it is known that the human population went through numerous bottlenecks (Marth, Schuler et al. 2003).

5-4 High-throughput experimental search for predicted point mutations in conjunction with the UT Southwestern PGA Project

5-4-1 Design of search strategy in partnership with the UT Southwestern Program in Genomic Applications (PGA) project and the UT Southwestern Reynolds Foundation

Over my graduate tenure another opportunity presented itself to test the predictive power of my point mutation mutability tables within an experimental setting. This came about through a Programs in Genomics Application project grant awarded to UTSW, where Dr. Garner was the a coinvestigator of one phase of the study and novel computational methods testing was chosen as one of his aims. The overall goal of the UTSW PGA was to discover mutations that may cause heart disease (particularly in an inflammation capacity) and to make any resources created over the course of the project available for public use. The UTSW PGA grant was split into four distinct projects, each run by a different investigator. In Project 1 genes whose alterations may be associated with heart disease were identified by both expert opinion and microarray-based gene expression analysis in mouse models of heart disease to point to new candidate genes. Additionally I was tasked with predicting point mutations within these candidate genes using my mutation spectra methods. In Project 2, a high throughput system to screen these genes for new SNPs was implemented,

with emphasis on mutations that would alter the protein product and potentially cause disease. My predicted point mutations for candidate genes were tested in this phase of the project. Next, association studies were designed to assess the clinical significance of these polymorphisms relative to heart disease (Project 3). Finally, Project 4 developed a high throughput system for creating antibodies to the protein encoded by these candidate genes. To facilitate Project 2, a DNA panel of 3,554 individuals from the Dallas area was available in collaboration with the Donald W. Reynolds Center for Cardiovascular Research and the Dallas Heart Study. The Dallas Heart Study has recruited over 6,000 individuals from Dallas County and measures a large range of heart disease-relevant phenotypes such as fasting lipoprotein levels, cholesterol statistics, and blood pressure. This then provides a basis for a future option of taking the discovery of interesting alleles within the Reynolds population to the next step—clinical verification.

Cardiac disease is a textbook multifactorial disorder as well as being one whose study greatly serves the public interest, for it is estimated that 64,400,000 individuals in the United States are afflicted with one or more aspects of cardiovascular disease, and it causes 38.5% of all deaths, making it the leading killer in the country (Association 2003). It is estimated to have a maximum heritability of 34% in whites and 53% in blacks, and a correlation of cardiac disease incidence has been found between spouses due to environmental factors such as smoking, obesity, and physical inactivity (Katzmarzyk, Perusse et al. 2000). Traditionally, pinpointing genetic lesions as relevant to aspects of heart disease has been a thorny challenge using conventional mutation discovery methods. This disorder is known to have a large, diverse environmental component, a wide range in heritability, numerous contributing genes,

and disagreement in the cardiology community about how to diagnose or even define heart disease (Desai and Jessup 2004).

Therefore, my novel allele search strategy was to choose candidate genes for the variety of cardiac disease phenotypes measured within the Reynolds population and make point mutation predictions for those genes utilizing a range of mutability scoring tables discussed in chapter 3. The Sequenom mass spectroscopy technology was purported to permit fast and accurate genotyping of the specific DNA bases requested, so they were contracted to do the genotyping. Twenty genes were chosen (table 5.5) based on literature research and counsel from PGA investigators. For each gene coding region, three distinct sets of nonsynonymous point mutation predictions were made as scored from the mutability tables from three separate spectra: M_{HGMD}, M_{TSC}, and M_{dbSNP}. At this point in time, these spectra were the most robust mutability tables calculated, for the M_{intron} and $M_{\mathrm{interspecific}}$ spectra were not generated until late 2003, which prohibited placing their representative predictions into the Sequenom pipeline. For each of these genes, the top prediction per mutability method (M_{HGMD}, M_{TSC}, or M_{dbSNP}) was flagged for genotyping in the full Dallas Heart Disease population. Preparation of these sixty mutations (table 5.6) required basic bioinformatics work where 100 genomic bases on either side of the target mutation site was pulled from GenBank and submitted to Sequenom.

Table 5.5: PGA genes chosen for mutation prediction experimental verification

Gene	**LocusLink Id**	**Name**
APOA1	335	apolipoprotein A-I
APOC1	341	apolipoprotein C-I
CD36	948	CD36 antigen
ENG	2022	endoglin
EPHB2	2048	ephrin receptor beta 2
FHL1	2273	four and a half LIM domains 1
GPRK2L	2868	G protein-coupled receptor kinase 4
HBB	3043	hemoglobin beta
HNF3A	3169	forkhead box A1
ITGB3	3690	integrin, beta 3
JMJ	3720	Jumonji, AT rich interactive domain 2
LCAT	3931	lecithin-cholesterol acyltransferase
LDHB	3945	lactate dehydrogenase B
LIPA	3988	lipase A, lysosomal acid, cholesterol esterase
MTP	4547	microsomal triglyceride transfer protein
NRP1	8829	neuropilin 1
PLA2G2A	5320	phospholipase A2, group IIA (platelets, synovial fluid)
RGS5	8490	regulator of G-protein signaling 5
UCP3	7352	uncoupling protein 3
VIM	7341	vimentin

It was possible that over the course of this genotyping survey that no validated predictions would be found. Given that twenty genes of the total gene set have known alleles causing heart disease, this is an unlikely extremely scenario for it would suggest that these published mutations have been characterized incorrectly. However, it is possible that given the multivariate nature of cardiac disease, a mutation may be associated with only one aspect of the disorder. The Reynolds population may be further stratified by the phenotypes originally used to choose members such as cholesterol levels, blood pressure, weight, left ventricular mass and geometry, measured subcutaneous and abdominal visceral fat, or

coronary calcification. The hope was that enthusiastic cooperation would be maintained from other members of the UTSW PGA project as to permit this type of analysis.

Because some genes already known to have association with cardiac disease were being examined, it was useful to make point mutation predictions for a few cardiac-related genes with known alleles to get a lower-bound estimate of this experiment's predictive power relative to random. Table 5.7 shows such results for four genes with an established relationship to heart disease. The M_{HGMD} spectrum was used to predict nonsynonymous point mutations for these genes, and the top 5% of these predictions were taken as the set to test for prediction accuracy. Nonsynonymous point mutation predictions were also made randomly as the null experiment. Given that I have a "total number of known mutations" statistic for each of these genes as taken from the HGMD, taking the product of the %accurate and %complete statistics would create a new value that describes each method's ability to a) predict accurately and b) find all known causative mutations. Initially, this may seem redundant, but it is possible (and sometimes the case) that the random method has better completeness statistics simply because it makes a larger number of predictions. For example, if one were to make a mutation prediction at every DNA base in a gene, it would have a completeness rate of 100% but the accuracy would be quite poor. Table 5.7 gives the results for four genes examined in this way. The statistics for "at random" predictions were generated for ten trials from randomized mutability tables. Any predicted polymorphic position that is known to be cardiac disease-associated will be scored as correct. The accuracy rates are remarkable given that these genes have not been sequenced in large populations. Most striking is the "ratio" column (ratio of HGMD-based

%complete*%accurate statistic to “at random” %complete*%accurate statistic) which shows that matrices predict mutations on average 21-fold better than random for these four genes.

Table 5.6: Computational benchmarking suggests that cSNP prediction is an order of magnitude more efficient at finding causative cSNPs than predicting mutations randomly

Gene	Total known HGMD alleles	Scoring matrix cSNP predictions			Random model predictions			Ratio[e]
		%accuracy[a]	%complete[c]	Product[d]	%accuracy[b]	%complete[c]	Product[d]	
MYH7	69	3.1 (7/228)	60.7	188.2	0.48	17.7	8.5	**22.1**
TNNT2	9	12.5 (5/40)	55.6	695.0	0.76	34.6	26.3	**26.4**
SCN5A	8	1.6 (5/308)	62.5	100.0	0.15	27.3	4.1	**24.3**
KCNQ1	44	12.3 (17/138)	38.6	474.9	2.16	16.9	36.5	**13.1**

MYH7=myosin heavy polypeptide 7; *TNNT2*=cardiac troponin T2; *SCN5A*=sodium channel voltage gated protein type V, alpha unit; *KCNQ1*=potassium voltage gated channel KQT-like subfamily member 1.

(a)Percentage of predicted point mutations that have been experimentally observed and are disease-causing according to the HGMD (May 2001 release); (# correct predictions / # predictions made)*100%.

(b)Calculated as in (a) but using the null model, which predicts mutations randomly across a gene.

(c)Percentage of known HGMD-detailed point mutations that were actually predicted (# correct predictions / # known mutations)*100%.

(d)Product of accuracy and completeness statistics.

(e)Ratio of (cSNP prediction product)/(null model prediction product).

5-4-2 Sequenom's MassArray technology

Matrix assisted laser desorption ionization time-of-flight mass spectroscopy (MALDI-TOF MS) has recently become a hot method to perform point mutation genotyping (Jackson, Scholl et al. 2000). The generalized experiment is shown in figure 5.2. While some methods may directly analyze the source genomic DNA, the more robust techniques detect SNPs as the product of allele discrimination reactions. This procedure calls for the amplification of a piece of the queried genomic DNA housing the SNP and then manipulation of the product to reduce mass fragment size during analysis. This is critical for MS analysis because both the signal intensity and mass resolution of the experiment decrease with increasing DNA size due to inadvertent fragmentation of the DNA phosphodiester backbone during the ionization process. Consequently all MS SNP detection methods must be able to employ an assay to decrease DNA fragment size. This is the basic principle upon which all MALDI-TOF genotyping methods vary. The MS experiment also is highly reproducible and sensitive, for it is able to reliably distinguish between the most subtle phenotypes, A/T heterozygotes, which is difficult because the mass difference between the two nucleotides is small and the peaks are only half as strong as those in the homozygotes.

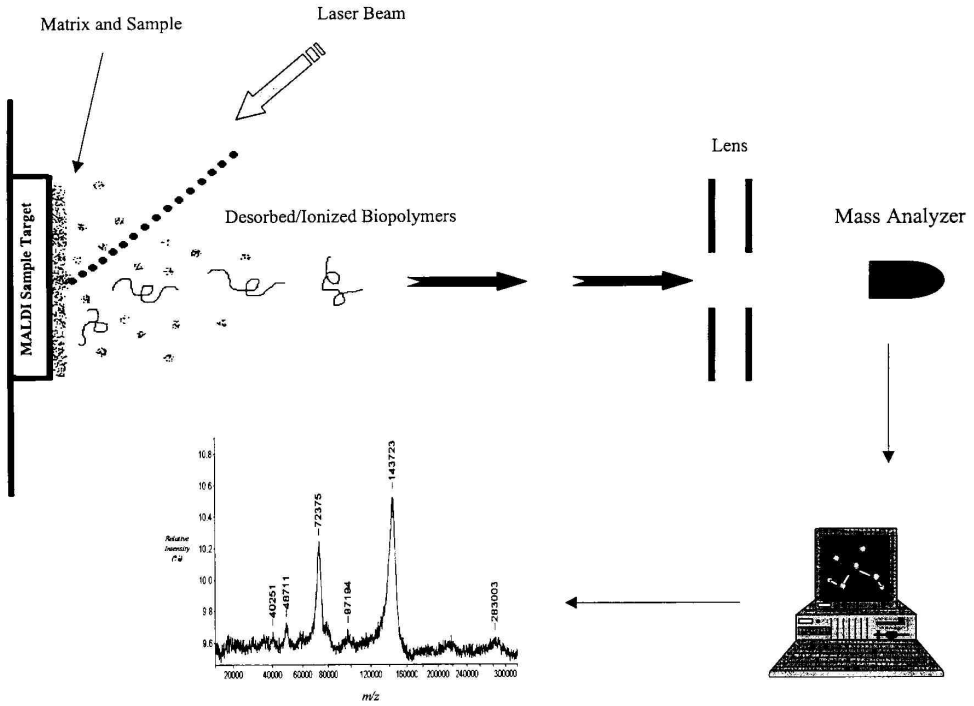


Figure 5.2. Schematic of a typical MALDI-TOF mass spectroscopy experiment (Guo 1999). A laser shoots ultraviolet photons at a target causing the desorption and ionization of biomolecules. These are focused through a vacuum tube under an electric potential so that their mass to charge ratio (m/z) may be calculated from their time-of-flight and deconvoluted into a mass spectrum.

The best system available for MS genotyping is the highly automated MassArray technology of Sequenom located in San Diego, CA (Buetow, Edmonson et al. 2001). There are three basic steps to genotype a single position in one individual: *1) Amplification of genomic DNA***.** Primers are designed to amplify the region queried with one amplification primer synthesized with a 5' biotin tag so that the PCR products may then be purified using streptavidin coated magnetic beads. *2) Single base extension reaction***.** A primer is annealed directly 3' of the queried position on the PCR product and added to an extension mixture of

ddNTPs, DNA polymerase, and Mg2+ (figure 5.2b). The extension reaction terminates as soon as one base is incorporated leaving products that may be discriminated based on size. *3) MS analysis.* Extension products (15-25mers) are denatured into ammonium hydroxide solution before piezoelectrically pipetting of 1-5 nl volumes onto a 96- or 384-element recyclable 2cm x 2cm silicon chip preloaded with 7 nl of crystalline matrix (typically 3-hydroxy-picolinic acid). The chip is then loaded into the MS instrument and SNPs are automatically called according to the precise mass of the unique primer extension products. A heterozygous individual at the queried position would show three peaks, one for each allele and one for the unextended primer (figure 5.3a).

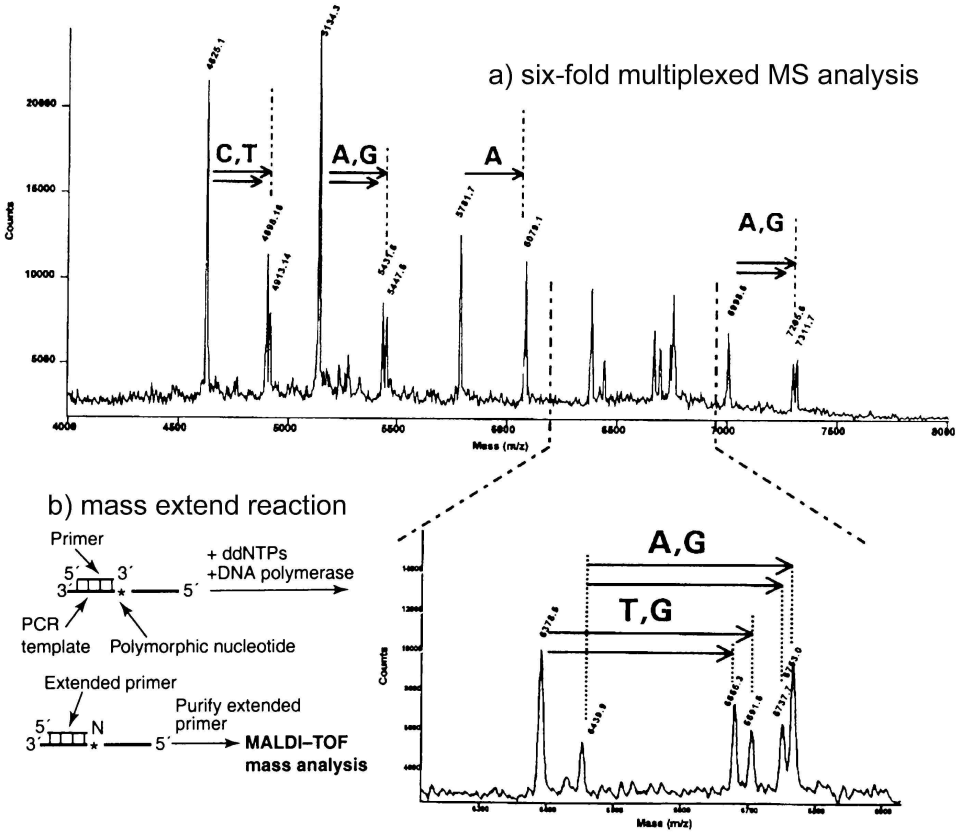


Figure 5.3. Detection of point mutations by MS analysis. a) Example of six-fold multiplexing of the extension reaction for MS genotyping. All possible alleles and unextended primer will have different resolvable masses allowing for the unambiguous discrimination of heterozygotes (5) and homozygotes (1). b) During the extension reaction, an extension primer will anneal directly 3' of the queried position on the template strand. A single ddNTP will be added and the masses of the different products can be used to determine the genotype of the individual (Ross, Hall et al. 1998; Griffin and Smith 2000).

The system is quite robust, for the average signal to noise ratio is 83:1. In a recent benchmarking of Sequenom's MassArray system, a detection limit of 0.2 fmol from a synthetic 36-mer was demonstrated (Tang, Fu et al. 1999). The system is more than sufficient to determine the mass difference between adenine and thymine. This entire

process may be multiplexed up to 7-fold given clever primer design and mass-tagging of the extension primers (Ross, Hall et al. 1998) so that each extension product corresponding to different queried SNP positions occupy completely separate mass windows during sample readout. Additionally, 10 384-element chips may be run in a single, unattended session permitting the analysis of up to 6,720 SNP genotyping events in a 1 hour period on one machine. Recently Sequenom announced the successful development of a new sample handling technology that performs the entire process from PCR amplification to MS analysis all on the silicon chip. This dramatically reduces costs by decreasing sample transfer and the volume of reagents required.

The drawback of MS genotyping is that despite its incredible throughput capability, it cannot be used alone for SNP discovery. This is because to create the MS assays, one must know the exact base in the queried DNA that is polymorphic. But given all possible polymorphic DNA bases are available as the human genome, the true challenge is to decide which of those bases are indeed polymorphic. Thus, there is a phenomenal need to identify promising polymorphic DNA bases to exploit technologies such as MALDI-TOF MS. Any identification method that performs better than randomly choosing bases from the human genome will allow effective discovery by MS. This challenge is unique to the post-genomic era, for one is provided with all the data from internet sources (the human DNA sequence) but must ask which portions of that data are useful i.e. mutable enough to cause disease.

5-4-3 Genotyping results

Genotyping was performed by "MALDI-on-a-Chip" mass spectroscopy analysis, as implemented by Sequenom's MassArray system (our collaborators in this work), whereby one can genotype 3,840 individuals at a single DNA position in a four hour period on one machine (Nelson, Marnellos et al. 2004). Using a population of 3,554 individuals, this experiment would be able to discover alleles as rare as 5/10,000 chromosomes 97% of the time. At the start of this study, it was understood and agreed that all verified point mutations, whether disease-associated or not, will be submitted to NCBI's dbSNP and presented to collaborators within the PGA for association analysis within the Reynolds collection of phenotypes. A full listing of all tested point mutations as well as the experimental verification information are shown in table 5.6.

Most of the alleles discovered in table 5.6 were reverified by in-house DNA sequencing in conjunction with the McDermott Center sequencing core. This was necessary because questions surrounding Sequenom's quality control processes arose from other PGA investigators over the course of the study. It was my preference to perform the DNA sequencing personally within the Garner lab as to gain more bench experience. However, the Reynolds Center has a policy of keeping all genomic DNAs internal given that samples have a limited supply and would be expended more quickly if they were constantly realiquoted to meet collaborators' requests. Therefore, resequencing was performed by the McDermott Center sequencing core facility at UT Southwestern by members of the Reynolds Center. Unfortunately, due to need to check all of Sequenom's SNP calls, funds for resequencing grew short, meaning that not every allele achieved reverification status which is indicated by

an asterisk in table 5.6. PGA funding was expected to be renewed for another four years beginning July 2004. Unfortunately, UT Southwestern was one of a number of institutions that did not win a PGA renewal. As a result, future work on this project remains at the discretion of Dr. Garner and our personal lab funds.

Table 5.7: Alleles investigated by mass spectroscopy genotyping

Gene	Amino acid mutation	codon change	Geno-types	# Sequenom miscalls	Mutation spectrum type	Minor allele freq.
Novel, rare missense mutations (*=alleles unverified by resequencing)						
APOC1	Arg54Cys*	CGC-->TGC	3409 CC 2 CT 143 undet	0	HGMD	0.0003
CD36	Arg5Trp	CGG-->TGG	3472 CC 1 CT 81 undet	0	HGMD	0.0001
ENG	Val483Ile	GTC-->ATC	3436 GG 1 GA 117 undet	0	dbSNP	0.0001
FHL1	Val121Ile	GTC-->ATC	3433 GG 1 GA 1 AA 119 undet	4	dbSNP	0.0004
FHL1	Arg199His	CGT-->CAT	3459 GG 1 AA 94 undet	4	TSC	0.0003
HNF3A	Thr55Met	ACG-->ATG	3377 CC 2 CT 175 undet	1	TSC	0.0003
HNF3A	Val468Ile*	GTC-->ATC	3424 GG 1 GA 129 undet	unknown	dbSNP	0.0001
ITGB3	Val520Ile	GTC-->ATC	3465 GG 1 GA 88 undet	1	dbSNP	0.0001
ITGB3	Arg750Stop	CGA-->TGA	3463 CC 3 CT 88 undet	1	HGMD	0.0004
JMJ	Thr343Met*	ACG-->ATG	3200 CC 182 CT 4 TT 168 undet	unknown	TSC	
LDHB	Val126Ile*	GTC-->ATC	3392 AA 1 AG 161 undet	unknown	TSC	0.0001
LDHB	Arg299Trp*	CGG-->TGG	3336 CC 3 CT 215 undet	unknown	HGMD	0.0004
MTP	Thr202Met*	ACG-->ATG	3344 CC 38 CT 4 TT 168 undet	unknown	dbSNP	0.0040
NRP1	Arg137Cys*	CGT-->TGT	3347 CC 21 CT 186 undet	unknown	TSC	0.0031

Table 5.7 Continued...

Gene	Amino acid mutation	codon change	Geno-types	# Sequenom miscalls	Mutation spectrum type	Minor allele freq.
Novel, rare missense mutations (*=alleles unverified by resequencing)						
PLA2G2A	Thr123Met*	ACG-->ATG	3240 CC 15 CT 299 undet	unknown	TSC	0.0023
UCP3	Arg95Cys*	CGT-->TGT	3345 CC 1 CT 205 undet	unknown	HGMD	0.0001
UCP3	Val292Ile*	GTA-->ATA	3238 GG 24 GA 292 undet	unknown	dbSNP	0.0068
Novel, common SNPs						
LIPA	Val206Ile*	GTC-->ATC	3122 GG 213 GA 40 AA 179 undet	unknown	dbSNP	0.043
EPHB2	Val268Ile*	GTC-->ATC	3056 GG 348 GA 150 undet	35	dbSNP	0.051
SNPs added as a control						
HBB	Gly70Ser*	GGT-->AGT	3425 GG 15 GA 114 undet	unknown	HGMD	0.0021
HBB	Val99Met	GTG-->ATG	3432 GG 31 AG 91 undet	unknown	TSC	0.0045
Monomorphic						
APOA1	Thr226Met	ACG-->ATG	n/a	2	TSC	n/a
CD36	Ser468Leu	TCG-->TTG	n/a	0	TSC	n/a
CD36	Thr92Met	ACG-->ATG	n/a	1	dbSNP	n/a
ENG	Arg205Trp	CGG-->TGG	n/a	0	HGMD	n/a
ENG	Ser50Leu	TCG-->TTG	n/a	0	TSC	n/a
EPHB2	Thr909Met*	ACG-->ATG	n/a	2	TSC	n/a
EPHB2	Arg353Stop	CGA-->TGA	n/a	1	HGMD	n/a
FHL1	Arg67Cys	CGC-->TGC	n/a	0	HGMD	n/a
GPRK2L	Arg312Stop	CGA-->TGA	n/a	0	HGMD	n/a
GPRK2L	Pro29Leu	CCG-->CTG	n/a	0	TSC	
HNF3A	Arg219Cys	CGC-->TGC	n/a	0	HGMD	n/a
ITGB3	Ser123Leu	TCG-->TTG	n/a		TSC	n/a
JMJ	Arg1176Stop	CGA-->TGA	n/a	0	HGMD	n/a
LCAT	Val49Ile	GTC-->ATC	n/a	0	dbSNP	n/a

Table 5.7 continued...

Gene	Amino acid mutation	codon change	Geno-types	# Sequenom miscalls	Mutation spectrum type	Minor allele freq.
Monomorphic						
LCAT	Arg123Cys	CGC-->TGC	n/a	0	HGMD	n/a
LCAT	Arg322Cys	CGT-->TGT	n/a	0	TSC	n/a
LDHB	Thr145met	ACG-->ATG	n/a	0	dbSNP	n/a
LIPA	Pro336Leu	CCG-->CTG	n/a	0	TSC	n/a
LIPA	Arg218Stop	CGA-->TGA	n/a	1	HGMD	n/a
MTP	Arg777Stop	CGA-->TGA	n/a	0	HGMD	n/a
MTP	Thr37Met	ACG-->ATG	n/a	16	TSC	n/a
NRP1	Arg405Stop	CGA-->TGA	n/a	1	HGMD	
PLA2G2A	Arg143Cys	CGT-->TGT	n/a	2	HGMD	n/a
RGS5	Thr123Met	ACG-->ATG	n/a	0	TSC	n/a
RGS5	Arg169Cys	CGC-->TGC	n/a	0	HGMD	n/a
VIM	Arg186Trp	CGG-->TGG	n/a	3	HGMD	n/a
Sequenom assay failed						
APOA1	Val245Met	GTG-->ATG	assay failed		dbSNP	n/a
APOA1	Arg239Cys	CGC-->TGC	assay failed		HGMD	n/a
APOC1	Val17Ile	GTC-->ATC	assay failed		dbSNP	n/a
APOC1	Ser15Leu	TCG-->TTG	assay failed		TSC	n/a
GPRK2L	Val140Ile	GTA-->ATA	assay failed		dbSNP	n/a
HBB	Val12Ile	GTT-->ATT	assay failed		dbSNP	n/a
HBB	Val12Ile	GTT-->ATT	assay failed		dbSNP	n/a
JMJ	Val894Ile	GTC-->ATC	assay failed		dbSNP	n/a
RGS5	Arg62His	CGT-->CAT	assay failed		dbSNP	n/a
NRP1	Val15Ile	GTC-->ATC	assay failed		dbSNP	n/a
PLA2G2A	Thr32Met	ACG-->ATG	assay failed		dbSNP	n/a
VIM	Val253Ile	GTC-->ATC	assay failed		dbSNP	n/a
VIM	Arg218Cys	CGT-->TGT	assay failed		TSC	n/a
UCP3	Thr52Met	ACG-->ATG	assay failed		TSC	n/a

Overall, 47 mutation predictions were successfully genotyped using Sequenom's Mass Array system. 21 of these were truly polymorphic, although extremely rare, and 26 were monomorphic. This represents a prediction success rate of 44.6%. This rate may be an overestimate, however, for not all Sequenom-validated mutations could be retested in house due to issues with obtaining genomic DNA from collaborators. Overall, the M_{dbSNP} model appears to be most accurate because 6/9 (66.6%) of the tested mutation predictions were indeed polymorphic. The M_{HGMD} and M_{TSC} models had accuracies of 27.8% (5/18) and 37.5% (6/16) respectively. With the exception of the hemoglobin variants, none of these mutations were known prior to this study. *ITGB3* Arg750X was not detailed in the HGMD before Sequenom testing, but has since been deposited in that database. According to a 1997 study it is strongly believed to cause a form of glanzmann thrombasthenia, a blood clotting disorder (Wang, Shattil et al. 1997). Contiguous mutations in the gene have been associated with a propensity for heart attacks. Mutations *CD36* Arg5Trp (CD36 antigen) and *MTP* Thr202Met (microsomal triglyceride transfer protein) are both predicted to be damaging to the protein product by both PolyPhen and SIFT mutation analysis codes, which are discussed in Chapter 2. *LIPA* Val206Ile is predicted to alter protein function according SIFT but not according to PolyPhen. These results cannot be conclusive about the utility of SNP prediction for causative allele discovery. Nine predicted mutations were called as polymorphic by Sequenom but could not be reverified by DNA resequencing. Future tests should be centered around the M_{intron} spectrum given that it is hypothesized to represent the true mutation rate in the absence of selection. The results, however, are provocative enough in my opinion to continue testing and potentially seek funding.

5-4-4 Implications for the multiequivalent risk model of complex disease

Analysis of SNP trends to infer locations of undiscovered mutations may present one solution to finding alleles contributing to multigenic diseases. To maximize chance of success in disease mapping, it is critical that experiments are competent to detect subtle genetic effects under a variety of genetic models. Most current disease association studies labor under confidence in the common disease/common variant (CDCV) model where experiments are sufficient to locate only frequent alleles. In such studies, nonsynonymous cSNPs, those most easily analyzed in terms of their contribution to a phenotype, are typically found at a rate of only 1 per 1000-1500 coding bases. Thus a mutation discovery approach biased against rare variants will miss precisely those alleles most likely to be functionally important. The CDCV-based approach lacks power to discover rare mutations that may exist, supportive of the multi-equivalent risk disease model, which holds that a large pool of rare alleles forms the genetic component of a complex phenotype where the cumulative allele frequency of the pool is large but each specific variant is rare. By using TMC matrices to genotype only those nucleotides having a sequence context historically prone to hypermutation, it now becomes practical to seek rare alleles supportive of the multi-equivalent risk model. Such an approach would not only test the appropriateness of this model for a complex disorder, but could be used as an invaluable screening strategy when a prohibitively large number of candidate genes exist.

CHAPTER SIX

CONTRASTING THE DIFFERENTIAL MUTATION PROPENSITIES UNDERLYING ENTIRE GENE CLASSES

6-1 Determination of individual gene mutability

The next and final phase of this study addresses specifically how distinct gene classes may be prone to or protected from extensive point mutation by virtue of their differing codon usages. My thesis is that a gene's propensity to mutate, as measured by the calculation of mutationally warm- and coolspots, alters the pool of alleles in a nonrandom fashion so as to either enhance or diminish the presence of radical amino acid substitutions. This is because the observation of a particular mutation in a population is shaped by two entangled forces: i) the natural mutational tendencies acting upon the genome that may predispose a spontaneous event to occur and ii) selective forces on this new variant. The ultimate goal of this study is to compare the effects of forces acting within separate point mutation datasets. This is accomplished by calculating their individual mutation spectra, which is then used to assign likelihoods of occurrence to all possible coding region point mutations in the genome. Each distinct mutation spectrum will impart a different set of likelihoods to the body of point mutations possible within a gene coding region. This body of likelihoods reflects which point mutation classes would be observed most often if a gene was subjected to the *same* selective forces as those acting upon the training SNP dataset, which was used to generate the likelihood scores.

6-1-1 Selection of mutation spectra for gene mutational load estimation

In chapter 3, table 3.2 details the various mutation spectra derived in this study. Given that such spectra can be used to rank the entire body of possible point mutations in a gene, this information can be thought of as describing that gene's total point mutation potential. Many of the spectra in table 3.2 are quite similar with only subtle cohort choice and genotyping methodology differences between them. Because the spectra are being used to predict gene mutability, the most complete versions with respect to genotyping methods should be chosen. That is, the spectra with the most number of training data SNPs are the best choices. Therefore, four mutation spectra are selected to predict four separate gene mutability values for all genes in the human genome: M_{intron}, $M_{\text{interspecific}}$, M_{dbSNP}, and M_{HGMD}. Figure 6.1 summarizes the characteristics of the databases creating these spectra, and the meaning to be gleaned if a gene scores highly for any of these metrics of mutability estimation. Since silent mutation does not alter the sequence of the encoded protein product of a gene, only the nonsynonymous mutation spectra was used to score gene sequences.

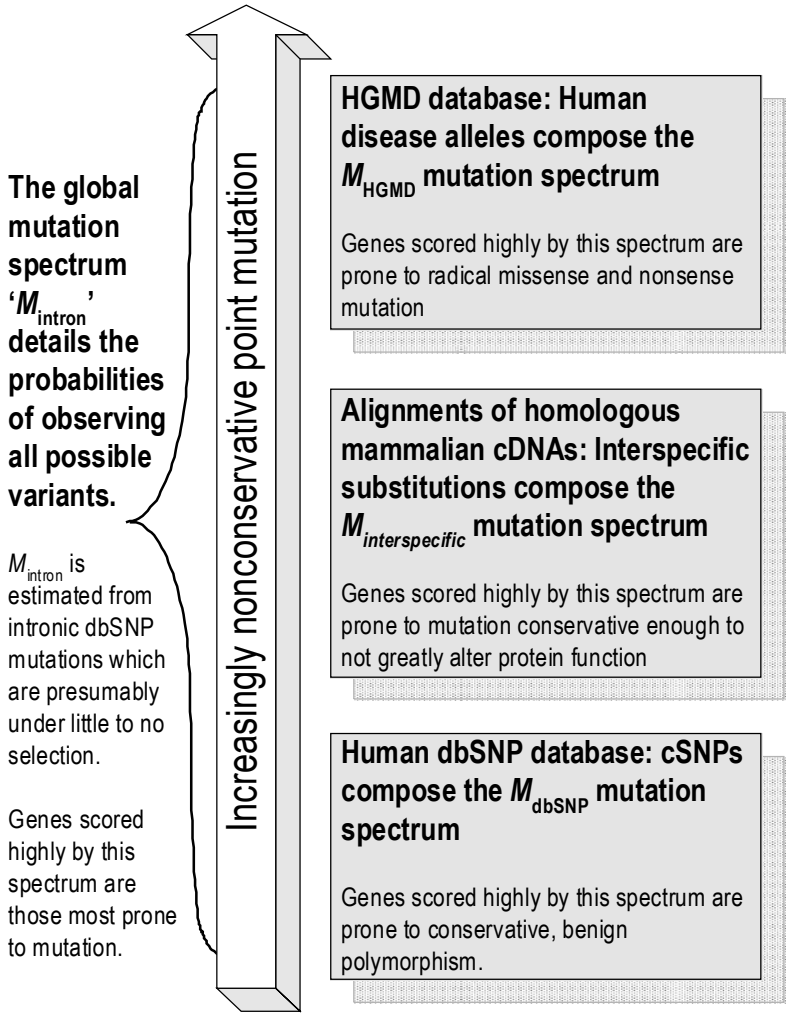


Figure 6.1. Summary of mutation spectra used to contrast gene mutability. M_{intron} is the mutation spectrum that should best represent the true mutational tendencies of DNA coding sequences, since the training dataset of intronic point mutations should reflect only the effects of mutation rate and not selection. The other metrics utilize gene coding region information and although they do reflect the same mutational trends as in M_{intron}, these spectra are biased by the effects of selection upon the DNA coding sequence.

6-1-2 Gene mutability calculation according to four distinct mutation spectra

The four mutation spectra are applied to a list of 12,865 human genes so that every point mutation that can occur in a gene is assigned a log odds score describing its likelihood of being observed. The gene list was obtained by retrieving each human protein encoding, non-pseudogene locus from LocusLink (Pruitt and Maglott 2001). The longest transcript corresponding to each locus from was then cross-referenced in RefSeq release 1 (Pruitt and Maglott 2001). In order to assign mutability values to each gene, the following process was performed four times per LocusLink gene, once for each different mutation spectrum. In a given RefSeq cDNA, all possible nonsynonymous mutations were compiled and assigned likelihoods. This process is illustrated in figure 6.1 for a ACG-**C**GA-TTA-ATG portion of a coding sequence where a C->T mutation occurs in the CGA codon and is scored using the M_{HGMD} scoring table. The trinucleotide mutation class of the point variant was evaluated for each frame (one coding and two noncoding), and then assigned a log odds score by referencing the appropriate frame-specific scoring table. The three log likelihood scores are summed to obtain the total likelihood of observing the C->T variant given its DNA sequence context.

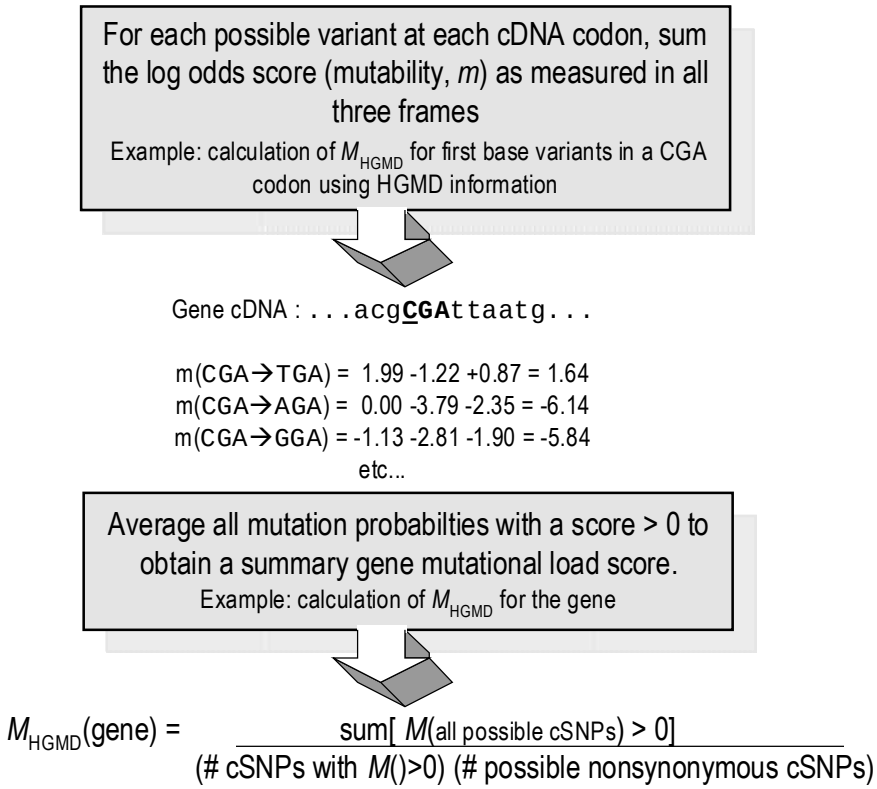


Figure 6.2. Method for obtaining mutability values for human gene sequences. The database log odds scores are applied to a gene coding sequence to attain a single summary mutational load score.

Once every possible nonsynonymous variant was scored for an entire coding sequence, the entire body of point mutations was ranked by descending log likelihood scores. This process was employed to derive M_{intron}–, M_{dbSNP}–, M_{HGMD}–, and $M_{\text{interspecific}}$–like mutation likelihoods for each gene in the LocusLink list. Note that as an exception, the M_{intron} spectrum only has one scoring table because the SNPs from the training dataset do not occur in coding sequences. Therefore, when this spectrum is applied to a gene, each of the

three TMCs corresponding to a particular mutation was summed from the single scoring table.

For any of the four metrics, the body of scores corresponding to a gene's mutational load is distributed so that there is a small number of highly likely nonsynonymous point mutations (positive log odds score) and a much larger number of unlikely variants (negative log odds score). Figure 6.3 shows this in the distribution of scores for four genes chosen at random from our gene list. In order to determine which gene classes were most mutable according to the four different metrics, each gene must have its body of mutation scores, such as those in figure 6.3, reduced to a single value.

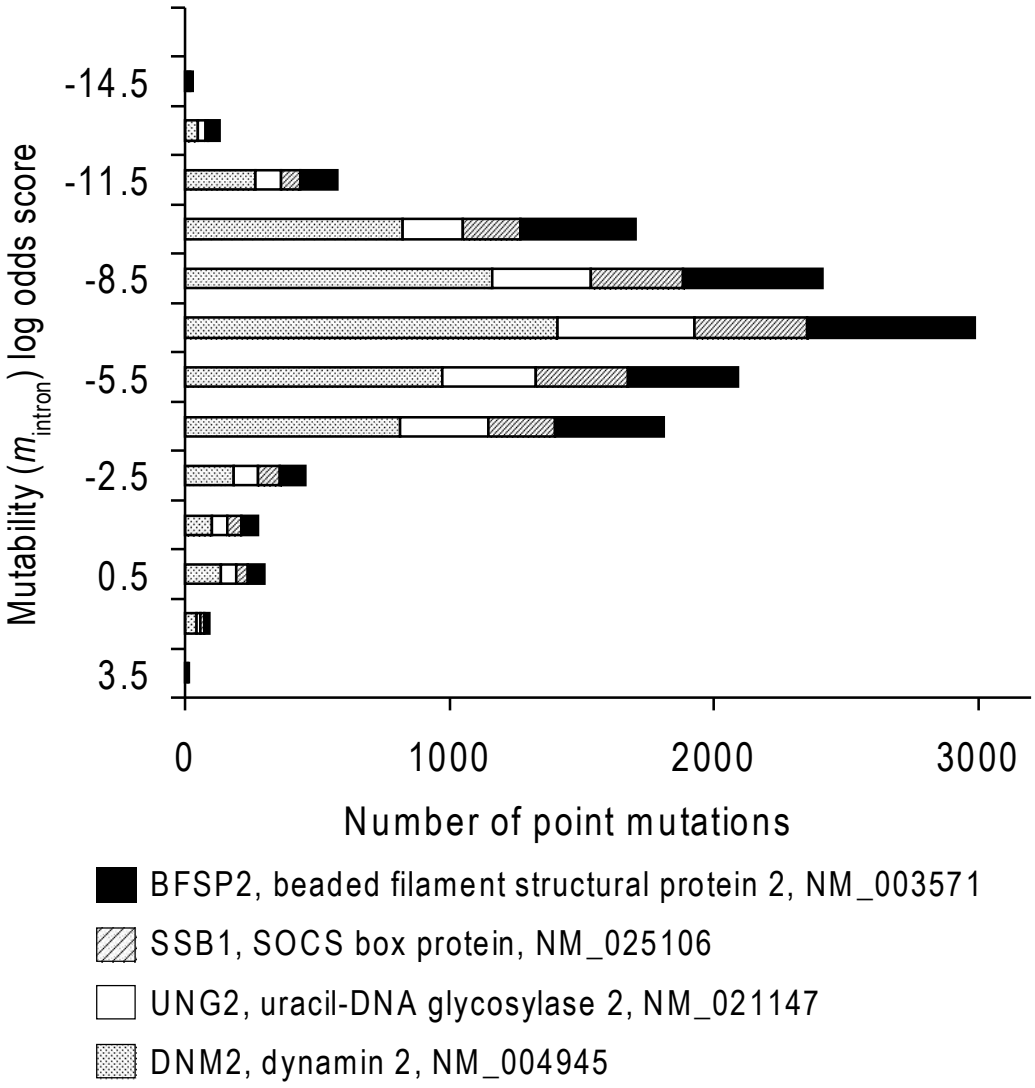


Figure 6.3. Distribution of mutation predictions across four genes. For any given gene there is a distinct set of potential point mutations that are far more likely to occur than expected given underlying codon usage of the gene. To illustrate this, shown is the resulting non-normal distribution of M_{intron} log odds scores applied to all possible nonsynonymous mutations for four genes selected randomly from this study. Extremely likely mutations will have a positive score while unlikely mutations will have a negative score.

It is clear that since the scores are not normally distributed, simply calculating the average score per gene would not appropriately summarize gene mutational load. What one really desires is to know is which genes have the longest, most positive tails with regard to the distribution of scores. To this end, a gene's mutational load score according to a particular metric is calculated as the average value of only those mutations having positive log odds scores. This value is then weighted by the total number of possible point mutations within the gene because the length of the tail is obviously correlated to gene length. This corresponds to taking the weighted average of the positive portion of the spectrum in figure 6.3. In the scenario where a gene does not have any predicted point mutations with a positive score, the mutational load score was set to zero. In this fashion the 12,865 human genes were each assigned four 'summary' mutational load values according to the M_{intron}, M_{HGMD}, $M_{\text{interspecific}}$, and M_{dbSNP} spectra. For the M_{intron} spectrum, which is assumed to represent the actual mutability of genes without the effects of selection (e.g. $M_{\text{intron}} = m_{\text{intron}}$), there is a 2800-fold difference in summary mutational load scores between the most (*HIST1H4A*, histone 1 H4a) and least (*MT1A*, metallothionein 1A) mutable genes.

6-1-3 Identification of the most and least mutable genes in the human genome

In order to show a sample of the most mutable human genes, table 6.1 lists twenty based on the M_{intron} metric of mutability. Since M_{intron} represents the consequences of natural human genome mutation without the confounding effects of selection, this list is the best possible estimate of the most mutable genes in the human genome.

Table 6.1: 20 most mutable human genes according to the M_{intron} mutability metric*

Gene	Locus ID	Protein product	Mutability (m_{intron})	Notes
HIST1H4A	8359	histone 1, H4a	3.16	Binds HIV-1 Tat protein, which recruits histone acetyltransferases to the HIV-1 LTR promoter, which then leads to histone acetylation
RPRM	50514	reprimo protein	3.13	TP53-dependant G2 arrest mediator
DEXI	28955	dexamethasone-induced transcript	3.03	Within Prader-Willi/Angelman syndrome chromosomal region; heavily methylated
HIST1H4F	8361	histone 1, H4f	3.01	Binds HIV-1 Tat protein
ADRAID	146	adrenergic receptor, alpha 1D	2.97	G-protein coupled receptor
TCF15	6939	transcription factor 15	2.92	expressed in early mesoderm and the developing somites, which pattern axial skeleton and skeleton muscle.
GTPBP6	8225	GTP binding protein 6	2.89	Involved in G-protein coupled receptor pathway
WARP	64856	von Willebrand factor A domain-related protein	2.88	extracellular matrix protein which may play a role in cartilage structure and function
HIST1H4B	8366	histone 1, H4b	2.85	Binds HIV-1 Tat protein
DRD4	1815	dopamine receptor D4	2.84	G-protein coupled receptor; promoter mutations have been associated with attention deficit disorder
RPS28	6234	ribosomal protein S28	2.83	multiple processed pseudogenes exist in the genome
HIST1H3C	8352	histone 1, H3c	2.79	Binds HIV-1 Tat protein
HIST1H4D	8360	histone 1, H4d	2.78	Binds HIV-1 Tat protein
CLDN5	7122	claudin 5	2.77	Tight junction protein. Exposure of brain microvascular endothelial cells to HIV-1 Tat results in a decrease of claudin-5 expression as well as cellular redistribution
SSTR5	6755	somatostatin receptor 5	2.76	G-protein coupled receptor binding somatostatin, which acts at many sites to inhibit the hormone and secretory protein release.
CLDN3	1365	claudin 3	2.73	Integral membrane protein that is a component of tight junction strands. It is also a low-affinity receptor for *Clostridium perfringens* enterotoxin.
GPR8	2832	G protein-coupled receptor 8	2.71	G-protein coupled receptor, structurally similar to opioid and somatostatin receptors. Product expressed primarily in brain frontal cortex.
U2AF1	7307	U2 small nuclear RNA auxiliary factor 1	2.68	This gene plays a role in both constitutive and enhancer-dependent RNA splicing.
PHLDA3	23612	pleckstrin homology-like domain, family A, member 3	2.68	fetal protein that mediates associations with membrane phosphatidyl inositol lipids
FGF22	27066	fibroblast growth factor 22	2.67	Putatively involved in hair development.

* This list excludes hypothetical genes, uncharacterized ORFs, and genes recently removed from LocusLink.

Interestingly, many of these genes are part of G-protein coupled receptor (GPCR) complexes or their associated pathways, which represent over 50% of all drug targets (Johnson and Lima 2003). Other highly mutable GPCRs not shown in table 1 include *HRH4* (histamine receptor H4). Histamine receptors are targets for ulcer and allergy drugs including famotidine (Pepcid®), ranitidine (Zantac®), loratadine (Claritin®), and fexofenadine (Allegra®). *HRH4* in particular is a recent, promising target and is the subject of much rational drug design given it is expected to modulate immune cell function (Repka-Ramirez 2003). These loci as well as those GPCRs in table 6.1 may encode good drug targets once their pharmacology is examined.

The dopamine D4 receptor has multiple known mutations that been associated with a variety of behavioral disorders including schizophrenia, attention deficit disorder, Parkinson's disease, and novelty seeking (NS) personality traits where thrill-seeking behaviors are displayed. Dopamine-binding genes have been of great interest to psychiatric geneticists given the role that dopamine plays in the brain's 'reward' system. A study of over 600 *DRD4* alleles in a worldwide cohort showed that this gene has numerous rare, recent mutations and the common alleles can be traced to a rare event that underwent positive selection (Ding, Chi et al. 2002).

There are also a fair number of histone proteins in this table that are targeted by HIV-1 tat protein, which causes recruitment of chromatin remodeling proteins to an HIV promoter and consequently poises that region for transcription. Although histone genes are thought to be slowly evolving, they are loci created by multiple instances of gene duplication (much like olfaction or HOX genes) and a recent study has shown that they may even be *selected* for

duplication (Braastad, Hovhannisyan et al. 2004). Gene duplication allows for more functional redundancy and the evolution of new features.

The *TCF15* gene may seem out of place in the context of the hypothesis that proteins critical to development processes should not be highly mutable. Closer analysis of the gene shows regions of simple gcc-runs which would account for the heightened mutability score. This suggests that during future applications of mutation spectra it may be a good idea to filter the queried DNA sequence for regions of low complexity.

6-2 Correlation of amino acid-altering gene mutability to gene ontology classification

6-2-1 Development of statistical methodology to correlate gene coding region mutability and gene class

In order to determine which gene classes were most mutable according to the four different metrics, each gene had its mutation scores reduced to a single 'summary' mutational load value by averaging only those nonsynonymous mutations having positive log odds scores. In the common scenario where a gene does not have any predicted point mutations with a positive score, the summary score was set to zero. In this fashion the 12,865 human genes were each assigned four summary mutational load values that describe mutation potential according to the M_{intron}, M_{HGMD}, $M_{\text{interspecific}}$, and M_{dbSNP} spectra. Within each list, the summarized mutational load scores for all 12,865 genes were normalized so that the scores for the same gene on different lists could be compared on the same scale. Then for each spectrum type, the gene list was sorted in descending order according to the summarized mutational load score.

To identify whether various gene ontology categories correlated to specific sections of the ranked gene list or whether scores were distributed randomly, a sliding window analysis was performed twice (window = 3000 genes, step = 1000; and window = 1500, step = 1000) to iteratively grab slices of each gene list and inspect for overrepresentation of gene ontology categories. The package EASE (Hosack, Dennis et al. 2003) was used with current annotation tables (04/05/2004) to query for overrepresentation of GO categories (Harris, Clark et al. 2004) and SwissProt keywords (Gasteiger, Jung et al. 2001) in the sliding window slices. A one-tailed Fisher exact test with a Bonferroni correction for multiple testing was employed to obtain the statistical significance of any correlation. A p-value <= 0.05 was deemed a significant correlation to the queried gene categories. Gene classes that correlated significantly to different slices of the four ranked lists are detailed in tables 6.2-6.7. Due to data redundancy, only the most significant correlation to a list slice per gene class is shown. GO categories have three primary branches: Biological process, cellular component, and molecular function. In this chapter, only correlations to the first two types of categories are reported and discussed. Correlations to molecular function classes were calculated, but since so many genes placed within identical classes are paralogs, any mutability correlations reflect sequence identity more than convergent use of nucleotide bias to reach a certain level of mutational load. This data, as well as complete data tables for all correlations calculated, are presented in the appendices for the curious reader.

In assigning a single summary mutational load value for each gene, I averaged only point mutations with positive log odds scores and then weighted the value by the total number of possible nonsynonymous mutations in the gene. To find the most mutationally

cold genes I repeated the analysis described above but averaged together only values from point mutations with negative log odds scores, that is, variants that are much less likely than expected to be observed. These negative scores describing the least mutable genes in the genome were ranked in ascending order so that the genes at the top of the list were those with the most negative scores, or in other words, least likely to mutate. The 1500 least mutable genes by each mutation spectrum were probed for GO category overlap using EASE. A detailed sliding window analysis was performed but was uninformative because in the ranked list, there is less than a two-fold difference between the most negatively and least negatively scored genes. Consequently, only the results from the first step of the sliding window analysis are shown in table 6.9.

6-2-2 Identification of hypermutable human gene groups

With the mutability scores computed for all 12,865 genes, one can infer what gene classes (if any) are inclined to mutate greatly in order to gain insight into how the human genome can respond to environmental challenges. The following methodology is employed to determine whether various gene ontology categories correlate to specific portions of a mutational load-ranked gene list, or whether such scores merely distribute genes randomly. For a given ranked gene list, a sliding window analysis is performed (window = 3000 genes, step = 1000) to iteratively grab 3000-gene slices of the list and inspect for overrepresentation of SwissProt, and GO category terms. The sliding window analysis is repeated with a window of 1500 genes and a step of 1000 in order to seek a finer set of correlations. The results reveal that specific gene groups are inundated with DNA motifs of high mutability

while others are programmed with the least labile sequences possible. Tables 6.2, 6.4, and 6.6 disclose results seen for the M_{intron}- and $M_{\text{interspecific}}$-ranked lists while correlation statistics for the gene list ranked by M_{dbSNP} and M_{HGMD} values are shown in tables 6.3, 6.5, and 6.7.

Any genes that score highly for M_{intron} –like mutability are those that mutate most freely to sample the most variety in the protein product possible. This is because the metric is putatively derived without the contamination of selection. Any genes that score highly for M_{HGMD} –like mutability can be thought of as being prone to radical variation which is a portion of the M_{intron} spectrum. The M_{dbSNP} spectrum is built upon cSNPs discovered when comparing datasets genotyped from only a few individuals, since much of the data emerged from large sequence surveying projects. Therefore, this metric may be considered as the most conservative with respect to impact on the protein product. As postulated, tables 6.2-6.7 clearly demonstrate that any categories involving responses to inflammation, immune systems, pathogens, external stress or other external stimuli are of the most mutable gene groups in the human genome and correlate to the top of all four ranked gene lists. These strong correlations are consistently repeated across both the SwissProt and GO categorization methods, and indicate that such genes are prone to both conservative and ‘risky’ mutation. By possessing nucleotide usage that enhances the ability to mutate rapidly, these gene classes foster an inherent ability to quickly respond to newly emerged pathogens.

Similarly, various types of signal transduction classes correlate to the upper portions of all gene lists. These are often gene categories targeted by pharmacogenetics groups since subtle changes are believed to underlie susceptibility to common disease and variable drug response (Charney and Manji 2004; Force, Kuida et al. 2004; Ziche, Donnini et al. 2004).

Interestingly, the greatest number of signal transduction classes correlate strongly to upper portions of the M_{dbSNP} list. Such genes may be involved in essential cell processes and biased towards only conservative mutation while others are redundant and therefore can pay the cost of sampling more diversity. Additionally, these groups are often part of large multidomain complexes which may cause selection to attenuate the pool of point mutations because the consequences of any single variant potentially ripple through many bound proteins. Given that multidomain protein complexes are often part of extensive pathway systems, this is consistent with a study that predicts increased evolutionary restraint on genes that are components of long pathways (Rutter and Zufall 2004). This view is also supported by the report that *Saccharomyces cerevisiae* proteins with more interactors have been seen to evolve more slowly because a greater portion of the protein is involved in a function than proteins with few interacting surfaces (Fraser, Hirsh et al. 2002). It is expected that such selective constraints would hold in higher mammals as well.

Olfaction genes also exhibit high mutability, for they strongly correlate to the top quarter of the M_{intron}-ranked list, which is consistent with studies that have found evidence of accelerated evolution occurring in these genes (Clark, Glanowski et al. 2003; Chuang and Li 2004). These genes are known to have been generated by rapid duplication events (Glusman, Yanai et al. 2001) and their high mutability reflects CpG context engendering innovation, that is, nucleotide bias most likely to give protein product variety for processes such as pheromone-induced behavior as well as food recognition. The same olfaction genes fall to a lower rank on the $M_{\mathrm{interspecific}}$ list because these genes do not have nucleotide bias specifically

tuned for more conservative mutability, and have a lower overall p-value as well ($p<4.15E-03$).

Ontology categories containing chromatin remodeling genes (biological process:"chromatin assembly/disassembly" and SwissProt:"acetylation") correlate strongly to the top of each of the four separate gene lists ranked by mutability. At first this may seem puzzling because such genes are often initially thought of as encoding housekeeping products. As discussed in section 6-1-3, many histone genes are members of duplicated clusters meaning there will be some histones that are redundant and therefore have paralogs more free to experience a wide variety of point mutation. Also, if duplicated loci end up being placed into a region of the genome that has a different isochore content (see discussion on isochores to follow), then the mutation biases could be dramatically altered. These 'fish out of water' loci could be put into areas where they are more mutable. Additionally, chromatin remodeling products are an underappreciated source of gene regulation, and therefore pockets of mutability in the system may be a desirable trait. Few of these proteins are absolutely required for the cell to function and many of the genes are duplicated. Such properties as combined with mutation-permissive nucleotide bias would provide a means to explore new functions without compromising the integrity of the system. Technological developments have been made that allow one in a high-throughput manner determine what genes are unwrapped from their chromatin structure and which are hidden from transcriptional machinery (Weil, Widlak et al. 2004). A wider analysis of chromatin remodelers and histones may reveal that there is more variation in efficacy of these proteins

than currently thought due to highly mutable gene portions as suggested by the correlations in table 6.2 and 6.3.

Classes falling at the bottom of the M_{intron}-ranked gene list exhibit a degree of mutational rigidity within humans. Table 6.2 shows that these groups are involved in essential gene processes such as transcription (p=3.89E-07) and metabolism (p=4.41E-10). If such genes did not display this rigidity and were intrinsically prone to extensive mutation of all impact types, a species could be genetically frail and therefore less viable. This is underscored in the correlation statistics seen in the $M_{\text{interspecific}}$-ranked list. In many cases these results reflect trends that are diametrically opposed to those seen in the M_{intron} correlations. 'Transcription' does not correlate to any portion of this ranked $M_{\text{interspecific}}$ list and 'development' correlates fairly highly (68.9-92.2 percentile) with a p-value <5.07E-05, as does the housekeeping gene class 'cell growth and/or maintenance' (88.3-100 percentile, p-value <4.75E-02). That is because $M_{\text{interspecific}}$ is derived from interspecific alignments meaning that patterns of conservative mutation will be overrepresented by this metric whereas M_{intron} does not include a selective component. Gene classes placed highly in the $M_{\text{interspecific}}$-ranked list but not in the $M_{\text{interspecific}}$-ranked list are prone to conservative mutations only.

All of these trends are recapitulated in the most general GO classification method—that by designation of a gene's "cellular component". Tables 6.4 and 6.5 indicate that extracellular and ribosome-related genes are the most freely mutable, that is, they undergo the most mutation according to any of the four metrics. Genes with products located in the

nucleus, where transcription occurs, or in the cytoskeleton are of the genes that are least mutable.

Taken as a whole, my results indicate that the ability for genes to mutate so that natural selection has an appropriate pool of variants on which to act in response to the environment is essentially programmed into many of those coding sequences. Consistent, clear, and indisputable correlations between non-homologous genes grouped by function and mutational load arise no matter which mutation spectrum is applied.

Table 6.2: Correlation of GO biological process categories to exhibition of M_{intron} –like and $M_{interspecific}$-like mutability

GO category	M_{intron}				$M_{interspecific}$			
biological process	**Gene list percentile (%)****	**# genes**	**# in category**	**p-value**	**Gene list percentile (%)****	**# genes**	**# in category**	**p-value**
cell-cell signaling	88.3-100.0	101	555	3.65E-05	76.7-100.0	186	555	3.33E-07
chromatin assembly/disassembly	88.3-100.0	26	93	4.76E-03	88.3-100.0	39	93	1.05E-11
cell growth and/or maintenance					88.3-100.0	429	3446	4.75E-02
humoral defense mechanism (sensu Vertebrata)	88.3-100.0	31	120	3.38E-03				
protein biosynthesis	88.3-100.0	95	558	2.42E-03				
response to chemical substance	88.3-100.0	56	223	7.09E-07	88.3-100.0	57	223	3.32E-07
DNA packaging					88.3-100.0	49	175	2.15E-07
homeostasis					88.3-100.0	31	98	2.61E-05
response to biotic stimulus	88.3-100.0	170	849	1.89E-14	76.7-100.0	276	849	3.86E-10
protein transport					80.6-92.2	76	417	2.54E-02
nuclear organization and biogenesis					88.3-100.0	49	194	9.99E-06
response to abiotic stimulus	88.3-100.0	63	285	1.42E-05	88.3-100.0	61	285	1.22E-04
response to stimulus	76.7-100.0	402	1320	5.05E-11	76.7-100.0	367	1320	2.41E-04
chromosome organization and biogenesis (sensu Eukarya)					88.3-100.0	49	190	4.71E-06
development					68.9-92.2	497	1776	5.07E-05
chemotaxis	76.7-100.0	61	118	4.04E-09	88.3-100.0	40	118	1.76E-08
defense response	76.7-100.0	296	787	1.29E-21	76.7-100.0	260	787	2.78E-10
immune response	76.7-100.0	268	713	3.45E-19	76.7-100.0	241	713	1.51E-10
inflammatory response	76.7-100.0	75	180	6.53E-06	88.3-100.0	49	180	6.33E-07
innate immune response	76.7-100.0	79	188	1.50E-06	88.3-100.0	50	188	1.02E-06
olfaction	76.7-100.0	30	48	5.61E-06	53.3-76.6	27	48	4.15E-03
response to external stimulus	76.7-100.0	350	1060	6.87E-15	76.7-100.0	313	1060	3.43E-06
response to pest/pathogen/parasite	76.7-100.0	180	447	1.75E-15	88.3-100.0	95	447	3.26E-08
response to stress	76.7-100.0	249	782	8.30E-08	76.7-100.0	226	782	5.46E-03
response to wounding	76.7-100.0	114	274	4.59E-10	88.3-100.0	70	274	2.64E-09
small GTPase mediated signal transduction	76.7-100.0	103	212	2.18E-14	68.9-92.2	107	212	2.21E-15
alcohol metabolism	61.1-84.4	83	228	1.48E-02				

Table 6.2 continued

G-protein coupled receptor protein signaling pathway	61.1-84.4	209	602	1.11E-07	53.3-76.6	242	602	2.72E-17
cell surface receptor linked signal transduction	53.3-76.6	306	974	7.62E-06	53.3-76.6	326	974	4.64E-10
cyclic-nucleotide-mediated signaling	53.3-76.6	54	109	9.76E-06	53.3-76.6	55	109	3.18E-06
second-messenger-mediated signaling	53.3-76.6	55	123	6.27E-04	53.3-76.6	58	123	2.94E-05
amino acid metabolism	41.7-53.3	55	261	4.51E-02				
carboxylic acid metabolism	41.7-53.3	84	423	3.61E-03	22.2-45.6	152	423	1.70E-05
organic acid metabolism	41.7-53.3	85	425	2.29E-03	22.2-45.6	152	425	2.46E-05
amine metabolism	30.0-53.3	117	341	2.21E-02	26.1-37.8	73	341	7.71E-04
carbohydrate metabolism	30.0-53.3	138	390	4.22E-04	30.0-53.3	147	390	1.42E-06
metabolism	30.0-53.3	1649	6240	4.41E-10	14.4-37.7	1612	6240	1.62E-06
catabolism	14.4-37.7	259	830	1.72E-03	14.4-37.7	254	830	4.01E-03
macromolecule catabolism	14.4-37.7	202	613	5.01E-04	6.7-30.0	191	613	3.02E-02
protein catabolism	14.4-37.7	195	587	4.33E-04				
proteolysis and peptidolysis	14.4-37.7	193	578	3.31E-04	6.7-30.0	184	578	1.01E-02
regulation of transcription	14.4-37.7	537	1786	2.31E-07				
transcription	14.4-37.7	565	1898	3.89E-07				
cation transport	6.7-30.0	131	383	9.32E-03				
homophilic cell adhesion	6.7-30.0	89	116	1.02E-29	6.7-30.0	91	116	3.26E-32
nucleobase\, nucleoside\, nucleotide and nucleic acid metabolism	6.7-30.0	749	2748	7.10E-03	30.0-53.3	165	500	5.21E-03
phosphate metabolism	6.7-30.0	235	681	3.28E-07	6.7-30.0	256	681	1.26E-13
phosphorylation	6.7-30.0	196	537	5.78E-08	6.7-30.0	198	537	5.38E-09
protein amino acid phosphorylation	6.7-30.0	193	491	2.62E-11	6.7-30.0	194	491	3.77E-12
protein metabolism	6.7-30.0	634	2296	1.00E-02	6.7-30.0	679	2296	5.97E-10
protein modification	6.7-30.0	324	975	1.38E-08	6.7-30.0	343	975	5.28E-14
cell adhesion	2.8-14.4	160	589	6.32E-22	6.7-30.0	235	589	6.37E-16
cell communication	2.8-14.4	432	2885	1.63E-05				

** 12,865 genes are ranked in order of descending summary mutational load scores. For example, a percentile of 88.3-100% indicates that the 1,500 most mutable genes according to the specified metric was evaluated for overrepresentation of GO categories. Blank spaces indicate that the listed class did not correlate with any portion of the ranked gene list.

Table 6.3: Correlation of GO biological process categories to exhibition of M_{HGMD} –like and M_{dbSNP}-like mutability

GO category	M_{HGMD}				M_{dbSNP}			
biological process	Gene list percentile (%)**	# genes	# in category	p-value	Gene list percentile (%)**	# genes	# in category	p-value
chemotaxis	88.3-100.0	45	118	3.08E-12				
chromatin assembly/disassembly	88.3-100.0	38	93	3.68E-11	88.3-100.0	27	93	3.65E-03
chromosome organization and biogenesis (sensu Eukarya)	88.3-100.0	47	190	2.41E-05				
defense response	88.3-100.0	136	787	1.94E-06	76.7-100.0	253	787	4.02E-07
DNA packaging	88.3-100.0	47	175	1.33E-06				
establishment and/or maintenance of chromatin architecture	88.3-100.0	44	163	3.95E-06				
inflammatory response	88.3-100.0	51	180	2.85E-08				
innate immune response	88.3-100.0	53	188	1.39E-08				
nuclear organization and biogenesis	88.3-100.0	47	194	4.87E-05				
organismal physiological process	88.3-100.0	190	1318	1.85E-03				
response to abiotic stimulus	88.3-100.0	66	285	5.45E-07				
response to biotic stimulus	88.3-100.0	142	849	7.65E-06	76.7-100.0	267	849	1.47E-06
response to chemical substance	88.3-100.0	61	223	1.49E-09				
response to pest/pathogen/parasite	88.3-100.0	94	447	2.83E-08				
response to wounding	88.3-100.0	69	274	3.42E-09				
cell-cell signaling	76.7-100.0	178	555	2.52E-05	76.7-100.0	176	555	9.45E-04
immune response	76.7-100.0	220	713	1.73E-05	76.7-100.0	232	713	7.30E-07
intracellular protein transport	76.7-100.0	122	395	4.67E-02				
macromolecule biosynthesis	76.7-100.0	255	869	1.65E-04				
protein biosynthesis	76.7-100.0	197	558	2.17E-10				
protein transport	76.7-100.0	134	417	1.67E-03				
regulation of cell proliferation	76.7-100.0	93	258	3.43E-04				
response to external stimulus	76.7-100.0	299	1060	6.96E-04	76.7-100.0	320	1060	4.07E-06
response to stimulus	76.7-100.0	359	1320	2.27E-03	76.7-100.0	369	1320	2.81E-03
response to stress	76.7-100.0	227	782	2.38E-03				

Table 6.3 continued

small GTPase mediated signal transduction	76.7-100.0	108	212	2.15E-17	76.7-100.0	90	212	2.47E-07
signal transduction					68.9-92.2	607	2214	2.55E-03
carbohydrate metabolism	45.6-68.9	147	390	2.55E-06	53.3-76.6	130	390	4.22E-02
cell surface receptor linked signal transduction	45.6-68.9	319	974	2.72E-07	61.1-84.4	315	974	1.82E-07
G-protein coupled receptor protein signaling pathway	45.6-68.9	233	602	5.22E-13	61.1-84.4	223	602	7.37E-11
hexose metabolism	45.6-68.9	50	107	8.95E-04				
amine metabolism	37.8-61.1	117	341	3.21E-02				
cyclic-nucleotide-mediated signaling	37.8-61.1	54	109	2.14E-05	72.8-84.4	38	109	5.94E-07
G-protein signaling\, coupled to cyclic nucleotide second messenger	37.8-61.1	53	103	5.23E-06	72.8-84.4	36	103	1.64E-06
neurophysiological process	37.8-61.1	98	276	3.49E-02	61.1-84.4	98	276	1.69E-02
second-messenger-mediated signaling	37.8-61.1	57	123	1.73E-04	61.1-84.4	40	123	2.18E-06
carboxylic acid metabolism	30.0-53.3	143	423	5.34E-03				
electron transport	30.0-53.3	130	381	9.38E-03				
lipid metabolism	30.0-53.3	164	500	7.44E-03				
metabolism	30.0-53.3	1655	6240	8.18E-11	22.2-45.6	1588	6240	8.81E-03
organic acid metabolism	30.0-53.3	143	425	7.19E-03				
catabolism	6.7-30.0	251	830	3.08E-02				
cation transport	6.7-30.0	130	383	1.51E-02				
homophilic cell adhesion	6.7-30.0	94	116	3.45E-35	22.2-45.6	61	116	6.34E-08
ion transport	6.7-30.0	180	531	2.32E-04				
macromolecule catabolism	6.7-30.0	201	613	6.66E-04				
neurogenesis	6.7-30.0	140	420	1.95E-02				
phosphate metabolism	6.7-30.0	239	681	2.82E-08	14.4-37.7	234	681	2.26E-07
phosphorus metabolism	6.7-30.0	239	681	2.82E-08				
phosphorylation	6.7-30.0	196	537	5.10E-08	22.2-45.6	181	537	2.48E-04
protein amino acid phosphorylation	6.7-30.0	190	491	2.20E-10	14.4-37.7	176	491	1.61E-06
protein catabolism	6.7-30.0	198	587	7.61E-05				
protein metabolism	6.7-30.0	641	2296	1.24E-03	22.2-45.6	645	2296	
protein modification	6.7-30.0	315	975	9.77E-07	22.2-45.6	310	975	

Table 6.3 continued

proteolysis and peptidolysis	6.7-30.0	196	578	5.68E-05				
transport	6.7-30.0	477	1653	1.07E-03				
muscle development					6.7-30.0	56	132	
cell adhesion	2.8-14.4	182	589	2.63E-33	6.7-30.0	212	589	
cell communication	2.8-14.4	468	2885	2.73E-12				

** 12,865 genes are ranked in order of descending summary mutational load scores

Table 6.4: Correlation of GO cellular component categories to exhibition of M_{intron} – like and $M_{\text{interspecific}}$-like mutability

GO category	M_{intron}				$M_{\text{interspecific}}$			
cellular component	Gene list percentile (%)**	# genes	# in category	p-value	Gene list percentile (%)**	# genes	# in category	p-value
extracellular	88.3-100.0	236	1230	2.38E-19	88.3-100.0	391	1230	1.25E-14
ribonucleoprotein complex	88.3-100.0	101	441	1.82E-11	2.8-14.4	148	441	1.41E-05
cytosol	76.7-100.0	132	388	4.21E-05				
extracellular space	76.7-100.0	143	410	1.51E-06	2.8-14.4	143	410	1.39E-06
large ribosomal subunit	76.7-100.0	35	55	9.15E-08	2.8-14.4	30	55	3.61E-04
nucleosome	76.7-100.0	42	79	3.26E-06	2.8-14.4	43	79	7.98E-19
ribosome	76.7-100.0	120	282	7.67E-12	76.7-100.0	112	282	1.50E-08
integral to membrane	61.1-84.4	766	2911	3.83E-03	76.7-100.0	784	2911	1.46E-02
integral to plasma membrane	53.3-76.6	368	1269	2.48E-03	45.6-68.9	366	1269	3.05E-02
intracellular	22.2-45.6	1755	6899	1.14E-03				
transcription factor complex	22.2-45.6	211	628	3.27E-05	2.8-14.4	192	628	4.16E-02
cytoskeleton	6.7-30.0	305	889	9.91E-10	2.8-14.4	171	889	2.20E-07
nucleus	6.7-30.0	815	2825	9.42E-09				
nucleoplasm	18.4-30.0	144	841	1.08E-02				
microtubule associated complex					30.0-53.3	29	103	1.97E-02
vacuole					30.0-53.3	55	130	9.16E-03
nuclear pore					37.8-61.1	21	52	5.33E-04
endoplasmic reticulum					76.7-100.0	135	401	1.46E-02
myosin					76.7-100.0	33	95	1.78E-05
lysosome					76.7-100.0	54	116	2.64E-04
extracellular matrix	2.8-14.4	76	302	2.65E-07	76.7-100.0	68	302	4.49E-04

** 12,865 genes are ranked in order of descending summary mutational load scores.

Table 6.5: Correlation of GO cellular component categories to exhibition of M_{HGMD} – like and M_{dbSNP}-like mutability

GO category	M_{HGMD}				M_{dbSNP}			
cellular component	Gene list percentile (%)**	# genes	# in category	p-value	Gene list percentile (%)**	# genes	# in category	p-value
extracellular	88.3-100.0	237	1230	4.83E-20	88.3-100.0	203	1230	1.07E-06
cytosol	76.7-100.0	124	388	2.86E-03	76.7-100.0	128	388	5.17E-03
extracellular space	76.7-100.0	142	410	1.31E-06	88.3-100.0	86	410	4.63E-06
ribonucleoprotein complex	76.7-100.0	190	441	1.28E-21	88.3-100.0	84	441	6.89E-04
ribosome	76.7-100.0	147	282	1.20E-26	88.3-100.0	67	282	1.50E-06
soluble fraction	76.7-100.0	91	241	1.80E-05				
membrane	2.8-22.2	629	4338	3.15E-07	61.1-84.4	1119	4338	2.14E-02
extracellular matrix	2.8-14.4	76	302	4.35E-07	6.7-30.0	136	302	6.80E-13
lysosome	30.0-53.3	52	116	2.42E-03				
microsome	30.0-53.3	61	117	1.45E-07				
vacuole	30.0-53.3	55	130	1.09E-02				
intracellular	22.2-45.6	1745	6899	2.71E-03				
membrane fraction	22.2-45.6	177	539	2.38E-03				
transcription factor complex	22.2-45.6	201	628	3.18E-03				
integral to membrane	6.7-30.0	812	2911	1.03E-05	61.1-84.4	808	2911	1.97E-06
integral to plasma membrane	6.7-30.0	371	1269	6.65E-03	61.1-84.4	388	1269	2.63E-06
plasma membrane	6.7-30.0	524	1812	1.37E-04	61.1-84.4	526	1812	9.55E-06

** 12,865 genes are ranked in order of descending summary mutational load scores. For example, a percentile

Table 6.6: Correlation of SwissProt Keywords to exhibition of M_{intron} –like and $M_{interspecific}$-like mutability

SwissProt Keyword	M_{intron} Gene list percentile (%)**	# genes	# in category	p-value	$M_{interspecific}$ Gene list percentile (%)**	# genes	# in category	p-value
Signal	88.3-100.0	244	1791	2.27E-02				
Lipid-binding	88.3-100.0	11	24	3.19E-02	76.7-100.0	17	24	2.40E-03
Inflammatory response	88.3-100.0	23	52	1.57E-06	88.3-100.0	22	52	2.71E-05
Cytokine	88.3-100.0	65	146	1.66E-22	76.7-100.0	84	146	2.38E-16
Chemotaxis	88.3-100.0	29	52	3.93E-12	88.3-100.0	30	52	1.51E-12
Ribosomal protein	76.7-100.0	58	88	1.52E-14	76.7-100.0	56	88	6.96E-13
Protein transport	76.7-100.0	69	189	3.43E-02				
Hormone	76.7-100.0	38	59	2.67E-08	88.3-100.0	34	59	1.78E-14
GTP-binding	76.7-100.0	69	158	1.04E-05	88.3-100.0	81	158	5.62E-11
Growth factor	76.7-100.0	46	109	1.38E-02	76.7-100.0	53	109	8.01E-06
Acetylation	76.7-100.0	106	178	4.53E-23	88.3-100.0	78	178	2.40E-25
Olfaction	68.9-92.2	32	44	1.43E-08	53.3-76.6	26	44	4.05E-03
Lipoprotein	68.9-92.2	122	317	1.05E-06	68.9-92.2	131	317	1.39E-09
Transmembrane	53.3-76.6	565	2026	1.60E-03				
G-protein coupled receptor	53.3-76.6	150	316	2.30E-17	53.3-76.6	174	316	7.05E-29
Homeobox	37.8-61.1	62	152	1.03E-02	68.9-92.2	101	152	2.15E-26
Developmental protein	37.8-61.1	107	261	1.72E-06	68.9-92.2	130	261	1.23E-17
Transferase	33.9-45.6	133	765	9.92E-03	14.4-37.7	246	765	1.04E-05
Transcription regulation	14.4-37.7	301	884	1.61E-10				
DNA-binding	14.4-37.7	327	974	8.87E-11				
Activator	14.4-37.7	95	260	4.73E-03				
Phosphorylation	6.7-30.0	336	997	1.07E-11	6.7-30.0	313	997	1.48E-07
Nuclear protein	6.7-30.0	529	1707	1.98E-12				
Metal-binding	6.7-30.0	155	479	1.14E-02	6.7-30.0	178	479	1.39E-09
Hydrolase	6.7-30.0	237	779	6.34E-03	6.7-30.0	243	779	6.56E-05
Cell adhesion	6.7-30.0	141	270	8.09E-22	6.7-30.0	143	270	5.07E-24
Calcium-binding	6.7-30.0	107	286	1.93E-04	18.4-30.0	74	286	2.48E-08
ATP-binding	6.7-30.0	290	670	6.06E-29	6.7-30.0	296	670	1.36E-33
Receptor	2.8-14.4	79	420	8.65E-03	10.6-22.3	91	420	1.40E-06
Extracellular matrix	2.8-14.4	42	124	5.85E-08				
Alternative splicing	2.8-14.4	330	1867	3.05E-17	6.7-30.0	549	1867	3.70E-10

** 12,865 genes are ranked in order of descending summary mutational load scores. For example, a percentile of 88.3-100% indicates that the 1,500 most mutable genes according to the specified metric was evaluated for overrepresentation of categories. Blank spaces indicate that the listed class did not correlate with any portion of the ranked gene list.

Table 6.7: Correlation of SwissProt Keywords to exhibition of M_{HGMD} –like and M_{dbSNP}-like mutability

SwissProt	M_{HGMD}				M_{dbSNP}			
Keyword	Gene list percentile (%)**	# genes	# in category	p-value	Gene list percentile (%)**	# genes	# in category	p-value
Chemotaxis	88.3-100.0	33	52	8.51E-17				
Inflammatory response	88.3-100.0	26	52	2.24E-09				
Acetylation	76.7-100.0	93	178	6.72E-16				
Cytokine	76.7-100.0	89	146	2.16E-21	88.3-100.0	39	146	5.93E-04
Growth factor	76.7-100.0	56	109	2.72E-08				
Hormone	76.7-100.0	47	59	1.06E-17				
Protein transport	76.7-100.0	72	189	6.22E-04				
Ribosomal protein	76.7-100.0	69	88	2.58E-26	76.7-100.0	50	88	6.33E-08
GTP-binding	68.9-92.2	76	158	3.07E-09	76.7-100.0	72	158	2.78E-06
Lipoprotein	68.9-92.2	119	317	1.82E-06	68.9-92.2	129	317	1.36E-07
Developmental protein	53.3-76.6	99	261	8.88E-04				
G-protein coupled receptor	45.6-68.9	170	316	4.54E-27	61.1-84.4	158	316	7.13E-21
Transferase	37.8-61.1	237	765	3.24E-02				
Activator	22.2-45.6	93	260	3.51E-02				
DNA-binding	22.2-45.6	310	974	5.08E-06				
Metal-binding	22.2-45.6	163	479	6.75E-04				
Transcription regulation	22.2-45.6	285	884	6.29E-06	37.8-61.1	265	884	4.22E-02
Nuclear protein	14.4-37.7	490	1707	5.03E-05	6.7-30.0	451	1707	3.66E-02
Alternative splicing	6.7-30.0	608	1867	3.34E-21	6.7-30.0	559	1867	2.43E-14
ATP-binding	6.7-30.0	286	670	4.76E-27	6.7-30.0	279	670	1.23E-27
Calcium-binding	6.7-30.0	111	286	1.15E-05				
Cell adhesion	6.7-30.0	156	270	3.90E-31	6.7-30.0	110	270	2.53E-08
Glycoprotein	6.7-30.0	580	2023	1.23E-06	61.1-84.4	565	2023	7.24E-03
Hydrolase	6.7-30.0	240	779	2.18E-03	22.2-45.6	241	779	9.20E-04
Phosphorylation	6.7-30.0	338	997	4.39E-12	6.7-30.0	314	997	5.53E-09
Transmembrane	6.7-30.0	576	2026	6.98E-06	61.1-84.4	593	2026	9.32E-07
Transport	6.7-30.0	176	498	2.05E-06				
Extracellular matrix	2.8-14.4	41	124	2.75E-07	6.7-30.0	63	124	1.37E-08
Magnesium	2.8-14.4	41	121	1.15E-07	14.4-37.7	62	121	4.09E-08
Signal	2.8-14.4	264	1791	2.26E-03				
Actin-binding					6.7-30.0	61	136	1.82E-05
Cytoskeleton					6.7-30.0	72	165	3.39E-06
Ionic channel					53.3-76.6	88	219	1.88E-04
Receptor					6.7-30.0	135	420	7.23E-03

** 12,865 genes are ranked in order of descending summary mutational load scores.

6-2-3 Identification of human gene groups avoiding nonsynonymous point mutation

Gene groups that are rigid with regard to any type of mutation predisposition will not be uncovered by our analysis thus far because when genes are ranked, their summary mutational load scores are constructed from only those variants with positive log odds scores. This analysis will only highlight which gene groups are the most mutationally hot in the genome. It is of interest, however, to also ask which gene groups are the most mutationally 'cold'. To find the most mutationally cold genes I repeated the analysis but by averaging together only the log odds scores of nonsynonymous mutations that were negative, that is, are much less likely to occur than expected, in order to assign a single mutational load value to each of the 12,865 genes. These negative scores describing the least mutable genes in the human genome were ranked in ascending order (most to least negative) so that the genes at the top of the list were those that were least likely to mutate. The 1500 least mutable genes according to each mutability spectrum are assigned their GO categories and statistics are again calculated detailing the overrepresentation of any specific GO category in that list. As expected, categories representing essential gene processes such as those related to cell cycle control, DNA repair, and DNA replication correlate strongly to the most negative portions of the M_{HGMD}, M_{dbSNP}, $M_{\mathrm{interspecific}}$, and M_{intron}–ranked gene lists (table 6.9). All metrics identified these gene classes as programmed by codon bias and context to avoid point mutation.

Table 6.8: Least mutable GO categories in the human genome

GO category* (#in category)	p-value (# genes)**			
biological process	M_{intron}	$M_{interspecific}$	$M_{HGMD***}$	M_{dbSNP}
cell proliferation (1021)	1.95E-04 (167)	3.61E-02 (158)	4.31E-08 (180)	none
cytokinesis (98)	3.06E-02 (26)	1.07E-04 (31)	2.25E-03 (28)	none
cell cycle (656)	3.47E-08 (130)	4.96E-05 (121)	1.98E-14 (146)	5.06E-04 (116)
cell cycle checkpoint (35)	3.00E-03 (15)	none	1.44E-02 (14)	none
mitotic cell cycle (160)	1.33E-06 (46)	4.67E-03 (39)	4.10E-12 (55)	1.25E-03 (40)
DNA replication and chromosome cycle (165)	3.97E-06 (46)	none	1.85E-11 (55)	1.09E-03 (41)
chromosome segregation (22)	1.25E-04 (13)	1.71E-03 (12)	2.74E-08 (16)	1.40E-04 (13)
M phase (161)	1.42E-07 (48)	none	5.58E-12 (55)	1.87E-04 (42)
nuclear division (155)	1.43E-06 (45)	none	7.35E-11 (52)	5.11E-04 (40)
mitosis (123)	1.57E-05 (37)	none	5.20E-10 (44)	6.39E-03 (32)
DNA metabolism (503)	1.63E-03 (92)	none	2.22E-13 (119)	none
DNA repair (176)	5.43E-03 (41)	none	7.62E-07 (49)	none

* successive indentation shows any encapsulation of GO categories
** 12,865 genes were ranked in order of descending summary mutational load scores according to each metric. The 1500 least-mutable genes according to the specified method were probed for overrepresentation of gene categories. 'None' indicates that the listed class did not correlate with any portion of the ranked gene list.
***additional categories correlating only to portions of the M_{HGMD}-ranked list are 'cell organization and biogenesis,' 'response to DNA damage stimulus,' 'response to endogenous stimulus,' 'DNA packaging,' 'nuclear organization and biogenesis,' 'chromosome organization and biogenesis,' 'nucleobase/ nucleoside/ nucleotide/ and nucleic acid metabolism,' 'chromatin assembly/disassembly,' 'establishment and/or maintenance of chromatin architecture,' 'DNA replication,' 'cell growth and/or maintenance,' 'DNA-dependent DNA replication,' 'cytoplasm organization and biogenesis,' 'double-strand break repair,' and 'DNA recombination.'

6-3 Relationship between mutational load and GC content

It is well known that large scale variation in GC content exists across human chromosomes. These regions, termed isochores, are fairly homogenous tracts of sequence that were discovered experimentally by Giorgio Bernardi using CsCl density gradient ultracentrifugation experiments (Bernardi 2001). There are five major isochore groups: the light tracks (L1 and L2, GC% <38% and 38-42%) as well as the heavy tracks (H1, H2, and H3, GC% of 42-47%, 47-52%, and >52%, respectively). Genes are typically found in the heavy isochores, although within gene GC content does not always match that of its surrounding isochore. The puzzle of this as well as isochore purpose and maintenance have been the subject of much heated debate. Since the rate of GC->AT transition mutations is an order of magnitude greater than any other point mutation, one would expect that the isochores are slowly disappearing and eventually will be homogenized as in the case of murine genomes (Duret, Semon et al. 2002). These conclusions, however, have recently been challenged for evidence has been uncovered suggesting that isochores are not vanishing, but are being maintained by some unknown mechanism (Alvarez-Valin, Clay et al. 2004). There are two major opinions of how this is occurring, either by mutational bias due to bias in DNA repair coupled with neutral evolution, or by way of selection for fixation of GC alleles (Eyre-Walker and Hurst 2001). The mutational bias hypothesis has been challenged by many groups. Lercher and Hurst found that the frequencies of noncoding SNPs from the Orchid SNP database highly favored a model of GC allele fixation (Lercher and Hurst 2002). Additionally, if selectively neutral processes determine the isochoric structure of chromosomes, one should not see any association between local gene

composition and gene function. This is clearly not the case, for gene expression breadth has been correlated to local GC content ($r = 0.24$, $P < 10\text{-}5$) which has led some to conclude that isochores are merely regions of comparable expression breadth (Lercher, Urrutia et al. 2003). Additionally, a detailed study by Jose Castresana has determined that chromosome 19 has an extremely high synonymous substitution rates compared to genes from other chromosomes. When chromosome 19 genes are sorted according to their ortholog's location on mouse chromosomes, the difference in the silent mutation rate disappears (Castresana 2002). Castresana surmises that since mouse orthologs are not GC rich, the mouse lineage has experienced homogenization of the isochores structure whereas human isochores have been maintained do to a strong selective pressure.

Although these issues and interpretations are continuously debated in correspondences, much of the community is accepting that a selective hypothesis may explain isochore maintenance. This then begs the question whether genes are placed within isochores randomly, or whether there is a systematic bias to where specific ontological classes map Given these selective forces it is possible that codon choice within genes is motivated in part by how much the gene sequence preserves the isochore. Anecdotally, it has been seen in human receptor tyrosine kinase genes that functionally important amino acids are statistically more often encoded by a GC base in the wobble position than an AT base (Epstein, Lin et al. 2000). Again, could this be so that those critical amino acids get the benefit of selection against GC->AT mutations within the isochore?

To address this, I have taken each of my 12,858 genes and found their corresponding RefSeq contig (NT_#, GenBank file) in order to calculate their local GC composition.

10,000 bases flanking each side of each gene were pulled. The GC% of these 20,000 bases was calculated for each gene and taken as an estimate of the composition of the host isochore. Genes were divided into groups based on their GC percentages, and the LocusLink ids pertaining to those genes were correlated to SwissProt terms and gene ontology categories using the package EASE (Hosack, Dennis et al. 2003). This is the same ontology correlation method used to correlate gene classes to high or low thresholds of mutability. These results are shown in table 6.9.

Table 6.9 shows that genes with extracellular products typically fall within light isochores while those that are intracellular, especially transcription factors, map to heavy isochores. Provocatively, this lends insight into the gene class vs. mutability correlations discussed in 6-2-3. Transcription factors and developmental proteins have high $M_{\text{interspecific}}$-like mutability meaning they employs a codon usage that causes the most likely mutations to be on average conservative in nature with respect to the protein product. Placing such gene types into regions of the genome that are hypothesized to be under some form of selection acting against the fixation of mutations to pyrimidines (e.g. heavy isochores) would give those genes a second line of defense against too much mutability. On the other hand, I have shown that all of the gene classes mapping to light isochores in table 6.9 have high M_{intron}-like mutability, meaning it is those classes that are most likely to undergo free mutation to sample as many variants possible.

Table 6.9: GC content of 20kb gene flanking sequence correlates to gene ontological class

Ontology method	Category	# genes	# genes in ontology	p-value
G+C = 30-40%				
GO Biological Process	G-protein coupled receptor protein signaling pathway	177	595	1.17E-09
GO Biological Process	homophilic cell adhesion	52	115	5.65E-08
GO Biological Process	olfaction	28	47	1.18E-06
GO Biological Process	cell surface receptor linked signal transduction	246	962	1.40E-06
GO Biological Process	cell communication	619	2847	2.19E-06
GO Biological Process	cell-cell adhesion	76	222	1.57E-05
GO Biological Process	cell adhesion	155	579	1.75E-04
GO Biological Process	signal transduction	464	2183	1.79E-02
GO Cellular Component	extracellular	291	1207	3.21E-05
GO Cellular Component	integral to membrane	603	2870	2.24E-03
GO Cellular Component	integral to plasma membrane	287	1251	5.84E-03
GO Cellular Component	extracellular matrix	85	300	2.75E-02
GO Molecular Function	calcium-dependent cell adhesion molecule activity	45	94	1.31E-07
GO Molecular Function	G-protein coupled receptor activity	125	402	2.19E-07
GO Molecular Function	rhodopsin-like receptor activity	106	324	2.60E-07
GO Molecular Function	signal transducer activity	474	2069	3.40E-07
GO Molecular Function	olfactory receptor activity	25	38	5.20E-07
GO Molecular Function	transmembrane receptor activity	213	838	4.84E-05
GO Molecular Function	GABA-A receptor activity	14	23	3.22E-02
GO Molecular Function	cell adhesion molecule activity	99	367	4.29E-02
GO Molecular Function	receptor binding	130	512	4.98E-02
SwissProt keyword	Glycoprotein	490	1989	3.58E-17
SwissProt keyword	Signal	421	1767	6.74E-11
SwissProt keyword	G-protein coupled receptor	105	310	6.60E-09
SwissProt keyword	Transmembrane	446	1999	7.37E-07
SwissProt keyword	Olfaction	26	43	2.15E-06
SwissProt keyword	Multigene family	229	963	5.66E-04
SwissProt keyword	Cell adhesion	80	264	1.02E-03
SwissProt keyword	Extracellular matrix	44	123	5.23E-03
SwissProt keyword	Antiviral	13	19	8.28E-03
G+C = 40-50%				
GO Molecular Function	MHC class I receptor activity	21	21	3.43E-02
GO Cellular Component	intracellular	3843	6817	3.93E-02
G+C = 50-60%				
SwissProt keyword	Homeobox	75	152	4.93E-06
SwissProt keyword	Developmental protein	105	258	1.44E-03
GO Molecular Function	transcription factor activity	261	800	2.78E-02

** No correlations for genes with G+C > 60%; no genes with G+C < 30%

Given the mutational needs of such gene groups, for an organism to be maximally fit, especially upon confronting pathogens, it would not be good to place such genes within areas of the genome that were under special selective pressures to avoid GC mutation. The failure to see housekeeping genes on this list does not weaken the hypothesis. Heavily transcribed sequences are known to be in a hypomethylated state, meaning they would not need the added protection of selection against transition mutations within heavy isochores. Inactivated genes are normally the ones methylated, but since this causes a mutational risk to these genes, which may be critical to an organism's future yet turned on only once in its lifetime, isochore protection is required. Although these conclusions are only preliminary and analyses of other mammals is needed before such ideas are ready for publication, the apparent relationships tying together GC content, gene ontology, and gene class mutability explain how the differential fitness needs of the human genome can be met in part by isochore placement.

6-4 Mutational load estimation provides insight into human evolution

To review, I have used a dataset of over one million unique intronic SNPs to calculate the frequency of point mutation within trinucleotides. This in turn is applied to 12,865 genes to obtain a list of all possible nonsynonymous mutations ranked by their likelihood of observance. The ranked gene list is then analyzed in light of gene ontology to determine which organismal processes consistently utilize areas of the genome that are prone to (or avoid) point mutation and are subject to various degrees of selection pressure. The mutability spectrum computed using only dbSNP intronic SNPs is employed as the most

accurate description possible of the true mutation rate of underlying DNA motifs without the confounding effects of natural selection. For point mutations occurring in gene coding regions, the effects of selection cannot be discounted, for it is estimated that 38% of amino acid-altering mutations would be eliminated by natural selection (Eyre-Walker and Keightley 1999) and 80% of all amino acid substitutions are deleterious (Fay, Wyckoff et al. 2001). Therefore, it is critical that when building a 'canonical' mutability description for gene coding regions that the SNP training data not originate from a coding region source; therefore, in this study intron SNPs are used.

Through contrasting the M_{dbSNP}, M_{HGMD}, and $M_{\mathrm{interspecific}}$ spectra against the 'canonical' mutability description, M_{intron}, of a gene, insight is gained into how the diversity repertoire is constrained by selective forces over the course of evolution. It may be hard to imagine that so much of gene mutability can be deduced from the perspective of simple sequence context. But when considering how the metrics were constructed, only strong, recurring, sequence-specific forces would become visible when such a large amount of SNP data is combined. More subtle mutation-inducing features such as polymerase arrest sites, proximity to palindromic sequences, and pockets of methylation-prone regions would be diluted by the collation of data. It is not just the presence of CpG dinucleotides that shapes the mutability of gene groups, but also their surrounding sequence context which in turn modulates DNA mutability to meet selective pressure needs. For the 43 trinucleotide mutation classes that have positive log odds scores in each of the four mutation spectra, 27 (86%) involve mutations at CpG dinucleotides yet there is up to a 24-fold difference in the magnitude of the log odds scores. If the concentration of CpG dinucleotides were the sole

factor in gene mutability, all metrics would have produced identical results when correlating their gene lists to ontologies after weighting the training data mutation frequencies by codon usage.

By calculating mutation spectra from the entire human genome, we see trends where genes involved in essential processes tend to overuse codons in a context so that if point mutation occurs, it is prone to be a conservative amino acid substitution. On the other hand, genes involved in pathogen defense and environmental response overuse codon contexts that permit extensive variety in the protein product. There are other reasons for codon selection in a gene, such as selection for translational efficiency or a sequence that minimizes the chemical differences between amino acids on the protein level upon a mutation. However, translational efficiency is not as important in mammalian systems as in prokaryotes, and it has been shown that the genetic code is far from being optimal with respect to error minimization (Archetti 2004). The genome's ability to tune mutation propensities according to gene ontology may be necessary or even mandated to ensure species fitness. As a result, coding region mutability can be correlated to both gene function and necessity. The gene classes that harbor the most mutational load according to any of the derived metrics are those that can be distinguished as the most mutable in the genome, although they are predetermined to be mutable in differing ways. The $M_{\text{interspecific}}$ spectrum shows that transcription-related gene groups are specifically prone to moderately conservative mutation, which means that their CpG usage specifically inclines them to avoid radical mutation (e.g. they also correlate to the bottom of the M_{HGMD}-ranked list). This analysis therefore allows

prediction of which genes have the greatest mutational load and therefore which genes potentially harbor the most variety to meet environmental challenges.

Many results reported here show the global applicability of genome-wide mutation spectra analysis and mirror what has been seen in studies that utilize multiple sequence alignments of homologs to calculate statistics that describe the type of selection individual genes have experienced throughout its history (Clark, Glanowski et al. 2003; Chuang and Li 2004). Here I have reached many of their conclusions with respect to the human system but without the laborious step of generating high quality multiple sequence alignments. Clark and coworkers were able to highlight biological processes that have undergone positive selection for mutation since human-chimpanzee divergence. In their table 1, they explain that categories ‘olfaction,’ ‘sensory perception,’ ‘cell surface receptor-mediated signal transduction,’ ‘chemosensory perception,’ ‘nuclear transport,’ ‘G-protein-mediated signaling,’ ‘signal transduction,’ ‘cell adhesion,’ ‘ion transport,’ ‘intracellular protein traffic,’ ‘transport,’ ‘amino acid metabolism,’ ‘cation transport,’ ‘developmental process,’ and ‘hearing’ underwent positive selection in humans (Clark, Glanowski et al. 2003). All categories except “hearing” have counterparts in my tables 6.2-6.7 that correlate to some metric of high mutability, which suggests that these categories were predisposed to mutate due to their underlying DNA sequence context manifested in codon usage. Based on these results of this study, I believe that variation is directed to specific places within gene sequences due to codon usage and context. The fact that these highly mutable genes, and regions therein, belong to specific classes of cellular processes suggests that this codon usage represents an element of foresight as to what genomic challenges the species is anticipated to

confront in its environment. That is, for such gene categories there will be a larger, broader pool of variants on which selection may act relative to immutable genes. The size of the variant pool is a consequence of the mutational load of the underlying gene sequence, which in turn is a function of codon context.

It is possible that the existence of gene regions predisposed to differential mutational load suggests that the evolvability of genes can itself evolve and thus be modulated over time. It has been stated that mutation must be random and cannot anticipate future needs because "selection lacks foresight, and no one has described a plausible way to provide it" (Dickinson and Seger 1999). Mutation is often considered to merely provide natural selection the raw material on which to act and is thought to be so ubiquitous that most any variant will appear in the gene pool somewhere. My results, however, challenge this line of thought. If the human genome arranges hierarchies of differentially mutable CpG dinucleotides and codon contexts so that the gene groups most likely to help a species respond to and survive in its changing environment have a large, biased pool of mutation in natural populations, selection is guided and somewhat forecast. CpG dinucleotides, once decayed into CpA's or TpG's, can subsequently mutate into other bases meaning that many sites in genes may have been involved in a CpG process at some point along its evolution. Therefore, by controlling the rate of mutation within a gene, DNA sequence context such as CpG dinucleotides potentially controls a gene's rate of evolution. Since an organism must rise to combat the recurring challenges implicit in pathogen defense, immune response, and the evolution of new genes in other species within its environment, the ability to somewhat pre-program the mutational needs of a gene into its DNA sequence could certainly provide a

selective advantage in maintaining fitness. Compatible genome hotspot studies in other organisms coupled with examination of gene group distribution across isochores are needed to further qualify this observation. But it is possible that a gene's isochoric location coupled with its codon bias may acutely predict what role it plays in species adaptation and the breadth of possible evolutionary futures. Additionally, in extended work it will be of interest to apply such mutability metrics to gene fragments in order to ask how mutational load is distributed across a coding sequence. Mutability may be heightened in specific protein domains, structural or binding motifs, pockets of amino acid residues contiguous in a three dimensional structure, or even change according to position in the mRNA transcript depending on how the gene is translated.

One way in which DNA mutability is modulated to meet the selective pressure needs of genes is by tuning codon context and codon usage bias as to reach an optimal level of mutation. Understanding inclinations towards point mutation throughout the human genome is invaluable in both choosing candidate genes for diseases or drug targets as well as forecasting how genes may respond to environmental changes in the future. For example, some GPCRs have a particularly high propensity for point mutation which would explain the range of effectiveness of drugs which target such products. This study suggests which human genes will provide selective forces the greatest protein product variety throughout evolution. Additionally, the combination of codon synonyms used to encode a particular protein's amino acids predicts how much mutational load is incorporated into the underlying gene. Examination of SNP differences between distinct human populations, all of whom have evolved recently to drive alleles conferring some selective benefit against local

pathogens to fixation, can disclose what genotypes are less susceptible to pathogenic disease. These differences also suggests areas that may make some molecular process, such as a protein function, change as to thwart the pathogen in the ever-enduring pathogen/host genetic ‘arms race’.

CHAPTER SEVEN
FUTURE WORK

7-1 Extension of mutational load prediction to other areas of the genome

Thus far, point mutation predictions have been made only for gene coding regions, and spectra have been developed based on intronic SNPs to a limited extent. These methods can be replicated to examine and rank the mutational load of other genome regions such as the UTRs, promoters, and intergenic spaces. To complete this, mutations in any sufficiently large SNP database (currently dbSNP, TSC, and JSNP) will be computationally mapped to their host gene in order to elucidate the surrounding genomic sequence context. Since the TMC-categorization method for non-coding SNPs is not frame-dependent as is the case for cSNP prediction, there is an inherent lack of sorting SNPs by impact. To add this component to the non-coding methods, SNPs predicted by TMC matrices will be ranked by their inclusion in known noncoding consensus sequences such as protein binding sites, transcriptional enhancers and silencers, splice sites, and sequence signaling motifs, to name a few. Noncoding SNP (intronic, UTR, promoter) prediction models will be subjected to rigorous computational testing (e.g. assessing the prediction accuracy of historically known SNPs as done for previous cSNP prediction) and then can be subjected to high-throughput genotyping.

Mutation could occur at a higher probability at a certain position because a mutation may be nonrandomly produced there by different mechanisms. Prediction methods could be augmented by other trends specifically investigated within databases of known point

mutations. The sequence TG(A/G)(A/G)(G/T)(A/C) is a known arrest sequence for polymerase α, (Krawczak and Cooper 1991) and could be sought out in the surrounding sequence of known SNPs to make additional prediction rules. Todorova and coworkers noticed during an extensive analysis of all dystrophin gene human SNPs that 60% of CpG dinucleotide transition mutations were surrounded by continuous runs of 2-6 A/T bases on the 5' end and 1-6 A/T bases at the 3' end (Todorova and Danieli 1997). Given these CpG's are located in a GC poor microenvironment, this observation suggests the possibility that 'lone' CpG's may be those that are preferentially methylated and therefore more prone to transition mutation. Future SNP mining could be done to search for and quantify this trend in order to later use it to improve mutation prediction. The calculation of SNP rates within genome regions specifically known to be heavily methylated would potentially shed much light on this idea.

7-2 Extension of mutation spectrum analysis to microorganisms

Comparative analysis of many highly related yet distinct species of microorganisms may disclose some point mutational paths that caused divergence. Akin to how the chimpanzee, mouse, and human coding region multiple alignments were utilized in this study to glean the path of evolution from a shared ancestor, high quality multiple alignments of microorganism trios can lend insight into the evolutionary history of species divergence as well as predict evolutionary futures. Such work could have great impact in public health fields, for as the world population becomes more integrated due to the ever-increasing multinational aspect of business the chance of shared pathogen nuisances increases as well.

Researchers at the University of Ottawa have noticed that within the *Saccharomyces cerevisia*e genome and *Plasmodium falciparum* chromosomes nucleotide bias affects protein sequence on a genome-wide scale, and have concluded that mutational bias inherent in nucleotide usage can have a major effect on the molecular evolution of proteins (Singer and Hickey 2000). In combination with the principles outlined by my study, their work suggests that calculations of genome mutability as derived from databases of point mutations could be applied to microorganisms such as pathogenic strains of bacteria or viruses. This would engender analysis of mutational proclivities inherent in those systems once the effects of nucleotide composition are taken into account. The first task would be to obtain a sufficiently large set of point mutations determined from a single strain and ask how nonrandomly those mutations are scattered across short DNA motifs. Such organisms mutate rapidly in concert with their high reproduction rate and any delineation of trends could aid researchers in forecasting what classes of strains are most likely to arise within populations next.

Microorganisms that replicate with great speed have the size of their population on their side for experimenting with genetic diversity. Because of this, these creatures may not need the elaborate tricks of mutational preinclination to ensure targeted mutation towards the parts of their genome most able to spur adaptation. Human systems have dramatic variations in GC content throughout the chromosomes which lends different mutational rates being seen for different areas. Viral and bacterial genomes are far too compact to utilize those mechanisms to control regions of mutation lability.

We are in a genetic arms race with the pathogens around us. Our human populations evolve to drive to fixation alleles that engender survival while the pathogens constantly evolve to maintain their survival by mutating to thwart human defense systems. A key question is whether the occurrence of such mutations within these pathogens is a mostly random process or whether there is some sort of targeting of the mutational forces to certain genomic spots, as is the case of mammals with their isochores and CpG dinucleotides. Also, could it be that some of the most virulent pathogens are those that either trick human mutational load raising mechanisms adeptly or have their own unique systems to target variation towards host-foiling genes? One could imagine that if it were proven that if isochore content indeed holds the blame for so much human mutational variation across chromosomes, dsDNA viruses that placed themselves in the regions that are not selected against mutation would have a mutational advantage.

7-3 Applying information concerning GC content

Since isochores may be under differing levels of selection due to an unknown mechanism that is related to GC content, it may be useful to make additional M_{intron} matrices for each of the separate isochore classes (three heavy, two light) that occur across chromosomes (see chapter six for an in-depth discussion on isochores). If heavy isochores are under selection as to reduce the number of point mutations that lower GC content, then it follows that separate mutability tables should be drawn up for each of these regions. As a result, when predicting point mutations in a gene for subsequent experimental testing, one

would want to use the mutability table that corresponds to the same isochore class in which the gene belongs (e.g. H1, H2, H3, L1, or L2).

APPENDIX A

LOG ODDS SCORING TABLES FOR MUTATION SPECTRA PUBLISHED OVER THE COURSE OF THIS STUDY

Table A.1: trinucleotide mutation classes composing the M_{HGMD} nonsynonymous mutation spectrum

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
1.991	CGA	TGA	0.473	TCG	TCA	1.186	CCG	CTG
0.628	CGG	TGG	0.189	ACG	ACA	1.169	TCG	TTG
0.597	CGC	TGC	0.172	CGT	TGT	1.143	ACG	ATG
0.525	CGT	TGT	0.167	CCG	CCA	0.866	GCG	GTG
0.319	CGT	CAT	0.130	CGA	TGA	0.711	CGG	CAG
0.278	ACG	ATG	0.069	TAC	TAT	0.401	CCG	CCA
0.211	CGC	CAC	0.057	CGG	TGG	0.269	TCG	TCA
0.191	CGG	CAG	-0.115	GCG	GCA	0.140	GCG	GCA
0.190	TGG	TAG	-0.139	GGT	AGT	0.130	CGA	CAA
0.163	CGA	CAA	-0.141	TCC	TCT	0.015	ACA	ATA
0.111	CAG	TAG	-0.241	GAC	GAT	-0.087	ACG	ACA
0.106	TGG	TGA	-0.253	GGA	AGA	-0.088	GTG	GTA
0.081	CCG	CTG	-0.262	ACC	ACT	-0.102	CGT	CAT
-0.010	CAA	TAA	-0.272	CGC	TGC	-0.112	ATG	ATA
-0.012	TCG	TTG	-0.293	CCC	CCT	-0.179	CTG	CTA
-0.176	TGT	TAT	-0.303	GGG	AGG	-0.232	ATG	ACG
-0.275	GGG	AGG	-0.321	GCC	GCT	-0.295	CCA	CTA
-0.305	TGT	CGT	-0.382	GTA	ATA	-0.309	GCA	GTA
-0.380	GGA	AGA	-0.386	TTC	TTT	-0.324	CGC	CAC
-0.409	GGT	GAT	-0.387	GTC	ATC	-0.373	TTG	TTA
-0.459	GCG	GTG	-0.393	CTC	CTT	-0.465	CTA	CTG
-0.486	GGT	AGT	-0.401	TGC	TGT	-0.468	TCA	TTA
-0.490	TAT	TGT	-0.407	GGC	GGT	-0.509	GTG	GCG
-0.545	TCG	TAG	-0.440	GGC	AGC	-0.642	AGG	AGA
-0.617	TGC	TAC	-0.456	AAC	AAT	-0.678	CGA	CTA
-0.631	TCA	TGA	-0.499	GCT	ACT	-0.683	GGG	GGA
-0.740	TGG	CGG	-0.517	GTG	ATG	-0.699	GTA	GTG
-0.760	TAC	TAA	-0.546	CAC	CAT	-0.727	TGG	TGA
-0.768	TAC	TAG	-0.558	GTC	GTT	-0.732	CTG	CCG
-0.773	TGC	CGC	-0.582	GTT	ATT	-0.861	TGG	TAG
-0.776	GGC	AGC	-0.619	AGC	AGT	-0.946	TCG	TCC
-0.796	GGA	GAA	-0.646	GAC	AAC	-0.957	GCT	GCC
-0.818	GGG	GAG	-0.673	GCA	ACA	-0.964	ATA	ATG
-0.834	GGC	GAC	-0.715	ATC	ATT	-0.990	AGG	AAG
-0.861	GGT	GTT	-0.740	GAG	AAG	-1.015	CAC	CAT

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-0.920	TGC	TGA	-0.744	GAT	AAT	-1.028	GCG	GCC
-0.939	CGT	CCT	-0.770	GCC	ACC	-1.044	CCC	CCT
-0.944	GAA	TAA	-0.890	GGT	GAT	-1.087	GCC	GCT
-0.947	CAT	CGT	-0.920	TTG	CTG	-1.091	CCT	CCC
-0.968	CTG	CCG	-1.000	GAA	AAA	-1.128	CGG	CGA
-0.971	GGT	TGT	-1.002	TAT	TAC	-1.148	CGG	CTG
-0.987	TAC	TGC	-1.022	ATG	GTG	-1.155	GGG	GAG
-0.995	CTA	CCA	-1.115	GCG	ACG	-1.187	CCG	CCC
-1.035	ATT	ACT	-1.127	CGT	AGT	-1.208	CGG	CCG
-1.039	GTG	ATG	-1.129	ATA	GTA	-1.215	GAC	GAT
-1.090	TAT	TAG	-1.151	TCG	TCT	-1.216	TCG	TCT
-1.126	CGA	CCA	-1.162	GTC	TTC	-1.242	TTG	TCG
-1.126	CGA	GGA	-1.173	GCG	GCT	-1.245	CAT	CAC
-1.188	CGC	CCC	-1.202	AGT	AGC	-1.268	TCT	TCC
-1.211	GAG	AAG	-1.211	CCG	CCT	-1.270	CTC	CTT
-1.216	GGT	CGT	-1.224	CGC	CGT	-1.297	ACT	ACC
-1.229	ATG	ACG	-1.275	TAG	CAG	-1.306	AGG	AGT
-1.230	GAG	TAG	-1.297	ACG	AGG	-1.349	GTC	GTT
-1.251	GAC	AAC	-1.297	TGT	CGT	-1.420	TGG	TGT
-1.259	TCA	TAA	-1.310	GGC	GAC	-1.450	ATG	AGG
-1.276	CGG	CCG	-1.320	TGG	CGG	-1.455	TCC	TCT
-1.282	CGC	CTC	-1.323	GGG	GAG	-1.456	TTG	TTT
-1.295	ATG	GTG	-1.328	TGT	TGC	-1.475	TGC	TGT
-1.319	TGC	TTC	-1.332	GAT	GAC	-1.484	TTC	TTG
-1.324	CTT	CCT	-1.337	ACG	AAG	-1.499	TAC	TAT
-1.333	TTA	TGA	-1.361	GAG	GAA	-1.506	CCG	CGG
-1.340	TAT	TAA	-1.377	ATC	GTC	-1.513	TAT	TAC
-1.364	AAT	AGT	-1.386	TCG	TCC	-1.527	CCG	CCT
-1.405	TCC	TTC	-1.390	CCT	TCT	-1.534	GTA	GCA
-1.426	CCC	CTC	-1.396	GCG	GAG	-1.545	GGG	GGT
-1.432	TTA	TAA	-1.404	ATT	GTT	-1.546	GGA	GTA
-1.447	CTC	CCC	-1.419	CAT	CAC	-1.556	TTG	TTC
-1.457	GGA	GTA	-1.421	TGA	CGA	-1.559	CTT	CTC
-1.473	GGA	TGA	-1.426	GGA	GAA	-1.560	AGA	ATA
-1.496	GCC	ACC	-1.458	CTG	CTA	-1.560	ACC	ACT
-1.503	TTA	TCA	-1.464	GTG	GTA	-1.575	ACG	ACC
-1.524	TGG	TGC	-1.464	GCT	TCT	-1.575	TTC	TTT
-1.524	TGG	TGT	-1.469	TCT	CCT	-1.581	CAA	CAG
-1.533	GGC	GTC	-1.471	ACG	ACT	-1.590	TGG	TTG
-1.534	TGT	TGA	-1.483	ACA	GCA	-1.595	TGA	TTA
-1.539	CGT	AGT	-1.499	TCT	TCC	-1.604	ATG	ATT
-1.557	GCG	ACG	-1.521	ACT	ACC	-1.607	GTG	GGG
-1.562	TGT	TTT	-1.532	CCC	TCC	-1.608	TTT	TTC
-1.590	CGA	CTA	-1.536	TGC	CGC	-1.616	GGC	GGT

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-1.611	GAA	AAA	-1.538	CCT	CCC	-1.631	AAC	AAT
-1.612	CAT	TAT	-1.538	ACT	GCT	-1.632	TTA	TTG
-1.621	TCG	CCG	-1.552	GTG	TTG	-1.637	GAT	GAC
-1.631	CTC	TTC	-1.553	GCT	GCC	-1.681	CCA	CCG
-1.641	TAC	CAC	-1.563	GGG	GGA	-1.681	CTC	CTG
-1.645	TTT	TCT	-1.570	TAT	GAT	-1.683	AGG	ACG
-1.662	TTG	TCG	-1.578	GCA	TCA	-1.687	ATC	ATG
-1.665	CGT	CTT	-1.612	CGG	AGG	-1.687	ATC	ATT
-1.687	CCT	CTT	-1.620	GAT	CAT	-1.693	TCG	TGG
-1.693	TCC	CCC	-1.624	CAA	GAA	-1.702	GTC	GCC
-1.694	ATA	ACA	-1.631	CGA	AGA	-1.713	GTG	GTT
-1.705	ACG	AGG	-1.666	ACC	AAC	-1.717	CTC	CCC
-1.710	GAT	GGT	-1.672	TAC	CAC	-1.719	CGT	CTT
-1.716	TGC	TGG	-1.682	ACG	GCG	-1.737	CCC	CTC
-1.731	GGG	CGG	-1.689	TTG	TTA	-1.756	CTA	CCA
-1.737	TGG	GGG	-1.695	TAT	CAT	-1.767	ACA	ACG
-1.741	GCA	ACA	-1.711	GAC	TAC	-1.770	ATC	ACC
-1.742	CAC	TAC	-1.726	CGA	GGA	-1.786	ATG	AAG
-1.743	TGC	GGC	-1.728	GGG	GGT	-1.787	TTC	TTA
-1.743	TGC	TCC	-1.735	GTT	TTT	-1.787	CAT	CGT
-1.750	TGT	TCT	-1.737	GGT	GGC	-1.790	GCG	GCT
-1.755	ACC	ATC	-1.748	TAA	CAA	-1.793	CGT	GGT
-1.765	CAC	CGC	-1.755	GCC	TCC	-1.795	CGA	AGA
-1.775	TAA	TAT	-1.758	TGG	TGA	-1.804	AAT	AAC
-1.784	CCT	TCT	-1.759	CCA	TCA	-1.808	GGG	GTG
-1.790	CCA	CTA	-1.760	AAT	AAC	-1.815	ATC	ATA
-1.793	TCT	CCT	-1.763	GAT	TAT	-1.838	CGC	CGT
-1.797	GCT	ACT	-1.764	AAG	AAA	-1.860	GGA	GAA
-1.808	CGT	GGT	-1.769	GGG	CGG	-1.869	CTG	CTC
-1.819	ACT	ATT	-1.770	ACT	AGT	-1.869	CTG	CTT
-1.832	CCC	TCC	-1.773	CTC	TTC	-1.880	ATA	ACA
-1.844	TCG	TGG	-1.796	CGC	AGC	-1.886	AGC	AGT
-1.847	GTT	TTT	-1.804	GCT	CCT	-1.904	GCG	GGG
-1.852	ATG	AGG	-1.809	AGG	AGA	-1.908	ATT	ATC
-1.852	GAT	AAT	-1.810	TTG	TTT	-1.911	ACG	ACT
-1.857	GCG	GAG	-1.820	CGT	CGC	-1.916	CTG	CGG
-1.868	AGT	AAT	-1.829	ATG	ATA	-1.940	CTC	CTA
-1.868	GTC	TTC	-1.834	CTA	TTA	-1.941	GTC	GTG
-1.881	CCG	CGG	-1.839	CAT	TAT	-1.952	TGC	TAC
-1.897	GAC	GGC	-1.841	ACG	ACC	-1.961	GTA	GGA
-1.899	TGT	GGT	-1.843	ACC	GCC	-1.964	GTG	GTC
-1.900	CGG	GGG	-1.865	CAG	GAG	-1.988	CGG	CGT
-1.922	ATG	ATA	-1.868	ATG	ATT	-1.992	TGA	TAA
-1.923	GTC	ATC	-1.876	TAA	GAA	-2.004	GCA	GCG

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-1.942	GCT	GTT	-1.886	GTA	TTA	-2.006	CGT	CGA
-1.944	GGC	CGC	-1.894	CAG	CAA	-2.013	AGC	AGA
-1.950	TTG	TAG	-1.896	AGT	GGT	-2.018	CGA	GGA
-1.956	AAA	TAA	-1.907	AAG	AAT	-2.020	CGC	CCC
-1.961	ATC	AAC	-1.908	CAC	TAC	-2.021	TAA	TGA
-1.964	GCT	CCT	-1.919	GGT	TGT	-2.021	GTC	GTA
-1.964	ACA	ATA	-1.925	GAG	CAG	-2.028	AGG	ATG
-1.974	GCC	GAC	-1.925	GAG	TAG	-2.031	ACT	ACG
-1.974	GCC	GTC	-1.926	CTG	TTG	-2.037	TGC	TGA
-1.981	ATC	ACC	-1.940	GAA	TAA	-2.052	TCA	TCG
-1.989	TGA	TGG	-1.950	CTT	TTT	-2.055	CGT	AGT
-1.989	TGA	TGT	-1.955	CAG	TAG	-2.056	GGG	GCG
-1.989	TGA	AGA	-1.961	CAA	CAG	-2.061	CAT	CAA
-1.989	TGA	GGA	-1.961	AGG	AGT	-2.062	GGT	GAT
-1.990	TTC	TCC	-1.968	TCC	CCC	-2.062	CAA	CGA
-1.999	CTT	CGT	-1.971	GGC	CGC	-2.084	CGG	GGG
-2.007	TGG	TCG	-1.989	TTA	CTA	-2.086	AGA	AAA
-2.010	GTC	GAC	-1.990	TGG	TGT	-2.090	TAA	TTA
-2.011	AAC	AAA	-1.994	ACC	AGC	-2.093	CGA	CGG
-2.034	AAC	AAG	-1.998	CTG	CTT	-2.096	CTT	CTG
-2.036	CGC	AGC	-2.005	TCA	TCG	-2.097	ATG	ATC
-2.040	AGA	TGA	-2.013	GTG	CTG	-2.113	TGG	TCG
-2.040	AGA	GGA	-2.020	GAG	GAT	-2.117	AGA	AGG
-2.040	TCT	TTT	-2.030	CCG	TCG	-2.125	CGC	AGC
-2.044	CGG	CTG	-2.060	CGG	CGT	-2.127	AGT	AGC
-2.058	AAC	AGC	-2.060	CGG	GGG	-2.127	AGT	AAT
-2.060	GGC	TGC	-2.062	CCA	CCG	-2.128	GCC	GTC
-2.063	GTA	ATA	-2.072	AGT	AGG	-2.131	TCC	TTC
-2.066	TCA	CCA	-2.084	CAT	AAT	-2.134	ACG	AGG
-2.067	GGG	TGG	-2.086	TAG	AAG	-2.135	CGG	AGG
-2.091	TAT	CAT	-2.094	GTA	CTA	-2.135	TTC	TCC
-2.122	TGT	TGG	-2.101	TAT	AAT	-2.136	GAG	GAA
-2.123	GTT	GAT	-2.109	TTT	CTT	-2.140	GCT	GCG
-2.126	GCA	GTA	-2.113	TTC	CTC	-2.147	GGC	GGA
-2.132	TAG	TGG	-2.115	GTG	GTT	-2.158	AGC	AAC
-2.132	TAG	CAG	-2.117	GCA	CCA	-2.160	CCC	CCG
-2.132	TAG	AAG	-2.118	ACA	AAA	-2.162	CTA	CTC
-2.132	CCG	TCG	-2.122	CTA	CTG	-2.167	AGT	AGG
-2.165	ACG	AAG	-2.139	TCG	TGG	-2.173	GGT	GTT
-2.168	AGG	GGG	-2.140	GCG	TCG	-2.175	GTG	GAG
-2.169	CGC	GGC	-2.143	CAA	AAA	-2.190	GCA	GCC
-2.171	CAT	CCT	-2.148	TCA	CCA	-2.196	CCT	CCG
-2.175	AAT	GAT	-2.153	TTG	TTC	-2.199	TGT	TAT
-2.182	CAG	CGG	-2.158	TTT	TTC	-2.213	GGC	GAC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.195	GAC	TAC	-2.162	TCT	GCT	-2.231	TGA	TGG
-2.206	GTA	GCA	-2.167	TCC	ACC	-2.233	TGC	TCC
-2.215	AAA	GAA	-2.168	GTC	CTC	-2.236	GTG	ATG
-2.230	TTC	CTC	-2.176	ACA	AGA	-2.237	GAT	GAA
-2.240	TCA	TTA	-2.179	GGC	TGC	-2.241	TAG	TGG
-2.241	GAT	GTT	-2.179	CGC	GGC	-2.258	GTT	GTC
-2.255	GGG	GTG	-2.182	CAG	AAG	-2.271	GTA	TTA
-2.265	CCC	CGC	-2.189	TGA	GGA	-2.274	TTA	TGA
-2.266	ATG	AAG	-2.191	ATG	AGG	-2.280	AAG	AAA
-2.269	GTT	GCT	-2.192	TAG	TAA	-2.287	CAG	CGG
-2.269	GTT	ATT	-2.192	TAG	TAT	-2.291	TAT	TAG
-2.282	TGC	AGC	-2.194	GCG	GCC	-2.298	CAT	CAG
-2.284	TGT	AGT	-2.196	ACA	ACG	-2.320	GGA	GGG
-2.285	TTT	TGT	-2.204	TCG	CCG	-2.323	TTT	TTG
-2.287	CTG	CGG	-2.213	CTT	CTC	-2.338	CAG	CAA
-2.293	AAG	GAG	-2.217	CCT	ACT	-2.343	CGT	CCT
-2.294	CTC	CGC	-2.226	CGT	CGA	-2.344	GCC	GCG
-2.300	ATA	AAA	-2.229	GGT	CGT	-2.350	GCG	GAG
-2.314	TAC	GAC	-2.229	GGT	GTT	-2.358	GTT	ATT
-2.322	CTT	TTT	-2.237	GTC	GAC	-2.367	GGT	GGA
-2.330	TGG	AGG	-2.244	GAG	GAC	-2.370	CAA	AAA
-2.346	GAT	TAT	-2.244	CCG	CCC	-2.391	GAT	GAG
-2.351	AGC	AAC	-2.248	TTG	GTG	-2.399	ACC	ATC
-2.373	GCA	GAA	-2.251	CAC	AAC	-2.404	GAC	AAC
-2.373	GTA	GGA	-2.251	CGG	CGA	-2.406	CGA	CCA
-2.373	CCG	CAG	-2.252	GCG	CCG	-2.414	GGC	GCC
-2.373	GAC	CAC	-2.259	ATA	ATG	-2.416	TAC	TGC
-2.384	AGC	AGG	-2.285	CAG	CAT	-2.416	TAC	TAA
-2.387	AAG	TAG	-2.303	ACT	ACG	-2.419	AGG	AGC
-2.391	CAC	CCC	-2.309	GAC	CAC	-2.430	CCC	CCA
-2.394	CAT	CTT	-2.312	GCG	GGG	-2.441	AGT	AGA
-2.394	TGA	TCA	-2.322	GTT	GTC	-2.447	ATT	ATG
-2.394	TGA	TGC	-2.322	GTT	GAT	-2.447	CAC	CCC
-2.394	GCC	CCC	-2.325	CCA	ACA	-2.452	CTG	CAG
-2.394	TGA	TTA	-2.325	GGT	GCT	-2.458	AAA	ATA
-2.397	GGA	CGA	-2.325	ATC	AAC	-2.462	TTT	TTA
-2.420	TAT	GAT	-2.328	TAA	AAA	-2.463	ACC	AAC
-2.430	TGG	TTG	-2.341	ACT	CCT	-2.469	GTT	GCT
-2.454	CTA	CGA	-2.347	TGC	GGC	-2.472	GTA	CTA
-2.455	GTG	GCG	-2.353	CAG	CAC	-2.473	TCA	TCC
-2.460	TTG	TTC	-2.355	CCC	ACC	-2.478	GAT	GTT
-2.468	TAA	GAA	-2.357	TGG	TGC	-2.486	CCG	CAG
-2.468	TAA	TAC	-2.359	GCA	GCG	-2.488	AGC	ACC
-2.468	TAA	TTA	-2.374	TAG	GAG	-2.494	AAA	AGA

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.484	ACC	AAC	-2.375	CTT	ATT	-2.494	CTG	GTG
-2.486	AGG	AAG	-2.379	CTC	CTG	-2.497	TTA	TCA
-2.490	GCT	GAT	-2.384	AAT	GAT	-2.497	AAT	AGT
-2.492	ATC	TTC	-2.409	CAT	GAT	-2.500	AAC	AGC
-2.497	GGT	GCT	-2.410	ATT	TTT	-2.521	CAA	CTA
-2.498	CCT	CGT	-2.411	TGT	AGT	-2.522	GGT	GGC
-2.500	AAC	GAC	-2.413	GCC	CCC	-2.524	GTG	CTG
-2.503	GAC	GTC	-2.418	CTG	CTC	-2.524	GAT	TAT
-2.522	CAG	CCG	-2.421	GGG	TGG	-2.530	GTT	GTA
-2.525	GGA	GCA	-2.428	GTT	CTT	-2.532	GGC	AGC
-2.526	ATC	ATG	-2.435	TTC	GTC	-2.539	CCC	ACC
-2.528	TAC	AAC	-2.436	ATC	AGC	-2.540	TGT	TTT
-2.528	CAT	GAT	-2.457	TAT	TAG	-2.545	CTA	CAA
-2.537	TAT	TCT	-2.460	TGT	GGT	-2.545	CTA	CGA
-2.551	AGA	AAA	-2.462	AGT	CGT	-2.554	TTG	TGG
-2.570	ACG	GCG	-2.462	AGT	ACT	-2.559	TAG	TTG
-2.575	ACA	GCA	-2.463	ACT	AAT	-2.563	CCT	CTT
-2.583	TTC	GTC	-2.470	TCT	ACT	-2.570	ACT	ATT
-2.585	GAA	GGA	-2.479	TGG	GGG	-2.572	GAA	GTA
-2.588	ATA	AGA	-2.485	ACC	ACG	-2.578	CCA	CCC
-2.588	ATA	GTA	-2.488	GCA	GAA	-2.578	TGT	TGC
-2.590	TCC	TGC	-2.493	GGC	GTC	-2.578	TGT	TGA
-2.596	CCA	TCA	-2.513	TCC	TCG	-2.590	GTA	ATA
-2.617	AAC	ATC	-2.515	TTT	GTT	-2.590	GTA	GAA
-2.620	AGA	ACA	-2.517	TAG	TAC	-2.594	GTT	GTG
-2.620	GGC	GCC	-2.528	TTA	TTG	-2.597	TCT	TGT
-2.631	CCT	ACT	-2.528	TTA	TTT	-2.601	ATT	ATA
-2.634	GAT	CAT	-2.529	CAA	TAA	-2.625	GAT	GGT
-2.636	CAC	CAG	-2.535	ATG	AAG	-2.629	GGG	GGC
-2.644	ACC	CCC	-2.541	GGA	CGA	-2.639	TAT	TAA
-2.647	GCA	CCA	-2.553	ATT	ATC	-2.645	CTT	CCT
-2.650	AGC	AGA	-2.554	ACT	TCT	-2.645	CTT	GTT
-2.654	CAA	CGA	-2.556	TGG	AGG	-2.650	CCT	GCT
-2.656	GCG	CCG	-2.567	AGA	AGT	-2.657	GAA	GGA
-2.663	TTC	TGC	-2.573	TCA	ACA	-2.667	AAC	AAG
-2.666	ACT	CCT	-2.573	TCA	GCA	-2.676	GCA	ACA
-2.673	GTC	GGC	-2.576	CAT	CAA	-2.686	ACC	ACA
-2.686	ACA	AGA	-2.576	CAT	CAG	-2.689	TGG	TGC
-2.686	ACA	AAA	-2.579	GTA	GTG	-2.693	TAG	TAA
-2.688	ATT	GTT	-2.587	ATG	TTG	-2.693	TAG	TAT
-2.704	ACG	CCG	-2.588	TAC	GAC	-2.704	GGT	AGT
-2.705	TTC	TTG	-2.595	TGA	AGA	-2.705	AAT	AAA
-2.705	GAG	GGG	-2.595	ATA	AAA	-2.713	CCT	ACT
-2.706	ATA	ATG	-2.596	GGG	GCG	-2.714	GTC	ATC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.727	GTG	GAG	-2.598	GCT	GAT	-2.715	TCT	TCG
-2.737	AGT	ATT	-2.603	ACT	ACA	-2.724	ACT	ACA
-2.737	AGT	AGG	-2.608	AAT	AAG	-2.733	GGA	GGT
-2.739	TAC	TCC	-2.609	TCG	GCG	-2.735	CAC	CAA
-2.746	ATT	AGT	-2.622	TGT	TGG	-2.735	CAC	CGC
-2.750	TTC	TTA	-2.631	CGT	GGT	-2.736	GCT	ACT
-2.750	CCA	CAA	-2.642	CGA	CGG	-2.737	TGA	TCA
-2.762	TTT	CTT	-2.642	CCC	CCG	-2.737	GTT	CTT
-2.776	CCC	CAC	-2.645	AGA	GGA	-2.741	GAC	GAA
-2.798	TCT	TGT	-2.647	AAG	AAC	-2.762	AAC	AAA
-2.808	CAA	AAA	-2.664	ATA	AGA	-2.764	AGA	ACA
-2.810	ACT	GCT	-2.664	ATA	ATT	-2.767	GCC	GAC
-2.825	GTA	GAA	-2.668	TGC	AGC	-2.768	CTA	GTA
-2.825	GTG	TTG	-2.674	GAA	GAG	-2.773	AAG	AGG
-2.836	ACC	GCC	-2.674	GAA	CAA	-2.774	GCT	GTT
-2.846	CTC	CAC	-2.678	GTC	GTA	-2.775	GGG	CGG
-2.846	CTC	GTC	-2.681	CCT	GCT	-2.796	GAT	CAT
-2.849	ATC	GTC	-2.682	GTG	GGG	-2.807	TGG	GGG
-2.850	GGG	GCG	-2.693	AGG	GGG	-2.812	CCG	ACG
-2.851	AAC	TAC	-2.703	TGC	TGG	-2.819	CGC	CGA
-2.861	AGG	AGT	-2.703	GGA	GTA	-2.819	TGA	GGA
-2.864	CAT	CAG	-2.704	ATC	ATG	-2.822	CAC	CAG
-2.866	TTG	TTT	-2.709	GCT	GGT	-2.829	GAG	AAG
-2.866	TTG	TGG	-2.719	GAT	GAG	-2.841	TGT	TGG
-2.871	ATT	AAT	-2.720	ATT	ATG	-2.846	CAT	CTT
-2.887	CAC	GAC	-2.720	ATT	AGT	-2.847	TAG	TAC
-2.897	GTG	GGG	-2.732	TCA	TCT	-2.853	CCC	CGC
-2.897	ATC	AGC	-2.738	TGA	TGG	-2.854	CGT	CGC
-2.904	AGT	GGT	-2.740	AAA	AAT	-2.854	CGT	CGG
-2.905	AAC	ACC	-2.756	TTT	TTG	-2.869	ACC	ACG
-2.909	TCC	TAC	-2.764	GGA	GGT	-2.872	AAT	AAG
-2.911	GTT	GGT	-2.764	ATG	ATC	-2.878	TCA	TGA
-2.912	TAT	AAT	-2.772	TTC	TTG	-2.881	CAA	GAA
-2.934	ATG	ATT	-2.777	CAC	GAC	-2.882	CGG	CGC
-2.938	CTG	CAG	-2.777	CAC	CAG	-2.885	CAG	GAG
-2.952	AGC	ATC	-2.788	AAA	AAG	-2.887	GAA	AAA
-2.956	AGG	ACG	-2.794	TAT	TAA	-2.889	ATT	AGT
-2.968	CCT	GCT	-2.808	CGC	CGG	-2.892	CCG	GCG
-2.988	CCC	ACC	-2.810	CTT	GTT	-2.892	TGT	AGT
-2.991	CAA	CCA	-2.815	CTA	CTT	-2.898	TCC	ACC
-2.999	AGT	CGT	-2.815	CTA	GTA	-2.907	CTT	CTA
-3.001	CCA	ACA	-2.817	TGT	TGA	-2.925	GAA	GAG
-3.014	ATT	ATG	-2.819	ATA	TTA	-2.930	GAT	AAT
-3.014	GTG	CTG	-2.829	GGA	GCA	-2.936	CGC	GGC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-3.037	CCT	CAT	-2.836	GCC	GGC	-2.936	CGC	CTC
-3.037	GAG	CAG	-2.841	ACC	CCC	-2.937	TAA	AAA
-3.043	GCC	TCC	-2.849	CTC	GTC	-2.941	TGA	TGT
-3.060	AAG	AAC	-2.856	GAA	GAT	-2.955	TAC	GAC
-3.061	AGA	AGT	-2.864	TTA	ATA	-2.964	GAC	CAC
-3.061	AGG	AGC	-2.864	TTA	GTA	-2.967	TTA	GTA
-3.076	AAT	AAA	-2.865	AGG	AGC	-2.967	TTA	TAA
-3.076	AAT	AAG	-2.868	GGG	GGC	-2.967	TTA	TTC
-3.087	TGA	CGA	-2.871	ATC	CTC	-2.967	CCA	GCA
-3.104	AGT	AGA	-2.876	TAC	AAC	-2.969	CCT	CCA
-3.108	GTC	GCC	-2.882	CCT	CCG	-2.972	GAG	GTG
-3.108	GTC	CTC	-2.889	AGA	AGG	-2.973	AAG	ATG
-3.113	CCC	GCC	-2.890	ACC	TCC	-2.974	TCG	TAG
-3.134	GTT	CTT	-2.891	TGT	TAT	-2.978	TCC	TCG
-3.147	CTA	GTA	-2.894	AAA	GAA	-2.978	ATA	AAA
-3.153	CAG	CAC	-2.898	GGA	TGA	-2.978	ATA	AGA
-3.156	CCA	CGA	-2.903	GCT	GCG	-2.991	CTG	ATG
-3.156	AAT	CAT	-2.915	TTC	ATC	-3.000	GTT	GAT
-3.157	ATG	CTG	-2.919	CGT	CGG	-3.002	AGA	AGT
-3.166	GAG	GAC	-2.920	ATT	AAT	-3.011	TGC	GGC
-3.169	CTT	GTT	-2.930	GTC	GTG	-3.024	CAA	CAT
-3.177	ACT	AGT	-2.942	AGT	AGA	-3.024	TAA	TAG
-3.181	ATT	TTT	-2.942	GAT	GAA	-3.043	GCG	ACG
-3.183	AAG	AGG	-2.946	AAG	GAG	-3.045	TAT	TGT
-3.183	AAG	AAT	-2.948	GCC	GAC	-3.048	ACA	AGA
-3.220	AGC	GGC	-2.954	GTG	GAG	-3.060	GTA	GTC
-3.221	CAT	CAA	-2.966	TCG	ACG	-3.064	TGT	GGT
-3.226	CTG	GTG	-2.972	GGA	GGG	-3.065	TCC	TAC
-3.254	GAC	GAA	-2.981	TCT	TCA	-3.070	CTT	CGT
-3.254	TTT	TTG	-2.981	TCT	TCG	-3.073	GCC	GCA
-3.254	TTT	GTT	-2.985	AGC	GGC	-3.083	AGT	ATT
-3.276	TTC	ATC	-2.986	CCG	GCG	-3.085	CTC	GTC
-3.281	CAG	GAG	-2.986	CCG	ACG	-3.085	CAG	CAT
-3.289	GCA	TCA	-2.997	ACA	TCA	-3.088	GCT	GCA
-3.291	AAA	AGA	-3.001	ATA	CTA	-3.089	CCC	GCC
-3.313	AGA	AGC	-3.013	AAT	TAT	-3.093	GGA	AGA
-3.350	TTT	ATT	-3.017	TAT	TGT	-3.105	GTT	GGT
-3.365	ATG	TTG	-3.018	CCA	GCA	-3.105	GTT	TTT
-3.398	AAC	CAC	-3.018	GGT	GGA	-3.114	TTG	GTG
-3.413	TTA	TTC	-3.018	GTG	GTC	-3.116	TGA	TGC
-3.421	CTT	CAT	-3.020	ATG	CTG	-3.130	CTT	ATT
-3.423	GAA	CAA	-3.022	GCA	GCT	-3.136	TGC	TTC
-3.427	TCT	TAT	-3.038	ATT	CTT	-3.139	GAG	GAT
-3.443	AAT	ATT	-3.039	CAC	CAA	-3.140	ATT	ACT

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-3.443	AAT	TAT	-3.039	GGC	GGA	-3.142	GCT	TCT
-3.447	CAC	CAA	-3.040	TGC	TGA	-3.142	GCT	GGT
-3.447	CAC	AAC	-3.047	TTG	ATG	-3.146	GAC	GGC
-3.467	AGG	ATG	-3.047	TCC	TCA	-3.146	GAC	TAC
-3.467	AGG	TGG	-3.049	GTA	GAA	-3.146	GAC	GCC
-3.481	CAG	AAG	-3.051	AAC	GAC	-3.153	ACA	ACT
-3.483	AGC	ACC	-3.058	TAC	TAA	-3.153	ACA	ACC
-3.488	GAA	GCA	-3.067	CTA	ATA	-3.155	GGG	AGG
-3.532	ATG	ATC	-3.072	ATC	ATA	-3.156	TTT	TGT
-3.532	GAT	GCT	-3.075	TTT	ATT	-3.159	CTC	ATC
-3.551	GCT	GGT	-3.076	AGC	AGG	-3.160	TCC	GCC
-3.551	GCT	TCT	-3.076	AGC	CGC	-3.179	CCA	ACA
-3.557	GAA	GTA	-3.084	GTC	GGC	-3.179	CCA	CAA
-3.559	TTG	GTG	-3.104	TCA	TCC	-3.185	TCT	ACT
-3.581	CAC	CTC	-3.108	TCC	GCC	-3.199	GGA	GCA
-3.588	AGC	CGC	-3.111	CTG	ATG	-3.203	CAG	CAC
-3.602	GAC	GAG	-3.111	CTG	GTG	-3.205	CAC	AAC
-3.643	GAG	GAT	-3.112	AGG	ACG	-3.214	CAT	AAT
-3.643	GAG	GTG	-3.125	AGC	AGA	-3.215	GGT	CGT
-3.643	GAG	GCG	-3.134	GGC	GCC	-3.215	GGT	GCT
-3.649	AGA	ATA	-3.142	GCC	GCG	-3.215	GCA	GGA
-3.672	GTA	TTA	-3.147	GCA	GGA	-3.226	AGC	AGG
-3.672	CCG	ACG	-3.162	TGA	TGT	-3.227	GCC	TCC
-3.672	CCG	GCG	-3.165	TTG	TGG	-3.251	GAC	GAG
-3.684	CAA	CAC	-3.177	TTC	TTA	-3.256	CCT	CGT
-3.687	ATA	CTA	-3.198	ACA	CCA	-3.265	TCC	TGC
-3.689	GAC	GCC	-3.198	ACA	ACT	-3.265	TCC	TCA
-3.694	GCC	GGC	-3.204	TCT	TGT	-3.266	ATA	CTA
-3.696	ATC	CTC	-3.216	GCA	GCC	-3.271	CGA	CGT
-3.701	TTA	TTT	-3.223	TGA	TGC	-3.273	AAC	ACC
-3.701	TAT	TTT	-3.223	TGA	TAA	-3.277	TGG	AGG
-3.732	GAT	GAA	-3.229	GAT	GTT	-3.290	GGC	GTC
-3.733	CAG	CAT	-3.239	GTT	GTG	-3.291	TCT	TCA
-3.755	GCG	GGG	-3.239	GCT	GCA	-3.291	TCT	TTT
-3.765	ACT	AAT	-3.240	GGC	GGG	-3.291	TAC	TAG
-3.781	CAT	AAT	-3.241	CCA	CCT	-3.302	GGT	GGG
-3.802	AAA	ATA	-3.241	GGT	GGG	-3.304	CCA	CGA
-3.849	AAT	ACT	-3.242	GGG	GTG	-3.320	TCA	TAA
-3.849	CCA	GCA	-3.247	ACC	ACA	-3.332	CAT	GAT
-3.910	ACA	CCA	-3.258	TGT	TTT	-3.337	CAA	CCA
-3.921	AAG	CAG	-3.261	AGA	ACA	-3.340	GAG	GAC
-3.956	CTG	ATG	-3.263	ACG	TCG	-3.340	GAG	GGG
-3.980	CTT	ATT	-3.272	GTA	GGA	-3.343	TAA	GAA
-3.988	ACT	TCT	-3.281	TAC	TAG	-3.367	TTC	ATC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-4.016	AAG	ACG	-3.287	AGC	ATC	-3.369	GCA	GAA
-4.043	TTT	TAT	-3.292	CCA	CCC	-3.373	CCA	CCT
-4.053	AAA	AAT	-3.299	CCC	CCA	-3.379	TTT	ATT
-4.053	AAA	ACA	-3.299	CCC	GCC	-3.379	AAG	AAC
-4.063	CTA	CAA	-3.301	AAT	CAT	-3.392	CAA	CAC
-4.078	GTA	CTA	-3.309	AAG	CAG	-3.402	TTG	TAG
-4.092	ATA	TTA	-3.316	TCC	TGC	-3.417	ACT	AGT
-4.106	TTA	GTA	-3.320	GAC	GTC	-3.418	CTT	CAT
-4.116	GAA	GAT	-3.320	GAC	GAA	-3.438	TCA	ACA
-4.138	GAT	GAG	-3.335	CGA	CGT	-3.438	TGT	TCT
-4.171	ACC	TCC	-3.357	TAA	TAG	-3.440	AGT	ACT
-4.195	CAA	GAA	-3.366	AGA	ATA	-3.456	GCA	TCA
-4.203	AGT	TGT	-3.380	ACA	ACC	-3.457	GAG	CAG
-4.203	AGT	ACT	-3.472	TGG	TAG	-3.476	TAA	TAT
-4.205	GCA	GGA	-3.478	ATC	TTC	-3.485	TCG	GCG
-4.207	AAA	CAA	-3.481	AGC	ACC	-3.490	GAT	GCT
-4.250	GAA	GAC	-3.481	AAA	CAA	-3.495	CCC	CAC
-4.292	CAG	CTG	-3.484	AGA	CGA	-3.512	TTT	GTT
-4.329	TCG	ACG	-3.497	AGG	CGG	-3.513	GGC	CGC
-4.329	TCG	GCG	-3.508	ATT	ATA	-3.514	GCC	GGC
-4.399	AGC	TGC	-3.510	TGT	TCT	-3.514	GCC	ACC
-4.407	TCT	ACT	-3.526	TCG	TAG	-3.521	ACG	AAG
-4.407	TCT	GCT	-3.549	GAA	GAC	-3.540	TAG	TCG
-4.454	TTC	TAC	-3.560	GTA	GTC	-3.540	TAG	AAG
-4.474	TAC	TTC	-3.561	AGT	ATT	-3.547	GCT	CCT
-4.527	AAG	ATG	-3.562	AAC	AAG	-3.561	AAG	AAT
-4.583	TCC	GCC	-3.570	CAA	CAT	-3.572	TCA	GCA
-4.583	TCC	ACC	-3.575	CCT	CCA	-3.618	GAA	GCA
-4.592	TCA	ACA	-3.580	TAA	TAC	-3.620	GGT	TGT
-4.600	CAA	CTA	-3.586	CGG	CGC	-3.629	CGC	CGG
-4.613	AAA	AAC	-3.591	GGA	GGC	-3.630	TAA	TAC
-4.658	TTG	ATG	-3.599	CTT	CTG	-3.643	CTA	ATA
-4.692	CTC	ATC	-3.608	GAC	GAG	-3.647	TGC	AGC
-4.799	TTA	ATA	-3.612	CTC	ATC	-3.660	AAT	TAT
-4.827	ACA	TCA	-3.691	AGA	AGC	-3.661	GGA	GGC
-4.853	GCG	TCG	-3.694	ATA	ATC	-3.672	ATA	ATT
-4.959	TTT	TTA	-3.701	GCC	GCA	-3.696	TCT	GCT
-5.087	ACC	AGC	-3.701	TCA	TGA	-3.698	GAA	GAC
-5.293	CAA	CAT	-3.705	AAA	ATA	-3.735	TGA	AGA
-5.579	ATT	CTT	-3.706	AAT	ACT	-3.736	GCG	CCG
			-3.745	AAC	TAC	-3.745	GAG	GCG
			-3.766	AGG	ATG	-3.774	GGG	TGG
			-3.789	CGC	CGA	-3.779	CAG	AAG
			-3.797	TCA	TAA	-3.785	GAA	GAT

Log odds score	TMC_{coding} wild type	TMC_{coding} mutant	Log odds score	TMC_{+1} wild type	TMC_{+1} mutant	Log odds score	TMC_{+2} wild type	TMC_{+2} mutant
			-3.822	ACG	CCG	-3.801	ATG	TTG
			-3.838	AAA	TAA	-3.802	CAT	CCT
*note, since TMCs in these spectra are based only on nonsynonymous mutations, the first coding frame can only have 476 possible mutation classes.			-3.868	TAA	TAT	-3.802	TAC	AAC
			-3.884	AAG	ACG	-3.807	TTG	ATG
			-3.897	TCT	TAT	-3.813	GTC	CTC
			-3.897	AGT	TGT	-3.813	GTC	TTC
			-3.916	TGA	TCA	-3.823	ACT	TCT
			-3.932	GTT	GGT	-3.834	CAC	GAC
			-3.954	AGA	TGA	-3.849	TTT	TCT
			-3.956	TGC	TTC	-3.855	GGA	CGA
			-3.965	GTA	GTT	-3.883	TTA	ATA
			-3.976	CAA	CAC	-3.890	TCG	ACG
			-4.028	AGG	TGG	-3.896	CAG	CTG
			-4.028	CGA	CGC	-3.898	TGC	TGG
			-4.035	GAG	GCG	-3.910	GTG	TTG
			-4.055	TTT	TGT	-3.918	GGC	TGC
			-4.055	TTT	TTA	-3.920	GCC	CCC
			-4.060	GAA	GCA	-3.927	TTC	TGC
			-4.070	TGG	TCG	-3.927	GAG	TAG
			-4.110	CTT	CTA	-3.945	TAG	GAG
			-4.174	AGC	TGC	-3.958	AGA	AGC
			-4.204	TGA	TTA	-3.964	ACA	AAA
			-4.244	TGC	TAC	-3.967	CAC	CTC
			-4.255	AAC	AAA	-3.976	GTA	GTT
			-4.255	AAC	ATC	-4.014	AGC	CGC
			-4.271	TGG	TTG	-4.014	AGC	ATC
			-4.274	TAA	TCA	-4.024	GCG	TCG
			-4.319	CTA	CTC	-4.026	TAT	GAT
			-4.381	TTC	TGC	-4.026	TAT	TTT
			-4.389	TAG	TGG	-4.030	CAG	CCG
			-4.389	TAG	TTG	-4.036	TAA	TCA
			-4.399	AAT	AAA	-4.062	GCA	GCT
			-4.399	AAT	ATT	-4.067	CCT	CAT
			-4.403	TAT	TTT	-4.072	AAG	ACG
			-4.472	AAG	TAG	-4.086	GCT	GAT
			-4.474	TTA	TGA	-4.100	GTC	GAC
			-4.625	GTT	GTA	-4.101	TCT	TAT
			-4.668	TAC	TGC	-4.131	TCA	TCT
			-4.668	TAC	TCC	-4.154	CTA	CTT
			-4.685	AAA	AAC	-4.171	AAT	ACT
			-4.729	GAG	GTG	-4.181	AAA	AAG
			-4.937	TGC	TCC	-4.181	AAA	AAC
			-4.967	TAA	TGA	-4.192	GGA	TGA
			-4.983	AAG	ATG	-4.206	GGC	GGG

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
			-5.082	TAG	TCG	-4.207	TAC	TCC
			-5.091	AAA	ACA	-4.213	ATC	TTC
			-5.187	TCC	TAC	-4.213	ATC	AAC
			-5.244	TTG	TAG	-4.237	GAA	CAA
			-5.354	AAC	ACC	-4.245	GCA	CCA
			-5.557	CTC	CTA	-4.287	AGT	CGT
						-4.287	AGT	TGT
						-4.350	GAC	GTC
						-4.359	TTT	TAT
						-4.365	ATA	GTA
						-4.365	ATA	ATC
						-4.365	ATA	TTA
						-4.393	ATT	GTT
						-4.404	AAA	AAT
						-4.404	AAA	CAA
						-4.404	AAA	ACA
						-4.431	TAT	AAT
						-4.477	AAG	TAG
						-4.478	ACC	AGC
						-4.574	GAA	TAA
						-4.577	TTA	TTT
						-4.577	AAT	ATT
						-4.626	CTC	CAC
						-4.657	ACA	CCA
						-4.657	ACA	TCA
						-4.657	CGA	CGC
						-4.861	AGG	CGG
						-4.874	AGA	CGA
						-4.906	ATC	AGC
						-4.906	ATC	GTC
						-4.931	AGC	TGC
						-5.065	AAC	CAC
						-5.065	AAC	ATC
						-5.086	ATT	AAT
						-5.086	ATT	TTT
						-5.086	ATT	CTT
						-5.124	TAT	TCT
						-5.170	AAG	CAG
						-5.171	ACC	TCC
						-5.199	GTC	GGC
						-5.209	ACT	AAT
						-5.270	AAT	CAT
						-5.313	TTC	TAC
						-5.351	ACA	GCA

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
						-5.554	AGG	TGG
						-5.554	AGG	GGG
						-5.724	CTC	CGC
						-5.790	AAA	GAA
						-5.790	AAA	TAA
						-6.260	AGA	TGA
						-6.260	AGA	GGA

Table A.2: Trinucleotide mutation classes composing the M_{dbSNP} nonsynonymous mutation spectrum

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
0.542	CGT	CAT	0.456	CGT	TGT	0.625	CGT	CAT
0.526	ACG	ATG	0.445	CGA	TGA	0.444	ACG	ATG
0.302	CCG	CTG	0.394	CCG	CCA	0.283	ACG	ACA
0.238	CGT	TGT	0.228	CGG	TGG	0.242	TCG	TCA
0.168	CGG	CAG	0.215	TCG	TCA	0.239	CGG	CAG
0.110	TCG	TTG	0.165	ACG	ACA	0.185	CCG	CCA
0.020	CGA	CAA	0.095	CGC	TGC	0.162	CGC	CAC
0.010	GCG	GTG	0.062	GCG	GCA	0.044	GCG	GCA
-0.082	CGC	CAC	-0.595	GTA	ATA	-0.093	CCG	CTG
-0.170	CGG	TGG	-0.612	CTA	CTG	-0.176	CGA	CAA
-0.284	CGC	TGC	-0.696	GGG	AGG	-0.308	TCG	TTG
-0.327	CGA	TGA	-0.716	GGT	AGT	-0.432	GCG	GTG
-0.332	GTA	ATA	-0.721	GTG	ATG	-0.517	ATG	ACG
-0.490	CAT	CGT	-0.741	GGC	AGC	-0.699	TAT	TGT
-0.490	GTC	ATC	-0.775	GAT	AAT	-0.716	CAC	CAT
-0.518	GTT	ATT	-0.775	GAA	AAA	-0.760	CAT	CGT
-0.574	ATA	ACA	-0.784	GTT	ATT	-0.799	AAC	AAT
-0.648	ATA	GTA	-0.800	GTC	ATC	-0.802	TAC	TAT
-0.768	TCT	TTT	-0.812	TAC	TAT	-0.872	TTG	TCG
-0.828	AGG	AAG	-0.822	GGA	AGA	-0.874	TAG	TGG
-0.860	GGT	AGT	-0.828	CAA	CAG	-0.893	GAC	GAT
-0.861	AGA	AAA	-0.846	TAG	CAG	-0.897	TAC	TGC
-0.868	ATA	ATG	-0.873	GAG	AAG	-0.940	CCC	CCT
-0.886	TGT	CGT	-0.911	GAC	GAT	-0.983	GTC	GTT
-0.896	ACG	GCG	-0.916	CAC	CAT	-0.989	CAG	CGG
-0.910	GTA	GCA	-0.918	CCA	CCG	-0.989	AAC	AGC
-0.960	TGG	CGG	-0.942	GAC	AAC	-0.996	ACC	ACT
-0.969	AGA	GGA	-0.952	TAC	CAC	-1.001	TTC	TTT
-0.975	ATG	GTG	-0.953	TAT	TAC	-1.011	AGT	AGC
-0.982	GTG	ATG	-0.960	GTA	GTG	-1.018	GTG	GCG
-0.983	GGC	AGC	-0.982	GCA	ACA	-1.025	GAG	GAA
-1.004	CTT	TTT	-1.000	ATA	GTA	-1.039	CCA	CCG
-1.019	GGG	AGG	-1.014	GCA	GCG	-1.049	AAG	AAA
-1.023	GTT	GCT	-1.032	ATG	GTG	-1.064	AAT	AGT
-1.025	CAC	CGC	-1.044	TAT	CAT	-1.066	AGG	AGA
-1.040	AGT	AAT	-1.060	AAC	AAT	-1.066	ACA	ACG
-1.045	GCA	ACA	-1.076	CTT	TTT	-1.066	CTC	CTT
-1.049	TAT	TGT	-1.077	TCC	TCT	-1.096	ACG	AGG
-1.051	AAT	AGT	-1.087	ATA	ATG	-1.099	CAC	CGC
-1.057	ACT	GCT	-1.101	ACA	ACG	-1.105	TAG	TAA
-1.057	ATC	GTC	-1.103	GCC	ACC	-1.107	ACC	ATC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-1.067	ATG	ACG	-1.111	CGA	GGA	-1.113	CAG	CAA
-1.077	ATT	GTT	-1.115	CTA	TTA	-1.114	GCC	GCT
-1.092	GCC	ACC	-1.148	CTG	TTG	-1.115	GCA	GCG
-1.105	CTA	CCA	-1.151	GAT	GAC	-1.121	TAT	TAC
-1.122	ACA	ATA	-1.163	CAT	CAC	-1.139	CAA	CAG
-1.124	GCT	ACT	-1.167	GGG	GGA	-1.157	ATC	ACC
-1.130	GGA	AGA	-1.200	CTC	CTT	-1.174	AGC	AGT
-1.132	AGG	GGG	-1.209	CTC	TTC	-1.175	CGC	CGT
-1.135	CGT	CCT	-1.211	CCC	CCT	-1.186	TGC	TGT
-1.136	ACC	GCC	-1.247	GCC	GCT	-1.188	GGT	GGC
-1.151	CCT	TCT	-1.249	TGT	CGT	-1.188	ATA	ACA
-1.153	ACT	ATT	-1.253	TTC	TTT	-1.200	GGC	GGT
-1.156	CAA	CGA	-1.262	ATT	GTT	-1.237	CTA	CTG
-1.164	CCC	TCC	-1.276	GCT	ACT	-1.238	TGT	TAT
-1.164	AGT	GGT	-1.279	ATC	GTC	-1.261	TCA	TCG
-1.167	TGC	CGC	-1.301	AAG	AAA	-1.275	GAT	GAC
-1.171	GGT	GAT	-1.318	GAA	GAG	-1.276	AGT	AAT
-1.184	ACA	GCA	-1.323	CCT	TCT	-1.307	CAT	CAC
-1.187	TGT	TAT	-1.330	ACC	ACT	-1.320	TCC	TTC
-1.237	CTC	TTC	-1.336	TCA	TCG	-1.326	AGA	AAA
-1.237	GAA	AAA	-1.336	GCG	ACG	-1.332	TGT	TGC
-1.253	CCT	CTT	-1.369	GGT	GGC	-1.333	ATC	ATT
-1.275	GCG	ACG	-1.387	CAG	CAA	-1.342	TCC	TCT
-1.326	GTC	GCC	-1.389	TTG	CTG	-1.347	ACA	ATA
-1.331	GAC	AAC	-1.394	TAA	CAA	-1.348	CCC	CTC
-1.336	TCT	CCT	-1.403	ACT	GCT	-1.358	AAT	AAC
-1.338	AGC	GGC	-1.404	CCC	TCC	-1.362	GAC	GGC
-1.343	GCT	GTT	-1.426	TGG	CGG	-1.372	TTA	TCA
-1.346	CAC	TAC	-1.435	TTA	TTG	-1.390	GCC	GTC
-1.350	GAG	AAG	-1.441	TCG	TCC	-1.394	CTG	CCG
-1.358	ACC	ATC	-1.446	GGC	GGT	-1.406	GGT	GAT
-1.379	AGC	AAC	-1.452	CGT	CGC	-1.406	AAG	AGG
-1.381	CAT	TAT	-1.456	CAG	TAG	-1.433	AGG	AAG
-1.403	CAG	CGG	-1.456	ACA	GCA	-1.435	GTC	GCC
-1.460	GTG	GCG	-1.457	AGG	AGA	-1.436	TCT	TCC
-1.463	GGG	GAG	-1.463	CGA	CGG	-1.448	ATA	ATG
-1.471	CCA	TCA	-1.477	GTT	TTT	-1.454	GGG	GAG
-1.479	GGA	GAA	-1.480	TGC	TGT	-1.464	GTA	GTG
-1.481	CCG	CGG	-1.483	GAG	GAA	-1.472	ATG	ATA
-1.527	ACG	AAG	-1.490	TAG	TAA	-1.482	CGC	CTC
-1.547	TCA	CCA	-1.491	AGT	AGC	-1.510	TAA	TGA
-1.550	TTG	TCG	-1.505	AGA	AGG	-1.510	ACT	ACC
-1.564	AAC	GAC	-1.507	GTT	CTT	-1.523	CGC	CCC
-1.569	CAC	CCC	-1.509	TGC	CGC	-1.524	GGG	GGA

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-1.570	ATT	ACT	-1.517	CGA	AGA	-1.537	CAA	CGA
-1.583	GCA	GTA	-1.518	AAA	AAG	-1.558	CGT	CCT
-1.594	AAC	AGC	-1.530	GTG	GTA	-1.560	AGC	AAC
-1.598	CGG	GGG	-1.541	AGT	GGT	-1.564	GGC	GAC
-1.598	AGT	ACT	-1.544	TTT	TTC	-1.569	TGC	TAC
-1.603	GAT	AAT	-1.546	CAC	TAC	-1.579	CGT	CTT
-1.603	GAT	GGT	-1.550	TCC	CCC	-1.583	TAG	TAC
-1.606	TAT	CAT	-1.562	ACG	ACC	-1.597	CCG	CCT
-1.610	AAA	GAA	-1.566	GTC	GTT	-1.613	GAA	GAG
-1.611	AAG	AGG	-1.580	CTT	CTC	-1.617	GAT	GGT
-1.611	CGT	GGT	-1.588	ATC	ATT	-1.645	CGT	CGC
-1.618	CCG	CAG	-1.592	CAT	TAT	-1.652	CGG	CCG
-1.624	TCC	CCC	-1.598	CTG	CTA	-1.653	GAG	GGG
-1.632	GCC	GTC	-1.600	AAT	AAC	-1.674	GTG	GTA
-1.636	CGC	CTC	-1.610	TTC	CTC	-1.680	AAA	AGA
-1.636	TCC	TTC	-1.613	GCG	GCC	-1.680	GGA	GAA
-1.636	AAA	AGA	-1.616	AGC	AGT	-1.687	CAC	CCC
-1.640	ATG	ATA	-1.639	AGC	GGC	-1.702	ACG	ACC
-1.642	TAC	CAC	-1.642	CGG	GGG	-1.702	ACG	ACT
-1.670	CCA	CTA	-1.655	CGC	CGT	-1.703	TTC	TCC
-1.675	GTT	TTT	-1.658	AGA	GGA	-1.709	TGG	TGA
-1.697	CTT	CCT	-1.667	GGA	GGG	-1.715	AAA	AAG
-1.700	GCG	GGG	-1.679	GCG	GCT	-1.729	TGG	TAG
-1.702	AAG	GAG	-1.682	CGG	AGG	-1.732	CTG	CTA
-1.702	CGT	CTT	-1.683	TGA	CGA	-1.743	TCG	TCC
-1.705	ACT	CCT	-1.696	ACG	GCG	-1.759	TAG	TAT
-1.707	TTA	TCA	-1.707	AAC	GAC	-1.764	GCG	GCT
-1.710	GTA	CTA	-1.711	GAT	TAT	-1.767	CCG	CGG
-1.710	CCG	TCG	-1.712	AGG	GGG	-1.772	TCT	TTT
-1.717	CCC	CTC	-1.720	AAA	GAA	-1.786	CCA	CTA
-1.720	GGC	GAC	-1.724	TCT	TCC	-1.798	CGA	CCA
-1.765	ACG	AGG	-1.729	TGG	TGA	-1.804	ACT	ATT
-1.765	AAT	GAT	-1.733	AAG	GAG	-1.804	CCG	CCC
-1.778	ACT	AGT	-1.733	ATG	ATA	-1.806	TCT	TGT
-1.782	AGA	ACA	-1.734	CAA	TAA	-1.813	CTC	CCC
-1.783	CGG	CTG	-1.743	ACT	ACC	-1.817	ACA	AGA
-1.789	CTG	CCG	-1.746	GCT	GCC	-1.819	AGA	AGG
-1.807	TTC	CTC	-1.753	CCC	CCG	-1.823	GGA	GGG
-1.808	TGC	TAC	-1.759	ACG	ACT	-1.825	CCT	CCC
-1.812	ATC	ACC	-1.777	TCT	CCT	-1.834	TTG	TTA
-1.816	TGG	TGA	-1.785	GTT	GTC	-1.872	CCT	CTT
-1.838	CAA	TAA	-1.789	CAC	CAG	-1.881	ATT	ATC
-1.856	CGT	AGT	-1.789	GAC	GAG	-1.884	CTC	CTG
-1.868	TAC	TGC	-1.794	CCT	CCC	-1.886	TTT	TTC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-1.872	CTA	GTA	-1.795	TTT	CTT	-1.889	TCA	TTA
-1.877	CGA	GGA	-1.799	ACC	GCC	-1.894	CTA	CCA
-1.889	GGT	TGT	-1.803	CCA	TCA	-1.911	GCT	GCC
-1.893	CAC	CAG	-1.808	ATT	ATC	-1.915	GCA	GTA
-1.896	GAC	GGC	-1.828	TAC	TAG	-1.918	GAA	GGA
-1.898	CGC	AGC	-1.837	TAT	TGT	-1.918	TCG	TGG
-1.905	ACC	CCC	-1.850	ACG	AGG	-1.925	CTT	CTC
-1.905	CAA	GAA	-1.850	TTG	TTA	-1.932	CGG	CGA
-1.905	CAA	AAA	-1.853	GTC	TTC	-1.935	TTG	TGG
-1.915	TTT	CTT	-1.860	CCG	TCG	-1.936	CAG	CAC
-1.920	TCA	TTA	-1.870	ACC	CCC	-1.978	AAC	ACC
-1.930	TCG	TGG	-1.871	GGT	TGT	-1.986	CCC	CCG
-1.946	CCC	GCC	-1.886	GTG	CTG	-1.991	CTT	CCT
-1.955	CCT	GCT	-1.887	GTC	CTC	-1.992	ACA	AAA
-1.958	CAG	CAC	-1.898	CCC	GCC	-1.992	CGG	CTG
-1.965	TGT	TCT	-1.911	CTC	CTG	-1.994	AAG	AAC
-1.969	GCG	TCG	-1.911	CGT	GGT	-1.996	GCC	GGC
-1.969	GCG	GAG	-1.915	GGT	CGT	-2.009	CCA	CCC
-1.971	CCA	ACA	-1.916	AAT	GAT	-2.013	GCA	GGA
-1.977	TCC	GCC	-1.924	GGC	GGG	-2.021	TAC	TAA
-1.989	CTC	CCC	-1.929	CCG	CCC	-2.024	GAG	GAC
-1.991	GTG	CTG	-1.949	AAC	ACC	-2.024	ACC	AGC
-1.995	AGA	ATA	-1.949	AAC	AAG	-2.025	TCA	TGA
-1.995	AGG	AGT	-1.953	AGA	ACA	-2.035	GGT	GTT
-2.004	CGC	CCC	-1.966	AGC	AGG	-2.042	AAC	AAA
-2.008	TCT	TGT	-1.967	TCA	CCA	-2.043	GAC	GAG
-2.010	CGA	CCA	-1.971	CCA	CCC	-2.052	GTA	GCA
-2.031	AGC	ACC	-1.992	TGA	TGG	-2.058	CAC	CAG
-2.034	TGG	TAG	-1.993	GTA	CTA	-2.064	CAG	CAT
-2.037	TTA	GTA	-1.994	GAG	CAG	-2.065	TTA	TTG
-2.037	AGT	ATT	-1.994	GAC	CAC	-2.065	CCC	CCA
-2.039	CTA	ATA	-1.995	TCG	CCG	-2.067	GCG	GGG
-2.044	GCT	CCT	-2.001	GCC	CCC	-2.076	AAC	AAG
-2.047	GAG	GGG	-2.008	GTT	GTG	-2.078	CTC	CGC
-2.053	GCA	TCA	-2.010	GGT	GGG	-2.081	GTC	GAC
-2.054	GAC	GAG	-2.013	TAA	TAG	-2.084	ACC	AAC
-2.056	CAA	CAC	-2.014	AGT	ACT	-2.101	AGT	ACT
-2.071	CAC	GAC	-2.028	TGT	TGC	-2.104	TGT	TCT
-2.082	TTA	TTC	-2.037	ATC	ATG	-2.109	TGA	TAA
-2.082	GGT	GTT	-2.044	AAC	AAA	-2.114	GGC	GTC
-2.087	TTG	GTG	-2.047	CGC	GGC	-2.116	CCC	CGC
-2.091	CGC	GGC	-2.050	GCT	TCT	-2.118	CGA	CGG
-2.091	CAG	TAG	-2.051	TAC	TGC	-2.128	TGA	TGG
-2.100	GTA	TTA	-2.056	CCG	CCT	-2.139	GAC	GAA

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.109	ACA	AAA	-2.068	GAT	CAT	-2.151	AGA	ACA
-2.120	CAG	GAG	-2.082	GCC	TCC	-2.159	CTG	CTC
-2.122	GTT	CTT	-2.091	CTT	GTT	-2.163	TAA	TAG
-2.138	TCG	CCG	-2.095	CTA	ATA	-2.167	GGG	GGT
-2.138	GAA	GGA	-2.097	TTC	TTG	-2.171	TTG	TTC
-2.139	CAT	CAG	-2.102	TAC	TAA	-2.178	TCC	TCA
-2.145	TGC	TCC	-2.103	CAA	CAC	-2.180	TGC	TGA
-2.145	GCC	TCC	-2.107	GCA	CCA	-2.183	ACC	ACA
-2.155	ACA	CCA	-2.110	CAG	CAC	-2.187	ACT	AGT
-2.157	CTT	GTT	-2.113	TCG	TCT	-2.207	GGG	GGC
-2.162	AAC	AAA	-2.114	GCT	CCT	-2.215	GTA	GTT
-2.166	AGA	AGT	-2.115	CTG	GTG	-2.215	GTT	GTC
-2.166	AGG	ATG	-2.137	CAC	AAC	-2.216	CGC	CGG
-2.182	GGC	GCC	-2.149	ACT	AGT	-2.218	ACC	CCC
-2.187	GAA	CAA	-2.169	GGG	CGG	-2.218	AGC	ATC
-2.188	AGC	AGG	-2.170	ACC	AAC	-2.233	GGC	GGG
-2.189	TTC	TCC	-2.175	CAT	AAT	-2.244	AGG	AGC
-2.199	AAC	AAG	-2.175	CGC	AGC	-2.247	GTT	TTT
-2.219	TTT	TCT	-2.176	ACA	AGA	-2.252	GAG	CAG
-2.222	CGA	CTA	-2.190	GAG	GAC	-2.254	TCG	TCT
-2.226	TCC	TGC	-2.195	GAC	GAA	-2.254	ACC	ACG
-2.230	GCT	GAT	-2.202	GGA	TGA	-2.258	TTC	TGC
-2.231	AGG	ACG	-2.203	TAG	TAC	-2.267	CAC	CTC
-2.235	GGG	CGG	-2.210	CGA	CGC	-2.282	GTC	GGC
-2.250	GCT	TCT	-2.213	TTA	CTA	-2.289	AGC	AGG
-2.252	CTC	GTC	-2.222	GAT	GAG	-2.290	CCG	CAG
-2.253	GAG	CAG	-2.223	CTC	GTC	-2.296	GTG	CTG
-2.254	GTA	GGA	-2.223	CCC	CCA	-2.306	CCT	CGT
-2.256	ACA	AGA	-2.234	GGT	GTT	-2.320	GTC	GTG
-2.259	TTG	TTT	-2.242	CGG	CGA	-2.322	CAC	CAA
-2.260	GAT	GAA	-2.245	GCG	CCG	-2.333	GCT	GGT
-2.260	CTG	GTG	-2.247	ATC	AAC	-2.335	TAC	TCC
-2.261	GCC	CCC	-2.252	GAC	TAC	-2.341	GGT	GCT
-2.267	AGC	AGA	-2.255	CAC	GAC	-2.343	GTG	GTT
-2.276	CCA	CAA	-2.259	GTC	GTG	-2.343	CAA	CCA
-2.296	TCT	GCT	-2.260	GGC	GGA	-2.362	ACA	ACC
-2.299	TGC	TTC	-2.270	TCC	TCG	-2.363	AGT	AGG
-2.299	CAA	CAT	-2.277	CTG	CTC	-2.369	GGC	GGA
-2.300	AGA	AGC	-2.278	CGG	CGT	-2.370	TCT	TAT
-2.300	AGG	AGC	-2.281	ATC	ATA	-2.371	AAG	AAT
-2.303	GTG	TTG	-2.281	GGG	TGG	-2.373	TGC	TCC
-2.305	TCG	GCG	-2.289	GGC	TGC	-2.377	GCC	GCG
-2.308	AGC	ATC	-2.307	CAA	AAA	-2.385	CGT	GGT
-2.312	CCT	CAT	-2.314	ACT	ACG	-2.388	ATG	ATT

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.320	GGT	CGT	-2.322	CGC	CGG	-2.390	AAT	AAG
-2.323	CCC	CAC	-2.327	AGG	AGC	-2.394	TTG	TTT
-2.325	ACC	AAC	-2.328	CCC	ACC	-2.396	TTC	TTG
-2.329	GTT	GGT	-2.328	GCG	TCG	-2.400	TAA	TCA
-2.335	AAG	AAC	-2.329	GTA	GTC	-2.401	GCC	GAC
-2.336	CCT	ACT	-2.329	ACG	AAG	-2.402	CCA	CAA
-2.337	TTT	TTG	-2.332	TTC	TTA	-2.406	GGC	GCC
-2.344	CCC	CGC	-2.333	ACC	ACG	-2.409	GGG	GCG
-2.347	TGG	TGC	-2.342	GCA	TCA	-2.419	AGC	ACC
-2.353	GTC	TTC	-2.344	CCT	GCT	-2.431	TGT	TTT
-2.362	GCC	GGC	-2.347	CTA	GTA	-2.432	TGG	TGC
-2.362	CAT	CAA	-2.347	CTG	ATG	-2.432	AGT	ATT
-2.362	AAC	ACC	-2.352	ATC	AGC	-2.437	AGG	AGT
-2.362	CAT	AAT	-2.353	GCC	GCA	-2.441	ATA	AAA
-2.366	GAC	GAA	-2.355	GGG	GGT	-2.446	TAC	TAG
-2.374	CAC	CAA	-2.359	AGT	ATT	-2.449	GCA	GAA
-2.376	CCA	GCA	-2.359	AGT	AGA	-2.450	GAG	GAT
-2.381	GCA	CCA	-2.360	GAA	TAA	-2.457	GCG	GCC
-2.384	GCT	GGT	-2.365	CAG	CAT	-2.462	CGG	GGG
-2.386	CCT	CGT	-2.367	GGT	GGA	-2.465	CCT	CAT
-2.389	CAG	CAT	-2.371	TGC	TGG	-2.469	ATT	ACT
-2.397	ACA	TCA	-2.384	GCA	GCC	-2.470	TTT	TCT
-2.400	TCA	TGA	-2.388	AGC	ACC	-2.470	TTT	TGT
-2.400	TCG	ACG	-2.404	GAT	GAA	-2.470	TTA	TTT
-2.402	AAA	CAA	-2.405	AGG	AGT	-2.474	TCC	TGC
-2.405	CAT	GAT	-2.405	ATT	ATG	-2.474	TCC	TAC
-2.409	CAA	CCA	-2.409	CTT	CTG	-2.475	ATG	ATC
-2.413	GGC	GTC	-2.416	TAG	TAT	-2.475	ATG	AAG
-2.415	CAG	AAG	-2.417	AAG	AAT	-2.479	GCC	GCA
-2.416	ACC	AGC	-2.422	ACT	ACA	-2.485	TGC	TGG
-2.419	GAG	GAT	-2.424	ACA	AAA	-2.487	GGG	GTG
-2.428	TTG	TTC	-2.429	ATC	CTC	-2.492	TGG	TGT
-2.430	TTC	TTG	-2.433	GAG	TAG	-2.495	GAC	GCC
-2.432	GAT	CAT	-2.441	GGT	GCT	-2.497	GTG	GTC
-2.434	GGA	GTA	-2.442	TTT	TTG	-2.501	GAA	CAA
-2.447	TGT	TTT	-2.450	CGT	AGT	-2.503	GTT	CTT
-2.452	TGG	TTG	-2.455	TTA	TTC	-2.509	TGC	TTC
-2.454	GCG	CCG	-2.455	TAA	TGA	-2.510	CGA	GGA
-2.458	ACG	TCG	-2.458	TCG	TGG	-2.512	CTG	CTT
-2.459	AAT	ACT	-2.462	CCT	ACT	-2.513	ACG	AAG
-2.469	GGG	TGG	-2.463	CAT	CAA	-2.520	CCA	CGA
-2.469	GGG	GCG	-2.465	TGG	TGC	-2.520	CTC	CAC
-2.478	GTC	CTC	-2.468	GCG	GGG	-2.529	TCA	TAA
-2.488	AAT	AAG	-2.474	AGC	ATC	-2.532	GAC	GTC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.488	GAG	GAC	-2.476	AAT	ATT	-2.537	AGC	AGA
-2.490	ACT	TCT	-2.481	ACC	AGC	-2.538	CCC	CAC
-2.490	ACT	AAT	-2.481	ACT	CCT	-2.539	CGT	CGA
-2.496	CAT	CCT	-2.483	CAA	GAA	-2.545	AGA	ATA
-2.496	ATG	ATT	-2.488	CCA	GCA	-2.552	GGT	GGG
-2.505	TTG	TGG	-2.492	AGC	AGA	-2.553	AAC	ATC
-2.505	ATC	ATG	-2.492	AGA	ATA	-2.558	AAA	ACA
-2.516	ACC	TCC	-2.494	GTG	GTT	-2.559	TAT	TTT
-2.519	GAT	TAT	-2.494	GTG	TTG	-2.563	GTA	GGA
-2.520	ATA	TTA	-2.501	ATT	AGT	-2.568	CGG	CGT
-2.520	ATA	CTA	-2.501	CTA	CTT	-2.568	CTT	CGT
-2.521	CTT	CAT	-2.502	CAC	CAA	-2.571	ATG	CTG
-2.522	AAA	AAC	-2.515	ACA	CCA	-2.576	ATC	ATG
-2.524	GGA	GCA	-2.520	AGG	ATG	-2.579	TGA	TCA
-2.526	CTC	ATC	-2.522	CAG	GAG	-2.582	GTC	GTA
-2.530	GGG	GTG	-2.525	AAA	ACA	-2.590	GTT	GAT
-2.531	GCA	GAA	-2.527	AAT	ACT	-2.592	ACT	AAT
-2.534	GGT	GCT	-2.550	GTG	GTC	-2.597	GGA	GCA
-2.534	TTA	TTT	-2.550	AGG	ACG	-2.604	AAT	ACT
-2.545	TAC	TTC	-2.552	GTA	TTA	-2.604	TGT	TGA
-2.547	GTC	GAC	-2.567	CCT	CCG	-2.613	TAT	TCT
-2.547	TCT	ACT	-2.569	CTC	ATC	-2.635	CAA	CAC
-2.548	TAT	TTT	-2.571	TCT	ACT	-2.636	GTT	GCT
-2.548	AAT	CAT	-2.573	TAA	TAC	-2.637	AGG	ATG
-2.549	AAG	CAG	-2.576	ACC	ACA	-2.638	CGA	CTA
-2.558	AAC	CAC	-2.576	TGT	TAT	-2.641	CTG	CAG
-2.558	TCC	ACC	-2.578	TTG	TTC	-2.645	ATC	AGC
-2.562	GAA	GAC	-2.579	CGG	CGC	-2.650	GCT	GTT
-2.564	CTG	ATG	-2.585	TAT	TAA	-2.650	GAA	GAC
-2.564	TGC	AGC	-2.590	CCA	ACA	-2.650	GTA	CTA
-2.565	GAT	GAG	-2.593	GAA	CAA	-2.650	AGA	AGC
-2.567	TCA	GCA	-2.602	TCT	GCT	-2.664	ATA	TTA
-2.579	CCG	GCG	-2.604	ATC	TTC	-2.677	GCA	GCC
-2.580	CCA	CGA	-2.606	ATT	CTT	-2.679	GTG	GGG
-2.584	CGG	CCG	-2.612	TGG	TGT	-2.685	GTT	GTG
-2.612	GGC	TGC	-2.615	TGT	TGG	-2.685	GTT	GGT
-2.614	ATG	CTG	-2.619	CTA	CTC	-2.687	CTA	CTC
-2.618	GAA	GCA	-2.627	AGT	TGT	-2.691	GAA	TAA
-2.620	GCC	GAC	-2.630	GCG	GAG	-2.691	CAG	GAG
-2.621	TCT	TAT	-2.632	AAG	CAG	-2.692	ATT	AGT
-2.631	AGT	CGT	-2.638	ATA	AGA	-2.710	TAA	TTA
-2.633	GCA	GGA	-2.642	AAC	CAC	-2.710	TAA	GAA
-2.642	CAC	AAC	-2.645	ATG	AGG	-2.722	GCT	GAT
-2.656	ATT	CTT	-2.645	ACT	AAT	-2.726	TGG	TTG

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.657	TGG	TCG	-2.648	TCC	TCA	-2.729	GAT	GCT
-2.657	TGG	GGG	-2.654	AAG	ATG	-2.735	AAA	AAC
-2.658	TGT	TGG	-2.654	CCT	CCA	-2.750	TCC	TCG
-2.658	TGT	AGT	-2.665	ATG	AAG	-2.756	ATT	AAT
-2.662	TCA	ACA	-2.671	CCG	ACG	-2.756	ATT	CTT
-2.666	ATG	TTG	-2.682	ACA	ACC	-2.758	TTA	TGA
-2.674	ATA	AGA	-2.685	ATG	ATT	-2.770	GAT	GTT
-2.676	GAC	CAC	-2.686	TGA	AGA	-2.780	GAG	GTG
-2.697	GTT	GAT	-2.688	TAG	TGG	-2.781	TGT	TGG
-2.701	GTC	GGC	-2.697	GGA	GGC	-2.783	GAC	CAC
-2.704	TGC	TGG	-2.697	CTT	ATT	-2.791	GTT	GTA
-2.718	TGT	TGA	-2.697	ACG	CCG	-2.791	CGT	AGT
-2.720	GAT	GTT	-2.697	AAA	CAA	-2.794	ATG	TTG
-2.720	ACG	CCG	-2.699	AAG	AAC	-2.794	CAT	CCT
-2.727	TAC	GAC	-2.700	TTT	GTT	-2.798	ATA	AGA
-2.729	TTG	ATG	-2.702	TAT	GAT	-2.798	ATA	ATT
-2.729	CCC	ACC	-2.704	GGT	GAT	-2.798	CTA	CTT
-2.731	AGT	AGA	-2.704	GAA	GAC	-2.800	ATC	AAC
-2.731	AGT	AGG	-2.705	TTG	TTT	-2.800	ATC	ATA
-2.741	GAA	TAA	-2.706	TAA	AAA	-2.801	GAA	GCA
-2.741	AAG	AAT	-2.707	AGT	AGG	-2.803	GGG	TGG
-2.747	GAC	GCC	-2.712	TTA	ATA	-2.805	GTC	TTC
-2.748	ATG	AAG	-2.717	TCA	TCC	-2.805	GTC	CTC
-2.756	TGC	GGC	-2.723	GGG	GGC	-2.807	CCA	CCT
-2.756	TGC	TGA	-2.724	ATT	AAT	-2.811	AGA	AGT
-2.761	AAA	ACA	-2.724	ATT	TTT	-2.813	GGA	GGC
-2.763	GAA	GAT	-2.744	TAC	GAC	-2.815	AAT	ATT
-2.771	GGA	CGA	-2.745	CAA	CAT	-2.815	ACT	ACA
-2.775	TTA	TAA	-2.750	AGT	CGT	-2.820	AGT	AGA
-2.775	TTA	ATA	-2.753	CTG	CTT	-2.822	TCT	TCA
-2.777	ATG	ATC	-2.755	CAG	AAG	-2.822	GTG	TTG
-2.780	TTC	TTA	-2.758	CTC	CTA	-2.825	ATT	ATG
-2.782	GTG	GAG	-2.763	AAT	AAG	-2.830	CGC	CGA
-2.783	TGT	GGT	-2.763	AAT	CAT	-2.830	CGC	GGC
-2.785	AGT	TGT	-2.771	GGC	CGC	-2.837	CCT	CCA
-2.793	TAT	TCT	-2.777	TGC	TGA	-2.842	AGG	ACG
-2.817	ATT	TTT	-2.785	GCT	GCG	-2.844	CCC	ACC
-2.827	CTA	CAA	-2.787	CGT	CGG	-2.848	CTT	CAT
-2.833	CAA	CTA	-2.787	CCA	CCT	-2.851	GTA	TTA
-2.836	TCC	TAC	-2.788	AAC	ATC	-2.852	TCA	GCA
-2.837	ATG	AGG	-2.790	TGT	TGA	-2.856	GCT	CCT
-2.848	TTT	ATT	-2.793	ATG	CTG	-2.857	GAT	GAA
-2.855	TAC	TAA	-2.794	TCC	ACC	-2.860	GGT	GGA
-2.855	TAC	TCC	-2.798	CGA	CGT	-2.872	CAT	CAG

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.855	CAC	CTC	-2.821	TGG	GGG	-2.876	TTA	TTC
-2.858	TGG	TGT	-2.823	AAG	ACG	-2.876	AAT	AAA
-2.860	AGA	TGA	-2.831	ACC	TCC	-2.876	AAT	CAT
-2.860	GGC	CGC	-2.836	AGC	CGC	-2.879	AAC	CAC
-2.868	CTC	CGC	-2.836	TAT	TCT	-2.884	CTC	GTC
-2.878	AAT	ATT	-2.839	TGT	AGT	-2.885	GCT	GCA
-2.878	AAT	AAA	-2.849	TCT	TCG	-2.893	CGG	AGG
-2.886	TCA	TAA	-2.849	ATA	AAA	-2.896	CGA	AGA
-2.902	GTG	GGG	-2.849	ATA	ATC	-2.901	GCA	TCA
-2.905	GAC	TAC	-2.850	TTG	GTG	-2.902	GTG	GAG
-2.906	CTC	CAC	-2.851	AGA	AGC	-2.903	TGA	TGC
-2.918	TTC	TAC	-2.856	GAG	GAT	-2.912	GCC	TCC
-2.918	GAG	GCG	-2.864	ATG	ATC	-2.912	GCC	CCC
-2.918	TTC	ATC	-2.864	GCC	GCG	-2.919	TTC	TTA
-2.918	TTC	TGC	-2.868	CAT	GAT	-2.920	GGG	CGG
-2.947	GTA	GAA	-2.882	TTG	ATG	-2.924	CTA	GTA
-2.947	CCG	ACG	-2.883	ACA	ACT	-2.925	TGA	TTA
-2.969	AAT	TAT	-2.890	TGT	GGT	-2.928	CAT	CTT
-2.983	CTG	CAG	-2.890	TGT	TTT	-2.936	GCG	GAG
-2.984	CAG	CTG	-2.893	TCG	ACG	-2.942	CCT	GCT
-2.995	TTC	GTC	-2.905	AAA	AAC	-2.943	GGT	CGT
-3.005	AAA	AAT	-2.910	ATA	ATT	-2.945	CGT	CGG
-3.007	GAT	GCT	-2.910	ATA	TTA	-2.946	TTT	TTG
-3.008	ATC	AAC	-2.915	GAT	GCT	-2.952	ATA	ATC
-3.030	TTT	TAT	-2.920	GGA	GGT	-2.952	GAT	CAT
-3.030	TTT	GTT	-2.923	TCA	GCA	-2.957	TGT	GGT
-3.032	CTT	ATT	-2.927	TCC	TGC	-2.968	TAA	TAC
-3.045	GGA	TGA	-2.927	TCC	GCC	-3.006	TCA	TCC
-3.048	ATC	AGC	-2.948	TAA	GAA	-3.012	CTC	CTA
-3.057	GAG	TAG	-2.955	TGG	AGG	-3.014	GCA	GCT
-3.062	ATT	AGT	-2.976	GCA	GAA	-3.018	TAT	GAT
-3.081	AAG	ACG	-2.976	TAG	GAG	-3.021	ACA	CCA
-3.081	TTT	TTA	-2.981	TTA	TTT	-3.021	CGA	CGC
-3.082	CAG	CCG	-2.981	TTA	TGA	-3.022	CAG	CTG
-3.089	ATC	CTC	-2.982	ACT	TCT	-3.022	CAG	AAG
-3.101	GAG	GTG	-2.985	AAA	AAT	-3.023	CAA	GAA
-3.114	AGC	CGC	-2.986	AAT	TAT	-3.028	TAA	AAA
-3.118	TAT	TAG	-2.990	TAT	TTT	-3.059	CCA	ACA
-3.118	TAT	AAT	-2.996	AGG	TGG	-3.063	CTT	CTG
-3.124	AAC	TAC	-2.998	AGA	AGT	-3.082	TAC	TTC
-3.131	ATC	TTC	-3.007	TGA	TGT	-3.089	GGA	GTA
-3.135	TTT	TGT	-3.010	CGT	CGA	-3.089	GGA	GGT
-3.173	TAC	AAC	-3.022	ATG	TTG	-3.093	TAA	TAT
-3.180	ATT	ATG	-3.023	TCT	TCA	-3.093	GTC	ATC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-3.183	TGG	AGG	-3.025	GGA	GCA	-3.095	CCA	GCA
-3.194	AAA	ATA	-3.025	GGA	CGA	-3.102	GTA	ATA
-3.199	TAT	GAT	-3.028	GAG	GTG	-3.102	GTA	GTC
-3.214	AGC	TGC	-3.043	ATA	CTA	-3.103	ACT	CCT
-3.227	TTA	TGA	-3.044	TCA	ACA	-3.115	GAT	TAT
-3.234	AGG	TGG	-3.050	CAT	CAG	-3.119	GAC	TAC
-3.238	TAC	TAG	-3.057	GTC	GTA	-3.124	CCT	CCG
-3.282	GAC	GTC	-3.058	GTT	GTA	-3.128	GAA	GAT
-3.311	GAA	GTA	-3.061	GTG	GAG	-3.129	TTT	GTT
-3.336	AAC	ATC	-3.065	ACA	TCA	-3.134	ATA	GTA
-3.342	CTT	CGT	-3.069	CGC	CGA	-3.146	CTT	GTT
-3.434	AAG	ATG	-3.072	TCT	TGT	-3.156	CGG	CGC
-3.486	TAT	TAA	-3.072	AAA	TAA	-3.157	GGC	CGC
-3.549	ATA	AAA	-3.073	AAT	AAA	-3.161	AGG	TGG
-3.563	CTG	CGG	-3.079	TGA	TGC	-3.162	ATT	ATA
-3.604	TCG	TAG	-3.097	GAT	GTT	-3.164	CGA	CGT
-3.744	CTA	CGA	-3.100	TAC	AAC	-3.175	CTG	GTG
-3.749	CAT	CTT	-3.118	TGC	TAC	-3.175	CTG	CGG
-3.807	AAA	TAA	-3.126	AGG	CGG	-3.188	GCA	ACA
-4.006	ATT	AAT	-3.139	GAC	GCC	-3.190	ACT	ACG
-4.127	AAG	TAG	-3.139	GAC	GTC	-3.190	CTT	CTA
-4.338	TTG	TAG	-3.140	GAA	GAT	-3.192	GCT	TCT
			-3.143	GCA	GCT	-3.198	TTT	TAT
			-3.146	TCA	TCT	-3.198	TGA	AGA
			-3.148	TTA	GTA	-3.198	GGC	TGC
			-3.148	GCT	GCA	-3.199	AGT	TGT
			-3.151	TTC	ATC	-3.201	TAT	TAG
			-3.151	TTC	TGC	-3.206	ACG	TCG
			-3.156	TGC	GGC	-3.219	CAG	CCG
			-3.158	TAG	AAG	-3.236	GGA	TGA
			-3.164	GCC	GGC	-3.244	ACA	ACT
			-3.169	AGA	CGA	-3.245	GGT	TGT
			-3.172	TAT	TAG	-3.247	AAG	ATG
			-3.177	GGC	GAC	-3.256	GTA	GAA
			-3.181	TTG	TGG	-3.257	TCA	TCT
			-3.190	ACG	TCG	-3.257	TCA	ACA
			-3.207	GGA	GTA	-3.259	CAA	CAT
			-3.211	TCG	GCG	-3.264	ATG	AGG
			-3.211	TCG	TAG	-3.264	CAT	CAA
			-3.215	GTG	GGG	-3.270	TTG	GTG
			-3.218	AAA	ATA	-3.271	CCG	GCG
			-3.239	TGA	TAA	-3.277	TCC	ACC
			-3.246	GTA	GGA	-3.278	TGC	GGC
			-3.246	GTA	GTT	-3.283	GTT	ATT

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
			-3.246	GGG	GAG	-3.288	GCA	CCA
			-3.248	AAC	TAC	-3.296	CAC	AAC
			-3.255	TAC	TCC	-3.302	TCT	GCT
			-3.269	GCC	GAC	-3.320	GCT	GCG
			-3.269	CCG	GCG	-3.328	CAA	AAA
			-3.282	GCT	GGT	-3.339	ATC	TTC
			-3.283	TGA	GGA	-3.343	CCC	GCC
			-3.289	GCA	GGA	-3.343	AGA	TGA
			-3.294	TTC	GTC	-3.347	CAC	GAC
			-3.308	GTC	GGC	-3.348	CCT	ACT
			-3.311	GGG	GCG	-3.351	GAG	AAG
			-3.331	GGC	GCC	-3.351	GAG	TAG
			-3.331	GGC	GTC	-3.352	TTT	TTA
			-3.340	GAA	GCA	-3.352	AAG	CAG
			-3.349	TGT	TCT	-3.352	TGA	GGA
			-3.364	AGC	TGC	-3.353	TCG	TAG
			-3.387	TCA	TGA	-3.382	TCT	ACT
			-3.390	CTT	CTA	-3.383	GAT	GAG
			-3.394	GTT	GGT	-3.385	ATT	TTT
			-3.396	TAT	AAT	-3.389	TTC	TAC
			-3.400	TAA	TCA	-3.391	ACT	TCT
			-3.416	AAG	TAG	-3.396	AAA	CAA
			-3.417	ATT	ATA	-3.408	GGA	CGA
			-3.418	TGC	AGC	-3.408	TAG	AAG
			-3.428	GTA	GAA	-3.419	TGG	TCG
			-3.445	TGG	TAG	-3.421	TGA	TGT
			-3.454	GGG	GTG	-3.427	ACA	GCA
			-3.457	AGA	TGA	-3.434	CTA	CGA
			-3.518	GGA	GAA	-3.434	CTA	CAA
			-3.551	TGG	TTG	-3.444	TCC	GCC
			-3.587	TTT	ATT	-3.448	CTT	ATT
			-3.600	TTA	TAA	-3.449	CGC	AGC
			-3.617	GTT	GAT	-3.461	AAA	ATA
			-3.628	TGG	TCG	-3.464	CTC	ATC
			-3.642	TGC	TTC	-3.470	AAG	TAG
			-3.645	GTC	GAC	-3.471	ACC	TCC
			-3.666	TGA	TTA	-3.493	ATC	GTC
			-3.670	TCT	TAT	-3.493	ATC	CTC
			-3.736	TAA	TAT	-3.513	GGT	AGT
			-3.890	TGA	TCA	-3.515	TGC	AGC
			-3.923	TTT	TTA	-3.518	AGC	CGC
			-3.923	TTT	TGT	-3.530	AAA	AAT
			-3.929	TGC	TCC	-3.530	TGG	GGG
			-3.959	TAA	TTA	-3.534	TTT	ATT

Log odds score	TMC_{coding} wild type	TMC_{coding} mutant	Log odds score	TMC_{+1} wild type	TMC_{+1} mutant	Log odds score	TMC_{+2} wild type	TMC_{+2} mutant
			-3.975	GCT	GAT	-3.555	GCG	CCG
			-4.009	GAG	GCG	-3.555	GCG	TCG
			-4.151	GAA	GTA	-3.557	TAT	AAT
			-4.249	TTC	TAC	-3.567	AGT	CGT
			-4.313	TCC	TAC	-3.569	TTA	GTA
			-4.353	TAC	TTC	-3.569	TTA	TAA
			-4.370	TTG	TAG	-3.569	AAT	TAT
			-4.768	TAG	TCG	-3.576	TCG	GCG
			-4.768	TAG	TTG	-3.591	TGG	AGG
			-4.869	TCA	TAA	-3.593	CAT	GAT
						-3.595	GTG	ATG
						-3.603	AAG	ACG
						-3.604	ACC	GCC
						-3.631	TAG	GAG
						-3.644	AGA	CGA
						-3.672	GAA	GTA
						-3.673	ATT	GTT
						-3.711	TAT	TAA
						-3.718	GAG	GCG
						-3.730	TGT	AGT
						-3.757	AAG	GAG
						-3.783	ACA	TCA
						-3.796	GGG	AGG
						-3.811	CTG	ATG
						-3.840	CTA	ATA
						-3.899	ACG	CCG
						-3.899	ACG	GCG
						-3.900	TTC	ATC
						-3.900	TTC	GTC
						-3.919	TAG	TCG
						-3.919	TAG	TTG
						-3.921	TCT	TCG
						-3.972	AAA	TAA
						-3.998	CAT	AAT
						-4.011	GCC	ACC
						-4.013	GCT	ACT
						-4.036	GAC	AAC
						-4.057	AGC	TGC
						-4.057	AAC	GAC
						-4.077	GAA	AAA
						-4.096	CAA	CTA
						-4.146	CCG	ACG
						-4.179	GGC	AGC
						-4.180	ATG	GTG

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
						-4.186	TTG	ATG
						-4.211	AGC	GGC
						-4.223	AAA	GAA
						-4.234	GGA	AGA
						-4.262	TTA	ATA
						-4.324	AGG	CGG
						-4.586	TAC	GAC
						-4.744	ATA	CTA
						-4.751	AAC	TAC
						-4.835	AGG	GGG
						-4.849	GAT	AAT
						-4.879	TTG	TAG
						-4.962	TCG	ACG
						-5.030	AGA	GGA
						-5.071	AGT	GGT
						-5.279	TAC	AAC
						-5.501	GCG	ACG
						-5.588	ACT	GCT
						-5.649	AAT	GAT

Table A.3: Trinucleotide mutation classes composing the nonsynonymous $M_{interspecific}$ mutation spectrum

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
0.954	ACG	ATG	1.020	CCG	CCA	1.189	CGT	CAT
0.774	GCG	GTG	0.887	CGA	TGA	0.888	CGC	CAC
0.769	CCG	CTG	0.813	CGT	TGT	0.523	CGG	CAG
0.582	CGG	CAG	0.799	CGG	TGG	0.504	CCG	CCA
0.467	TCG	TTG	0.687	GCG	GCA	0.350	TCG	TCA
0.415	GTA	ATA	0.671	CGC	TGC	0.257	ACG	ACA
0.385	CGT	CAT	0.570	TCG	TCA	0.226	GCG	GCA
0.253	GTC	ATC	0.385	ACG	ACA	0.144	CAT	CGT
0.252	CGC	CAC	-0.211	CAA	CAG	-0.009	CGA	CAA
0.235	ATA	GTA	-0.244	CTA	CTG	-0.053	TAT	TGT
0.112	CGA	CAA	-0.287	GTA	ATA	-0.159	AAT	AGT
-0.020	ATC	GTC	-0.306	GGG	AGG	-0.204	CAA	CAG
-0.058	GTT	ATT	-0.467	ACA	ACG	-0.301	GAC	GAT
-0.152	AAT	AGT	-0.477	CCA	CCG	-0.338	TAC	TAT
-0.249	ATT	GTT	-0.508	GCA	GCG	-0.350	CCA	CCG
-0.315	TGA	CGA	-0.529	GGC	AGC	-0.422	AAC	AAT
-0.380	CAT	CGT	-0.531	GCC	ACC	-0.494	GGC	GGT
-0.409	GCG	ACG	-0.570	ATC	GTC	-0.505	CAC	CGC
-0.437	ATG	GTG	-0.586	TCA	TCG	-0.509	CAG	CAA
-0.441	CGC	TGC	-0.600	GGT	AGT	-0.516	ACG	ATG
-0.458	GGC	AGC	-0.639	GGA	AGA	-0.521	CAC	CAT
-0.472	GCA	ACA	-0.677	ATA	GTA	-0.528	AAC	AGC
-0.515	AGT	AAT	-0.677	GAG	AAG	-0.536	TGT	TAT
-0.529	AGG	AAG	-0.709	ATA	ATG	-0.540	CAG	CGG
-0.538	ACC	GCC	-0.722	AAA	GAA	-0.566	CCG	CTG
-0.545	GTG	ATG	-0.735	GCT	ACT	-0.574	CCC	CCT
-0.583	CAC	CGC	-0.806	CAC	CAT	-0.649	ACC	ATC
-0.594	ACT	GCT	-0.818	GTC	ATC	-0.654	TAG	TAA
-0.623	GCC	ACC	-0.844	GTG	ATG	-0.674	TAC	TGC
-0.624	AGC	AAC	-0.848	GTT	ATT	-0.688	GCA	GCG
-0.632	AGT	GGT	-0.857	GTA	GTG	-0.697	AGC	AGT
-0.633	CAG	CGG	-0.877	ATG	GTG	-0.730	GGT	GAT
-0.642	ACG	GCG	-0.927	GAT	AAT	-0.744	AGT	AAT
-0.669	CCG	TCG	-0.942	CCC	TCC	-0.748	AGC	AAC
-0.673	GCT	ACT	-0.968	AGG	GGG	-0.766	ACA	ACG
-0.706	ACA	ATA	-0.984	TAA	TAG	-0.773	CCC	CTC
-0.706	ACA	GCA	-0.985	CAT	TAT	-0.780	GAG	GAA
-0.728	AAA	AGA	-0.999	ATT	GTT	-0.825	GAC	GGC
-0.743	CGT	TGT	-1.016	CTG	CTA	-0.839	TAA	TAG
-0.744	ACT	ATT	-1.041	CAG	TAG	-0.970	TAT	TAC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-0.794	CAA	CGA	-1.044	CCC	CCT	-0.978	GGT	GGC
-0.802	AGC	GGC	-1.052	GCA	ACA	-0.994	TGC	TGT
-0.805	TAG	CAG	-1.052	GAA	GAG	-1.002	CGC	CGT
-0.845	CGG	TGG	-1.069	CAC	TAC	-1.019	ATG	ACG
-0.850	AAC	AGC	-1.086	CAG	CAA	-1.036	TCA	TCG
-0.929	ATA	ACA	-1.086	TAG	CAG	-1.038	AAG	AAA
-0.943	ATG	ATA	-1.087	CGA	CGG	-1.059	GAT	GGT
-0.962	GGT	AGT	-1.099	ACG	GCG	-1.068	CTA	CTG
-0.971	AAG	AGG	-1.107	AAA	AAG	-1.069	GCC	GCT
-0.977	CCC	TCC	-1.113	AGG	AGA	-1.082	CTC	CTT
-1.008	TGA	TCA	-1.129	GAG	GAA	-1.087	AGT	AGC
-1.050	GCC	GTC	-1.136	TAT	CAT	-1.095	ACC	ACT
-1.063	CCT	TCT	-1.145	CTC	TTC	-1.116	AAA	AAG
-1.074	ACC	ATC	-1.147	CGG	CGA	-1.146	GCC	GTC
-1.077	CCA	TCA	-1.169	AGA	GGA	-1.159	GGC	GAC
-1.086	ATT	ACT	-1.170	GAA	AAA	-1.183	TTC	TTT
-1.089	GCA	GTA	-1.188	GGG	GGA	-1.200	AAT	AAC
-1.100	CAT	TAT	-1.224	GCC	GCT	-1.205	TCC	TCT
-1.104	CTT	TTT	-1.241	ACT	GCT	-1.220	AAG	AGG
-1.111	GTG	GCG	-1.254	CCG	TCG	-1.235	GAA	GAG
-1.135	AGA	AAA	-1.258	CCT	TCT	-1.242	GGG	GAG
-1.137	CAC	TAC	-1.268	TTA	TTG	-1.304	CAT	CAC
-1.201	ATG	ACG	-1.278	CTG	TTG	-1.325	TGC	TAC
-1.226	GGG	AGG	-1.297	CGT	CGC	-1.331	TGT	TGC
-1.235	GTC	GCC	-1.299	AGT	GGT	-1.333	GCG	GTG
-1.239	ATA	ATG	-1.300	ACC	GCC	-1.359	ATC	ATT
-1.265	GCT	GTT	-1.303	TTG	CTG	-1.375	GTC	GTT
-1.277	AAT	GAT	-1.307	GGC	GGT	-1.403	AAA	AGA
-1.289	TCT	CCT	-1.315	ACA	GCA	-1.432	AGG	AAG
-1.293	GAT	AAT	-1.316	TGG	CGG	-1.433	TAA	TGA
-1.356	GGA	AGA	-1.333	AAG	GAG	-1.442	CCC	CAC
-1.362	TCA	TTA	-1.341	GCG	ACG	-1.491	CCC	CGC
-1.363	CCG	CAG	-1.346	GTG	GTA	-1.493	AGG	AGA
-1.376	GAC	AAC	-1.347	ATG	ATA	-1.493	AGG	AGC
-1.389	GTT	GCT	-1.363	TGA	TGG	-1.562	TCC	TTC
-1.395	GTA	GCA	-1.367	GAC	AAC	-1.565	ACC	AAC
-1.402	TAT	TGT	-1.369	AAT	AAC	-1.570	TAG	TGG
-1.413	CCC	CTC	-1.374	TAC	TAT	-1.579	CTA	CCA
-1.427	GGG	GAG	-1.422	AGC	GGC	-1.587	GGG	GGA
-1.450	CAA	GAA	-1.426	AAG	AAA	-1.588	TTC	TGC
-1.513	TGT	TAT	-1.426	CTT	TTT	-1.596	TCG	TGG
-1.540	GAA	AAA	-1.438	TGT	CGT	-1.616	CTC	CCC
-1.541	AGA	GGA	-1.439	TGG	TGA	-1.619	GAT	GAC
-1.545	GGG	GCG	-1.448	ACC	ACT	-1.621	CGC	CTC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-1.564	CAC	CAG	-1.456	AAT	GAT	-1.677	CTC	CTG
-1.565	AGG	GGG	-1.459	TGC	CGC	-1.677	TGG	TAG
-1.588	TAC	CAC	-1.469	GGA	GGG	-1.687	TCT	TGT
-1.597	CCT	CTT	-1.498	TCC	TCT	-1.699	GGA	GAA
-1.608	TGT	TCT	-1.512	GTC	GTG	-1.716	ATC	ACC
-1.612	TCG	TGG	-1.514	GAT	GAC	-1.722	CGT	CGC
-1.613	CCC	ACC	-1.514	GAT	CAT	-1.724	TGG	TGA
-1.629	TTG	TCG	-1.557	AAC	GAC	-1.743	GTC	GGC
-1.632	GAG	AAG	-1.573	CTA	GTA	-1.750	CGA	CCA
-1.641	CTC	TTC	-1.585	CGT	AGT	-1.753	TAG	TAC
-1.652	GGC	GAC	-1.593	CTT	CTC	-1.773	CCA	CTA
-1.655	GGT	GAT	-1.597	CCA	TCA	-1.778	TCG	TTG
-1.698	AAC	GAC	-1.632	TGT	TAT	-1.788	ACC	ACG
-1.699	CCG	GCG	-1.641	CTC	CTG	-1.802	TTT	TTC
-1.700	AGT	ACT	-1.645	GTC	GTT	-1.814	GTT	GCT
-1.701	CCA	CTA	-1.655	GAC	GAT	-1.837	TTA	TCA
-1.702	TGA	TGG	-1.658	CAA	TAA	-1.878	TAC	TAG
-1.702	TGA	GGA	-1.659	CCC	GCC	-1.880	GCC	GGC
-1.722	CCT	GCT	-1.662	TAG	TAA	-1.898	TTG	TTA
-1.736	ACC	AAC	-1.715	CTC	CTT	-1.900	AGC	AGG
-1.743	AGC	ACC	-1.723	AGA	AGG	-1.907	GTA	GCA
-1.744	GCG	TCG	-1.731	TCC	CCC	-1.923	ACT	ACC
-1.744	CTA	CCA	-1.733	TCT	CCT	-1.935	GGG	GGC
-1.762	GGA	GCA	-1.735	ACG	AGG	-1.935	TGG	TGC
-1.797	CTT	GTT	-1.762	CAC	CAG	-1.941	GGC	GTC
-1.800	TCT	TGT	-1.780	ACT	ACC	-1.943	GAG	GGG
-1.801	TCC	GCC	-1.799	GTC	CTC	-1.956	ATT	ACT
-1.801	TCC	TTC	-1.813	GTA	CTA	-1.957	CGC	CGG
-1.821	TTT	CTT	-1.814	CCG	CCC	-1.957	CGC	CCC
-1.844	CCA	ACA	-1.829	TAT	TAC	-1.959	GGT	GCT
-1.866	CCT	ACT	-1.829	TAT	TAA	-1.974	CAA	CGA
-1.873	TCA	CCA	-1.834	ACC	ACA	-1.981	AGT	ACT
-1.888	TGC	CGC	-1.840	CGG	AGG	-1.994	TTC	TCC
-1.916	GGA	GAA	-1.844	TAC	CAC	-1.995	GAC	GTC
-1.925	GAT	GGT	-1.852	GCG	CCG	-2.026	GCG	GAG
-1.926	CTA	ATA	-1.894	ACC	ACG	-2.032	AAC	AAA
-1.930	ATC	ACC	-1.906	TGA	CGA	-2.042	CGG	CGT
-1.944	GTG	CTG	-1.909	TTG	TTA	-2.042	CGG	CGA
-1.951	GAC	GAG	-1.924	AAC	AAT	-2.052	ATC	AAC
-1.965	TGT	CGT	-1.947	CCC	CCA	-2.057	ACT	AGT
-1.967	GCG	GAG	-1.958	TGC	TGT	-2.058	TCA	TGA
-1.967	GCG	GGG	-1.959	ACC	AGC	-2.072	AGA	AAA
-1.986	GCA	TCA	-1.961	TTC	CTC	-2.083	TGA	TGG
-1.998	CTT	ATT	-1.966	AGC	AGT	-2.096	CTG	CCG

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.008	CAG	CAC	-1.975	AGA	ACA	-2.096	CTG	CTC
-2.011	CCA	GCA	-1.979	CTA	TTA	-2.098	ATA	ACA
-2.019	CCC	GCC	-1.985	CAC	CAA	-2.110	ATT	ATC
-2.023	ACC	AGC	-1.986	GGT	GGC	-2.113	GAC	GAG
-2.042	TCC	TGC	-1.987	ATC	ATT	-2.117	ATG	ATA
-2.045	CAT	AAT	-1.994	CGG	CGC	-2.127	TAA	TCA
-2.058	TAC	TGC	-1.996	GTG	CTG	-2.130	TGA	TAA
-2.061	GAA	GGA	-2.003	CGA	GGA	-2.136	AAG	AAC
-2.066	AAA	GAA	-2.016	CCC	ACC	-2.139	GTG	GCG
-2.067	GCC	TCC	-2.020	CGC	GGC	-2.143	CCT	CGT
-2.069	GAC	GGC	-2.020	CGC	CGT	-2.148	GTC	GCC
-2.082	GTT	CTT	-2.021	TCT	TCC	-2.153	ACA	ATA
-2.083	CGC	AGC	-2.022	CTG	GTG	-2.153	ACA	AGA
-2.088	TCT	TTT	-2.032	CCT	CCC	-2.166	GAG	GAC
-2.096	TAT	TTT	-2.067	TAC	TGC	-2.173	TCT	TCC
-2.096	TAT	CAT	-2.070	GCA	CCA	-2.176	TCA	TTA
-2.098	AAT	CAT	-2.074	TGT	TGC	-2.183	AGA	AGG
-2.098	ACT	AGT	-2.084	GCC	GCG	-2.184	CCC	CCG
-2.124	TCA	GCA	-2.090	CCC	CCG	-2.186	TTG	TCG
-2.128	TTC	CTC	-2.098	AGT	AGC	-2.186	TTG	TTC
-2.133	CTC	GTC	-2.098	AGT	ACT	-2.204	CCG	GCG
-2.144	CAA	CCA	-2.119	CCT	GCT	-2.205	GCT	GTT
-2.152	GAG	GAC	-2.120	ATC	ATA	-2.208	CCT	CTT
-2.153	AAG	GAG	-2.141	ACG	ACC	-2.246	GAC	GCC
-2.154	CTG	CCG	-2.147	TAA	TGA	-2.255	TCC	TAC
-2.179	TGG	CGG	-2.147	TAA	CAA	-2.255	TCC	TCA
-2.199	GAG	CAG	-2.148	AGC	AGG	-2.258	ACC	ACA
-2.206	GTA	TTA	-2.149	CAT	CAC	-2.258	ACC	AGC
-2.218	ATC	ATG	-2.151	GAA	CAA	-2.260	CGA	GGA
-2.221	GCT	GGT	-2.152	ATT	ATG	-2.264	TAG	GAG
-2.233	CCT	CGT	-2.165	ATA	CTA	-2.265	CGG	GGG
-2.238	GGG	CGG	-2.184	TTA	TGA	-2.266	CAC	CAA
-2.248	TTG	TTT	-2.184	TTA	CTA	-2.275	ATC	ATG
-2.250	GGC	GCC	-2.194	TTT	TTG	-2.283	CCA	CAA
-2.261	GCC	GGC	-2.194	TTT	CTT	-2.297	GCA	GAA
-2.266	CAG	AAG	-2.197	CAA	GAA	-2.306	AGC	ATC
-2.270	TTA	ATA	-2.212	AGG	ATG	-2.306	TCT	TTT
-2.270	TTA	TCA	-2.214	CCT	ACT	-2.306	TCT	TAT
-2.305	GCA	GGA	-2.244	AAT	CAT	-2.310	GGA	GCA
-2.305	TCG	GCG	-2.249	TAG	TAC	-2.328	CAT	CCT
-2.305	TCG	CCG	-2.257	GCG	GCT	-2.330	GTC	ATC
-2.334	GTC	CTC	-2.263	TCG	TCT	-2.331	CTT	CTC
-2.348	GGT	GCT	-2.274	GGT	CGT	-2.347	TTA	TTT
-2.353	GAG	GAT	-2.274	GGT	GGG	-2.351	CCT	CCC

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.355	CTG	GTG	-2.278	TCC	TGC	-2.362	GTG	CTG
-2.361	TCC	CCC	-2.284	CCG	GCG	-2.377	AGA	AGC
-2.371	ACA	TCA	-2.285	GCC	CCC	-2.393	ACT	ATT
-2.373	CGA	GGA	-2.287	GGG	GGC	-2.403	GCA	GGA
-2.381	CAT	CAG	-2.287	GGG	CGG	-2.407	CCC	GCC
-2.390	AAG	CAG	-2.307	CGC	AGC	-2.407	GAA	GGA
-2.414	CGG	GGG	-2.314	GTT	TTT	-2.409	TCC	TGC
-2.424	CCC	CAC	-2.314	GTT	CTT	-2.411	CAG	CAC
-2.437	CTA	GTA	-2.318	TTC	TTG	-2.415	CGT	CCT
-2.444	GCT	TCT	-2.324	GTA	GTC	-2.415	CGT	CTT
-2.448	CGT	GGT	-2.330	AAC	AAG	-2.419	GCC	GCA
-2.448	CGT	AGT	-2.330	AAC	AAA	-2.422	GGA	GGG
-2.450	TTT	TCT	-2.334	ATT	ATA	-2.437	AAC	AAG
-2.460	CAG	GAG	-2.337	AGG	ACG	-2.443	CGA	CGG
-2.461	TCA	ACA	-2.344	GCT	GCC	-2.454	GAG	CAG
-2.475	TGC	GGC	-2.349	AGT	CGT	-2.460	TGC	TGG
-2.480	CAA	AAA	-2.357	GGA	GGC	-2.460	TGC	TTC
-2.481	CAC	CAA	-2.372	AGC	AGA	-2.464	TTC	TAC
-2.493	TCT	GCT	-2.382	CTT	ATT	-2.464	TCA	TAA
-2.493	TCT	ACT	-2.382	CTT	GTT	-2.466	TGT	TCT
-2.520	GAA	CAA	-2.382	CTT	CTG	-2.467	GCT	GCC
-2.530	ATG	TTG	-2.385	CAG	CAC	-2.474	GGG	GCG
-2.538	ATA	CTA	-2.400	CGG	GGG	-2.487	GAA	GCA
-2.538	ATA	TTA	-2.406	AGG	AGC	-2.498	GGT	GTT
-2.541	GTG	TTG	-2.413	ACA	ACC	-2.504	ATA	AAA
-2.550	AAT	ACT	-2.431	GAT	GAA	-2.504	ATA	ATT
-2.561	GAT	GAA	-2.445	TCA	CCA	-2.507	GTT	ATT
-2.568	ACT	AAT	-2.454	AGA	AGC	-2.517	GGC	GGG
-2.570	AGC	CGC	-2.457	ATC	ATG	-2.531	AGA	ACA
-2.571	GGT	TGT	-2.472	TTC	TTT	-2.540	CCG	CGG
-2.578	GCT	CCT	-2.473	ACT	ACG	-2.563	ATC	AGC
-2.585	CTT	CCT	-2.480	AGG	AGT	-2.572	TAC	TAA
-2.615	CCG	CGG	-2.488	TGC	TGA	-2.573	GCC	GAC
-2.617	CCA	CGA	-2.488	CTC	GTC	-2.583	GAC	GAA
-2.639	AGA	AGT	-2.492	GTC	GTA	-2.600	GAT	GTT
-2.642	GTT	TTT	-2.500	ACA	AGA	-2.607	CTG	CTA
-2.649	TCC	ACC	-2.507	CCG	CCT	-2.613	CCT	CAT
-2.657	ACG	CCG	-2.511	GGA	CGA	-2.625	CAA	CCA
-2.657	ACG	AAG	-2.531	GGC	GGG	-2.636	AAA	AAC
-2.661	ATT	ATG	-2.531	GGC	GGA	-2.649	CTT	CGT
-2.667	TTC	TTG	-2.533	CAA	CAC	-2.654	GCA	GCC
-2.667	TTC	TCC	-2.536	GCC	GCA	-2.658	CTC	CAC
-2.667	TTC	GTC	-2.545	GCG	GCC	-2.669	TAG	TTG
-2.678	GAG	GCG	-2.551	TCG	CCG	-2.687	TTC	TTG

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.679	GAT	GAG	-2.551	TCG	TCC	-2.690	GCT	CCT
-2.703	GAA	GCA	-2.557	AAG	CAG	-2.705	AAT	ATT
-2.708	GGG	GTG	-2.557	AAG	AAC	-2.712	ACA	AAA
-2.708	ACA	CCA	-2.557	ATT	CTT	-2.718	CAC	CAG
-2.712	AGT	AGG	-2.557	ATT	ATC	-2.719	CCT	GCT
-2.712	AGT	TGT	-2.559	TGG	TAG	-2.725	AAC	ACC
-2.712	CCC	CGC	-2.559	TGG	TGC	-2.753	TTA	TGA
-2.717	ACC	TCC	-2.564	GAG	CAG	-2.753	TTA	TTG
-2.739	TGG	TGT	-2.564	GAG	GAT	-2.753	TTA	TAA
-2.740	CTC	CCC	-2.587	ACC	AAC	-2.756	GCC	TCC
-2.762	GAC	GAA	-2.589	TTA	TTC	-2.764	CCG	CCT
-2.777	CGC	GGC	-2.594	TGC	GGC	-2.768	GGC	CGC
-2.777	CGC	CCC	-2.596	TCC	TCG	-2.769	ACC	TCC
-2.777	CGC	CTC	-2.600	TGA	TAA	-2.776	AGC	ACC
-2.791	ACT	TCT	-2.654	TTC	TTA	-2.776	AGC	AGA
-2.791	ACT	CCT	-2.662	TGT	TGG	-2.789	CTG	CGG
-2.796	AAC	AAG	-2.671	CTC	CTA	-2.806	GAC	CAC
-2.796	AAC	CAC	-2.671	CTC	ATC	-2.808	GCA	GCT
-2.820	GCC	CCC	-2.673	CAG	GAG	-2.808	GCA	GTA
-2.821	CCT	CAT	-2.679	GGT	GGA	-2.811	GGG	CGG
-2.859	GGT	GTT	-2.680	ATC	AGC	-2.820	TAA	GAA
-2.865	AGG	AGC	-2.681	GCT	GCG	-2.822	CCA	CGA
-2.886	CAC	CTC	-2.681	GCT	TCT	-2.823	GTA	GTG
-2.892	CAT	CTT	-2.681	GCT	CCT	-2.823	GTA	GTT
-2.892	CAT	CAA	-2.683	CGT	GGT	-2.823	GTA	ATA
-2.899	TCT	TAT	-2.696	CGA	CGT	-2.829	AAG	ACG
-2.899	GTA	CTA	-2.696	CGA	CGC	-2.829	AAG	ATG
-2.901	AAG	AAC	-2.720	GTT	GTG	-2.834	GGT	CGT
-2.902	CAG	CTG	-2.724	ATA	TTA	-2.840	CTC	CGC
-2.902	CAG	CAT	-2.729	GTA	GTT	-2.841	GTC	GTG
-2.906	AGC	AGA	-2.733	ATG	AAG	-2.841	GTC	CTC
-2.906	AGC	AGG	-2.753	GAC	CAC	-2.841	GTC	TTC
-2.911	ATC	CTC	-2.753	GAC	GAA	-2.860	GAG	AAG
-2.920	TTT	TTG	-2.753	GAC	GAG	-2.866	TGC	TGA
-2.924	CGG	CTG	-2.755	AAT	ACT	-2.867	TAT	TAA
-2.931	GGG	TGG	-2.757	GGG	TGG	-2.873	CAC	CTC
-2.931	ACA	AGA	-2.761	TAC	TAA	-2.877	CCC	ACC
-2.935	ATG	CTG	-2.761	TAC	TAG	-2.879	AGG	ACG
-2.935	ATG	AGG	-2.777	AGC	ACC	-2.888	CAT	CTT
-2.941	GAG	GGG	-2.781	AAA	CAA	-2.893	GAA	CAA
-2.963	TTA	GTA	-2.810	ACT	AGT	-2.900	TTT	TGT
-2.967	GAT	CAT	-2.839	ATG	TTG	-2.904	ACT	ACG
-2.979	AAC	ACC	-2.844	GCA	GCC	-2.904	ACT	AAT
-2.986	TGC	TGG	-2.844	GAA	GAT	-2.922	CAG	CAT

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.986	TGC	TCC	-2.845	TGC	TGG	-2.922	CAG	GAG
-2.998	GCA	CCA	-2.845	ATT	AGT	-2.922	GGC	TGC
-3.000	AGT	CGT	-2.848	AGG	CGG	-2.937	CTT	CCT
-3.047	AAA	ACA	-2.851	TCA	TCC	-2.958	CGG	CTG
-3.048	CTG	ATG	-2.851	TGA	TGT	-2.972	AAA	AAT
-3.067	ATT	CTT	-2.875	ACC	CCC	-2.972	AAA	ACA
-3.067	ATT	TTT	-2.887	CTG	CTT	-2.977	CCA	GCA
-3.076	TAT	AAT	-2.887	CTG	CTC	-2.977	CCA	CCT
-3.076	TAT	TCT	-2.887	TTT	TTC	-2.977	TAC	TTC
-3.095	TTG	TGG	-2.887	TTT	TTA	-2.977	TAC	TCC
-3.095	TTG	TTC	-2.907	CCT	CCA	-2.977	TGA	TGC
-3.096	CTT	CGT	-2.907	CCT	CCG	-2.978	GCT	GAT
-3.110	AAT	TAT	-2.909	AGT	ATT	-2.978	GCT	GGT
-3.110	AAT	ATT	-2.911	AGC	ATC	-2.990	GCA	TCA
-3.110	AAT	AAG	-2.916	GGA	TGA	-3.001	ACG	AAG
-3.118	ATG	ATC	-2.916	GGA	GGT	-3.001	ACG	AGG
-3.122	ACC	CCC	-2.940	AGA	AGT	-3.005	GAT	CAT
-3.141	CGT	CTT	-2.950	GCG	GGG	-3.018	GTT	TTT
-3.145	CTC	ATC	-2.952	CCA	CCC	-3.018	GTT	GTC
-3.148	GTG	GGG	-2.956	ATG	AGG	-3.030	CAA	CAC
-3.174	CAC	AAC	-2.963	AAA	AAC	-3.034	GGG	TGG
-3.202	AAC	AAA	-2.966	CAC	AAC	-3.048	TGC	TCC
-3.214	GAA	GAC	-2.968	ATC	TTC	-3.051	CCG	CCC
-3.227	TTC	TTA	-2.980	GGG	GGT	-3.054	TGT	TTT
-3.229	AAA	CAA	-2.999	TGC	TAC	-3.055	GTG	GTA
-3.232	GAC	CAC	-3.026	GCA	GCT	-3.055	GTG	ATG
-3.262	TAC	AAC	-3.033	ACT	CCT	-3.055	GTG	GTC
-3.270	AGG	ACG	-3.042	GGC	CGC	-3.058	GGA	AGA
-3.270	AGG	ATG	-3.044	CAA	CAT	-3.080	AGT	ATT
-3.270	AGG	TGG	-3.065	CAT	CAG	-3.094	GAC	TAC
-3.286	GCA	GAA	-3.066	TCC	TCA	-3.095	TGA	GGA
-3.290	GCC	GAC	-3.070	TGG	TGT	-3.100	CCC	CCA
-3.298	CAT	GAT	-3.075	GAG	GAC	-3.102	TCC	TCG
-3.299	CTG	CGG	-3.077	CTA	ATA	-3.105	CAG	CCG
-3.302	GGA	CGA	-3.077	CTA	CTT	-3.105	GGC	GCC
-3.306	AAG	AAT	-3.085	CCA	CCT	-3.107	TAA	TAC
-3.306	AAG	ACG	-3.096	GCC	GGC	-3.119	CTT	CAT
-3.310	CCA	CAA	-3.138	TCA	TGA	-3.124	GCG	GCC
-3.330	CGG	CCG	-3.161	AAT	AAG	-3.124	GCG	GCT
-3.342	TCC	TAC	-3.166	TAG	TGG	-3.130	AAC	GAC
-3.350	ACG	TCG	-3.181	TGC	AGC	-3.147	GAG	GTG
-3.350	ACG	AGG	-3.181	TGC	TTC	-3.159	CCA	CCC
-3.353	GCG	CCG	-3.188	TGG	GGG	-3.174	ACC	GCC
-3.372	GAT	GCT	-3.196	AGT	AGG	-3.175	GGA	TGA

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-3.392	TGC	TAC	-3.200	CCG	ACG	-3.181	GAA	GAC
-3.392	TGC	TTC	-3.215	TAT	TAG	-3.188	TTT	TCT
-3.396	CAA	CAC	-3.239	ACG	AAG	-3.197	ATA	ATG
-3.396	CAA	CAT	-3.239	CCA	GCA	-3.197	ATA	ATC
-3.405	AGT	AGA	-3.239	CCA	ACA	-3.197	ATA	TTA
-3.417	AGC	ATC	-3.249	GAA	TAA	-3.208	ATT	ATG
-3.417	AGC	TGC	-3.249	GAA	GAC	-3.208	ATT	CTT
-3.425	GCT	GAT	-3.249	GAA	GTA	-3.216	ATG	ATC
-3.431	TTT	GTT	-3.251	ATT	AAT	-3.216	ATG	CTG
-3.432	TGG	TAG	-3.251	ATT	TTT	-3.216	ATG	AGG
-3.432	TGG	TGC	-3.261	CAG	CAT	-3.277	TGT	AGT
-3.453	CTG	CAG	-3.267	GGG	GCG	-3.278	CAC	GAC
-3.455	GAC	GTC	-3.289	ACA	AAA	-3.285	TTG	TTT
-3.455	GAC	GCC	-3.293	TTT	GTT	-3.285	TTG	TGG
-3.472	CGA	CCA	-3.320	ACT	ACA	-3.285	TTG	GTG
-3.482	TAT	GAT	-3.320	ACT	AAT	-3.285	TTG	TAG
-3.501	TTG	ATG	-3.357	CTG	ATG	-3.293	GAT	GCT
-3.501	TTG	GTG	-3.372	GGT	GCT	-3.293	GAT	GAG
-3.514	TTC	TAC	-3.372	GGT	GTT	-3.307	CCT	CCG
-3.536	CTA	CGA	-3.373	ATC	AAC	-3.309	GGA	CGA
-3.549	GGC	GTC	-3.383	GTG	GGG	-3.309	GGA	GGC
-3.579	CAC	GAC	-3.383	GTG	GTC	-3.310	ACT	CCT
-3.579	CAC	CCC	-3.384	GCC	TCC	-3.321	GGG	GTG
-3.592	GTA	GAA	-3.406	CGC	CGG	-3.321	GGG	GGT
-3.604	ATC	TTC	-3.406	CGC	CGA	-3.345	GGT	TGT
-3.624	ACA	AAA	-3.407	TCT	TCA	-3.351	CTC	GTC
-3.628	ATG	ATT	-3.407	TCT	TCG	-3.351	CTC	CTA
-3.628	ATG	AAG	-3.410	AGA	ATA	-3.359	CGA	AGA
-3.637	ATA	AAA	-3.413	GTT	GGT	-3.371	CTA	CTC
-3.637	ATA	AGA	-3.413	GTT	GTC	-3.371	CTA	CTT
-3.656	TTA	TTT	-3.413	TTG	TTT	-3.371	CTA	CGA
-3.656	TTA	TTC	-3.413	TTG	ATG	-3.380	TTC	TTA
-3.660	GAT	GTT	-3.413	TTG	TGG	-3.380	TTC	GTC
-3.668	TAC	TCC	-3.413	TTG	TTC	-3.382	TGG	AGG
-3.668	TAC	GAC	-3.418	ATA	ATT	-3.383	GCT	GCA
-3.708	GGA	GTA	-3.418	ATA	ATC	-3.388	TCG	TCT
-3.712	AAG	ATG	-3.418	ATA	AGA	-3.388	TCG	GCG
-3.713	CAG	CCG	-3.429	AAC	ATC	-3.388	TCG	TCC
-3.740	AAA	AAC	-3.429	AAC	ACC	-3.398	AAT	CAT
-3.743	GAC	TAC	-3.429	AAC	CAC	-3.398	AAT	ACT
-3.760	ATT	AAT	-3.446	GAC	GTC	-3.405	ACA	GCA
-3.803	AAT	AAA	-3.446	GAC	GGC	-3.405	ACA	ACT
-3.836	TTT	TGT	-3.446	GAC	TAC	-3.424	GTT	GTA
-3.836	TTT	TAT	-3.447	GGC	TGC	-3.424	GTT	GTG

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
-3.836	TTT	TTA	-3.454	TAC	TTC	-3.424	GTT	CTT
-3.911	TGT	GGT	-3.470	AGC	TGC	-3.424	GTT	GAT
-3.911	TGT	AGT	-3.473	AAG	AAT	-3.446	TTA	ATA
-3.920	TTC	ATC	-3.474	AAA	ACA	-3.446	TTA	TTC
-3.954	GGC	CGC	-3.475	TGG	TCG	-3.457	CCG	ACG
-3.958	GGT	CGT	-3.484	CAG	AAG	-3.483	AAA	TAA
-3.991	CAT	CCT	-3.512	ACA	TCA	-3.485	AGT	AGG
-4.026	AGA	ACA	-3.537	GCA	TCA	-3.499	GAC	AAC
-4.026	AGA	ATA	-3.568	ACC	TCC	-3.500	CAA	AAA
-4.026	AGA	TGA	-3.577	TCC	ACC	-3.500	CAA	GAA
-4.026	AGA	AGC	-3.578	TGT	GGT	-3.508	TCC	ACC
-4.028	GTT	GGT	-3.578	TGT	TGA	-3.516	GTA	GAA
-4.098	AGT	ATT	-3.597	GCT	GCA	-3.516	GTA	CTA
-4.125	TGG	TCG	-3.597	GCT	GAT	-3.553	GAG	GCG
-4.125	TGG	GGG	-3.633	AGA	CGA	-3.559	TCT	GCT
-4.125	TGG	TTG	-3.633	AGA	TGA	-3.559	TCT	TCG
-4.125	TGG	AGG	-3.649	TCG	TGG	-3.559	TGC	AGC
-4.146	AAA	AAT	-3.649	TCG	TAG	-3.559	TGC	GGC
-4.146	AAA	ATA	-3.650	ATG	ATT	-3.562	TCA	TCT
-4.195	CTT	CAT	-3.650	ATG	ATC	-3.565	TGT	TGG
-4.280	GTC	GAC	-3.650	ATG	CTG	-3.565	TGT	GGT
-4.280	GTC	GGC	-3.659	AGG	TGG	-3.566	CAC	CCC
-4.312	GAA	GTA	-3.698	TGA	TGC	-3.567	CGC	GGC
-4.312	GAA	GAT	-3.698	TGA	GGA	-3.567	CGC	AGC
-4.361	TAC	TTC	-3.756	TAA	TCA	-3.567	CGC	CGA
-4.361	TAC	TAG	-3.756	TAA	GAA	-3.572	AGG	AGT
-4.529	TTT	ATT	-3.756	TAA	TAC	-3.572	AGG	ATG
-4.613	TTC	TGC	-3.756	TAA	TAT	-3.572	AGG	CGG
-4.648	GGC	TGC	-3.756	TAA	AAA	-3.594	TTT	TTG
-4.702	ATC	AAC	-3.758	AGC	CGC	-3.615	CAG	AAG
-5.098	AAG	TAG	-3.758	CAT	GAT	-3.615	GGC	AGC
			-3.758	CAT	CAA	-3.615	GGC	GGA
			-3.761	AAG	ACG	-3.630	CTT	GTT
			-3.771	CTA	CTC	-3.651	CGG	CCG
			-3.786	CGG	CGT	-3.651	CGG	CGC
			-3.788	GTG	GAG	-3.662	ATC	ATA
			-3.788	GTG	TTG	-3.670	TGG	TGT
			-3.789	GCC	GAC	-3.670	TAC	GAC
			-3.831	TCA	TCT	-3.672	GCC	ACC
			-3.854	AAT	TAT	-3.705	CTG	ATG
			-3.854	AAT	ATT	-3.705	CTG	CTT
			-3.854	AAT	AAA	-3.727	GGG	AGG
			-3.859	TAG	AAG	-3.749	GTG	GGG
			-3.859	TAG	GAG	-3.751	GGT	GGG

Log odds score	TMC_{coding}		Log odds score	TMC_{+1}		Log odds score	TMC_{+2}	
	wild type	mutant		wild type	mutant		wild type	mutant
			-3.859	TAG	TAT	-3.801	CGT	CGG
			-3.879	AAA	AAT	-3.801	CGT	GGT
			-3.879	AAA	ATA	-3.801	CGT	CGA
			-3.932	ACG	TCG	-3.817	CCT	ACT
			-3.932	ACG	CCG	-3.818	GCG	GGG
			-3.932	ACG	ACT	-3.818	GCG	TCG
			-3.976	TTA	GTA	-3.823	AAC	ATC
			-3.991	CTT	CTA	-3.869	CAT	GAT
			-4.064	CAC	GAC	-3.869	CAT	CAG
			-4.106	TTG	GTG	-3.869	CAT	CAA
			-4.106	TTG	TAG	-3.888	AAA	GAA
			-4.140	GGC	GTC	-3.901	ATT	AAT
			-4.140	GGC	GCC	-3.901	ATT	AGT
			-4.143	CAA	AAA	-3.907	GCA	ACA
			-4.166	AAG	TAG	-3.918	AGA	AGT
			-4.169	TGG	TTG	-3.928	AAG	GAG
			-4.174	GAG	GCG	-3.928	AAG	CAG
			-4.174	GAG	GTG	-3.940	GTC	GTA
			-4.205	ACA	ACT	-3.940	GTC	GAC
			-4.205	ACA	CCA	-3.965	TAT	TTT
			-4.237	TCA	TAA	-3.965	TAT	TAG
			-4.264	TTC	TGC	-4.003	ACT	GCT
			-4.264	TTC	ATC	-4.003	ACT	TCT
			-4.264	TTC	GTC	-4.011	CAA	CAT
			-4.271	TGT	AGT	-4.021	CAG	CTG
			-4.271	TGT	TTT	-4.035	CTT	CTA
			-4.290	GCT	GGT	-4.035	CTT	ATT
			-4.295	AGT	TGT	-4.052	CGA	CGC
			-4.303	GGA	GTA	-4.064	CTA	CAA
			-4.366	GGG	GTG	-4.075	TGG	TCG
			-4.419	ACT	TCT	-4.075	TGG	TTG
			-4.506	TCT	TGT	-4.076	TGA	TGT
			-4.506	TCT	ACT	-4.076	TGA	AGA
			-4.572	AAA	TAA	-4.076	TGA	TCA
			-4.574	TGG	AGG	-4.076	GCT	TCT
			-4.635	GCA	GGA	-4.076	GCT	ACT
			-4.635	GCA	GAA	-4.091	AAT	TAT
			-4.676	TCC	TAC	-4.091	AAT	AAA
			-4.676	TCC	GCC	-4.098	ACA	ACC
			-4.791	TGC	TCC	-4.117	GTT	GGT
			-4.797	TGA	AGA	-4.150	CCG	CAG
			-4.797	TGA	TTA	-4.156	GGA	GGT
			-4.797	TGA	TCA	-4.178	AGT	CGT
			-4.860	AAG	ATG	-4.201	TCC	GCC

Log odds score	TMC_{coding} wild type	TMC_{coding} mutant	Log odds score	TMC_{+1} wild type	TMC_{+1} mutant	Log odds score	TMC_{+2} wild type	TMC_{+2} mutant
			-4.930	TCA	ACA	-4.206	TAA	TTA
			-4.930	TCA	GCA	-4.223	CCT	CCA
						-4.246	GAG	TAG
						-4.246	GAG	GAT
						-4.252	TCT	TCA
						-4.255	TCA	GCA
						-4.255	TCA	TCC
						-4.265	AGG	TGG
						-4.279	GAA	TAA
						-4.279	GAA	GTA
						-4.279	GAA	GAT
						-4.279	GAA	AAA
						-4.287	TTT	ATT
						-4.287	TTT	TAT
						-4.287	TTT	TTA
						-4.287	TTT	GTT
						-4.323	AGA	ATA
						-4.365	GCC	CCC
						-4.365	GCC	GCG
						-4.385	AGC	GGC
						-4.392	GAT	AAT
						-4.392	GAT	TAT
						-4.398	CTG	GTG
						-4.398	CTG	CAG
						-4.444	GGT	AGT
						-4.450	CTC	ATC
						-4.581	AAA	CAA
						-4.581	AAA	ATA
						-4.600	GCA	CCA
						-4.664	TGT	TGA
						-4.768	CCA	ACA
						-4.768	TGG	GGG
						-4.770	GCT	GCG
						-5.016	AGA	GGA
						-5.016	AGA	CGA
						-5.016	AGA	TGA
						-5.110	CAA	CTA
						-5.255	GGA	GTA

Table A.4: Trinucleotide mutation classes composing the M_{intron} mutation spectrum

Log odds score	TMC		Log odds score	TMC		Log odds score	TMC	
	wild type	mutant		wild type	mutant		wild type	mutant
0.976	CGT	CAT	-1.247	TGT	TAT	-1.491	TAT	TAC
0.874	ACG	ATG	-1.250	AAC	AAT	-1.493	GAG	GAA
0.727	CCG	CCA	-1.251	CAA	CAG	-1.493	AGC	AGT
0.687	CGG	TGG	-1.256	GGC	AGC	-1.506	ATT	GTT
0.679	TCG	TCA	-1.262	CTA	TTA	-1.517	CGC	CTC
0.662	GCG	GCA	-1.267	GGA	AGA	-1.517	TAG	CAG
0.642	CCG	CTG	-1.267	AGA	AAA	-1.525	AAT	AGT
0.642	ACG	ACA	-1.269	CCA	CCG	-1.531	ACG	ACC
0.635	CGC	TGC	-1.269	TTC	TTT	-1.533	CGA	CCA
0.632	CGT	TGT	-1.271	GAA	AAA	-1.536	ATC	ATT
0.629	CGA	TGA	-1.272	CCT	CTT	-1.539	GGG	GGA
0.618	CGG	CAG	-1.277	TCC	TCT	-1.542	ATG	ATA
0.564	CGC	CAC	-1.288	GCA	GCG	-1.543	GAC	AAC
0.559	GCG	GTG	-1.296	GGC	GGT	-1.543	CAG	CGG
0.468	TCG	TTG	-1.307	AGT	AAT	-1.545	TCG	TGG
0.414	CGA	CAA	-1.345	CTC	CTT	-1.547	TCG	TCC
-0.782	GTA	ATA	-1.345	AAG	AAA	-1.551	GTG	GCG
-0.892	GTG	ATG	-1.346	TCT	TTT	-1.555	CTC	TTC
-0.970	TAC	TAT	-1.348	ATC	GTC	-1.557	TCG	TCT
-0.972	CAT	CGT	-1.357	TGC	TGT	-1.559	TCA	TCG
-1.028	GTC	ATC	-1.385	GTA	GTG	-1.565	CGG	GGG
-1.048	CAC	CAT	-1.397	ACG	GCG	-1.565	TGT	CGT
-1.063	ATG	GTG	-1.399	ACA	ACG	-1.569	GTC	GTT
-1.078	GTT	ATT	-1.400	CAC	CGC	-1.569	GCG	GGG
-1.080	TAT	TGT	-1.401	GCT	ACT	-1.575	CGT	CTT
-1.085	CTA	CTG	-1.407	CTG	TTG	-1.581	ACG	AGG
-1.115	TAG	TAA	-1.409	CGT	CCT	-1.585	CGT	CGC
-1.140	GAC	GAT	-1.424	CAT	CAC	-1.588	CGA	GGA
-1.144	CCC	CCT	-1.437	CTT	TTT	-1.595	TAC	TGC
-1.151	GGT	AGT	-1.438	GAG	AAG	-1.596	ACG	ACT
-1.159	ATA	GTA	-1.441	CGC	CCC	-1.608	CTG	CCG
-1.165	GCG	ACG	-1.447	TGC	CGC	-1.608	TAC	CAC
-1.187	GCA	ACA	-1.458	CCG	CCC	-1.616	ACA	GCA
-1.190	GGG	AGG	-1.458	AGC	AAC	-1.622	ATA	ATG
-1.190	CGC	CGT	-1.461	ACA	ATA	-1.629	CCC	TCC
-1.196	ACC	ACT	-1.469	GAT	AAT	-1.630	CGT	GGT
-1.202	AGG	AAG	-1.474	ACC	GCC	-1.631	AGG	AGA
-1.223	GCC	GCT	-1.485	ACT	ATT	-1.637	GGT	GGC
-1.224	GCC	ACC	-1.488	ATA	ACA	-1.644	ACT	GCT
-1.237	CAG	CAA	-1.488	TGG	CGG	-1.646	GTG	GTA
-1.243	ATG	ACG	-1.491	GGT	GAT	-1.658	CTG	CTA

Log odds score	TMC		Log odds score	TMC		Log odds score	TMC	
	wild type	mutant		wild type	mutant		wild type	mutant
-1.662	GAT	GAC	-1.895	TTG	TTA	-2.158	TCG	CCG
-1.683	ACC	ATC	-1.896	GCA	GTA	-2.163	GGT	GGG
-1.687	GGC	GAC	-1.900	GCT	GCC	-2.184	TTA	TCA
-1.688	AAC	AGC	-1.910	TGT	TGC	-2.188	GAA	GGA
-1.692	CGT	AGT	-1.910	GCC	GTC	-2.199	ACC	CCC
-1.704	CCT	TCT	-1.910	TTG	TCG	-2.202	TTC	TCC
-1.705	CAT	TAT	-1.912	GTC	GCC	-2.207	CTC	CCC
-1.709	CAC	TAC	-1.915	ACT	ACC	-2.218	GGT	GTT
-1.714	GAC	GGC	-1.931	CGC	GGC	-2.220	TCT	TGT
-1.719	GGG	GAG	-1.941	AGA	AGG	-2.221	TTT	TTC
-1.722	TTG	CTG	-1.944	AAG	GAG	-2.230	TAA	TAG
-1.726	GCT	GTT	-1.946	TGG	TAG	-2.235	AAT	GAT
-1.726	AGG	GGG	-1.950	AAG	AGG	-2.240	TAA	CAA
-1.729	CGA	AGA	-1.950	CGG	CCG	-2.241	CTC	CTG
-1.737	CGG	CTG	-1.952	TTA	TTG	-2.259	TCA	CCA
-1.744	GAT	GGT	-1.956	TCT	TCC	-2.265	AAC	GAC
-1.747	TGA	CGA	-1.962	CCA	CTA	-2.275	AAC	AAA
-1.748	CCG	CCT	-1.970	CGG	CGA	-2.295	TAG	TAC
-1.751	GCG	GAG	-1.971	CGA	CGG	-2.299	TAG	TAT
-1.751	CCC	CTC	-1.972	CCG	CGG	-2.312	TTG	TTT
-1.767	ACG	AAG	-1.973	GAA	GAG	-2.319	TTT	CTT
-1.771	TAG	TGG	-1.975	CTT	CTC	-2.335	AAA	AGA
-1.782	CCG	CAG	-1.980	ATC	ACC	-2.338	GTC	CTC
-1.785	GCG	GCC	-1.980	CCA	TCA	-2.351	CTA	ATA
-1.798	CCT	CCC	-1.986	TCC	CCC	-2.363	GTG	GGG
-1.800	TCC	TTC	-1.998	CGA	CTA	-2.365	CAC	CCC
-1.805	CAA	CGA	-2.018	TCA	TTA	-2.382	TGT	TTT
-1.809	TGC	TAC	-2.024	CAA	TAA	-2.385	CTG	GTG
-1.814	CAG	TAG	-2.026	CGC	AGC	-2.420	AGA	ACA
-1.815	AGT	GGT	-2.027	TGA	TAA	-2.423	ATT	ATC
-1.818	TGG	TGA	-2.033	CTT	CCT	-2.430	GTT	GTC
-1.826	GCG	GCT	-2.046	CCG	TCG	-2.433	GGT	TGT
-1.834	GGA	GAA	-2.053	GTT	GCT	-2.436	ACC	AAC
-1.856	AGA	GGA	-2.055	TCT	CCT	-2.439	GTT	GGT
-1.858	AAT	AAC	-2.057	TCG	TAG	-2.449	CAA	AAA
-1.865	TAT	CAT	-2.073	TAA	TGA	-2.453	GTC	GTG
-1.866	CTA	CCA	-2.082	GTT	TTT	-2.455	GAC	GAG
-1.877	AGC	GGC	-2.083	TGA	TGG	-2.457	GTG	CTG
-1.885	AGT	AGC	-2.094	TTC	CTC	-2.461	ACT	AGT
-1.886	GGA	GGG	-2.123	AAA	GAA	-2.465	GGG	GGT
-1.886	ATT	ACT	-2.135	GAG	GGG	-2.468	GAG	CAG
-1.887	GTA	GCA	-2.143	AAA	AAG	-2.470	TTA	TTT
-1.892	CGG	AGG	-2.148	CTA	GTA	-2.474	CAG	CAC

Log odds score	TMC		Log odds score	TMC		Log odds score	TMC	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.491	CAC	CAG	-2.651	AGT	AGG	-2.755	TAC	TCC
-2.495	CCC	CCG	-2.653	ATT	TTT	-2.756	AGG	AGC
-2.498	GTA	TTA	-2.653	ACT	CCT	-2.760	AGC	AGA
-2.501	GTT	CTT	-2.656	TAA	TTA	-2.761	ACC	AGC
-2.507	GCT	CCT	-2.657	CCC	CCA	-2.764	CCT	ACT
-2.508	TTT	TCT	-2.659	AGC	ATC	-2.765	GGG	GGC
-2.510	TAA	AAA	-2.663	CAA	CCA	-2.775	GAA	TAA
-2.510	TTA	CTA	-2.668	GAC	GAA	-2.776	CCA	GCA
-2.511	GCT	TCT	-2.670	TTA	TAA	-2.778	TGT	GGT
-2.511	AGT	ACT	-2.670	TAT	TTT	-2.781	GCG	TCG
-2.521	GTC	TTC	-2.674	TGC	TTC	-2.784	TTC	TTA
-2.521	GCC	TCC	-2.679	AGA	ATA	-2.785	GAT	GTT
-2.528	ACC	ACA	-2.679	GGC	GGA	-2.789	CGG	CGT
-2.529	AAC	ACC	-2.679	TCC	TGC	-2.791	AAG	AAC
-2.534	GGC	GGG	-2.679	ACA	AGA	-2.791	CGC	CGG
-2.535	ATC	ATG	-2.680	CCT	CGT	-2.792	CCG	GCG
-2.540	GCC	CCC	-2.683	GTC	GGC	-2.803	GAG	GAC
-2.546	TGT	TCT	-2.684	ATC	ATA	-2.804	ATG	ATT
-2.550	GTG	TTG	-2.688	CCT	GCT	-2.805	TGC	TGA
-2.550	GTA	CTA	-2.689	CCA	CCC	-2.808	TCA	TGA
-2.554	AGC	AGG	-2.690	CCC	GCC	-2.812	CAC	CAA
-2.563	TAC	TAG	-2.691	AAT	AAA	-2.813	AGG	ATG
-2.569	GCT	GGT	-2.699	GGG	GTG	-2.823	ATA	AAA
-2.575	TTC	TTG	-2.704	GGC	GTC	-2.824	GCA	GCC
-2.582	AAC	AAG	-2.713	GCA	CCA	-2.826	ACA	ACC
-2.594	CTT	GTT	-2.716	GGT	GCT	-2.828	CCT	CAT
-2.596	ACA	AAA	-2.718	CCC	CAC	-2.833	TCA	TCC
-2.600	CCC	ACC	-2.719	CCC	CGC	-2.837	TTG	GTG
-2.604	GAT	TAT	-2.727	CAA	GAA	-2.838	TTT	TTG
-2.610	GAC	CAC	-2.729	CAG	CAT	-2.838	GCG	CCG
-2.611	GCA	TCA	-2.730	TCT	TAT	-2.844	GCT	GAT
-2.611	CTG	CTC	-2.731	TGC	TGG	-2.848	CGT	CGG
-2.617	TGT	TGG	-2.734	ACA	CCA	-2.849	AAG	AAT
-2.618	AGC	ACC	-2.734	TGG	TGT	-2.859	CTG	ATG
-2.621	CTC	GTC	-2.740	GGC	GCC	-2.862	AGG	ACG
-2.626	TTG	TGG	-2.742	AGT	ATT	-2.863	TGC	GGC
-2.632	GGG	TGG	-2.742	CAG	GAG	-2.864	TGC	TCC
-2.638	TGG	GGG	-2.744	GAC	GCC	-2.866	TGA	GGA
-2.642	AGG	AGT	-2.744	GTA	GGA	-2.876	ACG	CCG
-2.646	GGG	CGG	-2.750	GCC	GGC	-2.880	GGC	TGC
-2.649	GAT	CAT	-2.750	GAA	CAA	-2.883	GGG	GCG
-2.650	ACC	ACG	-2.752	GGT	CGT	-2.884	GCA	GAA
-2.651	TAC	TAA	-2.753	TTG	TTC	-2.886	TGG	TGC

Log odds score	TMC		Log odds score	TMC		Log odds score	TMC	
	wild type	mutant		wild type	mutant		wild type	mutant
-2.889	CCA	ACA	-3.018	AAA	TAA	-3.111	CAC	CTC
-2.894	GGA	GCA	-3.020	ATA	AGA	-3.114	ACT	ACG
-2.896	CAT	CCT	-3.022	CCA	CAA	-3.116	TAT	TAA
-2.897	TTT	TGT	-3.027	TAA	GAA	-3.118	TCA	TAA
-2.900	CTT	ATT	-3.029	GTC	GTA	-3.119	TCA	GCA
-2.902	TTT	GTT	-3.035	TTT	TTA	-3.119	GGC	CGC
-2.911	CAT	AAT	-3.037	TCC	TAC	-3.121	TTA	TGA
-2.912	TGA	TCA	-3.038	GAT	GAG	-3.125	AAA	AAT
-2.913	ACT	AAT	-3.039	TAT	TCT	-3.127	CAC	GAC
-2.913	TGG	TTG	-3.044	TCC	GCC	-3.129	GTG	GTC
-2.917	TCC	TCA	-3.045	GAG	TAG	-3.132	AGT	TGT
-2.925	CAA	CAC	-3.046	TCC	TCG	-3.133	CAC	AAC
-2.925	GCC	GAC	-3.048	AAA	ACA	-3.136	CTC	ATC
-2.926	CTA	CTC	-3.056	AAC	CAC	-3.140	AAC	ATC
-2.928	GCC	GCA	-3.060	AAA	AAC	-3.141	TAA	TAT
-2.928	ATA	TTA	-3.064	GAG	GAT	-3.143	GTG	GAG
-2.931	CGG	CGC	-3.065	CTC	CTA	-3.155	TTA	ATA
-2.933	ATT	ATG	-3.066	TTA	TTC	-3.159	AGA	AGT
-2.934	TTA	GTA	-3.067	TGA	TTA	-3.162	ATT	CTT
-2.937	GTT	GTG	-3.067	GGT	GGA	-3.162	ATA	ATC
-2.955	GCA	GGA	-3.069	TCT	GCT	-3.167	ATA	CTA
-2.961	GAC	TAC	-3.069	ATC	AAC	-3.171	TAA	TCA
-2.968	CAT	CTT	-3.073	CGA	CGC	-3.172	AGA	AGC
-2.968	GAC	GTC	-3.075	TAT	AAT	-3.172	ACA	ACT
-2.969	ACT	TCT	-3.077	TCG	GCG	-3.180	AAT	AAG
-2.969	GGA	TGA	-3.080	ATC	CTC	-3.185	AAT	TAT
-2.975	CTG	CTT	-3.081	TTC	TGC	-3.187	GGA	GGC
-2.976	AAT	ATT	-3.083	ACC	TCC	-3.188	ATT	ATA
-2.977	GGA	GTA	-3.083	GTC	GAC	-3.189	GTA	GTT
-2.981	ATC	TTC	-3.085	ATG	CTG	-3.193	GGA	CGA
-2.982	TAG	GAG	-3.085	TAT	GAT	-3.203	CAG	AAG
-2.984	GCC	GCG	-3.092	TAT	TAG	-3.210	TAC	TTC
-2.984	AAA	CAA	-3.092	TAA	TAC	-3.216	ATT	AGT
-2.987	ATG	TTG	-3.093	ATG	AAG	-3.219	ATC	AGC
-2.988	GTG	GTT	-3.096	CAT	GAT	-3.223	ACT	ACA
-2.991	ATA	ATT	-3.096	GAT	GAA	-3.223	TAC	GAC
-3.002	ATG	AGG	-3.097	TTT	ATT	-3.227	CTT	CGT
-3.002	AAT	CAT	-3.100	CAT	CAG	-3.228	GAA	GAC
-3.002	AGT	AGA	-3.102	ATG	ATC	-3.230	CCA	CGA
-3.009	ATT	AAT	-3.103	AGT	CGT	-3.231	TGT	AGT
-3.010	GTT	GAT	-3.105	TTC	GTC	-3.237	ACA	TCA
-3.013	CTT	CTG	-3.107	AAG	CAG	-3.237	TTG	ATG
-3.016	CGC	CGA	-3.107	CCG	ACG	-3.239	GAA	GCA

Log odds score	TMC		Log odds score	TMC	
	wild type	mutant		wild type	mutant
-3.242	CAT	CAA	-3.490	GAA	GAT
-3.253	CCT	CCG	-3.498	CTA	CGA
-3.253	TGA	TGT	-3.507	CTG	CAG
-3.255	GGA	GGT	-3.514	TAG	AAG
-3.256	TGA	TGC	-3.530	CTC	CAC
-3.260	CAA	CAT	-3.530	TTC	ATC
-3.261	TGG	TCG	-3.532	CCT	CCA
-3.264	TAC	AAC	-3.545	TAG	TCG
-3.266	TCA	TCT	-3.556	CTA	CAA
-3.267	TGT	TGA	-3.561	AGA	TGA
-3.271	CTG	CGG	-3.563	AGC	TGC
-3.276	CTC	CGC	-3.565	GCT	GCA
-3.283	CAG	CCG	-3.566	AGA	CGA
-3.306	GAT	GCT	-3.581	TCT	TCA
-3.318	GTA	GTC	-3.592	AAC	TAC
-3.339	TGA	AGA	-3.599	TCG	ACG
-3.345	AAT	ACT	-3.609	GTT	GTA
-3.349	CGT	CGA	-3.632	TTG	TAG
-3.355	GCA	GCT	-3.728	GAA	GTA
-3.358	TTT	TAT	-3.738	CAA	CTA
-3.361	GCT	GCG	-3.767	TTC	TAC
-3.363	AAG	ACG	-3.808	AAG	TAG
-3.367	CTT	CAT	-3.815	CTT	CTA
-3.367	AAG	ATG			
-3.369	ACG	TCG			
-3.369	TCT	ACT			
-3.375	TCC	ACC			
-3.377	AAA	ATA			
-3.386	CCA	CCT			
-3.386	GTA	GAA			
-3.388	CTA	CTT			
-3.403	GAG	GCG			
-3.407	GAG	GTG			
-3.410	AGG	CGG			
-3.410	TGG	AGG			
-3.414	TCA	ACA			
-3.416	TGC	AGC			
-3.418	CGA	CGT			
-3.430	TCT	TCG			
-3.438	TAG	TTG			
-3.476	AGC	CGC			
-3.479	CAG	CTG			
-3.483	AGG	TGG			

APPENDIX B

REPRESENTATIVE KEY SCRIPTS GENERATED TO COMPLETE THIS STUDY

B.1: predict_mutations_refseq.pl

```
### This script takes a file containing genbank cDNA records and uses a given
### set of log odds scores in order to predict every possible
### mutation that can occur in a gene coding sequence

### Bioperl must be installed for this script to work properly
###
### log odds score file must be in the form of
### <score> <wildtype triplet> <mutant triplet>
### such as
### 1.20 CAA TAA

#! /usr/bin/perl -w
use Bio::SeqIO;
use Bio::Seq;
use Bio::SimpleAlign;
use Bio::AlignIO;
use Bio::Root::IO;
use Bio::Perl ;
gencode();

if ($ARGV[0] eq "") {die
"input is <genbankfile> <0 if all snps to be predicted, 1 if exclude silent SNPs>
<log odds score file>\n";
                }
$syn_set = $ARGV[1]; ## is 0 if all snps; 1 if no syn snps
$matrix_file = $ARGV[2];

$in  = Bio::SeqIO->new(-file => $ARGV[0], '-format' => 'Genbank');


############### grab the next cdna
while ( $seq = $in->next_seq() ) {

      $locus = ""; $chromosome = ""; $version = ""; %exon_hash = ();
      $unique_biological_key = ""; $dna_string = ""; $cdna = "";
  $start = ""; $end = "";

########## get the species, only stick with humans;
      $species = $seq->species();
      $binomial_species = $species->binomial();
      next unless ($binomial_species eq 'Homo sapiens');

##### get gb and version
      $gi = $seq->primary_id;
      $unique_biological_key = $seq->accession_number;
      $version    = $seq->version();
      my @source =
grep { $_->primary_tag eq 'source' } $seq->get_SeqFeatures();
```

```
##### only select genes that have been successfully mapped to chromosomes
##### and are cdna's

	next if ($unique_biological_key =~ /NR|XR/);
	$test = $source[0]->has_tag('chromosome');
	if ($test) {
		my @tag = $source[0]->get_tag_values('chromosome');
		$chromosome = $tag[0];
		}
	else { next;}
	next if ($chromosome =~ /mitochondrion/);

##### get all the relevant annotation data
	my @cds =
grep { $_->primary_tag eq 'CDS' } $seq->get_SeqFeatures();
	foreach $feature ( @cds ) {
		$locus = 9999999999;
		my @tag = $feature->get_tag_values('db_xref');
		foreach $x (@tag) {if ($x =~ /LocusID/) {$locus = $x;}}
		$locus =~ s/LocusID://g;
		$featureseq = $feature->spliced_seq();
		$start = $feature->start;
		$end = $feature->end;
		$cdna = uc($featureseq->seq);
		$len = $featureseq->length()
	}

	$dna_string = uc($seq->seq); $dna_string = $dna_string."NNN";
	next if ($locus == 9999999999);  ## skip if no real Locus id


##### print out individual file of predictions for each genbank record
	$print_name = $unique_biological_key."_".$matrix_file.".preds";
	open(OUT,">PREDS/$print_name");
	calc_snps();
	close OUT;

system("sort -nr PREDS/$print_name > SORT/$print_name.sorted");
}


sub calc_snps{

$aa_pos = -1;
#$position = $start -3;
##### major iterative position loop
for ($i=$start; $i <= $end-2; $i=$i+3){   #use end-2 so dont run out ofDNA
		++$aa_pos;  if ($aa_pos == 0) {++$aa_pos;}
### gets rid of that zero problem

	if ($i == 1) {next;}  ### cannot get the flanking info
	if ($i == 2) {next;} ### cannot get the flanking info

	@nts = ("A","T","G","C"); $nt_byt = "";
  $codon = substr($dna_string,$i-1,3);

	NT: foreach $nt (@nts) {	##pick nucleotide to mutate to
	POS: for ($t = 0; $t<=2; ++$t) {  ##pick position to mutate

			$codon = substr($dna_string,$i-1,3);
			$nt_byt = substr($codon,$t,1);
			next if ($nt_byt eq $nt);
			%datahash = ();
			$position_real= $i; $hold_string = $dna_string;

			if ($t==0) { ### mutation is in position zero
```

```
            ###frame0 data is
                $codon = substr($hold_string,$i-1,3);
                $mutcodon_std = $codon;
  substr($mutcodon_std,0,1) = $nt;
                $match = $codon.$mutcodon_std;
  $file = $matrix_file.0;
                $temp = `grep $match $file`; chomp $temp;
  @holdit = split(/\s+/,$temp);
                $datahash{$nt} = $datahash{$nt}+$holdit[0];

            ###frame2 data is
                $codon2 = substr($hold_string,$i-1-1,3) ;
                $mutcodon = $codon2; substr($mutcodon,1,1) = $nt;
                $match = $codon2.$mutcodon; $file = $matrix_file.2;
                $temp = `grep $match $file`; chomp $temp;
  @holdit = split(/\s+/,$temp);
                $datahash{$nt} = $datahash{$nt}+$holdit[0];

            ###frame1 data is
                $codon1 = substr($hold_string,$i-1-2,3) ;
                $mutcodon = $codon1; substr($mutcodon,2,1) = $nt;
                $match = $codon1.$mutcodon; $file = $matrix_file.1;
                $temp = `grep $match $file`; chomp $temp;
  @holdit = split(/\s+/,$temp);
                $datahash{$nt} = $datahash{$nt}+$holdit[0];

$datahash{$nt} = $datahash{$nt};#/3*100;
if ($syn_set == 1) {next if ($gencode{$codon} eq $gencode{$mutcodon_std});}

print OUT "$datahash{$nt} $aa_pos $position_real $codon $mutcodon_std $gencode{$codon} $gencode{$mutcodon_std}\n";
next;}

if ($t==1) { ### mutation is in position one
            $position_real = $i+1;
            ###frame0 data is
            $codon = substr($hold_string,$i-1,3);
            $mutcodon_std = $codon; substr($mutcodon_std,1,1) = $nt;
            $match = $codon.$mutcodon_std; $file = $matrix_file.0;
            $temp = `grep $match $file`; chomp $temp;
   @holdit = split(/\s+/,$temp);
            $datahash{$nt} = $datahash{$nt}+$holdit[0];

            ###frame2 data is
            $codon2 = substr($hold_string,$i-1-1,3) ;
            $mutcodon = $codon2; substr($mutcodon,2,1) = $nt;
            $match = $codon2.$mutcodon; $file = $matrix_file.2;
            $temp = `grep $match $file`; chomp $temp;
   @holdit = split(/\s+/,$temp);
            $datahash{$nt} = $datahash{$nt}+$holdit[0];

            ###frame1 data is
            $codon1 = substr($hold_string,$i-1+1,3) ;
            $mutcodon = $codon1; substr($mutcodon,0,1) = $nt;
            $match = $codon1.$mutcodon; $file = $matrix_file.1;
            $temp = `grep $match $file`; chomp $temp;
   @holdit = split(/\s+/,$temp);
            $datahash{$nt} = $datahash{$nt}+$holdit[0];

$datahash{$nt} = $datahash{$nt};#/3*100;
if ($syn_set == 1){next if ($gencode{$codon} eq $gencode{$mutcodon_std});}

print OUT "$datahash{$nt} $aa_pos $position_real $codon $mutcodon_std $gencode{$codon} $gencode{$mutcodon_std}\n";
next;}

if ($t==2) { ### mutation is in position one
```

```
            $position_real = $i+2;
            ###frame0 data is
            $codon = substr($hold_string,$i-1,3);
            $mutcodon_std = $codon; substr($mutcodon_std,2,1) = $nt;
            $match = $codon.$mutcodon_std; $file = $matrix_file.0;
            $temp = `grep $match $file`; chomp $temp;
    @holdit = split(/\s+/,$temp);
            $datahash{$nt} = $datahash{$nt}+$holdit[0];

            ###frame2 data is
            $codon2 = substr($hold_string,$i+1,3) ;
            $mutcodon = $codon2; substr($mutcodon,0,1) = $nt;
            $match = $codon2.$mutcodon; $file = $matrix_file.2;
            $temp = `grep $match $file`; chomp $temp;
            @holdit = split(/\s+/,$temp);
            $datahash{$nt} = $datahash{$nt}+$holdit[0];

            ###frame1 data is
            $codon1 = substr($hold_string,$i,3) ;
            $mutcodon = $codon1; substr($mutcodon,1,1) = $nt;
            $match = $codon1.$mutcodon; $file = $matrix_file.1;
            $temp = `grep $match $file`; chomp $temp;
            @holdit = split(/\s+/,$temp);
            $datahash{$nt} = $datahash{$nt}+$holdit[0];

$datahash{$nt} = $datahash{$nt};
if ($syn_set == 1) {
next if ($gencode{$codon} eq $gencode{$mutcodon_std});}
print OUT "$datahash{$nt} $aa_pos $position_real $codon $mutcodon_std $gencode{$codon} $gencode{$mutcodon_std}\n";
next;}

}
}
}
return;
}

sub gencode {
%gencode = (
     'TTT', 'F',     'TCT', 'S',    'TAT', 'Y',    'TGT', 'C',
     'TTC', 'F',     'TCC', 'S',    'TAC', 'Y',    'TGC', 'C',
     'TTA', 'L',     'TCA', 'S',    'TAA', 'X',    'TGA', 'X',
     'TTG', 'L',     'TCG', 'S',    'TAG', 'X',    'TGG', 'W',
     'CTT', 'L',     'CCT', 'P',    'CAT', 'H',    'CGT', 'R',
     'CTC', 'L',     'CCC', 'P',    'CAC', 'H',    'CGC', 'R',
     'CTA', 'L',     'CCA', 'P',    'CAA', 'Q',    'CGA', 'R',
     'CTG', 'L',     'CCG', 'P',    'CAG', 'Q',    'CGG', 'R',
     'ATT', 'I',    'ACT', 'T',   'AAT', 'N',   'AGT', 'S',
     'ATC', 'I',     'ACC', 'T',    'AAC', 'N',    'AGC', 'S',
     'ATA', 'I',     'ACA', 'T',    'AAA', 'K',    'AGA', 'R',
     'ATG', 'M',     'ACG', 'T',    'AAG', 'K',    'AGG', 'R',
     'GTT', 'V',     'GCT', 'A',    'GAT', 'D',    'GGT', 'G',
     'GTC', 'V',     'GCC', 'A',    'GAC', 'D',    'GGC', 'G',
     'GTA', 'V',     'GCA', 'A',    'GAA', 'E',    'GGA', 'G',
     'GTG', 'V',     'GCG', 'A',    'GAG', 'E',    'GGG', 'G',
     'NNN', 'X',);

return;
}
```

B.2: grab_refseq_snps.pl

```
### This script takes a list of genbank (refseq cDNA) files
### and parses through all of them to glean coding region
### point mutations detailed in the record

### with this version of the script, all genbank cDNA records
### are assumed to be located in a specific location
### and there is one file in that directory per genbank
### record.

### This script can be easily modified to point from the
### genbank list to a relational database of genbank
### records, if desired.

### However, the directory of genbank files version
### of this script is shown because it is the most
### generic version

### The reader is asked to please excuse the primitive
### genbank parser, as this was one of my very first
### scripts, and I chose not to update a script
### that worked well.  Bioperl has been
### implemented in much of my other work

#! /usr/bin/perl
use POSIX;

%comp_hash = ('G','C', 'C','G', 'A','T', 'T','A');

%gencode = (
       'TTT', 'F',      'TCT', 'S',     'TAT', 'Y',     'TGT', 'C',
       'TTC', 'F',      'TCC', 'S',     'TAC', 'Y',     'TGC', 'C',
       'TTA', 'L',      'TCA', 'S',     'TAA', 'X',     'TGA', 'X',
       'TTG', 'L',      'TCG', 'S',     'TAG', 'X',     'TGG', 'W',

       'CTT', 'L',      'CCT', 'P',     'CAT', 'H',     'CGT', 'R',
       'CTC', 'L',      'CCC', 'P',     'CAC', 'H',     'CGC', 'R',
       'CTA', 'L',      'CCA', 'P',     'CAA', 'Q',     'CGA', 'R',
       'CTG', 'L',      'CCG', 'P',     'CAG', 'Q',     'CGG', 'R',

       'ATT', 'I',      'ACT', 'T',     'AAT', 'N',     'AGT', 'S',
       'ATC', 'I',      'ACC', 'T',     'AAC', 'N',     'AGC', 'S',
       'ATA', 'I',      'ACA', 'T',     'AAA', 'K',     'AGA', 'R',
       'ATG', 'M',      'ACG', 'T',     'AAG', 'K',     'AGG', 'R',

       'GTT', 'V',      'GCT', 'A',     'GAT', 'D',     'GGT', 'G',
       'GTC', 'V',      'GCC', 'A',     'GAC', 'D',     'GGC', 'G',
       'GTA', 'V',      'GCA', 'A',     'GAA', 'E',     'GGA', 'G',
       'GTG', 'V',      'GCG', 'A',     'GAG', 'E',     'GGG', 'G',

       'NNN', 'X',
);

open (ERROR, ">not_openable.log");  ### genbank files that cannot be found
open (LOG, ">grab_snps.log"); ### log file
open (FAIL,">non_snps.log"); ### SNPs with nonsensical mapping info
open (COD, ">SNPS.coding");
open (CODP, ">SNPS.phenotype_coding");

open (IN,"genbank_accession_list_uniq") ||
       die "cannot open genbank list file\n";

LOOP: while ($q = <IN>) {
print $q;
```

```
chomp $q;
undef %moocow;

#### loop through each genbank record of a locus
open (REC,"\/usr3/GENBANK_OCT/$q.gb") || print ERROR "$q not open-able\n";
undef $/; $record = <REC>; $/ = "\n";
if ($record !~ "   variation     ") {print LOG "$q no SNPs\n"; next;}
else {@parsed = get_genbank_info($record);
        $dna = get_genbank_dna($record);}
if (length($dna) < 20) {print LOG "ERROR, did not capture DNA $q\n"; next;}

$feature_ref = $parsed[0];          #create references to genbank data
$cds_frag_ref = $parsed[1];
$codonst_ref = $parsed[2];
$translat_ref = $parsed[4];
$variation_ref = $parsed[10];
$$feature_ref[2] =~ s/GI://g;

## ignore all things without codon_st of 1!
if($$codonst_ref[0] != 1) {
print LOG "codon start is not 1; skipping record $q\n"; next;}

#### print out the data about each SNP
#### loop through each "variation" section of a genbank file
LOOP: foreach $y (@$variation_ref)
        {
        $phenotype = "";
        $frequency = "NONE";
        $db_snp = "0000000";
        $mut = "";
        $all_ct = "";
        @allele = ();
        $complement = "";

          #### sort out the variation data and assign all annotation to the #### appropriate
          varaiables
        #### skip inappropriate records
        if ($y !~ "\/allele|\/replace") {
print FAIL "non-SNP mutation skipped in $iter $q $y\n"; next;}

        if ($y !~ "\/allele|\/replace") {
        print FAIL "non-SNP mutation skipped in $iter $q $y\n"; next;}
        if ($y =~ "\/replace=\"\"") {
print FAIL "deletion mutation skipped in $iter $q $y\n"; next;}
        if ($y =~ "evidence=not_experimental") {
print FAIL "ERROR: non-experimental SNP in $iter $q $y\n"; next;}

        #### grab the annotation data
        @bill = split(/\n/,$y);
        foreach $r (@bill) {
                if ($r =~ "     variation       ") {
                @hold = split(/\s+/,$r); $position = $hold[2]; }
                     if ($r =~ "\/phenotype") {@hold = split(/\"/,$r);
                                $phenotype = $hold[1]; }
                if ($r =~ "\/frequency") {@hold = split(/\"/,$r);
                                $frequency = $hold[1]; }
                if ($r =~ "\/db_xref") {@hold = split(/\"/,$r);
                                $db_snp = $hold[1]; $db_snp =~ s/dbSNP://g; next;}
                if ($r =~ "\/replace") {@hold = split(/\"/,$r);
                                $allele[$all_ct] = uc($hold[1]); ++$all_ct;}
                if ($r =~ "\/allele") {@hold = split(/\"/,$r);
                        $allele[$all_ct] = uc($hold[1]);
                                        ++$all_ct;}
                }
        ### fix complement records -- do the alleles listed make sense?
        if ($position =~ "complement") {
                $position_hold = $position;
```

```
                $position =~ s/complement\(//g; $position =~ s/\)//g;
                $wt = uc(substr($dna,$position-1,1));
                   $sensical_allele = 0;
        foreach $allele_loopydoopy (@allele) {
                if ($comp_hash{$allele_loopydoopy} eq $wt)
                      {++$sensical_allele }
                }
        $position = $position_hold;
        if ($sensical_allele != 1) {
               print LOG "alleles do not make sense; skip this one!\n";
                      next LOOP;}
        print LOG "sensical allele num is $sensical_allele\n";
        }
        $complement = "";

ALLELE: foreach $allele_loop (@allele) {
        $mut = $allele_loop;
        if (length($mut) != 1) {
               print FAIL "bad length SNP in $iter $q $y\n"; next;}
        if ($position =~ '\.\.') {
               print FAIL "$position weirdo non-SNP in $q $y\n"; next;}
        if ($mut eq "-") {
               print FAIL "non-SNP mutation skipped in $iter $q $y\n"; next;}

        ### fix complement records
        if ($position =~ "complement" || $complement eq "YES") {
                $position =~ s/complement\(//g; $position =~ s/\)//g;
          print LOG "MATCHER $comp_hash{$allele_loop} $allele_loop $wt\n";
                if ($comp_hash{$allele_loop} eq $wt) {
                      print LOG "$db_snp comp_hash equals wt $allele_loop\n";
                $complement = "YES"; next ALLELE;}
                $mut = $comp_hash{$allele_loop};
                $complement = "YES";
                }
        ### assign all the variables
else {

$wt = uc(substr($dna,$position-1,1));
        if ($allele_loop eq $wt) {next;}}
        print LOG "Ready to go for: $position $db_snp\n";
        $length = $$cds_frag_ref[0][0][1] - $$cds_frag_ref[0][0][0] +1;

        ### take care of coding SNPs
        if ($position <= $$cds_frag_ref[0][0][1] &&
                   $position >= $$cds_frag_ref[0][0][0]) {
                $codon_pos =
                      ceil(($position-$$cds_frag_ref[0][0][0] +1)/3);
                $mod = ($position-$$cds_frag_ref[0][0][0] +1)%3;
                $aa = substr($$translat_ref[0],$codon_pos-1,1);
                if ($mod == 0) {$codon = substr($dna,$position-3,3);
                      $mut_codon = $codon; substr($mut_codon,2,1) = $mut;}
                if ($mod == 1) {$codon = substr($dna,$position-1,3);
                      $mut_codon = $codon; substr($mut_codon,0,1) = $mut;}
                if ($mod == 2) {$codon = substr($dna,$position-2,3);
                      $mut_codon = $codon; substr($mut_codon,1,1) = $mut;}
                $translated = $gencode{$codon};
                $translated_mut = $gencode{$mut_codon};
                if ($codon =~ /N|X/) {next;}
                print "$codon $position $mut_codon $mut $mod\n";
                if ($aa ne $translated) {
               print LOG "ERROR TRANSLATION $position $q $iter\n"; next;}

####the following line makes the snp "unique"
next if ($dbsnp_hash{$db_snp} eq "GOT");
                       $dbsnp_hash{$db_snp} = "GOT";
                        $moocow = $codon.$mut_codon.$position;
                        next if ($moocow_hash{$moocow} eq "GOT");
```

```
                        $moocow_hash{$moocow} = "GOT";
                if ($phenotype eq "") {
                print COD
"$codon\t$mut_codon\t$q\t$db_snp\t$$feature_ref[4]\t$position\t$codon_pos
\t$aa\t$translated_mut\t$$cds_frag_ref[0][0][0]\t$$cds_frag_ref[0][0][1]\t$frequency\n";}
                else {
print CODP
"$codon\t$mut_codon\t$q\t$db_snp\t$$feature_ref[4]\t$position\t$codon_pos\t$aa\t$translated_m
ut\t$$cds_frag_ref[0][0][0]\t$$cds_frag_ref[0][0][1]\t$frequency\t$phenotype\n";}
next;           } ### end coding SNPs loop
        }  ## end $allele_loop
        }  ## end $y each SNP
} ## end each gb record

###################### FUNCTIONS ONLY FOLLOW
###################### function to grab only the dna of the record
sub get_genbank_dna {
my (@line,@data,$x,$flag,@data,$wholestring);
@data=(); $flag = 0;
@line = split(/\n/,$_[0]);

foreach $x (@line) {
unless ($x =~ "ORIGIN" || $flag == 1) {next;}
$flag = 1;
if ($x =~ /a|t|g|c|n/ && $x !~ "ORIGIN") {@data = split(/\s+/,$x);
  $wholestring =
"\U$wholestring"."\U$data[2]"."\U$data[3]"."\U$data[4]"."\U$data[5]"."\U$data[6]"."\U$data[7]
";}
}
unless (@data) {return("ERROR");}
return($wholestring);
}

####################
#################### function to get info from genbank feature table
sub get_genbank_info{

my
($variation_flag,@variation,$variation_count,@line,$x,@hold,@data,@feature_array,@cds_fragnum
,@codon_start,@cds_line,@cds_fragnum,@product,$cds_flag,$ident_flag,$count,$k,$hold,$l,@segac
c_fragnum,@complement,@hold2,@hold3,@translation,$gene_flag,$g_count,$gene_wait);

##### @feature_array is [0] = definition line; [1] is accession;
##### [2] is dna gi number; [3] is dna type (genomic or cDNA)
##### [4] = organism
##### intialize everything
$count = -1; $cds_flag = 0; $cds_wait = 0; $ident_flag = 0; $translation_flag = 0;@data  =
();
$product_flag = 0; $number_of_cds = 0; $variation_count = -1;
@line = split(/\n/,$_[0]);   # create array to iterate over


foreach $x (@line) {
if ($x =~ " CDS ") {++$number_of_cds;}  #### count the number of cdna reads in this record
if ($variation_flag > 0 && substr($x,0,21) ne  "                     ") {$variation_flag = 0;
}
        chomp $x;
        @data = split(/\s+/,$x);

if ($x =~ "  ORGANISM  ") {$feature_array[4] = "$data[2]"."_$data[3]"; next;}

# FOLLOWING IS THE FLAG BOX; USED MOSTLY FOR THE CDS LINE PARSING
#--------------------------------Flag Box--------------------------------
        if ($ident_flag == 1 && substr($x,0,12) eq "            ")
                {$hold = substr($x,12);
                $feature_array[0] = $feature_array[0]."_".$hold;next;}
        elsif ($ident_flag == 1) {$ident_flag = 0;}
```

```
        if($translation_flag == 1 && substr($x,0,20)
               eq "                    ") {
                $translation[$count] = "$translation[$count]"."$x";
                $translation[$count] =~ s/\s//g;
                $translation[$count] =~ s/"//g;
                $translation[$count] =~ s/=//g;
                next;}
        elsif ($translation_flag == 1) {$translation_flag = 0; }


##############################################  CDS LINE BOX

# following asks if the current line is just a continuation of a "CDS" #line
# if it is not, turn off the cds line flag and start to process the data
# genbank structures each CDS line so it describes the location of the dna #for each exon

        if ($cds_flag == 1 && substr($x,21,1) ne "/") {
            $cds_line[$count] = $cds_line[$count].$data[1];next;}

        elsif ($cds_flag == 1) {$cds_flag = 0;
                if ($cds_line[$count] =~ "order") {next LOOP;}
                if ($cds_line[$count] =~ "group") {next LOOP;}
                if ($cds_line[$count] =~ "one-of") {next LOOP;}
                if ($cds_line[$count] =~ "join") {next LOOP;}

################ MAKE ALLOWANCE FOR RECORDS CALLING COMPLEMENT DNA
####### $complement[$count] = 0 then no complement consids for ORF number $count in this
genbank record####### $complement[$count] = 1 then add complement consids for ORF number
$count in this genbank record


if ($cds_line[$count] =~ "complement") {
        $complement[$count] = 1;
        $cds_line[$count] =~ s/complement//g;}
else {$complement[$count] = 0;}

################ PROCESS SEGMENTED and NORMAL RECORDS
####### parse out cd_line into fragments
####### cds_fragnum array is [a][b][c]
####### a=cdna numb, which ORF;
####### b=which exon of the ORF; c[0] = start; c[1] = end;
####### segacc_fragnum array is [a][b]
####### a=cdna numb, which ORF; b=which exon of the ORF;
####### $segacc_fragnum[a][b] = the accession number containing the dna
####### still part of previous "elsif" loop

                # get rid of all of the crap
            $cds_line[$count] =~ s/>//g; $cds_line[$count] =~ s/<//g;
            $cds_line[$count] =~ s/join//g; $cds_line[$count] =~ s/\)//g;
            $cds_line[$count] =~ s/\(//g;

            @hold = split(/,/,$cds_line[$count]); $k = 0;
       $cds_wait = 1;# flag to be sure that only product taken is the cDNA
                #iterate through each exon cdsline contains
        foreach $l (@hold) {
                if ($l =~ ":") {
                        @hold3 = split(/:/,$l);
                        $segacc_fragnum[$count][$k]=$hold3[0];
                        @hold2 = split(/\../,$hold3[1]); # start/end tier
                        $cds_fragnum[$count][$k][0] = $hold2[0];
                        $cds_fragnum[$count][$k][1] = $hold2[1];
                        ++$k;}

                else{
                        @hold2 = split(/\../,$l); # start/end tier
```

```
                        $segacc_fragnum[$count][$k] = $feature_array[1];
                        $cds_fragnum[$count][$k][0] = $hold2[0];
                        $cds_fragnum[$count][$k][1] = $hold2[1];
                        ++$k;
                        }
        }

}    # terminates elsif referred to as LOOP:


#----------------------------end flag box-------------------------------------------


##############################################################
##############################################################
##### JUMP TO HERE IF USER IS READING THROUGH THE SCRIPT FOR FIRST TIME
        #get definition of dna; enter flag box to get entire definition on #  multiple lines
if ($data[0] eq "DEFINITION") {$feature_array[0]=substr($x,12);
        $ident_flag = 1; next;}

       if ($data[0] eq "ACCESSION") {$feature_array[1] = $data[1];
      print LOG "*********** GB:$feature_array[1] parsing initiated\n";next;}
        if ($data[0] eq "VERSION") {$feature_array[2] = $data[2];next;}

# start to grab the difficult CDS line
if ($data[1] eq  "CDS") {++$count;$cds_flag= 1;
        $cds_line[$count] = $data[2];next;}

        if ($data[1] =~ "/translation") {$translation_flag = 1;
                $translation[$count] = $x;
                $translation[$count] =~ s/\s//g;
                $translation[$count] =~ s/\/translation//g;
                $translation[$count] =~ s/"//g;
                $translation[$count] =~ s/=//g;
}
        if ($data[1] =~ "/pseudo") {$feature_array[1]="pseudo";}
        if ($data[1] =~ "/codon_start") {@hold = split(/=/,$data[1]);
                        $codon_start[$count] = $hold[1];next;}
        if ($cds_wait == 1 && $data[1] =~ "/gene=") {
                $gene_of_prot[$count] = substr($x,28);
                $gene_of_prot[$count] =~ s/"//g;
                        }
        if ($cds_wait == 1 && $data[1] =~ "/product=") {
                $cds_wait =0;
                $product[$count] = substr($x,30);
                $product[$count] =~ s/"//g;
                $product_flag = 1;next;
        }
        if ($x =~ "variation") {

                ++$variation_count; $variation[$variation_count] = $x;
                $variation_flag = 1; next}
        if ($variation_flag == 1)
        {
          $variation[$variation_count] = $variation[$variation_count]."\n"."$x"; }
        if ($data[0] eq "BASE COUNT") {last;}


}
$feature_array[0] =~ s/,//g; $feature_array[0] =~ s/\.//g;
@hold = split(/\s+/,$feature_array[0]); $feature_array[0] = "";
foreach $h (@hold) {$feature_array[0] = $feature_array[0]."$h"."_";}
return
\(@feature_array,@cds_fragnum,@codon_start,@product,@translation,@segacc_fragnum,@complement,
@gene_of_prot,@gene_ident,@gene_line,@variation);
}
```

B.3 make_logodds_scores.pl

```
#! /usr/bin/perl
use DBI;

### This script generates a set of log odds scores for
### a set of coding region SNPs
### user enters an output file and 'frame',
### 0 = coding frame, 1 = frame +1, 2=frame +2
### change one of the lines in the code in order
### to choose only silent, nonsynonymous or all
### SNPs

### this script prints out codon usage-weighted frequency based on
### what frame you choose (+1, +2, 0).  0 is the coding frame
### note that you must have the correct usage hash and that
### $ARGV[1] is required to designate frame;
### dont forget the exclude array if you need it...

### with this version of the script, all genbank cDNA records
### are assumed to be located in a specific location
### and there is one file in that directory per genbank
### record.

### This script can be easily modified to point from the
### genbank list to a relational database of genbank
### records, if desired.

### However, the directory of genbank files version
### of this script is shown because it is the most
### generic version

### input SNP file should be tab delimited
### [0] = wild_type trinucleotide
### [1] = mutant trinucleotide
### [2] = gene accesseion or an id
### [3] = snp uniq id, such as an rs#
### [5] = position in DNA sequence
### [6] = position in protein sequence

$out_file = $ARGV[0]; $frame = $ARGV[1];
if ($out_file eq "") {die "ERROR: input is <outputfile> <frame>\n";}
open (OUT,">$out_file");
$global_ct = 0;         ## global mutation counter;
open (SNP,"SNPS.coding") || die "cannot open SNP file\n";

##### GRAB THE CODON TRANSITION CLASS RAW NUMBERS FROM THE DATABASE
LOOP: while ($cycle = <SNP>) {
chomp $cycle; $cd_start = "YAY"; $cd_end = "YAY";
@data = split(/\s+/,$cycle);
$wild_type = $data[0];
$mutant = $data[1]; $rs_num =$data[3]; $gene_id = $data[2];
$position = $data[5]; $codon = $data[6];
$newwild_type=""; $newmutant = ""; undef @dna; $gene_sequence = "";

### correct if strings match in less than 2 places; excludes the mutation
$match_count = 0;
for ($i = 1; $i <=3; ++$i) {
        if (substr($wild_type,$i-1,1) eq substr($mutant,$i-1,1))
                {++$match_count; }
        else {$snpplace = $i; $allele = substr($mutant,$i-1,1)}
}
next if ($match_count != 2);

### grab the cdna
```

```
($whole_dna,$cd_start,$cd_end) = get_genbank_cdna($gene_id);
$gene_sequence = substr($whole_dna,$cd_start-1,$cd_end-$cd_start+1);
$old_codon = $codon;

if ($frame != 0 && $codon == 1 && $snpplace =~ /1|2/) {
               print "skipped "; next;}
if ($frame !=0  && $codon == $length) {print "skipped " ; next;}

if ($snpplace == 1) {
        if ($frame == 0) {
               $newwild_type = substr($gene_sequence,($codon*3-3),3);
            $newmutant = $newwild_type; substr($newmutant,0,1) = $allele;}

        if ($frame == 1) {
               $newwild_type = substr($gene_sequence,$codon*3-5,3);
            $newmutant = $newwild_type; substr($newmutant,2,1) = $allele;}

        if ($frame == 2) {
               $newwild_type = substr($gene_sequence,$codon*3-4,3);
            $newmutant = $newwild_type; substr($newmutant,1,1) = $allele;}
}

if ($snpplace == 2) {
        if ($frame == 0) {
               $newwild_type = substr($gene_sequence,$codon*3-3,3);
            $newmutant = $newwild_type;
            substr($newmutant,1,1) = $allele;}

        if ($frame == 1) {
               $newwild_type = substr($gene_sequence,$codon*3-2,3);
            $newmutant = $newwild_type; substr($newmutant,0,1) = $allele;}

        if ($frame == 2) {
               $newwild_type = substr($gene_sequence,$codon*3-4,3);
            $newmutant = $newwild_type; substr($newmutant,2,1) = $allele;}
}

if ($snpplace == 3) {
        if ($frame == 0) {
               $newwild_type = substr($gene_sequence,$codon*3-3,3);
            $newmutant = $newwild_type; substr($newmutant,2,1) = $allele;}

        if ($frame == 1) {
               $newwild_type = substr($gene_sequence,$codon*3-2,3);
            $newmutant = $newwild_type; substr($newmutant,1,1) = $allele;}

        if ($frame == 2) {
               $newwild_type = substr($gene_sequence,$codon*3-1,3);
            $newmutant = $newwild_type; substr($newmutant,0,1) = $allele;}
}

$sum = $newwild_type.$newmutant; next if ($sum =~ /N|X|M|W|Y|R|S|K|W/);
next if (length($newwild_type) != 3);
++$global_count;   ### global mutation counter; what is being analyzed
++$big_ass_hash{$sum};
}

$sum_tonormalize = 0; $sum_normal =0;
foreach $cycle (keys %big_ass_hash) {
$wild_type = substr($cycle,0,3);
$big_ass_hash{$cycle} = log($big_ass_hash{$cycle}/$global_count/usage_calc0($wild_type))
               if ($frame == 0);
$big_ass_hash{$cycle} = log($big_ass_hash{$cycle}/$global_count/usage_calc1($wild_type))
               if ($frame == 1);
$big_ass_hash{$cycle} = log($big_ass_hash{$cycle}/$global_count/usage_calc2($wild_type))
               if ($frame == 2);
$sum_tonormalize = $sum_tonormalize + $big_ass_hash{$cycle};
```

```
}

$correction = 1;
foreach $x(sort{$big_ass_hash{$b}<=>$big_ass_hash{$a}} keys %big_ass_hash)
{       normalized = $big_ass_hash{$x}*$correction;
        $start = substr($x,0,3); $end = substr($x,3,3); $startaa = gencode($start);
        $endaa = gencode($end);
        print OUT "$normalized $start$end $startaa $endaa\n";
        $sum_normal = $sum_normal + $normalized;
}
print "TOTAL: NORMALIZED = $sum_normal for $global_count mutations\n";

sub get_genbank_cdna {
        my (@line,@data,$x,$flag,@data,$string,
                $cd_start,$cd_end,$wholestring);
        @data=(); $flag = 0;

        open (GB,"\/usr3/GENBANK_OCT/$_[0].gb") ||
                die "cannot open $_[0]\n";
        undef $/; $string = <GB>; $/ = "\n"; close GB;

        @line = split(/\n/,$string);
        foreach $x (@line) {
        if ($x =~ /^                    \/codon_start=/)
                {if ($x ne /                    \/codon_start=1/)  {
            die "codon start is screwed!\n";}}
if ($x =~ /^     CDS             /) {$x =~ s/     CDS             //g;
        print "in subrout $x\n";
              $x =~ s/>//g; $x =~ s/<//g;
        @data = split(/\.\./,$x); $cd_start = $data[0];
                $cd_end = $data[1]; next;}

        unless ($x =~ "ORIGIN" || $flag == 1) {next;}
        $flag = 1;
        if ($x =~ /n|a|t|g|c/ && $x !~ "ORIGIN") {@data = split(/\s+/,$x);
           $wholestring =
        "\U$wholestring"."\U$data[2]"."\U$data[3]"."\U$data[4]"."\U$data[5]"."\U$data[6]"."\U$
        data[7]";}
}
unless (@data) {return("ERROR");}
return($wholestring,$cd_start,$cd_end);
}

sub gencode {
%gencode = (
       'TTT', 'F',      'TCT', 'S',     'TAT', 'Y',     'TGT', 'C',
       'TTC', 'F',      'TCC', 'S',     'TAC', 'Y',     'TGC', 'C',
       'TTA', 'L',      'TCA', 'S',     'TAA', 'X',     'TGA', 'X',
       'TTG', 'L',      'TCG', 'S',     'TAG', 'X',     'TGG', 'W',

       'CTT', 'L',      'CCT', 'P',     'CAT', 'H',     'CGT', 'R',
       'CTC', 'L',      'CCC', 'P',     'CAC', 'H',     'CGC', 'R',
       'CTA', 'L',      'CCA', 'P',     'CAA', 'Q',     'CGA', 'R',
       'CTG', 'L',      'CCG', 'P',     'CAG', 'Q',     'CGG', 'R',

       'ATT', 'I',      'ACT', 'T',     'AAT', 'N',     'AGT', 'S',
       'ATC', 'I',      'ACC', 'T',     'AAC', 'N',     'AGC', 'S',
       'ATA', 'I',      'ACA', 'T',     'AAA', 'K',     'AGA', 'R',
       'ATG', 'M',      'ACG', 'T',     'AAG', 'K',     'AGG', 'R',

       'GTT', 'V',      'GCT', 'A',     'GAT', 'D',     'GGT', 'G',
       'GTC', 'V',      'GCC', 'A',     'GAC', 'D',     'GGC', 'G',
       'GTA', 'V',      'GCA', 'A',     'GAA', 'E',     'GGA', 'G',
       'GTG', 'V',      'GCG', 'A',     'GAG', 'E',     'GGG', 'G',

       'NNN', 'X',
);
```

```
return($gencode{$_[0]});
}

sub usage_calc0 {
my(%c_use);
%c_use = ("TTT","16.9","TCT","14.6","TAT","12.0","TGT","9.9",
"TTC","20.4","TCC","17.4","TAC","15.6","TGC","12.2",
"TTA","7.2","TCA","11.7","TAA","0.7","TGA","1.3",
"TTG","12.5","TCG","4.5","TAG","0.5","TGG","12.8",
"CTT","12.7","CCT","17.3","CAT","10.4","CGT","4.7",
"CTC","19.4","CCC","20.0","CAC","14.9","CGC","10.9",
"CTA","6.9","CCA","16.7","CAA","11.8","CGA","6.4",
"CTG","40.3","CCG","7.0","CAG","34.7","CGG","11.9",
"ATT","15.7","ACT","12.8","AAT","16.7","AGT","11.9",
"ATC","21.5","ACC","19.2","AAC","19.5","AGC","19.3",
"ATA","7.1","ACA","14.8","AAA","23.9","AGA","11.4",
"ATG","22.3","ACG","6.2","AAG","32.9","AGG","11.4",
"GTT","10.9","GCT","18.6","GAT","22.3","GGT","10.8",
"GTC","14.6","GCC","28.6","GAC","26.1","GGC","22.8",
"GTA","7.0","GCA","15.9","GAA","29.1","GGA","16.3",
"GTG","29.0","GCG","7.6","GAG","40.8","GGG","16.4");
return($c_use{$_[0]}/1000);
}

sub usage_calc1 {
my(%c_use);
%c_use = ("GGT", "0.0061043110920918", "TGG", "0.0384508336392964", "TAT",
"0.00487778009431505", "TCT", "0.0176426147274011", "TGT", "0.0139741255286904", "ATA",
"0.0120015948555535", "ATC", "0.0115984978903521", "CTA", "0.00897230436027021", "CTC",
"0.0154721150844027", "GTA", "0.00629745867744748", "ATG", "0.0293690984609995", "GTC",
"0.00782437654019493", "CTG", "0.029491532406936", "GTG", "0.0183198000966322", "AAA",
"0.0194076797642255", "ACA", "0.0262986070435752", "AAC", "0.0126242839696124", "CAA",
"0.0127275784800863", "CAC", "0.0124689769474997", "ACC", "0.0214902256656323", "CCA",
"0.0305189256143649", "GAA", "0.0103858466358497", "ATT", "0.00996682449152291", "AAG",
"0.0261285436628233", "AGA", "0.0311223291904004", "TTA", "0.0104695634031786", "CCC",
"0.0226437053830174", "GAC", "0.0066064655935044", "CGA", "0.00335378428278326", "TTC",
"0.014317758751653", "AGC", "0.027164264716217", "CAG", "0.0269858734895004", "GCA",
"0.0208371472189305", "ACG", "0.0109167833417367", "CTT", "0.0109129846751563", "CCG",
"0.0118217426032291", "GCC", "0.0193429563297985", "CGC", "0.00791773607653553", "TTG",
"0.0226229588193861", "GGA", "0.0151539037070164", "AGG", "0.03351870339237", "GTT",
"0.006088970323209 55", "GAG", "0.0135082044623521", "GCG", "0.00964525275523862", "CGG",
"0.0107741872424121", "GGC", "0.0137162545089077", "GGG", "0.01680690807972", "TAA",
"0.00857037621555504", "AAT", "0.00971903454843424", "TAC", "0.00635502308639614", "ACT",
"0.0161257779413477", "CAT", "0.00863436913717818", "TCA", "0.0266000166264714", "TCC",
"0.0213721747965193", "CCT", "0.0191771299233093", "GAT", "0.00452465020490154", "AGT",
"0.0147033234095604", "TAG", "0.00961749326868977", "TTT", "0.0103354412523795", "TGA",
"0.0239821509423542", "CGT", "0.00331638202722271", "GCT", "0.0152355750384944", "TGC",
"0.0249549017920502", "TCG", "0.00811380571310688");
return($c_use{$_[0]}/1);
}

sub usage_calc2 {
my(%c_use);
%c_use = ("GGT", "0.0178361745613314", "TGG", "0.0237169401434754", "TAT",
"0.0100335787024855", "TCT", "0.0144306749475529", "TGT", "0.0204490689408763", "ATA",
"0.00469426841952471", "ATC", "0.00806324244647212", "CTA", "0.0114088400978477", "CTC",
"0.0182839782558799", "ATG", "0.00534456996746447", "GTA", "0.00636524096587093", "GTC",
"0.0108114275996392", "CTG", "0.0166347749592849", "GTG", "0.00893547345999725", "AAA",
"0.0195252637987458", "ACA", "0.0125801426163453", "CAA", "0.0319284764724816", "AAC",
"0.00945457510003993", "ACC", "0.0105163008580445", "CAC", "0.022081470467529", "CCA",
"0.0243749851158234", "CCC", "0.019665083745135", "TTA", "0.00580143200754706", "AAG",
"0.0105070964101631", "GAA", "0.0289259271326231", "ATT", "0.00965371259944274", "AGA",
"0.0312521687067181", "ACG", "0.00403622344717677", "TTC", "0.0121180208917591", "CGA",
"0.0125811653327766", "CAG", "0.0235083059914966", "GAC", "0.0138782619723203", "AGC",
"0.0165327955208527", "CTT", "0.0218532586010087", "GCA", "0.0208131559904081", "CGC",
```

```
"0.00900092731159847", "GCC", "0.0180510911142452", "CCG", "0.0129147169917176", "GAG",
"0.0151557808973226", "AGG", "0.0154260702398723", "GGA", "0.0394943865286623", "TTG",
"0.00537554366509719", "GTT", "0.011985944369778", "GCG", "0.0100156081137646", "CGG",
"0.00959278792060857", "GGC", "0.024026384915108", "GGG", "0.0181960246427907", "TAA",
"0.0135131522063573", "AAT", "0.01160417893622", "TAC", "0.00802204158452671", "ACT",
"0.0109223192812583", "CAT", "0.0267396516072646", "TCA", "0.0148657677378842", "TCC",
"0.01407374691304", "CCT", "0.0278953211745986", "AGT", "0.0130291151296724", "TAG",
"0.00617238586740306", "TGA", "0.0350267228498373", "TTT", "0.0140088774708281", "GAT",
"0.0156513600594461", "CGT", "0.0093224985780589", "TCG", "0.00584175625540853", "GCT",
"0.0248730480178514", "TGC", "0.0206067133736391");
return($c_use{$_[0]}/1);
}
```

APPENDIX C
COMPLETE LIST OF CORRELATIONS TO GENE ONTOLOGY CLASSES

Table C.1: Correlations to $M_{\mathbf{intron}}$-ranked gene list

System	Gene Category	Percentile (%)	# genes	# in category	p-value
GO Biological Process	defense response	88.3-100	163	787	3.35E-15
GO Biological Process	response to biotic stimulus	88.3-100	170	849	1.89E-14
GO Biological Process	response to pest/pathogen/parasite	88.3-100	105	447	1.28E-12
GO Biological Process	immune response	88.3-100	144	713	4.19E-12
GO Biological Process	response to external stimulus	88.3-100	188	1060	1.09E-10
GO Biological Process	taxis	88.3-100	41	118	2.87E-09
GO Biological Process	chemotaxis	88.3-100	41	118	2.87E-09
GO Biological Process	response to stimulus	88.3-100	213	1320	3.24E-08
GO Biological Process	response to wounding	88.3-100	66	274	1.30E-07
GO Biological Process	response to chemical substance	88.3-100	56	223	7.09E-07
GO Biological Process	biological_process unknown	88.3-100	148	863	1.03E-06
GO Biological Process	response to stress	88.3-100	137	782	1.16E-06
GO Biological Process	nucleosome assembly	88.3-100	25	61	1.30E-06
GO Biological Process	antimicrobial humoral response	88.3-100	31	89	1.49E-06
GO Biological Process	antimicrobial humoral response (sensu Vertebrata)	88.3-100	30	88	5.06E-06
GO Biological Process	organismal physiological process	88.3-100	203	1318	7.01E-06
GO Biological Process	response to abiotic stimulus	88.3-100	63	285	1.42E-05
GO Biological Process	cell-cell signaling	88.3-100	101	555	3.65E-05
GO Biological Process	inflammatory response	88.3-100	45	180	4.57E-05
GO Biological Process	innate immune response	88.3-100	46	188	6.50E-05
GO Biological Process	calcium ion homeostasis	88.3-100	15	31	2.59E-04
GO Biological Process	di-\, tri-valent inorganic cation homeostasis	88.3-100	21	59	5.59E-04
GO Biological Process	humoral immune response	88.3-100	40	168	1.12E-03
GO Biological Process	protein biosynthesis	88.3-100	95	558	2.42E-03
GO Biological Process	small GTPase mediated signal transduction	88.3-100	46	212	2.76E-03
GO Biological Process	humoral defense mechanism (sensu Vertebrata)	88.3-100	31	120	3.38E-03
GO Biological Process	metal ion homeostasis	88.3-100	21	65	3.52E-03
GO Biological Process	chromatin assembly/disassembly	88.3-100	26	93	4.76E-03
GO Biological Process	response to virus	88.3-100	12	29	3.67E-02
GO Biological Process	small GTPase mediated signal transduction	80.6-92.2	60	212	8.28E-09
GO Biological Process	olfaction	80.6-92.2	23	48	4.76E-07
GO Biological Process	defense response	76.7-100	296	787	1.29E-21
GO Biological Process	response to biotic stimulus	76.7-100	312	849	5.05E-21
GO Biological Process	immune response	76.7-100	268	713	3.45E-19
GO Biological Process	response to pest/pathogen/parasite	76.7-100	180	447	1.75E-15
GO Biological Process	response to external stimulus	76.7-100	350	1060	6.87E-15
GO Biological Process	small GTPase mediated signal transduction	76.7-100	103	212	2.18E-14
GO Biological Process	response to stimulus	76.7-100	402	1320	5.05E-11
GO Biological Process	response to wounding	76.7-100	114	274	4.59E-10
GO Biological Process	chemotaxis	76.7-100	61	118	4.04E-09
GO Biological Process	taxis	76.7-100	61	118	4.04E-09
GO Biological Process	response to stress	76.7-100	249	782	8.30E-08

GO Biological Process	biological_process unknown	76.7-100	267	863	4.45E-07
GO Biological Process	nucleosome assembly	76.7-100	36	61	1.34E-06
GO Biological Process	innate immune response	76.7-100	79	188	1.50E-06
GO Biological Process	organismal physiological process	76.7-100	378	1318	2.76E-06
GO Biological Process	olfaction	76.7-100	30	48	5.61E-06
GO Biological Process	inflammatory response	76.7-100	75	180	6.53E-06
GO Biological Process	humoral immune response	76.7-100	69	168	5.82E-05
GO Biological Process	antimicrobial humoral response	76.7-100	43	89	1.16E-04
GO Biological Process	response to chemical substance	76.7-100	84	223	2.01E-04
GO Biological Process	antimicrobial humoral response (sensu Vertebrata)	76.7-100	42	88	2.59E-04
GO Biological Process	cell-cell signaling	76.7-100	171	555	1.57E-03
GO Biological Process	cellular defense response	76.7-100	42	95	3.54E-03
GO Biological Process	protein-disulfide reduction	76.7-100	16	23	4.69E-03
GO Biological Process	response to abiotic stimulus	76.7-100	97	285	5.16E-03
GO Biological Process	chromatin assembly/disassembly	76.7-100	41	93	5.18E-03
GO Biological Process	gas transport	76.7-100	10	11	7.61E-03
GO Biological Process	oxygen transport	76.7-100	10	11	7.61E-03
GO Biological Process	protein biosynthesis	76.7-100	168	558	9.66E-03
GO Biological Process	ATP synthesis coupled electron transport	76.7-100	15	22	1.50E-02
GO Biological Process	ATP synthesis coupled electron transport (sensu Eukarya)	76.7-100	15	22	1.50E-02
GO Biological Process	humoral defense mechanism (sensu Vertebrata)	76.7-100	48	120	2.02E-02
GO Biological Process	di-\, tri-valent inorganic cation homeostasis	76.7-100	28	59	4.28E-02
GO Biological Process	calcium ion homeostasis	76.7-100	18	31	4.69E-02
GO Biological Process	immune response	72.8-84.4	134	713	6.58E-06
GO Biological Process	defense response	72.8-84.4	142	787	3.41E-05
GO Biological Process	response to biotic stimulus	72.8-84.4	150	849	5.37E-05
GO Biological Process	response to external stimulus	72.8-84.4	175	1060	4.73E-04
GO Biological Process	olfaction	72.8-84.4	19	48	1.51E-03
GO Biological Process	cell activation	72.8-84.4	23	70	4.83E-03
GO Biological Process	immune cell activation	72.8-84.4	23	70	4.83E-03
GO Biological Process	lymphocyte activation	72.8-84.4	20	59	1.33E-02
GO Biological Process	mitochondrial transport	72.8-84.4	10	18	2.03E-02
GO Biological Process	small GTPase mediated signal transduction	68.9-92.2	97	212	3.26E-10
GO Biological Process	immune response	68.9-92.2	241	713	1.04E-08
GO Biological Process	olfaction	68.9-92.2	33	48	5.56E-08
GO Biological Process	defense response	68.9-92.2	257	787	1.19E-07
GO Biological Process	response to biotic stimulus	68.9-92.2	273	849	1.76E-07
GO Biological Process	response to external stimulus	68.9-92.2	321	1060	7.25E-06
GO Biological Process	response to stimulus	68.9-92.2	374	1320	1.33E-03
GO Biological Process	antigen processing\, exogenous antigen via MHC class II	68.9-92.2	12	14	4.11E-03
GO Biological Process	antigen presentation\, exogenous antigen	68.9-92.2	12	14	4.11E-03
GO Biological Process	organismal physiological process	68.9-92.2	364	1318	2.90E-02
GO Biological Process	mitochondrial transport	68.9-92.2	13	18	4.32E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	65.0-76.7	114	602	2.96E-04
GO Biological Process	organismal physiological process	65.0-76.7	203	1318	4.29E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	61.1-84.4	209	602	1.11E-07
GO Biological Process	immune response	61.1-84.4	228	713	6.93E-05
GO Biological Process	defense response	61.1-84.4	241	787	1.54E-03
GO Biological Process	response to biotic stimulus	61.1-84.4	255	849	4.01E-03
GO Biological Process	amine biosynthesis	61.1-84.4	40	85	4.27E-03

GO Biological Process	response to external stimulus	61.1-84.4	309	1060	5.40E-03
GO Biological Process	alcohol metabolism	61.1-84.4	83	228	1.48E-02
GO Biological Process	organismal physiological process	61.1-84.4	371	1318	1.89E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	57.2-68.9	113	602	4.40E-04
GO Biological Process	G-protein coupled receptor protein signaling pathway	52.4-76.7	218	602	1.37E-09
GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	52.4-76.7	53	103	2.36E-06
GO Biological Process	cell surface receptor linked signal transduction	52.4-76.7	306	974	7.62E-06
GO Biological Process	cyclic-nucleotide-mediated signaling	52.4-76.7	54	109	9.76E-06
GO Biological Process	second-messenger-mediated signaling	52.4-76.7	55	123	6.27E-04
GO Biological Process	G-protein signaling\, coupled to IP3 second messenger (phospholipase C activating)	52.4-76.7	37	79	1.73E-02
GO Biological Process	G-protein signaling\, coupled to cAMP nucleotide second messenger	52.4-76.7	32	67	4.56E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	45.6-68.9	195	602	1.42E-03
GO Biological Process	organic acid metabolism	41.7-53.4	85	425	2.29E-03
GO Biological Process	carboxylic acid metabolism	41.7-53.4	84	423	3.61E-03
GO Biological Process	metabolism	41.7-53.4	815	6240	3.89E-02
GO Biological Process	amino acid metabolism	41.7-53.4	55	261	4.51E-02
GO Biological Process	metabolism	37.8-61.1	1598	6240	2.94E-04
GO Biological Process	carbohydrate metabolism	37.8-61.1	134	390	3.55E-03
GO Biological Process	pattern specification	37.8-61.1	26	48	2.05E-02
GO Biological Process	metabolism	33.9-45.6	823	6240	1.49E-03
GO Biological Process	metabolism	30.0-53.4	1649	6240	4.41E-10
GO Biological Process	carbohydrate metabolism	30.0-53.4	138	390	4.22E-04
GO Biological Process	organic acid metabolism	30.0-53.4	145	425	2.45E-03
GO Biological Process	carboxylic acid metabolism	30.0-53.4	144	423	3.10E-03
GO Biological Process	amine metabolism	30.0-53.4	117	341	2.21E-02
GO Biological Process	regulation of transcription\, DNA-dependent	30.0-53.4	489	1755	3.84E-02
GO Biological Process	transcription\, DNA-dependent	30.0-53.4	508	1831	4.13E-02
GO Biological Process	metabolism	26.2-37.8	828	6240	1.53E-03
GO Biological Process	metabolism	22.3-45.6	1640	6240	4.65E-09
GO Biological Process	regulation of transcription	22.3-45.6	532	1786	3.84E-07
GO Biological Process	regulation of transcription\, DNA-dependent	22.3-45.6	521	1755	1.15E-06
GO Biological Process	transcription	22.3-45.6	558	1898	1.22E-06
GO Biological Process	transcription\, DNA-dependent	22.3-45.6	539	1831	2.20E-06
GO Biological Process	phosphorus metabolism	22.3-45.6	217	681	1.61E-03
GO Biological Process	phosphate metabolism	22.3-45.6	217	681	1.61E-03
GO Biological Process	protein modification	22.3-45.6	295	975	2.67E-03
GO Biological Process	neurotransmitter transport	22.3-45.6	19	29	8.32E-03
GO Biological Process	nucleobase\, nucleoside\, nucleotide and nucleic acid metabolism	22.3-45.6	737	2748	3.80E-02
GO Biological Process	regulation of transcription	18.4-30.0	292	1786	3.14E-06
GO Biological Process	transcription	18.4-30.0	306	1898	5.28E-06
GO Biological Process	regulation of transcription\, DNA-dependent	18.4-30.0	284	1755	1.71E-05
GO Biological Process	transcription\, DNA-dependent	18.4-30.0	293	1831	2.95E-05
GO Biological Process	protein amino acid phosphorylation	18.4-30.0	95	491	2.65E-03
GO Biological Process	monovalent inorganic cation homeostasis	18.4-30.0	11	18	2.78E-03
GO Biological Process	hydrogen ion homeostasis	18.4-30.0	10	15	2.79E-03
GO Biological Process	homophilic cell adhesion	18.4-30.0	33	116	3.33E-03
GO Biological Process	protein modification	18.4-30.0	162	975	1.20E-02
GO Biological Process	regulation of pH	18.4-30.0	8	11	1.38E-02

GO Biological Process	phosphorylation	18.4-30.0	97	537	4.37E-02
GO Biological Process	regulation of transcription	14.5-37.8	537	1786	2.31E-07
GO Biological Process	homophilic cell adhesion	14.5-37.8	60	116	3.02E-07
GO Biological Process	transcription	14.5-37.8	565	1898	3.89E-07
GO Biological Process	regulation of transcription\, DNA-dependent	14.5-37.8	526	1755	6.79E-07
GO Biological Process	transcription\, DNA-dependent	14.5-37.8	543	1831	2.00E-06
GO Biological Process	metabolism	14.5-37.8	1620	6240	5.26E-05
GO Biological Process	protein modification	14.5-37.8	305	975	1.10E-04
GO Biological Process	phosphate metabolism	14.5-37.8	223	681	1.90E-04
GO Biological Process	phosphorus metabolism	14.5-37.8	223	681	1.90E-04
GO Biological Process	synaptogenesis	14.5-37.8	16	19	1.97E-04
GO Biological Process	synapse organization and biogenesis	14.5-37.8	16	19	1.97E-04
GO Biological Process	protein amino acid phosphorylation	14.5-37.8	169	491	2.23E-04
GO Biological Process	proteolysis and peptidolysis	14.5-37.8	193	578	3.31E-04
GO Biological Process	phosphorylation	14.5-37.8	181	537	4.07E-04
GO Biological Process	protein catabolism	14.5-37.8	195	587	4.33E-04
GO Biological Process	macromolecule catabolism	14.5-37.8	202	613	5.01E-04
GO Biological Process	catabolism	14.5-37.8	259	830	1.72E-03
GO Biological Process	nucleobase\, nucleoside\, nucleotide and nucleic acid metabolism	14.5-37.8	749	2748	8.14E-03
GO Biological Process	homophilic cell adhesion	10.6-22.3	59	116	2.14E-21
GO Biological Process	cell adhesion	10.6-22.3	151	589	1.17E-17
GO Biological Process	cell-cell adhesion	10.6-22.3	68	223	2.01E-10
GO Biological Process	extracellular matrix organization and biogenesis	10.6-22.3	15	31	1.69E-03
GO Biological Process	extracellular structure organization and biogenesis	10.6-22.3	15	31	1.69E-03
GO Biological Process	integrin-mediated signaling pathway	10.6-22.3	18	45	4.28E-03
GO Biological Process	synaptogenesis	10.6-22.3	11	19	5.58E-03
GO Biological Process	synapse organization and biogenesis	10.6-22.3	11	19	5.58E-03
GO Biological Process	proteolysis and peptidolysis	10.6-22.3	106	578	8.36E-03
GO Biological Process	protein catabolism	10.6-22.3	107	587	1.00E-02
GO Biological Process	protein amino acid phosphorylation	10.6-22.3	92	491	1.55E-02
GO Biological Process	macromolecule catabolism	10.6-22.3	109	613	2.60E-02
GO Biological Process	protein-nucleus import\, docking	10.6-22.3	10	18	2.74E-02
GO Biological Process	homophilic cell adhesion	6.7-30.0	89	116	1.02E-29
GO Biological Process	cell adhesion	6.7-30.0	241	589	1.94E-17
GO Biological Process	protein amino acid phosphorylation	6.7-30.0	193	491	2.62E-11
GO Biological Process	cell-cell adhesion	6.7-30.0	103	223	7.04E-10
GO Biological Process	protein modification	6.7-30.0	324	975	1.38E-08
GO Biological Process	phosphorylation	6.7-30.0	196	537	5.78E-08
GO Biological Process	phosphate metabolism	6.7-30.0	235	681	3.28E-07
GO Biological Process	phosphorus metabolism	6.7-30.0	235	681	3.28E-07
GO Biological Process	transcription	6.7-30.0	562	1898	1.01E-06
GO Biological Process	regulation of transcription	6.7-30.0	532	1786	1.31E-06
GO Biological Process	transcription\, DNA-dependent	6.7-30.0	543	1831	1.74E-06
GO Biological Process	regulation of transcription\, DNA-dependent	6.7-30.0	523	1755	1.80E-06
GO Biological Process	integrin-mediated signaling pathway	6.7-30.0	27	45	8.99E-04
GO Biological Process	synaptogenesis	6.7-30.0	15	19	2.38E-03
GO Biological Process	synapse organization and biogenesis	6.7-30.0	15	19	2.38E-03
GO Biological Process	nucleobase\, nucleoside\, nucleotide and nucleic acid metabolism	6.7-30.0	749	2748	7.10E-03
GO Biological Process	monovalent inorganic cation homeostasis	6.7-30.0	14	18	7.92E-03

GO Biological Process	cation transport	6.7-30.0	131	383	9.32E-03
GO Biological Process	protein metabolism	6.7-30.0	634	2296	1.00E-02
GO Biological Process	proteolysis and peptidolysis	6.7-30.0	185	578	1.33E-02
GO Biological Process	protein catabolism	6.7-30.0	187	587	1.64E-02
GO Biological Process	macromolecule catabolism	6.7-30.0	194	613	1.74E-02
GO Biological Process	hydrogen ion homeostasis	6.7-30.0	12	15	2.68E-02
GO Biological Process	extracellular matrix organization and biogenesis	6.7-30.0	19	31	3.58E-02
GO Biological Process	extracellular structure organization and biogenesis	6.7-30.0	19	31	3.58E-02
GO Biological Process	monovalent inorganic cation transport	6.7-30.0	90	250	3.67E-02
GO Biological Process	regulation of synapse	6.7-30.0	15	22	4.93E-02
GO Biological Process	cell adhesion	2.8-14.5	160	589	6.32E-22
GO Biological Process	homophilic cell adhesion	2.8-14.5	39	116	1.62E-06
GO Biological Process	integrin-mediated signaling pathway	2.8-14.5	22	45	2.75E-06
GO Biological Process	phosphorus metabolism	2.8-14.5	132	681	1.23E-05
GO Biological Process	phosphate metabolism	2.8-14.5	132	681	1.23E-05
GO Biological Process	cell communication	2.8-14.5	432	2885	1.63E-05
GO Biological Process	cellular process	2.8-14.5	796	5869	3.82E-05
GO Biological Process	protein amino acid phosphorylation	2.8-14.5	101	491	4.03E-05
GO Biological Process	enzyme linked receptor protein signaling pathway	2.8-14.5	36	119	1.67E-04
GO Biological Process	cell-matrix adhesion	2.8-14.5	25	67	1.87E-04
GO Biological Process	cell-cell adhesion	2.8-14.5	55	223	2.03E-04
GO Biological Process	transmembrane receptor protein tyrosine kinase signaling pathway	2.8-14.5	27	86	3.21E-03
GO Biological Process	phosphorylation	2.8-14.5	101	537	3.74E-03
GO Biological Process	protein modification	2.8-14.5	163	975	6.08E-03
GO Biological Process	nucleobase\, nucleoside\, nucleotide and nucleic acid metabolism	2.8-14.5	396	2748	8.82E-03
GO Biological Process	cyclic nucleotide metabolism	2.8-14.5	13	28	1.54E-02
GO Biological Process	transcription\, DNA-dependent	2.8-14.5	273	1831	4.15E-02
GO Biological Process	DNA replication and chromosome cycle	2.8-14.5	39	165	4.45E-02
GO Cellular Component	extracellular	88.3-100	236	1230	2.38E-19
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	88.3-100	33	69	8.04E-12
GO Cellular Component	ribosome	88.3-100	75	282	1.80E-11
GO Cellular Component	ribonucleoprotein complex	88.3-100	101	441	1.82E-11
GO Cellular Component	small ribosomal subunit	88.3-100	24	45	2.25E-09
GO Cellular Component	large ribosomal subunit	88.3-100	24	55	5.28E-07
GO Cellular Component	nucleosome	88.3-100	28	79	5.50E-06
GO Cellular Component	cytosolic small ribosomal subunit (sensu Eukarya)	88.3-100	16	29	7.49E-06
GO Cellular Component	eukaryotic 48S initiation complex	88.3-100	16	29	7.49E-06
GO Cellular Component	small nucleolar ribonucleoprotein complex	88.3-100	14	23	1.28E-05
GO Cellular Component	soluble fraction	88.3-100	54	241	8.61E-05
GO Cellular Component	cytosolic large ribosomal subunit (sensu Eukarya)	88.3-100	17	39	2.31E-04
GO Cellular Component	extracellular space	88.3-100	74	410	3.50E-03
GO Cellular Component	eukaryotic 43S preinitiation complex	88.3-100	17	46	3.89E-03
GO Cellular Component	hemoglobin complex	88.3-100	8	13	2.48E-02
GO Cellular Component	ribosome	80.6-92.2	60	282	7.21E-04
GO Cellular Component	mitochondrion	80.6-92.2	114	690	8.85E-03
GO Cellular Component	small ribosomal subunit	80.6-92.2	16	45	3.02E-02
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	76.7-100	52	69	9.81E-18
GO Cellular Component	extracellular	76.7-100	400	1230	6.10E-17
GO Cellular Component	small ribosomal subunit	76.7-100	37	45	3.05E-14

GO Cellular Component	ribosome	76.7-100	120	282	7.67E-12
GO Cellular Component	cytosolic small ribosomal subunit (sensu Eukarya)	76.7-100	26	29	3.51E-11
GO Cellular Component	eukaryotic 48S initiation complex	76.7-100	26	29	3.51E-11
GO Cellular Component	ribonucleoprotein complex	76.7-100	163	441	3.56E-10
GO Cellular Component	large ribosomal subunit	76.7-100	35	55	9.15E-08
GO Cellular Component	extracellular space	76.7-100	143	410	1.51E-06
GO Cellular Component	nucleosome	76.7-100	42	79	3.26E-06
GO Cellular Component	eukaryotic 43S preinitiation complex	76.7-100	28	46	3.81E-05
GO Cellular Component	cytosol	76.7-100	132	388	4.21E-05
GO Cellular Component	cytosolic large ribosomal subunit (sensu Eukarya)	76.7-100	25	39	5.18E-05
GO Cellular Component	mitochondrion	76.7-100	209	690	1.76E-04
GO Cellular Component	hemoglobin complex	76.7-100	12	13	3.93E-04
GO Cellular Component	soluble fraction	76.7-100	86	241	1.53E-03
GO Cellular Component	inner membrane	76.7-100	55	140	6.51E-03
GO Cellular Component	mitochondrial membrane	76.7-100	65	178	1.57E-02
GO Cellular Component	mitochondrial ribosome	76.7-100	17	27	1.57E-02
GO Cellular Component	organellar ribosome	76.7-100	17	27	1.57E-02
GO Cellular Component	integral to membrane	72.8-84.4	398	2911	2.25E-02
GO Cellular Component	mitochondrion	68.9-92.2	215	690	2.31E-04
GO Cellular Component	ribosome	68.9-92.2	102	282	4.81E-04
GO Cellular Component	mitochondrial membrane	68.9-92.2	70	178	1.32E-03
GO Cellular Component	extracellular space	68.9-92.2	133	410	8.40E-03
GO Cellular Component	proteasome core complex (sensu Eukarya)	68.9-92.2	14	19	1.21E-02
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	68.9-92.2	32	69	4.15E-02
GO Cellular Component	mitochondrial membrane	61.1-84.4	72	178	5.52E-04
GO Cellular Component	mitochondrion	61.1-84.4	213	690	2.74E-03
GO Cellular Component	integral to membrane	61.1-84.4	766	2911	3.83E-03
GO Cellular Component	extracellular space	61.1-84.4	133	410	2.45E-02
GO Cellular Component	mitochondrial inner membrane	61.1-84.4	51	125	2.64E-02
GO Cellular Component	cytoskeleton	6.7-30.0	305	889	9.91E-10
GO Cellular Component	nucleus	6.7-30.0	815	2825	9.42E-09
GO Cellular Component	cell	6.7-30.0	2501	10056	1.06E-08
GO Cellular Component	extracellular matrix	6.7-30.0	109	302	4.13E-03
GO Cellular Component	integrin complex	6.7-30.0	22	36	7.82E-03
GO Cellular Component	transcription factor complex	6.7-30.0	197	628	3.06E-02
GO Cellular Component	integral to plasma membrane	52.4-76.7	368	1269	2.48E-03
GO Cellular Component	integral to membrane	52.4-76.7	766	2911	4.74E-02
GO Cellular Component	intermediate filament	26.2-37.8	26	81	4.10E-03
GO Cellular Component	intermediate filament cytoskeleton	26.2-37.8	26	81	4.10E-03
GO Cellular Component	transcription factor complex	22.3-45.6	211	628	3.27E-05
GO Cellular Component	intracellular	22.3-45.6	1755	6899	1.14E-03
GO Cellular Component	cell	22.3-45.6	2461	10056	4.97E-03
GO Cellular Component	nucleus	22.3-45.6	765	2825	8.90E-03
GO Cellular Component	nucleoplasm	22.3-45.6	254	841	2.74E-02
GO Cellular Component	nucleus	18.4-30.0	430	2825	4.92E-06
GO Cellular Component	transcription factor complex	18.4-30.0	120	628	2.45E-04
GO Cellular Component	cell	18.4-30.0	1255	10056	2.63E-03
GO Cellular Component	nucleoplasm	18.4-30.0	144	841	1.08E-02
GO Cellular Component	nucleus	14.5-37.8	808	2825	2.69E-07
GO Cellular Component	cell	14.5-37.8	2498	10056	6.71E-06

GO Cellular Component	transcription factor complex	14.5-37.8	208	628	3.09E-04
GO Cellular Component	intermediate filament	14.5-37.8	40	81	2.43E-03
GO Cellular Component	intermediate filament cytoskeleton	14.5-37.8	40	81	2.43E-03
GO Cellular Component	intracellular	14.5-37.8	1762	6899	7.25E-03
GO Cellular Component	cytoskeleton	10.6-22.3	171	889	1.67E-07
GO Cellular Component	extracellular matrix	10.6-22.3	66	302	1.93E-03
GO Cellular Component	nucleus	10.6-22.3	409	2825	7.02E-03
GO Cellular Component	extracellular matrix	2.8-14.5	76	302	2.65E-07
GO Cellular Component	collagen	2.8-14.5	19	33	1.02E-06
GO Cellular Component	cytoskeleton	2.8-14.5	167	889	1.25E-06
GO Cellular Component	integrin complex	2.8-14.5	16	36	2.76E-03
GO Cellular Component	fibrillar collagen	2.8-14.5	7	9	2.45E-02
GO Molecular Function	receptor binding	88.3-100	145	522	3.13E-27
GO Molecular Function	G-protein-coupled receptor binding	88.3-100	38	50	1.69E-24
GO Molecular Function	cytokine activity	88.3-100	79	203	5.07E-24
GO Molecular Function	chemokine activity	88.3-100	36	48	9.56E-23
GO Molecular Function	chemokine receptor binding	88.3-100	36	48	9.56E-23
GO Molecular Function	chemoattractant activity	88.3-100	36	50	1.05E-21
GO Molecular Function	structural constituent of ribosome	88.3-100	76	210	1.13E-20
GO Molecular Function	hormone activity	88.3-100	38	99	3.13E-10
GO Molecular Function	hydrogen ion transporter activity	88.3-100	40	113	1.60E-09
GO Molecular Function	monovalent inorganic cation transporter activity	88.3-100	40	123	3.51E-08
GO Molecular Function	small monomeric GTPase activity	88.3-100	41	132	1.01E-07
GO Molecular Function	hematopoietin/interferon-class (D200-domain) cytokine receptor binding	88.3-100	21	43	4.35E-07
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	88.3-100	49	190	2.01E-06
GO Molecular Function	GTPase activity	88.3-100	52	217	9.99E-06
GO Molecular Function	structural molecule activity	88.3-100	122	711	1.53E-05
GO Molecular Function	cysteine protease inhibitor activity	88.3-100	13	21	3.02E-05
GO Molecular Function	molecular_function unknown	88.3-100	101	584	2.26E-04
GO Molecular Function	antiviral response protein activity	88.3-100	15	33	5.87E-04
GO Molecular Function	interferon-alpha/beta receptor binding	88.3-100	8	11	3.57E-03
GO Molecular Function	cation transporter activity	88.3-100	43	200	5.11E-03
GO Molecular Function	RAB small monomeric GTPase activity	88.3-100	18	53	6.89E-03
GO Molecular Function	growth factor activity	88.3-100	35	151	7.55E-03
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors	88.3-100	10	20	2.12E-02
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors\, oxygen as acceptor	88.3-100	10	20	2.12E-02
GO Molecular Function	cytochrome-c oxidase activity	88.3-100	10	20	2.12E-02
GO Molecular Function	heme-copper terminal oxidase activity	88.3-100	10	20	2.12E-02
GO Molecular Function	pre-mRNA splicing factor activity	88.3-100	20	68	2.40E-02
GO Molecular Function	antifungal peptide activity	88.3-100	7	10	2.53E-02
GO Molecular Function	RNA binding	88.3-100	80	484	2.93E-02
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, other acceptor	88.3-100	14	38	2.97E-02
GO Molecular Function	enzyme inhibitor activity	88.3-100	41	200	2.99E-02
GO Molecular Function	small monomeric GTPase activity	80.6-92.2	52	132	4.96E-14
GO Molecular Function	structural constituent of ribosome	80.6-92.2	62	210	3.44E-10
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	80.6-92.2	54	190	6.81E-08
GO Molecular Function	olfactory receptor activity	80.6-92.2	21	39	1.55E-07

GO Molecular Function	Rho small monomeric GTPase activity	80.6-92.2	17	27	3.61E-07
GO Molecular Function	GTPase activity	80.6-92.2	56	217	1.71E-06
GO Molecular Function	RAB small monomeric GTPase activity	80.6-92.2	20	53	9.09E-04
GO Molecular Function	GTP binding	80.6-92.2	54	244	9.49E-04
GO Molecular Function	guanyl nucleotide binding	80.6-92.2	54	251	2.44E-03
GO Molecular Function	nucleobase\, nucleoside\, nucleotide kinase activity	80.6-92.2	17	44	4.67E-03
GO Molecular Function	structural molecule activity	80.6-92.2	118	711	6.65E-03
GO Molecular Function	receptor binding	76.7-100	231	522	9.18E-29
GO Molecular Function	small monomeric GTPase activity	76.7-100	90	132	4.53E-27
GO Molecular Function	structural constituent of ribosome	76.7-100	119	210	2.95E-25
GO Molecular Function	cytokine activity	76.7-100	116	203	5.25E-25
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	76.7-100	101	190	3.22E-18
GO Molecular Function	G-protein-coupled receptor binding	76.7-100	40	50	6.93E-15
GO Molecular Function	chemoattractant activity	76.7-100	40	50	6.93E-15
GO Molecular Function	GTPase activity	76.7-100	104	217	1.67E-14
GO Molecular Function	chemokine receptor binding	76.7-100	38	48	9.65E-14
GO Molecular Function	chemokine activity	76.7-100	38	48	9.65E-14
GO Molecular Function	hydrogen ion transporter activity	76.7-100	64	113	1.60E-12
GO Molecular Function	RAB small monomeric GTPase activity	76.7-100	39	53	2.98E-12
GO Molecular Function	hormone activity	76.7-100	56	99	1.18E-10
GO Molecular Function	monovalent inorganic cation transporter activity	76.7-100	64	123	3.66E-10
GO Molecular Function	growth factor activity	76.7-100	71	151	1.06E-08
GO Molecular Function	GTP binding	76.7-100	99	244	4.35E-08
GO Molecular Function	guanyl nucleotide binding	76.7-100	99	251	2.98E-07
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, other acceptor	76.7-100	27	38	3.25E-07
GO Molecular Function	NADH dehydrogenase activity	76.7-100	26	36	4.03E-07
GO Molecular Function	olfactory receptor activity	76.7-100	27	39	8.37E-07
GO Molecular Function	Rho small monomeric GTPase activity	76.7-100	21	27	2.60E-06
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH	76.7-100	36	63	3.00E-06
GO Molecular Function	hematopoietin/interferon-class (D200-domain) cytokine receptor binding	76.7-100	28	43	3.47E-06
GO Molecular Function	NADH dehydrogenase (ubiquinone) activity	76.7-100	25	36	3.53E-06
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, quinone or similar compound as acceptor	76.7-100	26	40	1.46E-05
GO Molecular Function	sodium ion transporter activity	76.7-100	25	40	9.41E-05
GO Molecular Function	glutathione transferase activity	76.7-100	16	21	4.93E-04
GO Molecular Function	protein transporter activity	76.7-100	93	262	4.94E-04
GO Molecular Function	RAS small monomeric GTPase activity	76.7-100	19	29	1.54E-03
GO Molecular Function	structural molecule activity	76.7-100	206	711	4.64E-03
GO Molecular Function	cysteine protease inhibitor activity	76.7-100	15	21	4.89E-03
GO Molecular Function	oxygen transporter activity	76.7-100	10	11	6.56E-03
GO Molecular Function	cation transporter activity	76.7-100	72	200	6.83E-03
GO Molecular Function	molecular_function unknown	76.7-100	172	584	1.22E-02
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors\, oxygen as acceptor	76.7-100	14	20	1.63E-02
GO Molecular Function	cytochrome-c oxidase activity	76.7-100	14	20	1.63E-02
GO Molecular Function	heme-copper terminal oxidase activity	76.7-100	14	20	1.63E-02
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors	76.7-100	14	20	1.63E-02
GO Molecular Function	antimicrobial peptide activity	76.7-100	20	35	1.71E-02
GO Molecular Function	interleukin-1 receptor binding	76.7-100	9	10	2.77E-02

GO Molecular Function	RNA binding	76.7-100	144	484	4.30E-02
GO Molecular Function	olfactory receptor activity	72.8-84.4	19	39	2.37E-05
GO Molecular Function	rhodopsin-like receptor activity	72.8-84.4	71	331	2.52E-04
GO Molecular Function	HLA-C specific inhibitory MHC class I receptor activity	72.8-84.4	7	7	7.30E-04
GO Molecular Function	small monomeric GTPase activity	72.8-84.4	34	132	1.01E-02
GO Molecular Function	G-protein coupled receptor activity	72.8-84.4	76	410	3.38E-02
GO Molecular Function	small monomeric GTPase activity	68.9-92.2	81	132	3.75E-18
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	68.9-92.2	90	190	1.49E-10
GO Molecular Function	olfactory receptor activity	68.9-92.2	31	39	2.96E-10
GO Molecular Function	structural constituent of ribosome	68.9-92.2	96	210	3.31E-10
GO Molecular Function	rhodopsin-like receptor activity	68.9-92.2	132	331	3.11E-09
GO Molecular Function	GTPase activity	68.9-92.2	93	217	7.67E-08
GO Molecular Function	Rho small monomeric GTPase activity	68.9-92.2	22	27	5.56E-07
GO Molecular Function	GTP binding	68.9-92.2	99	244	7.52E-07
GO Molecular Function	guanyl nucleotide binding	68.9-92.2	99	251	4.72E-06
GO Molecular Function	RAB small monomeric GTPase activity	68.9-92.2	32	53	1.42E-05
GO Molecular Function	G-protein coupled receptor activity	68.9-92.2	141	410	9.17E-05
GO Molecular Function	receptor binding	68.9-92.2	168	522	7.73E-04
GO Molecular Function	growth factor activity	68.9-92.2	61	151	2.66E-03
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH	68.9-92.2	31	63	1.26E-02
GO Molecular Function	MHC class II receptor activity	68.9-92.2	12	16	4.85E-02
GO Molecular Function	rhodopsin-like receptor activity	65.0-76.7	90	331	1.93E-11
GO Molecular Function	G-protein coupled receptor activity	65.0-76.7	97	410	1.71E-08
GO Molecular Function	peptide receptor activity	65.0-76.7	36	109	1.12E-05
GO Molecular Function	peptide receptor activity\, G-protein coupled	65.0-76.7	36	109	1.12E-05
GO Molecular Function	receptor activity	65.0-76.7	214	1306	2.82E-04
GO Molecular Function	transmembrane receptor activity	65.0-76.7	148	852	1.18E-03
GO Molecular Function	peptide binding	65.0-76.7	38	143	2.61E-03
GO Molecular Function	rhodopsin-like receptor activity	61.1-84.4	157	331	4.14E-19
GO Molecular Function	G-protein coupled receptor activity	61.1-84.4	170	410	1.87E-13
GO Molecular Function	peptide receptor activity	61.1-84.4	56	109	5.13E-07
GO Molecular Function	peptide receptor activity\, G-protein coupled	61.1-84.4	56	109	5.13E-07
GO Molecular Function	purinergic nucleotide receptor activity	61.1-84.4	22	29	1.11E-05
GO Molecular Function	nucleotide receptor activity	61.1-84.4	22	29	1.11E-05
GO Molecular Function	purinergic nucleotide receptor activity\, G-protein coupled	61.1-84.4	22	29	1.11E-05
GO Molecular Function	nucleotide receptor activity\, G-protein coupled	61.1-84.4	22	29	1.11E-05
GO Molecular Function	peptide binding	61.1-84.4	63	143	9.90E-05
GO Molecular Function	transmembrane receptor activity	61.1-84.4	262	852	3.48E-04
GO Molecular Function	MHC class I receptor activity	61.1-84.4	16	21	1.60E-03
GO Molecular Function	receptor activity	61.1-84.4	374	1306	3.13E-03
GO Molecular Function	oxidoreductase activity\, acting on the CH-OH group of donors\, NAD or NADP as acceptor	61.1-84.4	42	92	6.57E-03
GO Molecular Function	olfactory receptor activity	61.1-84.4	23	39	6.57E-03
GO Molecular Function	S-adenosylmethionine-dependent methyltransferase activity	61.1-84.4	37	78	8.69E-03
GO Molecular Function	oxidoreductase activity\, acting on CH-OH group of donors	61.1-84.4	42	96	2.52E-02
GO Molecular Function	rhodopsin-like receptor activity	57.2-68.9	73	331	1.78E-04
GO Molecular Function	G-protein coupled receptor activity	57.2-68.9	81	410	4.48E-03
GO Molecular Function	carbon-oxygen lyase activity	57.2-68.9	18	49	1.63E-02

GO Molecular Function	rhodopsin-like receptor activity	52.4-76.7	158	331	6.64E-19
GO Molecular Function	G-protein coupled receptor activity	52.4-76.7	171	410	3.47E-13
GO Molecular Function	peptide receptor activity\, G-protein coupled	52.4-76.7	66	109	5.07E-13
GO Molecular Function	peptide receptor activity	52.4-76.7	66	109	5.07E-13
GO Molecular Function	peptide binding	52.4-76.7	70	143	1.02E-07
GO Molecular Function	transmembrane receptor activity	52.4-76.7	275	852	4.00E-06
GO Molecular Function	receptor activity	52.4-76.7	390	1306	5.90E-05
GO Molecular Function	GABA-A receptor activity	52.4-76.7	16	23	1.47E-02
GO Molecular Function	cytokine binding	52.4-76.7	29	57	2.60E-02
GO Molecular Function	DNA N-glycosylase activity	52.4-76.7	8	8	3.70E-02
GO Molecular Function	nucleotide receptor activity\, G-protein coupled	52.4-76.7	18	29	4.16E-02
GO Molecular Function	purinergic nucleotide receptor activity	52.4-76.7	18	29	4.16E-02
GO Molecular Function	nucleotide receptor activity	52.4-76.7	18	29	4.16E-02
GO Molecular Function	purinergic nucleotide receptor activity\, G-protein coupled	52.4-76.7	18	29	4.16E-02
GO Molecular Function	neurotransmitter receptor activity	49.5-61.1	22	50	3.36E-05
GO Molecular Function	neurotransmitter binding	49.5-61.1	22	51	5.25E-05
GO Molecular Function	GABA-A receptor activity	49.5-61.1	12	23	8.36E-03
GO Molecular Function	GABA receptor activity	49.5-61.1	12	25	2.56E-02
GO Molecular Function	extracellular ligand-gated ion channel activity	49.5-61.1	20	60	2.93E-02
GO Molecular Function	rhodopsin-like receptor activity	49.5-61.1	66	331	3.32E-02
GO Molecular Function	rhodopsin-like receptor activity	45.6-68.9	134	331	1.83E-08
GO Molecular Function	neurotransmitter receptor activity	45.6-68.9	31	50	3.28E-05
GO Molecular Function	peptide receptor activity\, G-protein coupled	45.6-68.9	53	109	4.42E-05
GO Molecular Function	peptide receptor activity	45.6-68.9	53	109	4.42E-05
GO Molecular Function	G-protein coupled receptor activity	45.6-68.9	147	410	4.65E-05
GO Molecular Function	neurotransmitter binding	45.6-68.9	31	51	6.47E-05
GO Molecular Function	GABA-A receptor activity	45.6-68.9	16	23	1.62E-02
GO Molecular Function	peptide binding	45.6-68.9	58	143	2.00E-02
GO Molecular Function	solute\:cation symporter activity	41.7-53.4	17	43	1.10E-02
GO Molecular Function	neurotransmitter receptor activity	37.8-61.1	33	50	9.94E-07
GO Molecular Function	neurotransmitter binding	37.8-61.1	33	51	2.18E-06
GO Molecular Function	solute\:cation symporter activity	37.8-61.1	26	43	1.11E-03
GO Molecular Function	GABA-A receptor activity	37.8-61.1	16	23	1.66E-02
GO Molecular Function	acetylcholine receptor activity	37.8-61.1	14	19	2.35E-02
GO Molecular Function	porter activity	37.8-61.1	70	184	3.38E-02
GO Molecular Function	electrochemical potential-driven transporter activity	37.8-61.1	70	185	4.19E-02
GO Molecular Function	organic cation transporter activity	37.8-61.1	15	22	4.92E-02
GO Molecular Function	neurotransmitter transporter activity	33.9-45.6	10	18	2.83E-02
GO Molecular Function	solute\:cation symporter activity	33.9-45.6	16	43	4.79E-02
GO Molecular Function	solute\:cation symporter activity	30.0-53.4	32	43	1.35E-08
GO Molecular Function	monooxygenase activity	30.0-53.4	42	72	1.42E-06
GO Molecular Function	solute\:sodium symporter activity	30.0-53.4	23	32	4.94E-05
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	30.0-53.4	45	87	6.14E-05
GO Molecular Function	oxygen binding	30.0-53.4	18	24	7.00E-04
GO Molecular Function	symporter activity	30.0-53.4	39	78	1.77E-03
GO Molecular Function	transcription factor activity	30.0-53.4	251	809	2.88E-03
GO Molecular Function	electrochemical potential-driven transporter activity	30.0-53.4	73	185	5.51E-03
GO Molecular Function	neurotransmitter transporter activity	30.0-53.4	14	18	8.50E-03
GO Molecular Function	porter activity	30.0-53.4	72	184	9.34E-03

GO Molecular Function	amine/polyamine transporter activity	30.0-53.4	31	60	1.09E-02
GO Molecular Function	protein serine/threonine kinase activity	22.3-45.6	131	355	7.70E-05
GO Molecular Function	transcription factor activity	22.3-45.6	259	809	1.44E-04
GO Molecular Function	transcription regulator activity	22.3-45.6	341	1117	2.65E-04
GO Molecular Function	neurotransmitter transporter activity	22.3-45.6	15	18	7.17E-04
GO Molecular Function	monooxygenase activity	22.3-45.6	36	72	5.39E-03
GO Molecular Function	solute\:cation symporter activity	22.3-45.6	25	43	6.35E-03
GO Molecular Function	adenyl nucleotide binding	22.3-45.6	320	1071	6.58E-03
GO Molecular Function	neurotransmitter\:sodium symporter activity	22.3-45.6	13	16	8.72E-03
GO Molecular Function	ATP binding	22.3-45.6	316	1059	8.75E-03
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	22.3-45.6	189	598	2.80E-02
GO Molecular Function	DNA binding	22.3-45.6	565	2039	3.01E-02
GO Molecular Function	nucleic acid binding	22.3-45.6	737	2729	4.29E-02
GO Molecular Function	adenyl nucleotide binding	18.4-30.0	202	1071	2.67E-08
GO Molecular Function	ATP binding	18.4-30.0	198	1059	1.01E-07
GO Molecular Function	calcium-dependent cell adhesion molecule activity	18.4-30.0	31	94	2.27E-04
GO Molecular Function	purine nucleotide binding	18.4-30.0	218	1303	2.42E-04
GO Molecular Function	transcription regulator activity	18.4-30.0	191	1117	3.63E-04
GO Molecular Function	nucleotide binding	18.4-30.0	219	1317	3.70E-04
GO Molecular Function	DNA binding	18.4-30.0	314	2039	1.19E-03
GO Molecular Function	protein serine/threonine kinase activity	18.4-30.0	75	355	1.80E-03
GO Molecular Function	transcription factor activity	18.4-30.0	143	809	2.50E-03
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	18.4-30.0	111	598	4.37E-03
GO Molecular Function	nucleic acid binding	18.4-30.0	399	2729	6.73E-03
GO Molecular Function	protein kinase activity	18.4-30.0	96	508	1.04E-02
GO Molecular Function	calcium-dependent cell adhesion molecule activity	14.5-37.8	56	94	7.10E-10
GO Molecular Function	adenyl nucleotide binding	14.5-37.8	354	1071	6.19E-09
GO Molecular Function	ATP binding	14.5-37.8	349	1059	1.41E-08
GO Molecular Function	transcription regulator activity	14.5-37.8	356	1117	9.89E-07
GO Molecular Function	DNA binding	14.5-37.8	596	2039	7.34E-06
GO Molecular Function	transcription factor activity	14.5-37.8	266	809	9.23E-06
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	14.5-37.8	206	598	1.05E-05
GO Molecular Function	protein kinase activity	14.5-37.8	178	508	3.25E-05
GO Molecular Function	nucleic acid binding	14.5-37.8	765	2729	8.70E-05
GO Molecular Function	metal ion binding	14.5-37.8	347	1135	3.01E-04
GO Molecular Function	protein serine/threonine kinase activity	14.5-37.8	128	355	7.02E-04
GO Molecular Function	cysteine-type peptidase activity	14.5-37.8	53	117	1.40E-03
GO Molecular Function	kinase activity	14.5-37.8	230	725	2.98E-03
GO Molecular Function	peptidase activity	14.5-37.8	169	508	3.91E-03
GO Molecular Function	cell adhesion molecule activity	14.5-37.8	129	374	9.76E-03
GO Molecular Function	calcium ion binding	14.5-37.8	186	576	1.02E-02
GO Molecular Function	purine nucleotide binding	14.5-37.8	380	1303	1.58E-02
GO Molecular Function	nucleotide binding	14.5-37.8	382	1317	2.63E-02
GO Molecular Function	protein-tyrosine kinase activity	14.5-37.8	88	241	3.24E-02
GO Molecular Function	binding	14.5-37.8	1699	6657	3.41E-02
GO Molecular Function	cysteine-type endopeptidase activity	14.5-37.8	44	100	3.42E-02
GO Molecular Function	calcium-dependent cell adhesion molecule activity	10.6-22.3	55	94	8.93E-24
GO Molecular Function	cell adhesion molecule activity	10.6-22.3	122	374	2.24E-23
GO Molecular Function	metal ion binding	10.6-22.3	216	1135	9.41E-10
GO Molecular Function	calcium ion binding	10.6-22.3	126	576	1.30E-08

GO Molecular Function	adenyl nucleotide binding	10.6-22.3	198	1071	1.68E-07
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	10.6-22.3	28	64	5.09E-07
GO Molecular Function	ATP binding	10.6-22.3	194	1059	6.14E-07
GO Molecular Function	transmembrane receptor protein kinase activity	10.6-22.3	29	76	1.17E-05
GO Molecular Function	metallopeptidase activity	10.6-22.3	47	169	4.55E-05
GO Molecular Function	binding	10.6-22.3	891	6657	1.11E-04
GO Molecular Function	protein-tyrosine kinase activity	10.6-22.3	59	241	1.24E-04
GO Molecular Function	peptidase activity	10.6-22.3	101	508	3.57E-04
GO Molecular Function	ephrin receptor activity	10.6-22.3	10	15	2.74E-03
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	10.6-22.3	111	598	3.21E-03
GO Molecular Function	ubiquitin thiolesterase activity	10.6-22.3	18	45	4.53E-03
GO Molecular Function	protein kinase activity	10.6-22.3	96	508	7.85E-03
GO Molecular Function	kinase activity	10.6-22.3	127	725	1.24E-02
GO Molecular Function	purine nucleotide binding	10.6-22.3	207	1303	1.57E-02
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	10.6-22.3	32	119	1.86E-02
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	10.6-22.3	32	119	1.86E-02
GO Molecular Function	nucleotide binding	10.6-22.3	207	1317	3.25E-02
GO Molecular Function	DNA binding	10.6-22.3	302	2039	4.13E-02
GO Molecular Function	ATPase activity\, coupled	10.6-22.3	51	236	4.78E-02
GO Molecular Function	calcium-dependent cell adhesion molecule activity	6.7-30.0	79	94	5.22E-31
GO Molecular Function	adenyl nucleotide binding	6.7-30.0	430	1071	1.01E-30
GO Molecular Function	ATP binding	6.7-30.0	424	1059	6.92E-30
GO Molecular Function	cell adhesion molecule activity	6.7-30.0	190	374	3.77E-26
GO Molecular Function	purine nucleotide binding	6.7-30.0	455	1303	4.27E-17
GO Molecular Function	nucleotide binding	6.7-30.0	459	1317	4.27E-17
GO Molecular Function	metal ion binding	6.7-30.0	400	1135	2.54E-15
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	6.7-30.0	45	64	2.76E-11
GO Molecular Function	ATPase activity\, coupled	6.7-30.0	111	236	3.61E-11
GO Molecular Function	calcium ion binding	6.7-30.0	220	576	4.07E-11
GO Molecular Function	protein kinase activity	6.7-30.0	199	508	4.49E-11
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	6.7-30.0	224	598	2.58E-10
GO Molecular Function	ATPase activity	6.7-30.0	113	249	4.48E-10
GO Molecular Function	protein-tyrosine kinase activity	6.7-30.0	109	241	1.61E-09
GO Molecular Function	transmembrane receptor protein kinase activity	6.7-30.0	48	76	1.95E-09
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	6.7-30.0	63	119	3.97E-08
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	6.7-30.0	63	119	3.97E-08
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	6.7-30.0	116	275	8.20E-08
GO Molecular Function	binding	6.7-30.0	1758	6657	1.01E-07
GO Molecular Function	kinase activity	6.7-30.0	248	725	7.79E-07
GO Molecular Function	DNA binding	6.7-30.0	596	2039	1.64E-05
GO Molecular Function	transcription regulator activity	6.7-30.0	348	1117	5.08E-05
GO Molecular Function	P-P-bond-hydrolysis-driven transporter activity	6.7-30.0	63	137	6.38E-05
GO Molecular Function	GTPase regulator activity	6.7-30.0	93	231	1.27E-04
GO Molecular Function	small GTPase regulatory/interacting protein activity	6.7-30.0	70	163	3.24E-04
GO Molecular Function	ATP-binding cassette (ABC) transporter activity	6.7-30.0	39	74	3.91E-04
GO Molecular Function	transferase activity\, transferring phosphorus-containing groups	6.7-30.0	242	750	4.65E-04

GO Molecular Function	nucleic acid binding	6.7-30.0	759	2729	1.07E-03
GO Molecular Function	ubiquitin thiolesterase activity	6.7-30.0	27	45	1.12E-03
GO Molecular Function	metallopeptidase activity	6.7-30.0	70	169	1.74E-03
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of ions\, phosphorylative mechanism	6.7-30.0	26	44	2.83E-03
GO Molecular Function	peptidase activity	6.7-30.0	169	508	5.08E-03
GO Molecular Function	protein serine/threonine kinase activity	6.7-30.0	125	355	5.14E-03
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of ions	6.7-30.0	26	46	9.14E-03
GO Molecular Function	metalloendopeptidase activity	6.7-30.0	43	93	9.53E-03
GO Molecular Function	glutamate receptor activity	6.7-30.0	21	34	1.21E-02
GO Molecular Function	magnesium ion binding	6.7-30.0	56	133	1.29E-02
GO Molecular Function	cysteine-type peptidase activity	6.7-30.0	50	117	2.56E-02
GO Molecular Function	cation\:cation antiporter activity	6.7-30.0	11	13	2.76E-02
GO Molecular Function	mannosidase activity	6.7-30.0	11	13	2.76E-02
GO Molecular Function	ephrin receptor activity	6.7-30.0	12	15	3.01E-02
GO Molecular Function	thiolester hydrolase activity	6.7-30.0	28	54	3.62E-02
GO Molecular Function	cation\:chloride symporter activity	6.7-30.0	8	8	4.15E-02
GO Molecular Function	cell adhesion molecule activity	2.8-14.5	121	374	1.09E-22
GO Molecular Function	adenyl nucleotide binding	2.8-14.5	245	1071	2.02E-22
GO Molecular Function	ATP binding	2.8-14.5	241	1059	1.23E-21
GO Molecular Function	GTPase regulator activity	2.8-14.5	79	231	1.40E-15
GO Molecular Function	small GTPase regulatory/interacting protein activity	2.8-14.5	60	163	5.14E-13
GO Molecular Function	nucleotide binding	2.8-14.5	255	1317	6.05E-13
GO Molecular Function	purine nucleotide binding	2.8-14.5	252	1303	1.10E-12
GO Molecular Function	ATPase activity	2.8-14.5	75	249	2.90E-11
GO Molecular Function	ATPase activity\, coupled	2.8-14.5	72	236	4.82E-11
GO Molecular Function	guanyl-nucleotide exchange factor activity	2.8-14.5	39	95	1.72E-09
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	2.8-14.5	31	64	1.91E-09
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	2.8-14.5	44	119	4.50E-09
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	2.8-14.5	44	119	4.50E-09
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	2.8-14.5	75	275	7.84E-09
GO Molecular Function	binding	2.8-14.5	914	6657	7.38E-08
GO Molecular Function	transmembrane receptor protein kinase activity	2.8-14.5	31	76	4.81E-07
GO Molecular Function	P-P-bond-hydrolysis-driven transporter activity	2.8-14.5	44	137	9.72E-07
GO Molecular Function	ATP-binding cassette (ABC) transporter activity	2.8-14.5	30	74	1.13E-06
GO Molecular Function	transmembrane receptor protein tyrosine phosphatase activity	2.8-14.5	14	19	2.04E-06
GO Molecular Function	transmembrane receptor protein phosphatase activity	2.8-14.5	14	19	2.04E-06
GO Molecular Function	metal ion binding	2.8-14.5	200	1135	1.16E-05
GO Molecular Function	protein kinase activity	2.8-14.5	103	508	1.05E-04
GO Molecular Function	GTPase activator activity	2.8-14.5	34	108	1.65E-04
GO Molecular Function	calcium-dependent cell adhesion molecule activity	2.8-14.5	31	94	1.90E-04
GO Molecular Function	protein-tyrosine kinase activity	2.8-14.5	58	241	3.18E-04
GO Molecular Function	epidermal growth factor receptor activity	2.8-14.5	7	7	8.96E-04
GO Molecular Function	guanylate cyclase activity	2.8-14.5	11	17	1.14E-03
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	2.8-14.5	112	598	2.00E-03
GO Molecular Function	extracellular matrix structural constituent	2.8-14.5	28	89	2.38E-03
GO Molecular Function	DNA binding	2.8-14.5	311	2039	2.62E-03

GO Molecular Function	cation\:chloride symporter activity	2.8-14.5	7	8	6.41E-03
GO Molecular Function	cell adhesion receptor activity	2.8-14.5	17	42	7.00E-03
GO Molecular Function	enzyme regulator activity	2.8-14.5	102	549	9.16E-03
GO Molecular Function	phosphorus-oxygen lyase activity	2.8-14.5	11	20	1.09E-02
GO Molecular Function	calcium ion binding	2.8-14.5	105	576	1.59E-02
GO Molecular Function	protein binding	2.8-14.5	240	1548	1.93E-02
GO Molecular Function	helicase activity	2.8-14.5	33	125	2.06E-02
GO Molecular Function	magnesium ion binding	2.8-14.5	34	133	3.20E-02
GO Molecular Function	collagen	2.8-14.5	6	7	4.69E-02
SwissProt keyword	Cytokine	88.3-100	65	146	1.66E-22
SwissProt keyword	Acetylation	88.3-100	58	178	3.83E-12
SwissProt keyword	Chemotaxis	88.3-100	29	52	3.93E-12
SwissProt keyword	Ribosomal protein	88.3-100	36	88	4.19E-10
SwissProt keyword	Nucleosome core	88.3-100	19	30	1.33E-08
SwissProt keyword	Hormone	88.3-100	26	59	1.33E-07
SwissProt keyword	Inflammatory response	88.3-100	23	52	1.57E-06
SwissProt keyword	Amidation	88.3-100	16	32	7.38E-05
SwissProt keyword	Prenylation	88.3-100	29	91	8.53E-05
SwissProt keyword	3D-structure	88.3-100	165	1067	5.77E-04
SwissProt keyword	Chromosomal protein	88.3-100	20	54	7.30E-04
SwissProt keyword	Thiol protease inhibitor	88.3-100	10	15	8.18E-04
SwissProt keyword	Cleavage on pair of basic residues	88.3-100	20	56	1.45E-03
SwissProt keyword	Pyrrolidone carboxylic acid	88.3-100	18	52	8.81E-03
SwissProt keyword	Antibiotic	88.3-100	10	19	1.68E-02
SwissProt keyword	Antiviral	88.3-100	10	19	1.68E-02
SwissProt keyword	Vitamin A	88.3-100	6	7	2.20E-02
SwissProt keyword	Signal	88.3-100	244	1791	2.27E-02
SwissProt keyword	Lipocalin	88.3-100	8	13	3.10E-02
SwissProt keyword	Lipid-binding	88.3-100	11	24	3.19E-02
SwissProt keyword	Defensin	88.3-100	5	5	3.23E-02
SwissProt keyword	Fungicide	88.3-100	7	10	3.32E-02
SwissProt keyword	CF(0)	88.3-100	7	10	3.32E-02
SwissProt keyword	Acetylation	80.6-92.2	53	178	2.21E-08
SwissProt keyword	Olfaction	80.6-92.2	22	44	4.79E-07
SwissProt keyword	Prenylation	80.6-92.2	31	91	1.45E-05
SwissProt keyword	Ribosomal protein	80.6-92.2	27	88	1.57E-03
SwissProt keyword	Tight junction	80.6-92.2	13	26	2.73E-03
SwissProt keyword	GTP-binding	80.6-92.2	37	158	2.23E-02
SwissProt keyword	Acetylation	76.7-100	106	178	4.53E-23
SwissProt keyword	Cytokine	76.7-100	89	146	5.59E-20
SwissProt keyword	Ribosomal protein	76.7-100	58	88	1.52E-14
SwissProt keyword	Prenylation	76.7-100	59	91	2.57E-14
SwissProt keyword	Nucleosome core	76.7-100	27	30	2.41E-11
SwissProt keyword	Chemotaxis	76.7-100	36	52	3.64E-09
SwissProt keyword	3D-structure	76.7-100	334	1067	8.57E-09
SwissProt keyword	Hormone	76.7-100	38	59	2.67E-08
SwissProt keyword	Chromosomal protein	76.7-100	34	54	8.47E-07
SwissProt keyword	Olfaction	76.7-100	29	44	3.67E-06
SwissProt keyword	Ubiquinone	76.7-100	25	36	1.01E-05
SwissProt keyword	GTP-binding	76.7-100	69	158	1.04E-05

SwissProt keyword	Lipoprotein	76.7-100	115	317	6.00E-05
SwissProt keyword	Cleavage on pair of basic residues	76.7-100	32	56	8.64E-05
SwissProt keyword	Pyrrolidone carboxylic acid	76.7-100	30	52	1.75E-04
SwissProt keyword	Inflammatory response	76.7-100	29	52	8.02E-04
SwissProt keyword	Mitochondrion	76.7-100	131	392	1.58E-03
SwissProt keyword	Amidation	76.7-100	20	32	5.25E-03
SwissProt keyword	Lipocalin	76.7-100	11	13	1.33E-02
SwissProt keyword	Growth factor	76.7-100	46	109	1.38E-02
SwissProt keyword	Thiol protease inhibitor	76.7-100	12	15	1.39E-02
SwissProt keyword	Oxygen transport	76.7-100	8	8	2.34E-02
SwissProt keyword	Protein transport	76.7-100	69	189	3.43E-02
SwissProt keyword	CF(0)	76.7-100	9	10	4.22E-02
SwissProt keyword	Olfaction	72.8-84.4	19	44	5.63E-04
SwissProt keyword	Lipoprotein	72.8-84.4	69	317	1.05E-03
SwissProt keyword	Prenylation	72.8-84.4	26	91	4.17E-02
SwissProt keyword	Prenylation	68.9-92.2	53	91	1.80E-09
SwissProt keyword	Olfaction	68.9-92.2	32	44	1.43E-08
SwissProt keyword	Acetylation	68.9-92.2	80	178	2.69E-07
SwissProt keyword	Lipoprotein	68.9-92.2	122	317	1.05E-06
SwissProt keyword	G-protein coupled receptor	68.9-92.2	117	316	3.56E-05
SwissProt keyword	Ribosomal protein	68.9-92.2	44	88	1.12E-04
SwissProt keyword	GTP-binding	68.9-92.2	67	158	1.66E-04
SwissProt keyword	Threonine protease	68.9-92.2	13	16	5.33E-03
SwissProt keyword	MHC II	68.9-92.2	11	13	1.79E-02
SwissProt keyword	Tight junction	68.9-92.2	17	26	1.90E-02
SwissProt keyword	Mitogen	68.9-92.2	18	29	2.95E-02
SwissProt keyword	G-protein coupled receptor	65.0-76.7	85	316	1.68E-10
SwissProt keyword	Transmembrane	65.0-76.7	299	2026	6.43E-03
SwissProt keyword	G-protein coupled receptor	61.1-84.4	146	316	1.49E-15
SwissProt keyword	Lipoprotein	61.1-84.4	124	317	1.35E-06
SwissProt keyword	Transmembrane	61.1-84.4	561	2026	3.77E-03
SwissProt keyword	Olfaction	61.1-84.4	24	44	3.83E-02
SwissProt keyword	Palmitate	61.1-84.4	53	130	4.09E-02
SwissProt keyword	G-protein coupled receptor	57.2-68.9	70	316	3.86E-04
SwissProt keyword	G-protein coupled receptor	52.4-76.7	150	316	2.30E-17
SwissProt keyword	Palmitate	52.4-76.7	62	130	7.08E-06
SwissProt keyword	Transmembrane	52.4-76.7	565	2026	1.60E-03
SwissProt keyword	Postsynaptic membrane	49.5-61.1	22	62	3.97E-03
SwissProt keyword	Developmental protein	49.5-61.1	56	261	2.30E-02
SwissProt keyword	G-protein coupled receptor	45.6-68.9	126	316	2.16E-07
SwissProt keyword	Palmitate	45.6-68.9	54	130	1.76E-02
SwissProt keyword	Postsynaptic membrane	45.6-68.9	31	62	2.31E-02
SwissProt keyword	Monooxygenase	41.7-53.4	20	56	8.41E-03
SwissProt keyword	Developmental protein	37.8-61.1	107	261	1.72E-06
SwissProt keyword	Postsynaptic membrane	37.8-61.1	33	62	2.33E-03
SwissProt keyword	Homeobox	37.8-61.1	62	152	1.03E-02
SwissProt keyword	Transferase	37.8-61.1	235	765	1.57E-02
SwissProt keyword	Transferase	33.9-45.6	133	765	9.92E-03
SwissProt keyword	Monooxygenase	30.0-53.4	37	56	6.70E-08
SwissProt keyword	Microsome	30.0-53.4	41	82	6.89E-04

SwissProt keyword	Transcription regulation	30.0-53.4	267	884	4.83E-03
SwissProt keyword	DNA-binding	30.0-53.4	290	974	6.19E-03
SwissProt keyword	Developmental protein	30.0-53.4	95	261	6.95E-03
SwissProt keyword	DNA-binding	26.2-37.8	172	974	3.34E-05
SwissProt keyword	Transcription regulation	26.2-37.8	157	884	1.14E-04
SwissProt keyword	Nuclear protein	26.2-37.8	259	1707	7.77E-03
SwissProt keyword	Zinc-finger	26.2-37.8	103	581	3.68E-02
SwissProt keyword	DNA-binding	22.3-45.6	318	974	2.67E-07
SwissProt keyword	Transcription regulation	22.3-45.6	292	884	4.57E-07
SwissProt keyword	Zinc-finger	22.3-45.6	196	581	1.34E-04
SwissProt keyword	Serine/threonine-protein kinase	22.3-45.6	90	226	3.15E-04
SwissProt keyword	Activator	22.3-45.6	96	260	7.28E-03
SwissProt keyword	Nuclear protein	22.3-45.6	484	1707	1.15E-02
SwissProt keyword	Intermediate filament	22.3-45.6	25	44	1.28E-02
SwissProt keyword	Monooxygenase	22.3-45.6	29	56	2.65E-02
SwissProt keyword	ATP-binding	22.3-45.6	209	670	3.05E-02
SwissProt keyword	Keratin	22.3-45.6	20	33	3.28E-02
SwissProt keyword	Alternative splicing	22.3-45.6	520	1867	4.38E-02
SwissProt keyword	DNA-binding	18.4-30.0	179	974	1.86E-06
SwissProt keyword	ATP-binding	18.4-30.0	132	670	5.96E-06
SwissProt keyword	Repeat	18.4-30.0	317	1993	6.31E-06
SwissProt keyword	Transcription regulation	18.4-30.0	163	884	1.02E-05
SwissProt keyword	Nuclear protein	18.4-30.0	277	1707	1.50E-05
SwissProt keyword	Alternative splicing	18.4-30.0	289	1867	8.27E-04
SwissProt keyword	Activator	18.4-30.0	58	260	5.00E-03
SwissProt keyword	Zinc-finger	18.4-30.0	107	581	7.63E-03
SwissProt keyword	Phosphorylation	18.4-30.0	164	997	2.22E-02
SwissProt keyword	Nuclear protein	14.5-37.8	523	1707	7.89E-11
SwissProt keyword	DNA-binding	14.5-37.8	327	974	8.87E-11
SwissProt keyword	Transcription regulation	14.5-37.8	301	884	1.61E-10
SwissProt keyword	Repeat	14.5-37.8	591	1993	9.70E-10
SwissProt keyword	ATP-binding	14.5-37.8	233	670	1.82E-08
SwissProt keyword	Phosphorylation	14.5-37.8	315	997	1.37E-06
SwissProt keyword	Alternative splicing	14.5-37.8	532	1867	3.93E-05
SwissProt keyword	Zinc-finger	14.5-37.8	190	581	4.78E-04
SwissProt keyword	Activator	14.5-37.8	95	260	4.73E-03
SwissProt keyword	Serine/threonine-protein kinase	14.5-37.8	84	226	8.91E-03
SwissProt keyword	Tyrosine-protein kinase	14.5-37.8	44	98	9.59E-03
SwissProt keyword	SH2 domain	14.5-37.8	35	74	2.44E-02
SwissProt keyword	Metalloprotease	14.5-37.8	41	92	2.63E-02
SwissProt keyword	Cell adhesion	14.5-37.8	95	270	3.04E-02
SwissProt keyword	Intermediate filament	14.5-37.8	24	44	3.35E-02
SwissProt keyword	Repeat	10.6-22.3	363	1993	1.03E-21
SwissProt keyword	Cell adhesion	10.6-22.3	84	270	3.19E-15
SwissProt keyword	ATP-binding	10.6-22.3	138	670	2.56E-09
SwissProt keyword	Phosphorylation	10.6-22.3	177	997	8.32E-07
SwissProt keyword	Tyrosine-protein kinase	10.6-22.3	34	98	3.17E-06
SwissProt keyword	Alternative splicing	10.6-22.3	286	1867	2.04E-05
SwissProt keyword	Nuclear protein	10.6-22.3	264	1707	4.19E-05
SwissProt keyword	Integrin	10.6-22.3	13	25	2.15E-03

SwissProt keyword	Metalloprotease	10.6-22.3	26	92	2.36E-02
SwissProt keyword	Calcium transport	10.6-22.3	9	15	2.42E-02
SwissProt keyword	Receptor	10.6-22.3	78	420	2.47E-02
SwissProt keyword	DNA-binding	10.6-22.3	153	974	4.42E-02
SwissProt keyword	Repeat	6.7-30.0	700	1993	2.29E-39
SwissProt keyword	ATP-binding	6.7-30.0	290	670	6.06E-29
SwissProt keyword	Cell adhesion	6.7-30.0	141	270	8.09E-22
SwissProt keyword	Alternative splicing	6.7-30.0	588	1867	2.23E-16
SwissProt keyword	Tyrosine-protein kinase	6.7-30.0	61	98	7.58E-13
SwissProt keyword	Nuclear protein	6.7-30.0	529	1707	1.98E-12
SwissProt keyword	Phosphorylation	6.7-30.0	336	997	1.07E-11
SwissProt keyword	DNA-binding	6.7-30.0	318	974	7.90E-09
SwissProt keyword	Transcription regulation	6.7-30.0	290	884	5.02E-08
SwissProt keyword	Integrin	6.7-30.0	20	25	1.32E-05
SwissProt keyword	Coiled coil	6.7-30.0	149	416	1.45E-05
SwissProt keyword	Calcium-binding	6.7-30.0	107	286	1.93E-04
SwissProt keyword	Zinc-finger	6.7-30.0	191	581	2.07E-04
SwissProt keyword	Calcium transport	6.7-30.0	13	15	1.48E-03
SwissProt keyword	Magnesium	6.7-30.0	52	121	5.23E-03
SwissProt keyword	Hydrolase	6.7-30.0	237	779	6.34E-03
SwissProt keyword	Activator	6.7-30.0	94	260	7.33E-03
SwissProt keyword	Metalloprotease	6.7-30.0	42	92	8.07E-03
SwissProt keyword	Metal-binding	6.7-30.0	155	479	1.14E-02
SwissProt keyword	Kringle	6.7-30.0	11	13	2.03E-02
SwissProt keyword	SH2 domain	6.7-30.0	35	74	2.10E-02
SwissProt keyword	EGF-like domain	6.7-30.0	49	119	4.39E-02
SwissProt keyword	Receptor	6.7-30.0	136	420	4.52E-02
SwissProt keyword	Repeat	2.8-14.5	439	1993	2.42E-55
SwissProt keyword	ATP-binding	2.8-14.5	165	670	1.51E-20
SwissProt keyword	Cell adhesion	2.8-14.5	92	270	2.23E-20
SwissProt keyword	Alternative splicing	2.8-14.5	330	1867	3.05E-17
SwissProt keyword	Extracellular matrix	2.8-14.5	42	124	5.85E-08
SwissProt keyword	Tyrosine-protein kinase	2.8-14.5	36	98	1.13E-07
SwissProt keyword	Connective tissue	2.8-14.5	19	32	2.11E-07
SwissProt keyword	Integrin	2.8-14.5	16	25	1.38E-06
SwissProt keyword	Coiled coil	2.8-14.5	89	416	3.63E-06
SwissProt keyword	Guanine-nucleotide releasing factor	2.8-14.5	23	52	5.48E-06
SwissProt keyword	Collagen	2.8-14.5	21	46	1.36E-05
SwissProt keyword	Phosphorylation	2.8-14.5	170	997	2.26E-05
SwissProt keyword	Helicase	2.8-14.5	24	62	6.07E-05
SwissProt keyword	Basement membrane	2.8-14.5	13	22	2.37E-04
SwissProt keyword	Nuclear protein	2.8-14.5	258	1707	2.39E-04
SwissProt keyword	Zinc-finger	2.8-14.5	106	581	7.99E-04
SwissProt keyword	Calcium	2.8-14.5	28	91	1.26E-03
SwissProt keyword	Magnesium	2.8-14.5	33	121	2.75E-03
SwissProt keyword	Calmodulin-binding	2.8-14.5	26	86	4.65E-03
SwissProt keyword	Hydroxylation	2.8-14.5	19	52	4.93E-03
SwissProt keyword	Receptor	2.8-14.5	79	420	8.65E-03
SwissProt keyword	Bromodomain	2.8-14.5	12	27	3.84E-02

Table C.2: Correlations to $M_{interspecitic}$-ranked gene list

System	Gene Category	Percentile (%)	# genes	# in category	p-value
GO Biological Process	nucleosome assembly	88.3-100	39	61	2.82E-20
GO Biological Process	chromatin assembly/disassembly	88.3-100	39	93	1.05E-11
GO Biological Process	di-\, tri-valent inorganic cation homeostasis	88.3-100	29	59	2.28E-10
GO Biological Process	metal ion homeostasis	88.3-100	30	65	7.08E-10
GO Biological Process	response to wounding	88.3-100	70	274	2.64E-09
GO Biological Process	chemotaxis	88.3-100	40	118	1.76E-08
GO Biological Process	taxis	88.3-100	40	118	1.76E-08
GO Biological Process	biological_process unknown	88.3-100	155	863	2.06E-08
GO Biological Process	response to pest/pathogen/parasite	88.3-100	95	447	3.26E-08
GO Biological Process	defense response	88.3-100	143	787	6.16E-08
GO Biological Process	response to biotic stimulus	88.3-100	150	849	1.67E-07
GO Biological Process	immune response	88.3-100	131	713	2.14E-07
GO Biological Process	DNA packaging	88.3-100	49	175	2.15E-07
GO Biological Process	response to chemical substance	88.3-100	57	223	3.32E-07
GO Biological Process	cation homeostasis	88.3-100	30	80	4.28E-07
GO Biological Process	establishment and/or maintenance of chromatin architecture	88.3-100	46	163	6.06E-07
GO Biological Process	inflammatory response	88.3-100	49	180	6.33E-07
GO Biological Process	innate immune response	88.3-100	50	188	1.02E-06
GO Biological Process	ion homeostasis	88.3-100	30	86	3.29E-06
GO Biological Process	cell ion homeostasis	88.3-100	30	86	3.29E-06
GO Biological Process	cell-cell signaling	88.3-100	105	555	3.44E-06
GO Biological Process	chromosome organization and biogenesis (sensu Eukarya)	88.3-100	49	190	4.71E-06
GO Biological Process	cell homeostasis	88.3-100	31	94	8.43E-06
GO Biological Process	nuclear organization and biogenesis	88.3-100	49	194	9.99E-06
GO Biological Process	organismal physiological process	88.3-100	202	1318	0.000026
GO Biological Process	homeostasis	88.3-100	31	98	0.0000261
GO Biological Process	calcium ion homeostasis	88.3-100	16	31	0.0000329
GO Biological Process	response to abiotic stimulus	88.3-100	61	285	0.000122
GO Biological Process	response to stress	88.3-100	127	782	0.000886
GO Biological Process	copper ion homeostasis	88.3-100	9	13	0.00173
GO Biological Process	response to external stimulus	88.3-100	161	1060	0.00184
GO Biological Process	transition metal ion homeostasis	88.3-100	13	28	0.0036
GO Biological Process	response to stimulus	88.3-100	187	1320	0.0238
GO Biological Process	heavy metal sensitivity/resistance	88.3-100	7	10	0.029
GO Biological Process	cell growth and/or maintenance	88.3-100	429	3446	0.0475
GO Biological Process	small GTPase mediated signal transduction	80.6-92.2	67	212	1.54E-12
GO Biological Process	antigen presentation\, exogenous antigen	80.6-92.2	10	14	4.89E-04
GO Biological Process	antigen processing\, exogenous antigen via MHC class II	80.6-92.2	10	14	4.89E-04
GO Biological Process	cell-cell signaling	80.6-92.2	97	555	1.02E-02
GO Biological Process	immune response	80.6-92.2	118	713	1.45E-02
GO Biological Process	response to biotic stimulus	80.6-92.2	136	849	1.59E-02
GO Biological Process	defense response	80.6-92.2	127	787	2.28E-02
GO Biological Process	protein transport	80.6-92.2	76	417	2.54E-02

GO Biological Process	nucleosome assembly	76.7-100	44	61	2.61E-13
GO Biological Process	small GTPase mediated signal transduction	76.7-100	99	212	2.67E-12
GO Biological Process	immune response	76.7-100	241	713	1.51E-10
GO Biological Process	defense response	76.7-100	260	787	2.78E-10
GO Biological Process	response to biotic stimulus	76.7-100	276	849	3.86E-10
GO Biological Process	cell-cell signaling	76.7-100	186	555	3.33E-07
GO Biological Process	response to external stimulus	76.7-100	313	1060	3.43E-06
GO Biological Process	chromatin assembly/disassembly	76.7-100	47	93	4.03E-06
GO Biological Process	response to wounding	76.7-100	103	274	7.23E-06
GO Biological Process	di-\, tri-valent inorganic cation homeostasis	76.7-100	34	59	0.0000096
GO Biological Process	response to pest/pathogen/parasite	76.7-100	151	447	0.0000112
GO Biological Process	metal ion homeostasis	76.7-100	36	65	0.0000145
GO Biological Process	taxis	76.7-100	54	118	0.000024
GO Biological Process	chemotaxis	76.7-100	54	118	0.000024
GO Biological Process	biological_process unknown	76.7-100	258	863	0.0000354
GO Biological Process	innate immune response	76.7-100	74	188	0.000148
GO Biological Process	response to stimulus	76.7-100	367	1320	0.000241
GO Biological Process	inflammatory response	76.7-100	71	180	0.00025
GO Biological Process	organismal physiological process	76.7-100	362	1318	0.00121
GO Biological Process	calcium ion homeostasis	76.7-100	20	31	0.0014
GO Biological Process	DNA packaging	76.7-100	67	175	0.00211
GO Biological Process	establishment and/or maintenance of chromatin architecture	76.7-100	63	163	0.00308
GO Biological Process	cation homeostasis	76.7-100	37	80	0.00393
GO Biological Process	response to stress	76.7-100	226	782	0.00546
GO Biological Process	development	76.7-100	465	1776	0.00962
GO Biological Process	ATP synthesis coupled electron transport (sensu Eukarya)	76.7-100	15	22	0.0146
GO Biological Process	ATP synthesis coupled electron transport	76.7-100	15	22	0.0146
GO Biological Process	cell homeostasis	76.7-100	40	94	0.0197
GO Biological Process	chromosome organization and biogenesis (sensu Eukarya)	76.7-100	68	190	0.0275
GO Biological Process	nuclear organization and biogenesis	76.7-100	69	194	0.0302
GO Biological Process	ion homeostasis	76.7-100	37	86	0.0321
GO Biological Process	cell ion homeostasis	76.7-100	37	86	0.0321
GO Biological Process	protein-disulfide reduction	76.7-100	15	23	0.0335
GO Biological Process	small GTPase mediated signal transduction	72.8-84.4	55	212	1.05E-05
GO Biological Process	frizzled-2 signaling pathway	72.8-84.4	10	18	2.03E-02
GO Biological Process	small GTPase mediated signal transduction	68.9-82.2	107	212	2.21E-15
GO Biological Process	development	68.9-82.2	497	1776	5.07E-05
GO Biological Process	antigen presentation	68.9-82.2	20	27	7.66E-05
GO Biological Process	response to biotic stimulus	68.9-82.2	254	849	1.21E-03
GO Biological Process	pattern specification	68.9-82.2	27	48	1.87E-03
GO Biological Process	immune response	68.9-82.2	217	713	2.35E-03
GO Biological Process	defense response	68.9-82.2	236	787	2.68E-03
GO Biological Process	antigen presentation\, exogenous antigen	68.9-82.2	12	14	3.82E-03
GO Biological Process	antigen processing\, exogenous antigen via MHC class II	68.9-82.2	12	14	3.82E-03
GO Biological Process	cell-cell signaling	68.9-82.2	174	555	4.20E-03
GO Biological Process	antigen processing	68.9-82.2	18	27	4.92E-03
GO Biological Process	response to external stimulus	68.9-82.2	299	1060	3.17E-02
GO Biological Process	frizzled-2 signaling pathway	68.9-82.2	13	18	4.02E-02

GO Biological Process	G-protein coupled receptor protein signaling pathway	65.0-76.7	115	602	1.29E-04
GO Biological Process	frizzled-2 signaling pathway	65.0-76.7	11	18	2.34E-03
GO Biological Process	lymphocyte proliferation	65.0-76.7	10	15	2.38E-03
GO Biological Process	regulation of lymphocyte proliferation	65.0-76.7	8	10	3.69E-03
GO Biological Process	cell surface receptor linked signal transduction	65.0-76.7	160	974	8.88E-03
GO Biological Process	regulation of T-cell proliferation	65.0-76.7	7	9	2.54E-02
GO Biological Process	positive regulation of lymphocyte proliferation	65.0-76.7	7	9	2.54E-02
GO Biological Process	frizzled-2 signaling pathway	61.1-84.4	16	18	2.53E-05
GO Biological Process	G-protein coupled receptor protein signaling pathway	61.1-84.4	199	602	4.79E-05
GO Biological Process	response to external stimulus	61.1-84.4	315	1060	8.65E-04
GO Biological Process	response to biotic stimulus	61.1-84.4	256	849	3.74E-03
GO Biological Process	defense response	61.1-84.4	238	787	7.27E-03
GO Biological Process	cell surface receptor linked signal transduction	61.1-84.4	286	974	1.01E-02
GO Biological Process	immune response	61.1-84.4	216	713	1.95E-02
GO Biological Process	lymphocyte proliferation	61.1-84.4	12	15	2.07E-02
GO Biological Process	development	61.1-84.4	486	1776	2.71E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	57.2-68.9	129	602	1.03E-08
GO Biological Process	cell surface receptor linked signal transduction	57.2-68.9	168	974	2.96E-04
GO Biological Process	organismal physiological process	57.2-68.9	210	1318	3.36E-03
GO Biological Process	response to external stimulus	57.2-68.9	169	1060	4.28E-02
GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	57.2-68.9	28	103	4.53E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	53.4-76.7	242	602	2.72E-17
GO Biological Process	cell surface receptor linked signal transduction	53.4-76.7	326	974	4.64E-10
GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	53.4-76.7	55	103	1.91E-07
GO Biological Process	cyclic-nucleotide-mediated signaling	53.4-76.7	55	109	3.18E-06
GO Biological Process	second-messenger-mediated signaling	53.4-76.7	58	123	2.94E-05
GO Biological Process	olfaction	53.4-76.7	27	48	4.15E-03
GO Biological Process	G-protein coupled receptor protein signaling pathway	49.5-61.1	116	602	2.25E-04
GO Biological Process	G-protein coupled receptor protein signaling pathway	45.6-68.9	230	602	2.55E-12
GO Biological Process	cell surface receptor linked signal transduction	45.6-68.9	311	974	6.32E-06
GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	45.6-68.9	46	103	1.13E-02
GO Biological Process	mesoderm cell fate commitment	45.6-68.9	8	8	4.23E-02
GO Biological Process	carbohydrate metabolism	41.7-53.4	82	390	5.89E-04
GO Biological Process	energy derivation by oxidation of organic compounds	41.7-53.4	34	130	2.23E-02
GO Biological Process	carbohydrate metabolism	37.8-61.1	141	390	0.0000737
GO Biological Process	G-protein coupled receptor protein signaling pathway	37.8-61.1	193	602	0.00579
GO Biological Process	carboxylic acid metabolism	33.9-45.6	84	423	1.93E-03
GO Biological Process	organic acid metabolism	33.9-45.6	84	425	2.37E-03
GO Biological Process	carbohydrate metabolism	33.9-45.6	74	390	4.90E-02
GO Biological Process	carbohydrate metabolism	30.0-53.4	147	390	1.42E-06
GO Biological Process	metabolism	30.0-53.4	1602	6240	7.61E-04
GO Biological Process	carboxylic acid metabolism	30.0-53.4	146	423	1.18E-03
GO Biological Process	organic acid metabolism	30.0-53.4	146	425	1.63E-03
GO Biological Process	lipid metabolism	30.0-53.4	165	500	5.21E-03
GO Biological Process	amine metabolism	26.2-37.8	73	341	7.71E-04
GO Biological Process	metabolism	26.2-37.8	820	6240	3.11E-03
GO Biological Process	amino acid and derivative metabolism	26.2-37.8	63	302	1.27E-02
GO Biological Process	metabolism	22.3-45.6	1604	6240	1.27E-05

GO Biological Process	carboxylic acid metabolism	22.3-45.6	152	423	1.70E-05
GO Biological Process	organic acid metabolism	22.3-45.6	152	425	2.46E-05
GO Biological Process	amine metabolism	22.3-45.6	120	341	2.64E-03
GO Biological Process	carbohydrate metabolism	22.3-45.6	132	390	8.32E-03
GO Biological Process	protein metabolism	22.3-45.6	624	2296	1.92E-02
GO Biological Process	amino acid and derivative metabolism	22.3-45.6	105	302	2.43E-02
GO Biological Process	homophilic cell adhesion	18.4-30.0	56	116	5.24E-19
GO Biological Process	cell-cell adhesion	18.4-30.0	63	223	4.22E-08
GO Biological Process	protein modification	18.4-30.0	173	975	2.32E-05
GO Biological Process	regulation of synapse	18.4-30.0	14	22	3.59E-05
GO Biological Process	protein amino acid phosphorylation	18.4-30.0	98	491	2.16E-04
GO Biological Process	synapse organization and biogenesis	18.4-30.0	12	19	4.67E-04
GO Biological Process	synaptogenesis	18.4-30.0	12	19	4.67E-04
GO Biological Process	phosphate metabolism	18.4-30.0	124	681	9.16E-04
GO Biological Process	phosphorus metabolism	18.4-30.0	124	681	9.16E-04
GO Biological Process	protein metabolism	18.4-30.0	341	2296	1.28E-03
GO Biological Process	cell adhesion	18.4-30.0	108	589	3.94E-03
GO Biological Process	phosphorylation	18.4-30.0	99	537	8.65E-03
GO Biological Process	extracellular structure organization and biogenesis	18.4-30.0	14	31	1.06E-02
GO Biological Process	extracellular matrix organization and biogenesis	18.4-30.0	14	31	1.06E-02
GO Biological Process	homophilic cell adhesion	14.5-37.8	82	116	2.26E-23
GO Biological Process	cell-cell adhesion	14.5-37.8	97	223	1.28E-07
GO Biological Process	protein modification	14.5-37.8	315	975	2.48E-07
GO Biological Process	protein amino acid phosphorylation	14.5-37.8	178	491	3.40E-07
GO Biological Process	metabolism	14.5-37.8	1612	6240	1.62E-06
GO Biological Process	phosphorus metabolism	14.5-37.8	229	681	2.58E-06
GO Biological Process	phosphate metabolism	14.5-37.8	229	681	2.58E-06
GO Biological Process	protein metabolism	14.5-37.8	648	2296	2.33E-05
GO Biological Process	phosphorylation	14.5-37.8	184	537	2.82E-05
GO Biological Process	catabolism	14.5-37.8	254	830	4.01E-03
GO Biological Process	synaptogenesis	14.5-37.8	14	19	2.19E-02
GO Biological Process	synapse organization and biogenesis	14.5-37.8	14	19	2.19E-02
GO Biological Process	amine metabolism	14.5-37.8	115	341	4.53E-02
GO Biological Process	regulation of synapse	14.5-37.8	15	22	4.58E-02
GO Biological Process	homophilic cell adhesion	10.6-22.3	46	116	4.98E-11
GO Biological Process	cell adhesion	10.6-22.3	131	589	7.30E-10
GO Biological Process	phosphorus metabolism	10.6-22.3	139	681	1.10E-07
GO Biological Process	phosphate metabolism	10.6-22.3	139	681	1.10E-07
GO Biological Process	cell-cell adhesion	10.6-22.3	61	223	5.23E-07
GO Biological Process	protein amino acid phosphorylation	10.6-22.3	105	491	2.19E-06
GO Biological Process	phosphorylation	10.6-22.3	108	537	4.50E-05
GO Biological Process	protein modification	10.6-22.3	168	975	5.58E-04
GO Biological Process	cyclic nucleotide metabolism	10.6-22.3	14	28	2.38E-03
GO Biological Process	transmembrane receptor protein tyrosine kinase signaling pathway	10.6-22.3	26	86	1.20E-02
GO Biological Process	cyclic nucleotide biosynthesis	10.6-22.3	12	24	1.47E-02
GO Biological Process	enzyme linked receptor protein signaling pathway	10.6-22.3	32	119	1.63E-02
GO Biological Process	integrin-mediated signaling pathway	10.6-22.3	17	45	2.04E-02
GO Biological Process	homophilic cell adhesion	6.7-30.0	91	116	3.26E-32
GO Biological Process	cell adhesion	6.7-30.0	235	589	6.37E-16

GO Biological Process	cell-cell adhesion	6.7-30.0	111	223	4.20E-14
GO Biological Process	protein modification	6.7-30.0	343	975	5.28E-14
GO Biological Process	phosphate metabolism	6.7-30.0	256	681	1.26E-13
GO Biological Process	phosphorus metabolism	6.7-30.0	256	681	1.26E-13
GO Biological Process	protein amino acid phosphorylation	6.7-30.0	194	491	3.77E-12
GO Biological Process	protein metabolism	6.7-30.0	679	2296	5.97E-10
GO Biological Process	phosphorylation	6.7-30.0	198	537	5.38E-09
GO Biological Process	enzyme linked receptor protein signaling pathway	6.7-30.0	54	119	5.55E-04
GO Biological Process	cyclic nucleotide metabolism	6.7-30.0	19	28	3.26E-03
GO Biological Process	transmembrane receptor protein tyrosine kinase signaling pathway	6.7-30.0	41	86	3.50E-03
GO Biological Process	metabolism	6.7-30.0	1583	6240	5.07E-03
GO Biological Process	protein catabolism	6.7-30.0	187	587	7.97E-03
GO Biological Process	proteolysis and peptidolysis	6.7-30.0	184	578	1.01E-02
GO Biological Process	synaptogenesis	6.7-30.0	14	19	2.12E-02
GO Biological Process	synapse organization and biogenesis	6.7-30.0	14	19	2.12E-02
GO Biological Process	protein amino acid dephosphorylation	6.7-30.0	53	128	2.24E-02
GO Biological Process	macromolecule catabolism	6.7-30.0	191	613	3.02E-02
GO Biological Process	cyclic nucleotide biosynthesis	6.7-30.0	16	24	3.48E-02
GO Biological Process	dephosphorylation	6.7-30.0	53	130	3.83E-02
GO Biological Process	protein-nucleus import\, docking	2.8-14.5	15	18	2.32E-08
GO Biological Process	cell adhesion	2.8-14.5	127	589	4.14E-08
GO Biological Process	phosphorus metabolism	2.8-14.5	131	681	4.81E-05
GO Biological Process	phosphate metabolism	2.8-14.5	131	681	4.81E-05
GO Biological Process	cell-matrix adhesion	2.8-14.5	26	67	5.27E-05
GO Biological Process	protein amino acid phosphorylation	2.8-14.5	99	491	2.91E-04
GO Biological Process	protein modification	2.8-14.5	170	975	4.78E-04
GO Biological Process	protein metabolism	2.8-14.5	348	2296	1.03E-03
GO Biological Process	striated muscle contraction	2.8-14.5	14	29	4.50E-03
GO Biological Process	integrin-mediated signaling pathway	2.8-14.5	18	45	4.73E-03
GO Biological Process	chromosome segregation	2.8-14.5	12	22	4.77E-03
GO Biological Process	phosphorylation	2.8-14.5	100	537	1.15E-02
GO Biological Process	nucleobase biosynthesis	2.8-14.5	10	18	2.88E-02
GO Cellular Component	nucleosome	88.3-100	43	79	7.98E-19
GO Cellular Component	chromatin	88.3-100	57	171	8.3E-13
GO Cellular Component	extracellular	88.3-100	210	1230	1.14E-10
GO Cellular Component	small nucleolar ribonucleoprotein complex	88.3-100	15	23	8.9E-07
GO Cellular Component	chromosome	88.3-100	64	279	1.95E-06
GO Cellular Component	small ribosomal subunit	88.3-100	20	45	0.0000117
GO Cellular Component	ribonucleoprotein complex	88.3-100	86	441	0.0000146
GO Cellular Component	ribosome	88.3-100	61	282	0.0000515
GO Cellular Component	obsolete cellular component	88.3-100	84	444	0.0000927
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	88.3-100	24	69	0.00012
GO Cellular Component	organellar ribosome	88.3-100	14	27	0.000212
GO Cellular Component	mitochondrial ribosome	88.3-100	14	27	0.000212
GO Cellular Component	mitochondrion	88.3-100	115	690	0.00045
GO Cellular Component	soluble fraction	88.3-100	50	241	0.00351
GO Cellular Component	extracellular space	88.3-100	72	410	0.0149
GO Cellular Component	cytosolic small ribosomal subunit (sensu Eukarya)	88.3-100	12	29	0.0345
GO Cellular Component	eukaryotic 48S initiation complex	88.3-100	12	29	0.0345

GO Cellular Component	cellular_component unknown	88.3-100	92	575	0.0429
GO Cellular Component	small ribosomal subunit	80.6-92.2	19	45	2.13E-04
GO Cellular Component	extracellular	80.6-92.2	191	1230	6.21E-04
GO Cellular Component	ribosome	80.6-92.2	58	282	4.60E-03
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	80.6-92.2	22	69	6.63E-03
GO Cellular Component	extracellular space	80.6-92.2	76	410	8.47E-03
GO Cellular Component	extracellular	76.7-100	391	1230	1.25E-14
GO Cellular Component	nucleosome	76.7-100	48	79	2.35E-10
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	76.7-100	43	69	1.34E-09
GO Cellular Component	small ribosomal subunit	76.7-100	32	45	6.52E-09
GO Cellular Component	ribosome	76.7-100	112	282	1.5E-08
GO Cellular Component	chromatin	76.7-100	75	171	2.37E-07
GO Cellular Component	mitochondrion	76.7-100	221	690	2.72E-07
GO Cellular Component	extracellular space	76.7-100	143	410	1.39E-06
GO Cellular Component	ribonucleoprotein complex	76.7-100	148	441	0.0000141
GO Cellular Component	soluble fraction	76.7-100	90	241	0.0000688
GO Cellular Component	large ribosomal subunit	76.7-100	30	55	0.000361
GO Cellular Component	cytosolic small ribosomal subunit (sensu Eukarya)	76.7-100	19	29	0.00166
GO Cellular Component	eukaryotic 48S initiation complex	76.7-100	19	29	0.00166
GO Cellular Component	cytosolic large ribosomal subunit (sensu Eukarya)	76.7-100	23	39	0.00169
GO Cellular Component	mitochondrial ribosome	76.7-100	18	27	0.00227
GO Cellular Component	organellar ribosome	76.7-100	18	27	0.00227
GO Cellular Component	small nucleolar ribonucleoprotein complex	76.7-100	15	23	0.0299
GO Cellular Component	obsolete cellular component	76.7-100	135	444	0.0329
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	72.8-84.4	25	69	2.14E-04
GO Cellular Component	cytosolic large ribosomal subunit (sensu Eukarya)	72.8-84.4	15	39	3.72E-02
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	68.9-82.2	40	69	9.26E-07
GO Cellular Component	extracellular	68.9-82.2	366	1230	2.74E-06
GO Cellular Component	extracellular space	68.9-82.2	146	410	3.87E-06
GO Cellular Component	small ribosomal subunit	68.9-82.2	27	45	2.80E-04
GO Cellular Component	ribosome	68.9-82.2	101	282	1.07E-03
GO Cellular Component	large ribosomal subunit	68.9-82.2	29	55	4.14E-03
GO Cellular Component	cytosolic large ribosomal subunit (sensu Eukarya)	68.9-82.2	22	39	1.89E-02
GO Cellular Component	transcription factor complex	61.1-84.4	192	628	4.16E-02
GO Cellular Component	integral to membrane	45.6-68.9	784	2911	1.46E-02
GO Cellular Component	integral to plasma membrane	45.6-68.9	366	1269	3.05E-02
GO Cellular Component	nicotinic acetylcholine-gated receptor-channel complex	37.8-61.1	11	12	0.00516
GO Cellular Component	lytic vacuole	30.0-53.4	54	116	2.64E-04
GO Cellular Component	lysosome	30.0-53.4	54	116	2.64E-04
GO Cellular Component	vacuole	30.0-53.4	55	130	9.16E-03
GO Cellular Component	endoplasmic reticulum	30.0-53.4	135	401	1.46E-02
GO Cellular Component	unlocalized	30.0-53.4	67	174	3.94E-02
GO Cellular Component	cell	26.2-37.8	1236	10056	6.70E-03
GO Cellular Component	cell	22.3-45.6	2454	10056	1.53E-03
GO Cellular Component	lysosome	22.3-45.6	49	116	3.07E-02
GO Cellular Component	lytic vacuole	22.3-45.6	49	116	3.07E-02
GO Cellular Component	cell	14.5-37.8	2469	10056	1.50E-07
GO Cellular Component	cytoskeleton	10.6-22.3	154	889	9.51E-04
GO Cellular Component	cell	6.7-30.0	2469	10056	9.58E-07
GO Cellular Component	cytoskeleton	6.7-30.0	278	889	2.08E-04

GO Cellular Component	integrin complex	6.7-30.0	22	36	6.82E-03
GO Cellular Component	cytoskeleton	2.8-14.5	171	889	2.20E-07
GO Cellular Component	striated muscle thick filament	2.8-14.5	12	15	7.51E-06
GO Cellular Component	myosin	2.8-14.5	33	95	1.78E-05
GO Cellular Component	extracellular matrix	2.8-14.5	68	302	4.49E-04
GO Cellular Component	pore complex	2.8-14.5	21	52	5.33E-04
GO Cellular Component	nuclear pore	2.8-14.5	21	52	5.33E-04
GO Cellular Component	collagen	2.8-14.5	15	33	4.95E-03
GO Cellular Component	kinesin complex	2.8-14.5	11	19	5.86E-03
GO Cellular Component	microtubule associated complex	2.8-14.5	29	103	1.97E-02
GO Molecular Function	receptor binding	88.3-100	145	522	3.3E-27
GO Molecular Function	G-protein-coupled receptor binding	88.3-100	39	50	5.72E-26
GO Molecular Function	chemokine activity	88.3-100	37	48	3.42E-24
GO Molecular Function	chemokine receptor binding	88.3-100	37	48	3.42E-24
GO Molecular Function	chemoattractant activity	88.3-100	37	50	4.37E-23
GO Molecular Function	hormone activity	88.3-100	51	99	1.67E-21
GO Molecular Function	hydrogen ion transporter activity	88.3-100	47	113	8.56E-15
GO Molecular Function	monovalent inorganic cation transporter activity	88.3-100	47	123	4.8E-13
GO Molecular Function	structural constituent of ribosome	88.3-100	64	210	9.59E-13
GO Molecular Function	cytokine activity	88.3-100	62	203	2.49E-12
GO Molecular Function	NADH dehydrogenase activity	88.3-100	19	36	5.58E-07
GO Molecular Function	NADH dehydrogenase (ubiquinone) activity	88.3-100	19	36	5.58E-07
GO Molecular Function	cation transporter activity	88.3-100	51	200	1.38E-06
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, other acceptor	88.3-100	19	38	1.87E-06
GO Molecular Function	antimicrobial peptide activity	88.3-100	18	35	0.0000029
GO Molecular Function	sodium ion transporter activity	88.3-100	19	40	5.68E-06
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, quinone or similar compound as acceptor	88.3-100	19	40	5.68E-06
GO Molecular Function	antifungal peptide activity	88.3-100	9	10	0.000026
GO Molecular Function	neuropeptide hormone activity	88.3-100	13	22	0.0000646
GO Molecular Function	molecular_function unknown	88.3-100	102	584	0.000117
GO Molecular Function	cytochrome-c oxidase activity	88.3-100	12	20	0.000177
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors\, oxygen as acceptor	88.3-100	12	20	0.000177
GO Molecular Function	heme-copper terminal oxidase activity	88.3-100	12	20	0.000177
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors	88.3-100	12	20	0.000177
GO Molecular Function	primary active transporter activity	88.3-100	46	202	0.00039
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH	88.3-100	21	63	0.00146
GO Molecular Function	ion transporter activity	88.3-100	53	261	0.00227
GO Molecular Function	growth factor activity	88.3-100	36	151	0.00264
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	88.3-100	42	190	0.00306
GO Molecular Function	small monomeric GTPase activity	88.3-100	32	132	0.00697
GO Molecular Function	metal ion transporter activity	88.3-100	23	82	0.0127
GO Molecular Function	cyclin-dependent protein kinase inhibitor activity	88.3-100	6	7	0.0167
GO Molecular Function	retinoid binding	88.3-100	7	10	0.0244
GO Molecular Function	isoprenoid binding	88.3-100	7	10	0.0244
GO Molecular Function	copper/cadmium binding	88.3-100	5	5	0.0254
GO Molecular Function	structural molecule activity	88.3-100	108	711	0.0463
GO Molecular Function	small monomeric GTPase activity	80.6-92.2	56	132	3.42E-17

GO Molecular Function	receptor binding	80.6-92.2	119	522	4.97E-12
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	80.6-92.2	59	190	9.47E-11
GO Molecular Function	structural constituent of ribosome	80.6-92.2	60	210	3.35E-09
GO Molecular Function	GTPase activity	80.6-92.2	61	217	4.73E-09
GO Molecular Function	RAB small monomeric GTPase activity	80.6-92.2	24	53	6.58E-07
GO Molecular Function	hormone activity	80.6-92.2	34	99	1.28E-06
GO Molecular Function	GTP binding	80.6-92.2	60	244	2.70E-06
GO Molecular Function	guanyl nucleotide binding	80.6-92.2	60	251	8.67E-06
GO Molecular Function	cytokine activity	80.6-92.2	50	203	6.28E-05
GO Molecular Function	MHC class II receptor activity	80.6-92.2	11	16	1.76E-04
GO Molecular Function	protein transporter activity	80.6-92.2	56	262	1.63E-03
GO Molecular Function	Rho small monomeric GTPase activity	80.6-92.2	13	27	3.68E-03
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH	80.6-92.2	21	63	4.55E-03
GO Molecular Function	growth factor activity	80.6-92.2	36	151	1.25E-02
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, quinone or similar compound as acceptor	80.6-92.2	15	40	2.67E-02
GO Molecular Function	NADH dehydrogenase (ubiquinone) activity	80.6-92.2	14	36	3.22E-02
GO Molecular Function	receptor binding	76.7-100	249	522	2.7E-38
GO Molecular Function	structural constituent of ribosome	76.7-100	116	210	2.62E-23
GO Molecular Function	G-protein-coupled receptor binding	76.7-100	45	50	1.83E-21
GO Molecular Function	hormone activity	76.7-100	68	99	3.03E-20
GO Molecular Function	chemokine receptor binding	76.7-100	43	48	3.26E-20
GO Molecular Function	chemokine activity	76.7-100	43	48	3.26E-20
GO Molecular Function	chemoattractant activity	76.7-100	44	50	5.11E-20
GO Molecular Function	small monomeric GTPase activity	76.7-100	81	132	1.29E-19
GO Molecular Function	cytokine activity	76.7-100	107	203	3.88E-19
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	76.7-100	96	190	2.91E-15
GO Molecular Function	hydrogen ion transporter activity	76.7-100	67	113	1.04E-14
GO Molecular Function	monovalent inorganic cation transporter activity	76.7-100	67	123	4.21E-12
GO Molecular Function	GTPase activity	76.7-100	99	217	5.51E-12
GO Molecular Function	RAB small monomeric GTPase activity	76.7-100	36	53	1.86E-09
GO Molecular Function	growth factor activity	76.7-100	71	151	9.3E-09
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, other acceptor	76.7-100	27	38	3.01E-07
GO Molecular Function	NADH dehydrogenase activity	76.7-100	26	36	3.75E-07
GO Molecular Function	NADH dehydrogenase (ubiquinone) activity	76.7-100	26	36	3.75E-07
GO Molecular Function	molecular_function unknown	76.7-100	191	584	4.11E-07
GO Molecular Function	GTP binding	76.7-100	95	244	1.42E-06
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, quinone or similar compound as acceptor	76.7-100	27	40	0.0000019
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH	76.7-100	36	63	2.76E-06
GO Molecular Function	guanyl nucleotide binding	76.7-100	95	251	8.09E-06
GO Molecular Function	neuropeptide hormone activity	76.7-100	18	22	9.66E-06
GO Molecular Function	sodium ion transporter activity	76.7-100	26	40	0.0000136
GO Molecular Function	Rho small monomeric GTPase activity	76.7-100	20	27	0.0000272
GO Molecular Function	antimicrobial peptide activity	76.7-100	22	35	0.000525
GO Molecular Function	structural molecule activity	76.7-100	208	711	0.00161
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors\, oxygen as acceptor	76.7-100	15	20	0.00166
GO Molecular Function	cytochrome-c oxidase activity	76.7-100	15	20	0.00166

GO Molecular Function	oxidoreductase activity\, acting on heme group of donors	76.7-100	15	20	0.00166
GO Molecular Function	heme-copper terminal oxidase activity	76.7-100	15	20	0.00166
GO Molecular Function	cation transporter activity	76.7-100	73	200	0.00288
GO Molecular Function	glutathione transferase activity	76.7-100	15	21	0.00464
GO Molecular Function	kinase inhibitor activity	76.7-100	16	25	0.019
GO Molecular Function	small protein conjugating enzyme activity	76.7-100	28	58	0.0196
GO Molecular Function	protein translocase activity	76.7-100	13	18	0.02
GO Molecular Function	antifungal peptide activity	76.7-100	9	10	0.0265
GO Molecular Function	retinoid binding	76.7-100	9	10	0.0265
GO Molecular Function	isoprenoid binding	76.7-100	9	10	0.0265
GO Molecular Function	protein kinase inhibitor activity	76.7-100	15	23	0.0268
GO Molecular Function	kinase regulator activity	76.7-100	30	65	0.0289
GO Molecular Function	ubiquitin conjugating enzyme activity	76.7-100	27	56	0.0294
GO Molecular Function	small monomeric GTPase activity	72.8-84.4	43	132	2.34E-07
GO Molecular Function	GTP binding	72.8-84.4	59	244	4.36E-05
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	72.8-84.4	49	190	8.28E-05
GO Molecular Function	guanyl nucleotide binding	72.8-84.4	59	251	1.30E-04
GO Molecular Function	RAB small monomeric GTPase activity	72.8-84.4	21	53	3.71E-04
GO Molecular Function	GTPase activity	72.8-84.4	51	217	1.06E-03
GO Molecular Function	structural constituent of ribosome	72.8-84.4	49	210	2.24E-03
GO Molecular Function	transcription factor activity	72.8-84.4	136	809	5.50E-03
GO Molecular Function	receptor binding	72.8-84.4	94	522	1.37E-02
GO Molecular Function	RAS small monomeric GTPase activity	72.8-84.4	13	29	1.83E-02
GO Molecular Function	small monomeric GTPase activity	68.9-82.2	90	132	1.44E-25
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	68.9-82.2	100	190	3.78E-16
GO Molecular Function	receptor binding	68.9-82.2	202	522	6.42E-14
GO Molecular Function	GTPase activity	68.9-82.2	105	217	1.27E-13
GO Molecular Function	structural constituent of ribosome	68.9-82.2	102	210	2.60E-13
GO Molecular Function	RAB small monomeric GTPase activity	68.9-82.2	40	53	1.43E-12
GO Molecular Function	GTP binding	68.9-82.2	109	244	3.70E-11
GO Molecular Function	guanyl nucleotide binding	68.9-82.2	109	251	3.80E-10
GO Molecular Function	Rho small monomeric GTPase activity	68.9-82.2	22	27	4.92E-07
GO Molecular Function	RAS small monomeric GTPase activity	68.9-82.2	22	29	5.87E-06
GO Molecular Function	growth factor activity	68.9-82.2	66	151	2.02E-05
GO Molecular Function	cytokine activity	68.9-82.2	82	203	2.82E-05
GO Molecular Function	MHC class II receptor activity	68.9-82.2	14	16	2.33E-04
GO Molecular Function	transcription factor activity	68.9-82.2	245	809	3.40E-04
GO Molecular Function	hematopoietin/interferon-class (D200-domain) cytokine receptor binding	68.9-82.2	25	43	1.79E-03
GO Molecular Function	hormone activity	68.9-82.2	44	99	3.89E-03
GO Molecular Function	glutathione transferase activity	68.9-82.2	15	21	9.20E-03
GO Molecular Function	rhodopsin-like receptor activity	65.0-76.7	94	331	1.45E-13
GO Molecular Function	G-protein coupled receptor activity	65.0-76.7	100	410	7.20E-10
GO Molecular Function	transmembrane receptor activity	65.0-76.7	153	852	5.28E-05
GO Molecular Function	transferase activity\, transferring sulfur-containing groups	65.0-76.7	20	44	6.60E-05
GO Molecular Function	sulfotransferase activity	65.0-76.7	18	40	4.05E-04
GO Molecular Function	receptor activity	65.0-76.7	209	1306	1.55E-03
GO Molecular Function	peptide receptor activity\, G-protein coupled	65.0-76.7	30	109	1.44E-02

GO Molecular Function	peptide receptor activity	65.0-76.7	30	109	1.44E-02
GO Molecular Function	rhodopsin-like receptor activity	61.1-84.4	152	331	1.03E-16
GO Molecular Function	G-protein coupled receptor activity	61.1-84.4	159	410	1.86E-09
GO Molecular Function	small monomeric GTPase activity	61.1-84.4	61	132	1.73E-05
GO Molecular Function	transcription factor activity	61.1-84.4	256	809	2.83E-05
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	61.1-84.4	77	190	2.40E-04
GO Molecular Function	peptide receptor activity\, G-protein coupled	61.1-84.4	50	109	5.59E-04
GO Molecular Function	peptide receptor activity	61.1-84.4	50	109	5.59E-04
GO Molecular Function	GTP binding	61.1-84.4	92	244	7.53E-04
GO Molecular Function	transcription regulator activity	61.1-84.4	326	1117	2.62E-03
GO Molecular Function	guanyl nucleotide binding	61.1-84.4	92	251	3.27E-03
GO Molecular Function	GTPase activity	61.1-84.4	80	217	1.29E-02
GO Molecular Function	peptide binding	61.1-84.4	57	143	2.20E-02
GO Molecular Function	RAS small monomeric GTPase activity	61.1-84.4	18	29	3.19E-02
GO Molecular Function	transmembrane receptor activity	61.1-84.4	250	852	3.86E-02
GO Molecular Function	rhodopsin-like receptor activity	57.2-68.9	102	331	9.76E-18
GO Molecular Function	G-protein coupled receptor activity	57.2-68.9	111	410	1.07E-14
GO Molecular Function	transmembrane receptor activity	57.2-68.9	159	852	1.72E-06
GO Molecular Function	receptor activity	57.2-68.9	216	1306	7.38E-05
GO Molecular Function	peptide receptor activity\, G-protein coupled	57.2-68.9	34	109	1.36E-04
GO Molecular Function	peptide receptor activity	57.2-68.9	34	109	1.36E-04
GO Molecular Function	signal transducer activity	57.2-68.9	315	2102	1.03E-03
GO Molecular Function	sialyltransferase activity	57.2-68.9	10	16	6.01E-03
GO Molecular Function	transferase activity\, transferring other glycosyl groups	57.2-68.9	10	16	6.01E-03
GO Molecular Function	peptide binding	57.2-68.9	37	143	6.51E-03
GO Molecular Function	rhodopsin-like receptor activity	53.4-76.7	187	331	1.19E-35
GO Molecular Function	G-protein coupled receptor activity	53.4-76.7	204	410	1.53E-28
GO Molecular Function	transmembrane receptor activity	53.4-76.7	298	852	8.14E-12
GO Molecular Function	peptide receptor activity\, G-protein coupled	53.4-76.7	63	109	4.84E-11
GO Molecular Function	peptide receptor activity	53.4-76.7	63	109	4.84E-11
GO Molecular Function	receptor activity	53.4-76.7	411	1306	5.26E-09
GO Molecular Function	peptide binding	53.4-76.7	68	143	8.52E-07
GO Molecular Function	olfactory receptor activity	53.4-76.7	27	39	6.70E-06
GO Molecular Function	signal transducer activity	53.4-76.7	598	2102	1.18E-05
GO Molecular Function	transferase activity\, transferring sulfur-containing groups	53.4-76.7	27	44	3.32E-04
GO Molecular Function	sulfotransferase activity	53.4-76.7	25	40	6.00E-04
GO Molecular Function	transcription factor activity	53.4-76.7	247	809	4.32E-03
GO Molecular Function	nucleotide receptor activity\, G-protein coupled	53.4-76.7	18	29	3.93E-02
GO Molecular Function	nucleotide receptor activity	53.4-76.7	18	29	3.93E-02
GO Molecular Function	purinergic nucleotide receptor activity	53.4-76.7	18	29	3.93E-02
GO Molecular Function	purinergic nucleotide receptor activity\, G-protein coupled	53.4-76.7	18	29	3.93E-02
GO Molecular Function	G-protein coupled receptor activity	49.5-61.1	95	410	2.55E-07
GO Molecular Function	rhodopsin-like receptor activity	49.5-61.1	79	331	2.41E-06
GO Molecular Function	receptor activity	49.5-61.1	214	1306	1.06E-03
GO Molecular Function	peptide receptor activity	49.5-61.1	31	109	7.87E-03
GO Molecular Function	peptide receptor activity\, G-protein coupled	49.5-61.1	31	109	7.87E-03
GO Molecular Function	dopamine receptor activity	49.5-61.1	6	6	8.80E-03
GO Molecular Function	C-C chemokine receptor activity	49.5-61.1	9	15	3.80E-02

GO Molecular Function	C-C chemokine binding	49.5-61.1	9	15	3.80E-02
GO Molecular Function	rhodopsin-like receptor activity	45.6-68.9	167	331	7.92E-23
GO Molecular Function	G-protein coupled receptor activity	45.6-68.9	189	410	4.26E-20
GO Molecular Function	receptor activity	45.6-68.9	422	1306	4.25E-10
GO Molecular Function	transmembrane receptor activity	45.6-68.9	289	852	1.45E-08
GO Molecular Function	peptide receptor activity\, G-protein coupled	45.6-68.9	59	109	3.39E-08
GO Molecular Function	peptide receptor activity	45.6-68.9	59	109	3.39E-08
GO Molecular Function	peptide binding	45.6-68.9	63	143	3.12E-04
GO Molecular Function	signal transducer activity	45.6-68.9	587	2102	4.61E-03
GO Molecular Function	coreceptor activity	45.6-68.9	16	23	1.80E-02
GO Molecular Function	C-C chemokine binding	45.6-68.9	12	15	2.91E-02
GO Molecular Function	C-C chemokine receptor activity	45.6-68.9	12	15	2.91E-02
GO Molecular Function	pepsin A activity	45.6-68.9	8	8	4.18E-02
GO Molecular Function	G-protein coupled receptor activity	37.8-61.1	155	410	5.07E-07
GO Molecular Function	rhodopsin-like receptor activity	37.8-61.1	129	331	1.98E-06
GO Molecular Function	nicotinic acetylcholine-activated cation-selective channel activity	37.8-61.1	14	16	0.000581
GO Molecular Function	receptor activity	37.8-61.1	385	1306	0.00247
GO Molecular Function	neurotransmitter binding	37.8-61.1	28	51	0.00756
GO Molecular Function	neurotransmitter receptor activity	37.8-61.1	27	50	0.0176
GO Molecular Function	catalytic activity	33.9-45.6	579	4200	3.65E-03
GO Molecular Function	monooxygenase activity	33.9-45.6	24	72	4.34E-03
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	33.9-45.6	26	87	1.62E-02
GO Molecular Function	catalytic activity	30.0-53.4	1149	4200	2.73E-07
GO Molecular Function	monooxygenase activity	30.0-53.4	39	72	1.28E-04
GO Molecular Function	neurotransmitter binding	30.0-53.4	30	51	4.25E-04
GO Molecular Function	neurotransmitter receptor activity	30.0-53.4	29	50	1.06E-03
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	30.0-53.4	42	87	2.63E-03
GO Molecular Function	solute\:sodium symporter activity	30.0-53.4	21	32	2.75E-03
GO Molecular Function	unspecific monooxygenase activity	30.0-53.4	17	24	6.52E-03
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen\, reduced flavin or flavoprotein as one donor\, and incorporation of one atom of oxygen	30.0-53.4	17	24	6.52E-03
GO Molecular Function	solute\:cation symporter activity	30.0-53.4	25	43	6.70E-03
GO Molecular Function	oxidoreductase activity	30.0-53.4	174	544	3.30E-02
GO Molecular Function	oxygen binding	30.0-53.4	16	24	4.55E-02
GO Molecular Function	catalytic activity	26.2-37.8	609	4200	1.18E-07
GO Molecular Function	porter activity	26.2-37.8	43	184	2.56E-02
GO Molecular Function	electrochemical potential-driven transporter activity	26.2-37.8	43	185	2.96E-02
GO Molecular Function	transferase activity	26.2-37.8	206	1327	4.40E-02
GO Molecular Function	catalytic activity	22.3-45.6	1179	4200	8.15E-13
GO Molecular Function	transferase activity\, transferring hexosyl groups	22.3-45.6	56	118	5.93E-05
GO Molecular Function	monooxygenase activity	22.3-45.6	39	72	9.97E-05
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	22.3-45.6	43	87	6.54E-04
GO Molecular Function	transferase activity	22.3-45.6	392	1327	6.64E-04
GO Molecular Function	amine/polyamine transporter activity	22.3-45.6	32	60	2.84E-03
GO Molecular Function	protein serine/threonine kinase activity	22.3-45.6	124	355	3.99E-03
GO Molecular Function	calcium-dependent cell adhesion molecule activity	22.3-45.6	43	94	9.44E-03

GO Molecular Function	diacylglycerol binding	22.3-45.6	21	34	1.02E-02
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	22.3-45.6	189	598	1.81E-02
GO Molecular Function	porter activity	22.3-45.6	71	184	1.82E-02
GO Molecular Function	acyl-CoA or acyl binding	22.3-45.6	23	40	1.93E-02
GO Molecular Function	electrochemical potential-driven transporter activity	22.3-45.6	71	185	2.28E-02
GO Molecular Function	symporter activity	22.3-45.6	36	78	4.80E-02
GO Molecular Function	calcium-dependent cell adhesion molecule activity	18.4-30.0	56	94	5.68E-25
GO Molecular Function	adenyl nucleotide binding	18.4-30.0	210	1071	4.01E-11
GO Molecular Function	ATP binding	18.4-30.0	207	1059	9.02E-11
GO Molecular Function	cell adhesion molecule activity	18.4-30.0	94	374	1.50E-09
GO Molecular Function	calcium ion binding	18.4-30.0	128	576	1.55E-09
GO Molecular Function	metal ion binding	18.4-30.0	212	1135	5.02E-09
GO Molecular Function	purine nucleotide binding	18.4-30.0	234	1303	1.96E-08
GO Molecular Function	nucleotide binding	18.4-30.0	235	1317	3.44E-08
GO Molecular Function	catalytic activity	18.4-30.0	588	4200	2.88E-04
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	18.4-30.0	113	598	7.06E-04
GO Molecular Function	protein kinase activity	18.4-30.0	98	508	1.74E-03
GO Molecular Function	kinase activity	18.4-30.0	127	725	8.72E-03
GO Molecular Function	transferase activity\, transferring phosphorus-containing groups	18.4-30.0	130	750	1.16E-02
GO Molecular Function	transferase activity	18.4-30.0	209	1327	1.54E-02
GO Molecular Function	calcium-dependent cell adhesion molecule activity	14.5-37.8	76	94	6.54E-28
GO Molecular Function	adenyl nucleotide binding	14.5-37.8	369	1071	3.85E-13
GO Molecular Function	ATP binding	14.5-37.8	365	1059	5.48E-13
GO Molecular Function	catalytic activity	14.5-37.8	1177	4200	3.92E-12
GO Molecular Function	metal ion binding	14.5-37.8	377	1135	9.89E-11
GO Molecular Function	cell adhesion molecule activity	14.5-37.8	154	374	1.33E-10
GO Molecular Function	calcium ion binding	14.5-37.8	209	576	1.56E-08
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	14.5-37.8	212	598	1.34E-07
GO Molecular Function	purine nucleotide binding	14.5-37.8	406	1303	3.17E-07
GO Molecular Function	nucleotide binding	14.5-37.8	409	1317	4.38E-07
GO Molecular Function	transferase activity	14.5-37.8	408	1327	2.03E-06
GO Molecular Function	protein kinase activity	14.5-37.8	179	508	8.43E-06
GO Molecular Function	transferase activity\, transferring phosphorus-containing groups	14.5-37.8	247	750	1.21E-05
GO Molecular Function	kinase activity	14.5-37.8	239	725	2.02E-05
GO Molecular Function	metalloendopeptidase activity	14.5-37.8	47	93	7.59E-05
GO Molecular Function	protein serine/threonine kinase activity	14.5-37.8	126	355	1.36E-03
GO Molecular Function	metallopeptidase activity	14.5-37.8	69	169	2.43E-03
GO Molecular Function	transferase activity\, transferring hexosyl groups	14.5-37.8	51	118	9.77E-03
GO Molecular Function	hydrolase activity	14.5-37.8	480	1699	1.12E-02
GO Molecular Function	symporter activity	14.5-37.8	37	78	1.71E-02
GO Molecular Function	porter activity	14.5-37.8	71	184	1.95E-02
GO Molecular Function	electrochemical potential-driven transporter activity	14.5-37.8	71	185	2.43E-02
GO Molecular Function	cell adhesion molecule activity	10.6-22.3	110	374	1.04E-16
GO Molecular Function	adenyl nucleotide binding	10.6-22.3	227	1071	5.93E-16
GO Molecular Function	ATP binding	10.6-22.3	223	1059	3.04E-15
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	10.6-22.3	33	64	3.66E-11
GO Molecular Function	transmembrane receptor protein kinase activity	10.6-22.3	36	76	6.98E-11
GO Molecular Function	purine nucleotide binding	10.6-22.3	244	1303	1.69E-10
GO Molecular Function	nucleotide binding	10.6-22.3	246	1317	1.75E-10

GO Molecular Function	calcium-dependent cell adhesion molecule activity	10.6-22.3	38	94	6.49E-09
GO Molecular Function	glutamate receptor activity	10.6-22.3	21	34	2.30E-08
GO Molecular Function	metal ion binding	10.6-22.3	209	1135	8.30E-08
GO Molecular Function	protein kinase activity	10.6-22.3	112	508	1.56E-07
GO Molecular Function	calcium ion binding	10.6-22.3	121	576	6.74E-07
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	10.6-22.3	124	598	9.50E-07
GO Molecular Function	glutamate-gated ion channel activity	10.6-22.3	13	16	1.15E-06
GO Molecular Function	glutamate channel activity	10.6-22.3	13	16	1.15E-06
GO Molecular Function	monocarboxylate channel activity	10.6-22.3	13	16	1.15E-06
GO Molecular Function	inotropic glutamate receptor activity	10.6-22.3	13	17	4.36E-06
GO Molecular Function	protein-tyrosine kinase activity	10.6-22.3	62	241	8.24E-06
GO Molecular Function	guanylate cyclase activity	10.6-22.3	12	17	8.47E-05
GO Molecular Function	kainate selective glutamate receptor activity	10.6-22.3	8	8	1.20E-04
GO Molecular Function	kinase activity	10.6-22.3	135	725	2.02E-04
GO Molecular Function	binding	10.6-22.3	890	6657	2.88E-04
GO Molecular Function	transferase activity\, transferring phosphorus-containing groups	10.6-22.3	138	750	3.01E-04
GO Molecular Function	phosphorus-oxygen lyase activity	10.6-22.3	12	20	1.22E-03
GO Molecular Function	ephrin receptor activity	10.6-22.3	10	15	2.90E-03
GO Molecular Function	ATP dependent helicase activity	10.6-22.3	30	103	6.63E-03
GO Molecular Function	metalloendopeptidase activity	10.6-22.3	28	93	6.96E-03
GO Molecular Function	metallopeptidase activity	10.6-22.3	40	169	4.62E-02
GO Molecular Function	adenyl nucleotide binding	6.7-30.0	440	1071	7.29E-36
GO Molecular Function	ATP binding	6.7-30.0	434	1059	5.62E-35
GO Molecular Function	calcium-dependent cell adhesion molecule activity	6.7-30.0	81	94	6.58E-34
GO Molecular Function	cell adhesion molecule activity	6.7-30.0	194	374	4.69E-29
GO Molecular Function	purine nucleotide binding	6.7-30.0	478	1303	7.59E-25
GO Molecular Function	nucleotide binding	6.7-30.0	480	1317	3.25E-24
GO Molecular Function	metal ion binding	6.7-30.0	409	1135	8.10E-19
GO Molecular Function	protein kinase activity	6.7-30.0	205	508	1.11E-13
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	6.7-30.0	46	64	2.26E-12
GO Molecular Function	calcium ion binding	6.7-30.0	222	576	2.44E-12
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	6.7-30.0	228	598	4.00E-12
GO Molecular Function	transmembrane receptor protein kinase activity	6.7-30.0	50	76	3.64E-11
GO Molecular Function	ATPase activity\, coupled	6.7-30.0	107	236	9.93E-10
GO Molecular Function	protein-tyrosine kinase activity	6.7-30.0	108	241	1.96E-09
GO Molecular Function	kinase activity	6.7-30.0	252	725	2.35E-08
GO Molecular Function	transferase activity\, transferring phosphorus-containing groups	6.7-30.0	259	750	2.58E-08
GO Molecular Function	ATPase activity	6.7-30.0	107	249	6.19E-08
GO Molecular Function	metalloendopeptidase activity	6.7-30.0	52	93	1.10E-07
GO Molecular Function	GTPase regulator activity	6.7-30.0	99	231	4.27E-07
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	6.7-30.0	113	275	4.89E-07
GO Molecular Function	small GTPase regulatory/interacting protein activity	6.7-30.0	76	163	5.65E-07
GO Molecular Function	glutamate receptor activity	6.7-30.0	25	34	5.03E-06
GO Molecular Function	guanylate cyclase activity	6.7-30.0	16	17	5.36E-06
GO Molecular Function	binding	6.7-30.0	1724	6657	6.13E-06
GO Molecular Function	metallopeptidase activity	6.7-30.0	75	169	1.20E-05
GO Molecular Function	glutamate-gated ion channel activity	6.7-30.0	15	16	2.12E-05
GO Molecular Function	monocarboxylate channel activity	6.7-30.0	15	16	2.12E-05

GO Molecular Function	glutamate channel activity	6.7-30.0	15	16	2.12E-05
GO Molecular Function	magnesium ion binding	6.7-30.0	62	133	2.77E-05
GO Molecular Function	ATP dependent helicase activity	6.7-30.0	51	103	4.84E-05
GO Molecular Function	phosphorus-oxygen lyase activity	6.7-30.0	17	20	5.14E-05
GO Molecular Function	guanyl-nucleotide exchange factor activity	6.7-30.0	48	95	5.63E-05
GO Molecular Function	catalytic activity	6.7-30.0	1125	4200	8.59E-05
GO Molecular Function	inotropic glutamate receptor activity	6.7-30.0	15	17	1.40E-04
GO Molecular Function	protein serine/threonine kinase activity	6.7-30.0	127	355	8.27E-04
GO Molecular Function	hydrolase activity	6.7-30.0	488	1699	1.45E-03
GO Molecular Function	transferase activity	6.7-30.0	390	1327	2.19E-03
GO Molecular Function	helicase activity	6.7-30.0	54	125	5.26E-03
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	6.7-30.0	52	119	5.39E-03
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	6.7-30.0	52	119	5.39E-03
GO Molecular Function	cell adhesion receptor activity	6.7-30.0	24	42	1.42E-02
GO Molecular Function	cation\:chloride symporter activity	6.7-30.0	8	8	3.92E-02
GO Molecular Function	kainate selective glutamate receptor activity	6.7-30.0	8	8	3.92E-02
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of ions\, phosphorylative mechanism	6.7-30.0	24	44	4.24E-02
GO Molecular Function	RNA helicase activity	6.7-30.0	18	29	4.86E-02
GO Molecular Function	ATP binding	2.8-14.5	258	1059	4.11E-28
GO Molecular Function	adenyl nucleotide binding	2.8-14.5	260	1071	4.21E-28
GO Molecular Function	purine nucleotide binding	2.8-14.5	271	1303	3.30E-18
GO Molecular Function	nucleotide binding	2.8-14.5	273	1317	3.74E-18
GO Molecular Function	ATPase activity\, coupled	2.8-14.5	73	236	1.95E-11
GO Molecular Function	ATPase activity	2.8-14.5	74	249	1.30E-10
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	2.8-14.5	76	275	3.69E-09
GO Molecular Function	ATP dependent helicase activity	2.8-14.5	37	103	8.38E-07
GO Molecular Function	cell adhesion molecule activity	2.8-14.5	87	374	1.92E-06
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	2.8-14.5	27	64	3.42E-06
GO Molecular Function	motor activity	2.8-14.5	38	113	4.29E-06
GO Molecular Function	protein kinase activity	2.8-14.5	106	508	1.85E-05
GO Molecular Function	helicase activity	2.8-14.5	39	125	2.92E-05
GO Molecular Function	GTPase regulator activity	2.8-14.5	59	231	3.16E-05
GO Molecular Function	transmembrane receptor protein kinase activity	2.8-14.5	28	76	6.28E-05
GO Molecular Function	protein-tyrosine kinase activity	2.8-14.5	58	241	3.99E-04
GO Molecular Function	small GTPase regulatory/interacting protein activity	2.8-14.5	44	163	4.17E-04
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	2.8-14.5	35	119	8.26E-04
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	2.8-14.5	35	119	8.26E-04
GO Molecular Function	epidermal growth factor receptor activity	2.8-14.5	7	7	9.78E-04
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	2.8-14.5	113	598	1.53E-03
GO Molecular Function	binding	2.8-14.5	888	6657	2.62E-03
GO Molecular Function	calmodulin binding	2.8-14.5	31	105	3.52E-03
GO Molecular Function	ubiquitin thiolesterase activity	2.8-14.5	18	45	4.98E-03
GO Molecular Function	magnesium ion binding	2.8-14.5	36	133	5.11E-03
GO Molecular Function	glutamate-gated ion channel activity	2.8-14.5	10	16	6.85E-03
GO Molecular Function	monocarboxylate channel activity	2.8-14.5	10	16	6.85E-03
GO Molecular Function	glutamate channel activity	2.8-14.5	10	16	6.85E-03

GO Molecular Function	inotropic glutamate receptor activity	2.8-14.5	10	17	1.48E-02
GO Molecular Function	thiolester hydrolase activity	2.8-14.5	19	54	2.43E-02
GO Molecular Function	P-P-bond-hydrolysis-driven transporter activity	2.8-14.5	35	137	2.87E-02
GO Molecular Function	ATP-binding cassette (ABC) transporter activity	2.8-14.5	23	74	3.14E-02
SwissProt keyword	Acetylation	88.3-100	78	178	2.4E-25
SwissProt keyword	Chromosomal protein	88.3-100	37	54	9.07E-20
SwissProt keyword	Nucleosome core	88.3-100	27	30	1.43E-19
SwissProt keyword	Hormone	88.3-100	34	59	1.78E-14
SwissProt keyword	Chemotaxis	88.3-100	30	52	1.51E-12
SwissProt keyword	Amidation	88.3-100	22	32	6.11E-11
SwissProt keyword	Cytokine	88.3-100	51	146	7.3E-11
SwissProt keyword	Cleavage on pair of basic residues	88.3-100	27	56	1.38E-08
SwissProt keyword	Mitochondrion	88.3-100	87	392	5.61E-07
SwissProt keyword	Ubiquinone	88.3-100	19	36	0.0000027
SwissProt keyword	Antibiotic	88.3-100	13	19	0.0000149
SwissProt keyword	3D-structure	88.3-100	179	1067	0.0000183
SwissProt keyword	Inflammatory response	88.3-100	22	52	0.0000271
SwissProt keyword	Fungicide	88.3-100	9	10	0.0000595
SwissProt keyword	Metal-thiolate cluster	88.3-100	9	10	0.0000595
SwissProt keyword	Ribosomal protein	88.3-100	29	88	0.000117
SwissProt keyword	Cadmium	88.3-100	7	8	0.00374
SwissProt keyword	Mitogen	88.3-100	13	29	0.0125
SwissProt keyword	Pyrrolidone carboxylic acid	88.3-100	18	52	0.0182
SwissProt keyword	Vitamin A	88.3-100	6	7	0.029
SwissProt keyword	Defensin	88.3-100	5	5	0.0405
SwissProt keyword	Lipocalin	88.3-100	8	13	0.0446
SwissProt keyword	Prenylation	80.6-92.2	42	91	2.94E-13
SwissProt keyword	Homeobox	80.6-92.2	52	152	2.63E-10
SwissProt keyword	GTP-binding	80.6-92.2	49	158	8.24E-08
SwissProt keyword	Ribosomal protein	80.6-92.2	33	88	5.70E-07
SwissProt keyword	Developmental protein	80.6-92.2	64	261	5.66E-06
SwissProt keyword	Growth factor	80.6-92.2	32	109	9.07E-04
SwissProt keyword	Lipoprotein	80.6-92.2	67	317	1.20E-03
SwissProt keyword	Mitogen	80.6-92.2	14	29	2.57E-03
SwissProt keyword	MHC II	80.6-92.2	9	13	4.17E-03
SwissProt keyword	Hormone	80.6-92.2	20	59	1.28E-02
SwissProt keyword	Mitochondrion	80.6-92.2	75	392	1.37E-02
SwissProt keyword	Acetylation	76.7-100	104	178	1.47E-21
SwissProt keyword	Prenylation	76.7-100	63	91	1.14E-17
SwissProt keyword	Cytokine	76.7-100	84	146	2.38E-16
SwissProt keyword	Chromosomal protein	76.7-100	43	54	3.49E-15
SwissProt keyword	Hormone	76.7-100	44	59	1.21E-13
SwissProt keyword	Ribosomal protein	76.7-100	56	88	6.96E-13
SwissProt keyword	Nucleosome core	76.7-100	28	30	7.79E-13
SwissProt keyword	Chemotaxis	76.7-100	39	52	5.6E-12
SwissProt keyword	Amidation	76.7-100	27	32	7.67E-10
SwissProt keyword	Mitochondrion	76.7-100	149	392	8.49E-09
SwissProt keyword	Cleavage on pair of basic residues	76.7-100	37	56	1.7E-08
SwissProt keyword	Homeobox	76.7-100	72	152	5.31E-08
SwissProt keyword	Ubiquinone	76.7-100	26	36	1.27E-06

SwissProt keyword	GTP-binding	76.7-100	71	158	1.47E-06
SwissProt keyword	Lipoprotein	76.7-100	120	317	1.73E-06
SwissProt keyword	3D-structure	76.7-100	321	1067	6.55E-06
SwissProt keyword	Growth factor	76.7-100	53	109	8.01E-06
SwissProt keyword	Antibiotic	76.7-100	16	19	0.0000776
SwissProt keyword	Lipocalin	76.7-100	12	13	0.000638
SwissProt keyword	Inflammatory response	76.7-100	29	52	0.000832
SwissProt keyword	Lipid-binding	76.7-100	17	24	0.0024
SwissProt keyword	Mitogen	76.7-100	18	29	0.0207
SwissProt keyword	Inner membrane	76.7-100	27	54	0.0346
SwissProt keyword	Metal-thiolate cluster	76.7-100	9	10	0.0421
SwissProt keyword	CF(0)	76.7-100	9	10	0.0421
SwissProt keyword	Fungicide	76.7-100	9	10	0.0421
SwissProt keyword	Homeobox	72.8-84.4	58	152	6.00E-14
SwissProt keyword	Developmental protein	72.8-84.4	65	261	4.17E-06
SwissProt keyword	Lipoprotein	72.8-84.4	70	317	2.06E-04
SwissProt keyword	GTP-binding	72.8-84.4	41	158	1.46E-03
SwissProt keyword	Prenylation	72.8-84.4	27	91	8.45E-03
SwissProt keyword	Myristate	72.8-84.4	17	45	1.63E-02
SwissProt keyword	Homeobox	68.9-82.2	101	152	2.15E-26
SwissProt keyword	Developmental protein	68.9-82.2	130	261	1.23E-17
SwissProt keyword	Prenylation	68.9-82.2	61	91	2.79E-15
SwissProt keyword	GTP-binding	68.9-82.2	81	158	5.62E-11
SwissProt keyword	Ribosomal protein	68.9-82.2	53	88	4.33E-10
SwissProt keyword	Lipoprotein	68.9-82.2	131	317	1.39E-09
SwissProt keyword	Growth factor	68.9-82.2	51	109	2.23E-04
SwissProt keyword	MHC II	68.9-82.2	11	13	1.92E-02
SwissProt keyword	Selenocysteine	68.9-82.2	13	17	1.94E-02
SwissProt keyword	Mitogen	68.9-82.2	18	29	3.34E-02
SwissProt keyword	G-protein coupled receptor	65.0-76.7	84	316	4.26E-09
SwissProt keyword	Developmental protein	65.0-76.7	61	261	9.53E-04
SwissProt keyword	Homeobox	65.0-76.7	41	152	1.73E-03
SwissProt keyword	Homeobox	61.1-84.4	84	152	2.99E-13
SwissProt keyword	G-protein coupled receptor	61.1-84.4	135	316	2.79E-10
SwissProt keyword	Developmental protein	61.1-84.4	113	261	1.16E-08
SwissProt keyword	Lipoprotein	61.1-84.4	127	317	3.31E-07
SwissProt keyword	GTP-binding	61.1-84.4	64	158	1.08E-02
SwissProt keyword	G-protein coupled receptor	57.2-68.9	96	316	1.80E-14
SwissProt keyword	Palmitate	57.2-68.9	37	130	2.18E-03
SwissProt keyword	G-protein coupled receptor	53.4-76.7	174	316	7.05E-29
SwissProt keyword	Palmitate	53.4-76.7	65	130	9.21E-07
SwissProt keyword	Olfaction	53.4-76.7	26	44	4.05E-03
SwissProt keyword	Lipoprotein	53.4-76.7	113	317	1.66E-02
SwissProt keyword	Glycoprotein	53.4-76.7	569	2023	3.92E-02
SwissProt keyword	G-protein coupled receptor	49.5-61.1	82	316	2.48E-09
SwissProt keyword	Palmitate	49.5-61.1	36	130	1.37E-03
SwissProt keyword	G-protein coupled receptor	45.6-68.9	160	316	9.29E-22
SwissProt keyword	Palmitate	45.6-68.9	62	130	1.35E-05
SwissProt keyword	Intermediate filament	45.6-68.9	25	44	1.32E-02
SwissProt keyword	Glycoprotein	45.6-68.9	561	2023	3.59E-02

SwissProt keyword	G-protein coupled receptor	37.8-61.1	127	316	6.62E-08
SwissProt keyword	Palmitate	37.8-61.1	54	130	0.0151
SwissProt keyword	Monooxygenase	37.8-61.1	29	56	0.017
SwissProt keyword	Monooxygenase	33.9-45.6	22	56	2.81E-04
SwissProt keyword	Oxidoreductase	33.9-45.6	71	351	4.00E-03
SwissProt keyword	Microsome	33.9-45.6	24	82	3.49E-02
SwissProt keyword	Monooxygenase	30.0-53.4	36	56	2.71E-07
SwissProt keyword	Oxidoreductase	30.0-53.4	125	351	2.83E-04
SwissProt keyword	Transferase	30.0-53.4	236	765	1.18E-03
SwissProt keyword	Serine/threonine-protein kinase	30.0-53.4	86	226	1.37E-03
SwissProt keyword	Symport	30.0-53.4	30	60	2.31E-02
SwissProt keyword	Microsome	30.0-53.4	37	82	4.23E-02
SwissProt keyword	Serine/threonine-protein kinase	22.3-45.6	91	226	4.70E-05
SwissProt keyword	Monooxygenase	22.3-45.6	32	56	2.63E-04
SwissProt keyword	Alternative splicing	22.3-45.6	525	1867	4.72E-04
SwissProt keyword	Transferase	22.3-45.6	240	765	5.42E-04
SwissProt keyword	Hydrolase	22.3-45.6	239	779	4.25E-03
SwissProt keyword	Microsome	22.3-45.6	39	82	6.49E-03
SwissProt keyword	Phorbol-ester binding	22.3-45.6	19	29	7.37E-03
SwissProt keyword	Oxidoreductase	22.3-45.6	119	351	1.95E-02
SwissProt keyword	Zinc	22.3-45.6	102	294	3.07E-02
SwissProt keyword	Repeat	22.3-45.6	542	1993	3.35E-02
SwissProt keyword	ATP-binding	18.4-30.0	142	670	2.12E-10
SwissProt keyword	Repeat	18.4-30.0	322	1993	2.80E-09
SwissProt keyword	Calcium-binding	18.4-30.0	74	286	2.48E-08
SwissProt keyword	Cell adhesion	18.4-30.0	63	270	7.68E-05
SwissProt keyword	Sodium transport	18.4-30.0	15	36	1.30E-02
SwissProt keyword	Transferase	18.4-30.0	126	765	4.59E-02
SwissProt keyword	Repeat	14.5-37.8	628	1993	4.98E-19
SwissProt keyword	ATP-binding	14.5-37.8	245	670	1.79E-12
SwissProt keyword	Cell adhesion	14.5-37.8	111	270	1.17E-07
SwissProt keyword	Alternative splicing	14.5-37.8	541	1867	2.16E-07
SwissProt keyword	Zinc-finger	14.5-37.8	201	581	4.12E-07
SwissProt keyword	Calcium-binding	14.5-37.8	112	286	3.18E-06
SwissProt keyword	Metal-binding	14.5-37.8	167	479	9.58E-06
SwissProt keyword	Transferase	14.5-37.8	246	765	1.04E-05
SwissProt keyword	Hydrolase	14.5-37.8	241	779	6.73E-04
SwissProt keyword	Metalloprotease	14.5-37.8	44	92	8.11E-04
SwissProt keyword	Phorbol-ester binding	14.5-37.8	20	29	8.89E-04
SwissProt keyword	Serine/threonine-protein kinase	14.5-37.8	86	226	1.35E-03
SwissProt keyword	FAD	14.5-37.8	29	58	3.21E-02
SwissProt keyword	Phosphorylation	14.5-37.8	287	997	4.51E-02
SwissProt keyword	Repeat	10.6-22.3	387	1993	2.45E-31
SwissProt keyword	ATP-binding	10.6-22.3	148	670	3.17E-13
SwissProt keyword	Cell adhesion	10.6-22.3	80	270	3.60E-13
SwissProt keyword	Tyrosine-protein kinase	10.6-22.3	36	98	1.32E-07
SwissProt keyword	Receptor	10.6-22.3	91	420	1.40E-06
SwissProt keyword	Metal-binding	10.6-22.3	98	479	7.60E-06
SwissProt keyword	Zinc-finger	10.6-22.3	110	581	7.29E-05
SwissProt keyword	Magnesium	10.6-22.3	32	121	9.54E-03

SwissProt keyword	Integrin	10.6-22.3	12	25	1.63E-02
SwissProt keyword	EGF-like domain	10.6-22.3	31	119	1.90E-02
SwissProt keyword	Calcium-binding	10.6-22.3	58	286	2.06E-02
SwissProt keyword	cAMP biosynthesis	10.6-22.3	6	7	3.87E-02
SwissProt keyword	Repeat	6.7-30.0	730	1993	4.17E-55
SwissProt keyword	ATP-binding	6.7-30.0	296	670	1.36E-33
SwissProt keyword	Cell adhesion	6.7-30.0	143	270	5.07E-24
SwissProt keyword	Tyrosine-protein kinase	6.7-30.0	61	98	2.57E-13
SwissProt keyword	Zinc-finger	6.7-30.0	210	581	2.29E-10
SwissProt keyword	Alternative splicing	6.7-30.0	549	1867	3.70E-10
SwissProt keyword	Metal-binding	6.7-30.0	178	479	1.39E-09
SwissProt keyword	Phosphorylation	6.7-30.0	313	997	1.48E-07
SwissProt keyword	Calcium-binding	6.7-30.0	114	286	2.26E-07
SwissProt keyword	Magnesium	6.7-30.0	59	121	1.38E-06
SwissProt keyword	Metalloprotease	6.7-30.0	48	92	3.38E-06
SwissProt keyword	Receptor	6.7-30.0	148	420	1.23E-05
SwissProt keyword	Hydrolase	6.7-30.0	243	779	6.56E-05
SwissProt keyword	Integrin	6.7-30.0	18	25	9.81E-04
SwissProt keyword	Helicase	6.7-30.0	32	62	2.98E-03
SwissProt keyword	Antiport	6.7-30.0	12	14	4.45E-03
SwissProt keyword	Ligase	6.7-30.0	43	95	4.53E-03
SwissProt keyword	cGMP biosynthesis	6.7-30.0	9	9	6.55E-03
SwissProt keyword	Calcium	6.7-30.0	41	91	9.24E-03
SwissProt keyword	Leucine-rich repeat	6.7-30.0	38	84	1.90E-02
SwissProt keyword	Transferase	6.7-30.0	226	765	2.24E-02
SwissProt keyword	Phorbol-ester binding	6.7-30.0	18	29	2.68E-02
SwissProt keyword	Repeat	2.8-14.5	417	1993	8.07E-50
SwissProt keyword	ATP-binding	2.8-14.5	174	670	7.03E-27
SwissProt keyword	Cell adhesion	2.8-14.5	72	270	6.38E-10
SwissProt keyword	Coiled coil	2.8-14.5	89	416	5.65E-07
SwissProt keyword	Myosin	2.8-14.5	18	32	1.32E-06
SwissProt keyword	Calmodulin-binding	2.8-14.5	31	86	1.86E-06
SwissProt keyword	Receptor	2.8-14.5	88	420	2.21E-06
SwissProt keyword	Thick filament	2.8-14.5	12	15	2.56E-06
SwissProt keyword	Tyrosine-protein kinase	2.8-14.5	33	98	4.03E-06
SwissProt keyword	Phosphorylation	2.8-14.5	168	997	4.42E-06
SwissProt keyword	Helicase	2.8-14.5	24	62	3.11E-05
SwissProt keyword	Integrin	2.8-14.5	14	25	1.26E-04
SwissProt keyword	Hypothetical protein	2.8-14.5	37	133	1.45E-04
SwissProt keyword	Alternative splicing	2.8-14.5	270	1867	2.02E-04
SwissProt keyword	Alkylation	2.8-14.5	9	11	2.67E-04
SwissProt keyword	Magnesium	2.8-14.5	34	121	3.85E-04
SwissProt keyword	Connective tissue	2.8-14.5	15	32	9.13E-04
SwissProt keyword	Basement membrane	2.8-14.5	12	22	1.74E-03
SwissProt keyword	Chromosome partition	2.8-14.5	7	9	1.47E-02
SwissProt keyword	EGF-like domain	2.8-14.5	30	119	2.37E-02
SwissProt keyword	ANK repeat	2.8-14.5	21	70	3.07E-02
SwissProt keyword	Collagen	2.8-14.5	16	46	4.31E-02

Table C.3: Correlations to M_{HGMD}-ranked gene list

System	Gene Category	Percentile (%)	# genes	# in category	p-value
GO Biological Process	nucleosome assembly	88.3-100	36	61	4.08E-17
GO Biological Process	taxis	88.3-100	45	118	3.08E-12
GO Biological Process	chemotaxis	88.3-100	45	118	3.08E-12
GO Biological Process	chromatin assembly/disassembly	88.3-100	38	93	3.68E-11
GO Biological Process	response to chemical substance	88.3-100	61	223	1.49E-09
GO Biological Process	response to wounding	88.3-100	69	274	3.42E-09
GO Biological Process	protein biosynthesis	88.3-100	112	558	6.63E-09
GO Biological Process	innate immune response	88.3-100	53	188	1.39E-08
GO Biological Process	response to pest/pathogen/parasite	88.3-100	94	447	2.83E-08
GO Biological Process	inflammatory response	88.3-100	51	180	2.85E-08
GO Biological Process	response to abiotic stimulus	88.3-100	66	285	5.45E-07
GO Biological Process	DNA packaging	88.3-100	47	175	1.33E-06
GO Biological Process	defense response	88.3-100	136	787	1.94E-06
GO Biological Process	establishment and/or maintenance of chromatin architecture	88.3-100	44	163	3.95E-06
GO Biological Process	biological_process unknown	88.3-100	144	863	6.88E-06
GO Biological Process	response to biotic stimulus	88.3-100	142	849	7.65E-06
GO Biological Process	chromosome organization and biogenesis (sensu Eukarya)	88.3-100	47	190	2.41E-05
GO Biological Process	calcium ion homeostasis	88.3-100	16	31	2.59E-05
GO Biological Process	immune response	88.3-100	122	713	2.86E-05
GO Biological Process	nuclear organization and biogenesis	88.3-100	47	194	4.87E-05
GO Biological Process	protein-disulfide reduction	88.3-100	13	23	1.51E-04
GO Biological Process	di-\, tri-valent inorganic cation homeostasis	88.3-100	21	59	4.78E-04
GO Biological Process	ATP synthesis coupled electron transport	88.3-100	12	22	8.32E-04
GO Biological Process	ATP synthesis coupled electron transport (sensu Eukarya)	88.3-100	12	22	8.32E-04
GO Biological Process	response to external stimulus	88.3-100	160	1060	9.49E-04
GO Biological Process	cell-cell signaling	88.3-100	95	555	1.28E-03
GO Biological Process	organismal physiological process	88.3-100	190	1318	1.85E-03
GO Biological Process	macromolecule biosynthesis	88.3-100	134	869	2.90E-03
GO Biological Process	response to stress	88.3-100	123	782	2.95E-03
GO Biological Process	oxidative phosphorylation	88.3-100	13	28	3.00E-03
GO Biological Process	metal ion homeostasis	88.3-100	21	65	3.03E-03
GO Biological Process	mitochondrial translocation	88.3-100	10	17	3.19E-03
GO Biological Process	mitochondrial electron transport\, NADH to ubiquinone	88.3-100	10	19	1.24E-02
GO Biological Process	antimicrobial humoral response	88.3-100	24	89	1.97E-02
GO Biological Process	response to stimulus	88.3-100	183	1320	3.74E-02
GO Biological Process	small GTPase mediated signal transduction	80.6-92.2	73	212	3.38E-16
GO Biological Process	protein biosynthesis	80.6-92.2	102	558	6.94E-04
GO Biological Process	regulation of cell proliferation	80.6-92.2	54	258	9.14E-03
GO Biological Process	protein transport	80.6-92.2	77	417	1.43E-02
GO Biological Process	regulation of biological process	80.6-92.2	90	523	4.51E-02
GO Biological Process	small GTPase mediated signal transduction	76.7-100	108	212	2.15E-17

GO Biological Process	protein biosynthesis	76.7-100	197	558	2.17E-10
GO Biological Process	nucleosome assembly	76.7-100	38	61	3.57E-08
GO Biological Process	defense response	76.7-100	241	787	6.53E-06
GO Biological Process	response to biotic stimulus	76.7-100	256	849	1.00E-05
GO Biological Process	response to wounding	76.7-100	102	274	1.18E-05
GO Biological Process	biological_process unknown	76.7-100	259	863	1.32E-05
GO Biological Process	immune response	76.7-100	220	713	1.73E-05
GO Biological Process	cell-cell signaling	76.7-100	178	555	2.52E-05
GO Biological Process	taxis	76.7-100	53	118	6.03E-05
GO Biological Process	chemotaxis	76.7-100	53	118	6.03E-05
GO Biological Process	macromolecule biosynthesis	76.7-100	255	869	1.65E-04
GO Biological Process	response to pest/pathogen/parasite	76.7-100	146	447	1.66E-04
GO Biological Process	regulation of cell proliferation	76.7-100	93	258	3.43E-04
GO Biological Process	protein-disulfide reduction	76.7-100	17	23	4.76E-04
GO Biological Process	response to external stimulus	76.7-100	299	1060	6.96E-04
GO Biological Process	innate immune response	76.7-100	71	188	1.47E-03
GO Biological Process	chromatin assembly/disassembly	76.7-100	42	93	1.47E-03
GO Biological Process	ATP synthesis coupled electron transport	76.7-100	16	22	1.62E-03
GO Biological Process	ATP synthesis coupled electron transport (sensu Eukarya)	76.7-100	16	22	1.62E-03
GO Biological Process	protein transport	76.7-100	134	417	1.67E-03
GO Biological Process	response to stimulus	76.7-100	359	1320	2.27E-03
GO Biological Process	response to stress	76.7-100	227	782	2.38E-03
GO Biological Process	inflammatory response	76.7-100	68	180	2.52E-03
GO Biological Process	oxidative phosphorylation	76.7-100	18	28	5.35E-03
GO Biological Process	response to chemical substance	76.7-100	79	223	6.73E-03
GO Biological Process	mitochondrial electron transport\, NADH to ubiquinone	76.7-100	14	19	6.77E-03
GO Biological Process	mitochondrial translocation	76.7-100	13	17	8.01E-03
GO Biological Process	organismal physiological process	76.7-100	353	1318	1.39E-02
GO Biological Process	response to abiotic stimulus	76.7-100	94	285	2.56E-02
GO Biological Process	calcium ion homeostasis	76.7-100	18	31	4.23E-02
GO Biological Process	intracellular protein transport	76.7-100	122	395	4.67E-02
GO Biological Process	small GTPase mediated signal transduction	72.8-84.4	48	212	1.05E-02
GO Biological Process	positive regulation of cell proliferation	72.8-84.4	31	115	1.39E-02
GO Biological Process	small GTPase mediated signal transduction	68.9-92.2	109	212	2.40E-16
GO Biological Process	antigen processing\, exogenous antigen via MHC class II	61.1-84.4	12	14	5.90E-03
GO Biological Process	antigen presentation\, exogenous antigen	61.1-84.4	12	14	5.90E-03
GO Biological Process	mitochondrial transport	61.1-84.4	14	18	6.61E-03
GO Biological Process	G-protein coupled receptor protein signaling pathway	57.2-68.9	118	602	4.15E-05
GO Biological Process	frizzled-2 signaling pathway	57.2-68.9	11	18	2.74E-03
GO Biological Process	cell surface receptor linked signal transduction	57.2-68.9	160	974	2.01E-02
GO Biological Process	glucose metabolism	57.2-68.9	24	78	2.04E-02
GO Biological Process	carbohydrate metabolism	57.2-68.9	76	390	2.13E-02
GO Biological Process	monosaccharide metabolism	57.2-68.9	30	110	2.38E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	53.4-76.7	217	602	1.05E-08
GO Biological Process	frizzled-2 signaling pathway	53.4-76.7	17	18	1.33E-06
GO Biological Process	cell surface receptor linked signal transduction	53.4-76.7	299	974	6.03E-04

GO Biological Process	hexose catabolism	53.4-76.7	32	61	4.60E-03
GO Biological Process	alcohol metabolism	53.4-76.7	86	228	5.77E-03
GO Biological Process	monosaccharide catabolism	53.4-76.7	32	62	7.35E-03
GO Biological Process	alcohol catabolism	53.4-76.7	32	62	7.35E-03
GO Biological Process	monosaccharide metabolism	53.4-76.7	48	110	1.32E-02
GO Biological Process	hexose metabolism	53.4-76.7	46	107	3.40E-02
GO Biological Process	carbohydrate metabolism	53.4-76.7	130	390	3.55E-02
GO Biological Process	frizzled-2 signaling pathway	49.5-61.1	14	18	7.50E-07
GO Biological Process	G-protein coupled receptor protein signaling pathway	49.5-61.1	124	602	1.97E-06
GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	49.5-61.1	33	103	2.04E-04
GO Biological Process	G-protein signaling\, coupled to cAMP nucleotide second messenger	49.5-61.1	25	67	2.85E-04
GO Biological Process	cell surface receptor linked signal transduction	49.5-61.1	171	974	3.26E-04
GO Biological Process	cAMP-mediated signaling	49.5-61.1	25	69	5.55E-04
GO Biological Process	carbohydrate metabolism	49.5-61.1	82	390	7.13E-04
GO Biological Process	cyclic-nucleotide-mediated signaling	49.5-61.1	33	109	9.13E-04
GO Biological Process	second-messenger-mediated signaling	49.5-61.1	33	123	1.79E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	45.6-68.9	233	602	5.22E-13
GO Biological Process	frizzled-2 signaling pathway	45.6-68.9	18	18	2.69E-08
GO Biological Process	cell surface receptor linked signal transduction	45.6-68.9	319	974	2.72E-07
GO Biological Process	carbohydrate metabolism	45.6-68.9	147	390	2.55E-06
GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	45.6-68.9	51	103	6.26E-05
GO Biological Process	cyclic-nucleotide-mediated signaling	45.6-68.9	52	109	2.21E-04
GO Biological Process	monosaccharide metabolism	45.6-68.9	52	110	3.22E-04
GO Biological Process	hexose metabolism	45.6-68.9	50	107	8.95E-04
GO Biological Process	second-messenger-mediated signaling	45.6-68.9	53	123	9.15E-03
GO Biological Process	glucose metabolism	45.6-68.9	37	78	2.10E-02
GO Biological Process	G-protein signaling\, coupled to IP3 second messenger (phospholipase C activating)	45.6-68.9	37	79	3.04E-02
GO Biological Process	alcohol metabolism	45.6-68.9	84	228	3.37E-02
GO Biological Process	glycolysis	45.6-68.9	24	44	4.96E-02
GO Biological Process	metabolism	41.7-53.4	832	6240	2.70E-03
GO Biological Process	metabolism	37.8-61.1	1637	6240	8.67E-07
GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	37.8-61.1	53	103	5.23E-06
GO Biological Process	carbohydrate metabolism	37.8-61.1	145	390	8.31E-06
GO Biological Process	G-protein coupled receptor protein signaling pathway	37.8-61.1	207	602	8.37E-06
GO Biological Process	cyclic-nucleotide-mediated signaling	37.8-61.1	54	109	2.14E-05
GO Biological Process	gamma-aminobutyric acid signaling pathway	37.8-61.1	15	17	1.57E-04
GO Biological Process	second-messenger-mediated signaling	37.8-61.1	57	123	1.73E-04
GO Biological Process	G-protein signaling\, coupled to cAMP nucleotide second messenger	37.8-61.1	36	67	5.55E-04
GO Biological Process	frizzled-2 signaling pathway	37.8-61.1	15	18	7.30E-04
GO Biological Process	cAMP-mediated signaling	37.8-61.1	36	69	1.46E-03
GO Biological Process	alcohol metabolism	37.8-61.1	87	228	4.01E-03
GO Biological Process	G-protein signaling\, coupled to IP3 second messenger (phospholipase C activating)	37.8-61.1	38	79	9.75E-03
GO Biological Process	cell surface receptor linked signal transduction	37.8-61.1	292	974	1.67E-02
GO Biological Process	amine metabolism	37.8-61.1	117	341	3.21E-02

GO Biological Process	neurophysiological process	37.8-61.1	98	276	3.49E-02
GO Biological Process	carboxylic acid metabolism	37.8-61.1	140	423	3.76E-02
GO Biological Process	organic acid metabolism	37.8-61.1	140	425	4.94E-02
GO Biological Process	metabolism	33.9-45.6	823	6240	1.97E-04
GO Biological Process	metabolism	30.0-53.4	1655	6240	8.18E-11
GO Biological Process	physiological process	30.0-53.4	2288	9258	1.81E-03
GO Biological Process	carboxylic acid metabolism	30.0-53.4	143	423	5.34E-03
GO Biological Process	organic acid metabolism	30.0-53.4	143	425	7.19E-03
GO Biological Process	lipid metabolism	30.0-53.4	164	500	7.44E-03
GO Biological Process	electron transport	30.0-53.4	130	381	9.38E-03
GO Biological Process	metabolism	22.3-45.6	1636	6240	6.68E-09
GO Biological Process	physiological process	22.3-45.6	2272	9258	1.69E-02
GO Biological Process	carboxylic acid metabolism	22.3-45.6	140	423	2.03E-02
GO Biological Process	organic acid metabolism	22.3-45.6	140	425	2.68E-02
GO Biological Process	organic acid transport	22.3-45.6	28	55	4.24E-02
GO Biological Process	carboxylic acid transport	22.3-45.6	28	55	4.24E-02
GO Biological Process	peptide cross-linking	18.4-30.0	7	7	1.01E-03
GO Biological Process	neurotransmitter transport	18.4-30.0	14	29	4.48E-03
GO Biological Process	peptide cross-linking	18.4-30.0	7	7	1.01E-03
GO Biological Process	neurotransmitter transport	18.4-30.0	14	29	4.48E-03
GO Biological Process	neurotransmitter transport	14.5-37.8	21	29	1.60E-04
GO Biological Process	frizzled signaling pathway	14.5-37.8	14	17	2.24E-03
GO Biological Process	proteolysis and peptidolysis	14.5-37.8	187	578	3.41E-03
GO Biological Process	protein catabolism	14.5-37.8	189	587	4.27E-03
GO Biological Process	protein amino acid phosphorylation	14.5-37.8	162	491	5.07E-03
GO Biological Process	protein modification	14.5-37.8	293	975	5.51E-03
GO Biological Process	macromolecule catabolism	14.5-37.8	195	613	7.17E-03
GO Biological Process	phosphorylation	14.5-37.8	174	537	7.43E-03
GO Biological Process	phosphate metabolism	14.5-37.8	211	681	2.06E-02
GO Biological Process	phosphorus metabolism	14.5-37.8	211	681	2.06E-02
GO Biological Process	ion transport	14.5-37.8	170	531	2.19E-02
GO Biological Process	homophilic cell adhesion	10.6-22.3	53	116	2.23E-16
GO Biological Process	cell adhesion	10.6-22.3	141	589	1.07E-13
GO Biological Process	synaptogenesis	10.6-22.3	15	19	8.22E-08
GO Biological Process	synapse organization and biogenesis	10.6-22.3	15	19	8.22E-08
GO Biological Process	cellular process	10.6-22.3	809	5869	1.42E-07
GO Biological Process	cell-cell adhesion	10.6-22.3	61	223	4.78E-07
GO Biological Process	extracellular matrix organization and biogenesis	10.6-22.3	18	31	2.46E-06
GO Biological Process	extracellular structure organization and biogenesis	10.6-22.3	18	31	2.46E-06
GO Biological Process	protein-nucleus import\, docking	10.6-22.3	12	18	1.92E-04
GO Biological Process	cell communication	10.6-22.3	423	2885	2.74E-04
GO Biological Process	protein amino acid phosphorylation	10.6-22.3	98	491	2.85E-04
GO Biological Process	protein catabolism	10.6-22.3	109	587	2.47E-03
GO Biological Process	proteolysis and peptidolysis	10.6-22.3	107	578	3.63E-03
GO Biological Process	macromolecule catabolism	10.6-22.3	112	613	3.95E-03
GO Biological Process	integrin-mediated signaling pathway	10.6-22.3	18	45	3.96E-03
GO Biological Process	regulation of synapse	10.6-22.3	12	22	4.23E-03
GO Biological Process	phosphorylation	10.6-22.3	100	537	6.21E-03

GO Biological Process	phosphate metabolism	10.6-22.3	121	681	6.40E-03
GO Biological Process	phosphorus metabolism	10.6-22.3	121	681	6.40E-03
GO Biological Process	frizzled signaling pathway	10.6-22.3	10	17	1.30E-02
GO Biological Process	monovalent inorganic cation homeostasis	10.6-22.3	10	18	2.62E-02
GO Biological Process	hydrogen ion homeostasis	10.6-22.3	9	15	3.20E-02
GO Biological Process	cation transport	10.6-22.3	74	383	3.57E-02
GO Biological Process	ion transport	10.6-22.3	96	531	3.57E-02
GO Biological Process	protein metabolism	10.6-22.3	332	2296	4.65E-02
GO Biological Process	protein modification	10.6-22.3	158	975	4.73E-02
GO Biological Process	homophilic cell adhesion	6.7-30.0	94	116	3.45E-35
GO Biological Process	cell adhesion	6.7-30.0	257	589	7.64E-24
GO Biological Process	cell-cell adhesion	6.7-30.0	113	223	7.95E-15
GO Biological Process	protein amino acid phosphorylation	6.7-30.0	190	491	2.20E-10
GO Biological Process	cellular process	6.7-30.0	1556	5869	2.46E-08
GO Biological Process	phosphorus metabolism	6.7-30.0	239	681	2.82E-08
GO Biological Process	phosphate metabolism	6.7-30.0	239	681	2.82E-08
GO Biological Process	phosphorylation	6.7-30.0	196	537	5.10E-08
GO Biological Process	protein modification	6.7-30.0	315	975	9.77E-07
GO Biological Process	synapse organization and biogenesis	6.7-30.0	17	19	1.00E-05
GO Biological Process	synaptogenesis	6.7-30.0	17	19	1.00E-05
GO Biological Process	cell-matrix adhesion	6.7-30.0	38	67	2.97E-05
GO Biological Process	integrin-mediated signaling pathway	6.7-30.0	29	45	3.19E-05
GO Biological Process	cell communication	6.7-30.0	802	2885	3.80E-05
GO Biological Process	proteolysis and peptidolysis	6.7-30.0	196	578	5.68E-05
GO Biological Process	protein catabolism	6.7-30.0	198	587	7.61E-05
GO Biological Process	ion transport	6.7-30.0	180	531	2.32E-04
GO Biological Process	protein-nucleus import\, docking	6.7-30.0	15	18	6.61E-04
GO Biological Process	macromolecule catabolism	6.7-30.0	201	613	6.66E-04
GO Biological Process	enzyme linked receptor protein signaling pathway	6.7-30.0	54	119	7.77E-04
GO Biological Process	extracellular matrix organization and biogenesis	6.7-30.0	21	31	1.07E-03
GO Biological Process	extracellular structure organization and biogenesis	6.7-30.0	21	31	1.07E-03
GO Biological Process	transport	6.7-30.0	477	1653	1.07E-03
GO Biological Process	protein metabolism	6.7-30.0	641	2296	1.24E-03
GO Biological Process	frizzled signaling pathway	6.7-30.0	14	17	2.32E-03
GO Biological Process	transmembrane receptor protein tyrosine kinase signaling pathway	6.7-30.0	41	86	4.59E-03
GO Biological Process	regulation of synapse	6.7-30.0	16	22	6.72E-03
GO Biological Process	cation transport	6.7-30.0	130	383	1.51E-02
GO Biological Process	neurogenesis	6.7-30.0	140	420	1.95E-02
GO Biological Process	catabolism	6.7-30.0	251	830	3.08E-02
GO Biological Process	anion transport	6.7-30.0	46	107	3.65E-02
GO Biological Process	neurotransmitter transport	6.7-30.0	18	29	4.92E-02
GO Biological Process	cell adhesion	2.8-14.5	182	589	2.63E-33
GO Biological Process	homophilic cell adhesion	2.8-14.5	66	116	4.10E-28
GO Biological Process	cell-cell adhesion	2.8-14.5	81	223	3.97E-18
GO Biological Process	cell communication	2.8-14.5	468	2885	2.73E-12
GO Biological Process	cellular process	2.8-14.5	823	5869	2.24E-09
GO Biological Process	enzyme linked receptor protein signaling pathway	2.8-14.5	41	119	2.70E-07

GO Biological Process	phosphorus metabolism	2.8-14.5	137	681	5.24E-07
GO Biological Process	phosphate metabolism	2.8-14.5	137	681	5.24E-07
GO Biological Process	protein amino acid phosphorylation	2.8-14.5	105	491	2.47E-06
GO Biological Process	transmembrane receptor protein tyrosine kinase signaling pathway	2.8-14.5	31	86	1.47E-05
GO Biological Process	phosphorylation	2.8-14.5	105	537	3.59E-04
GO Biological Process	integrin-mediated signaling pathway	2.8-14.5	19	45	7.55E-04
GO Biological Process	cyclic nucleotide biosynthesis	2.8-14.5	13	24	1.71E-03
GO Biological Process	cyclic nucleotide metabolism	2.8-14.5	14	28	2.28E-03
GO Biological Process	protein-nucleus import\, docking	2.8-14.5	11	18	2.49E-03
GO Biological Process	cell-matrix adhesion	2.8-14.5	23	67	3.64E-03
GO Biological Process	protein modification	2.8-14.5	164	975	4.37E-03
GO Cellular Component	extracellular	88.3-100	237	1230	4.83E-20
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	88.3-100	41	69	5.38E-20
GO Cellular Component	ribosome	88.3-100	88	282	5.10E-19
GO Cellular Component	ribonucleoprotein complex	88.3-100	114	441	8.56E-18
GO Cellular Component	nucleosome	88.3-100	39	79	4.89E-15
GO Cellular Component	large ribosomal subunit	88.3-100	31	55	1.08E-13
GO Cellular Component	cytosolic large ribosomal subunit (sensu Eukarya)	88.3-100	25	39	1.66E-12
GO Cellular Component	small ribosomal subunit	88.3-100	25	45	1.88E-10
GO Cellular Component	chromatin	88.3-100	49	171	4.12E-08
GO Cellular Component	eukaryotic 48S initiation complex	88.3-100	16	29	6.86E-06
GO Cellular Component	cytosolic small ribosomal subunit (sensu Eukarya)	88.3-100	16	29	6.86E-06
GO Cellular Component	soluble fraction	88.3-100	55	241	2.70E-05
GO Cellular Component	mitochondrion	88.3-100	118	690	4.34E-05
GO Cellular Component	small nucleolar ribonucleoprotein complex	88.3-100	13	23	1.46E-04
GO Cellular Component	extracellular space	88.3-100	78	410	1.56E-04
GO Cellular Component	obsolete cellular component	88.3-100	82	444	2.73E-04
GO Cellular Component	inner membrane	88.3-100	36	140	4.14E-04
GO Cellular Component	organellar ribosome	88.3-100	13	27	1.71E-03
GO Cellular Component	mitochondrial ribosome	88.3-100	13	27	1.71E-03
GO Cellular Component	eukaryotic 43S preinitiation complex	88.3-100	17	46	3.58E-03
GO Cellular Component	chromosome	88.3-100	55	279	4.41E-03
GO Cellular Component	cytosol	88.3-100	67	388	4.14E-02
GO Cellular Component	ribosome	80.6-92.2	71	282	4.06E-08
GO Cellular Component	ribonucleoprotein complex	80.6-92.2	85	441	5.15E-04
GO Cellular Component	large ribosomal subunit	80.6-92.2	19	55	9.24E-03
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	80.6-92.2	21	69	2.82E-02
GO Cellular Component	mitochondrial ribosome	80.6-92.2	12	27	3.05E-02
GO Cellular Component	organellar ribosome	80.6-92.2	12	27	3.05E-02
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	76.7-100	60	69	1.40E-27
GO Cellular Component	ribosome	76.7-100	147	282	1.20E-26
GO Cellular Component	ribonucleoprotein complex	76.7-100	190	441	1.28E-21
GO Cellular Component	small ribosomal subunit	76.7-100	39	45	5.49E-17
GO Cellular Component	large ribosomal subunit	76.7-100	43	55	1.50E-15
GO Cellular Component	extracellular	76.7-100	391	1230	1.88E-15
GO Cellular Component	cytosolic large ribosomal subunit (sensu Eukarya)	76.7-100	33	39	2.34E-13
GO Cellular Component	cytosolic small ribosomal subunit (sensu Eukarya)	76.7-100	26	29	2.58E-11

GO Cellular Component	eukaryotic 48S initiation complex	76.7-100	26	29	2.58E-11
GO Cellular Component	nucleosome	76.7-100	43	79	5.10E-07
GO Cellular Component	mitochondrion	76.7-100	218	690	5.50E-07
GO Cellular Component	extracellular space	76.7-100	142	410	1.31E-06
GO Cellular Component	eukaryotic 43S preinitiation complex	76.7-100	29	46	4.76E-06
GO Cellular Component	soluble fraction	76.7-100	91	241	1.80E-05
GO Cellular Component	organellar ribosome	76.7-100	20	27	2.62E-05
GO Cellular Component	mitochondrial ribosome	76.7-100	20	27	2.62E-05
GO Cellular Component	cytosol	76.7-100	124	388	2.86E-03
GO Cellular Component	chromatin	76.7-100	64	171	4.64E-03
GO Cellular Component	organellar small ribosomal subunit	76.7-100	10	11	6.16E-03
GO Cellular Component	mitochondrial small ribosomal subunit	76.7-100	10	11	6.16E-03
GO Cellular Component	inner membrane	76.7-100	53	140	2.47E-02
GO Cellular Component	small nucleolar ribonucleoprotein complex	76.7-100	15	23	2.61E-02
GO Cellular Component	proteasome core complex (sensu Eukarya)	72.8-84.4	12	19	2.99E-04
GO Cellular Component	proteasome complex (sensu Eukarya)	72.8-84.4	18	45	2.18E-03
GO Cellular Component	proteasome core complex (sensu Eukarya)	68.9-92.2	17	19	4.19E-06
GO Cellular Component	ribosome	68.9-92.2	99	282	4.15E-03
GO Cellular Component	proteasome complex (sensu Eukarya)	68.9-92.2	24	45	2.73E-02
GO Cellular Component	exosome (RNase complex)	68.9-92.2	9	10	4.57E-02
GO Cellular Component	proteasome complex (sensu Eukarya)	61.1-84.4	25	45	1.11E-02
GO Cellular Component	proteasome core complex (sensu Eukarya)	61.1-84.4	14	19	1.76E-02
GO Cellular Component	microsome	37.8-61.1	61	117	1.80E-07
GO Cellular Component	vesicular fraction	37.8-61.1	61	119	4.50E-07
GO Cellular Component	lysosome	37.8-61.1	52	116	2.84E-03
GO Cellular Component	lytic vacuole	37.8-61.1	52	116	2.84E-03
GO Cellular Component	vacuole	37.8-61.1	54	130	3.00E-02
GO Cellular Component	microsome	33.9-45.6	38	117	8.28E-06
GO Cellular Component	vesicular fraction	33.9-45.6	38	119	1.41E-05
GO Cellular Component	cell	33.9-45.6	1236	10056	1.64E-02
GO Cellular Component	microsome	30.0-53.4	61	117	1.45E-07
GO Cellular Component	cell	30.0-53.4	2494	10056	3.29E-07
GO Cellular Component	vesicular fraction	30.0-53.4	61	119	3.64E-07
GO Cellular Component	lytic vacuole	30.0-53.4	52	116	2.42E-03
GO Cellular Component	lysosome	30.0-53.4	52	116	2.42E-03
GO Cellular Component	intracellular	30.0-53.4	1756	6899	6.05E-03
GO Cellular Component	vacuole	30.0-53.4	55	130	1.09E-02
GO Cellular Component	transcription factor complex	30.0-53.4	197	628	3.31E-02
GO Cellular Component	cell	22.3-45.6	2479	10056	1.23E-08
GO Cellular Component	membrane fraction	22.3-45.6	177	539	2.38E-03
GO Cellular Component	intracellular	22.3-45.6	1745	6899	2.71E-03
GO Cellular Component	transcription factor complex	22.3-45.6	201	628	3.18E-03
GO Cellular Component	nicotinic acetylcholine-gated receptor-channel complex	22.3-45.6	11	12	5.02E-03
GO Cellular Component	microsome	22.3-45.6	49	117	4.08E-02
GO Cellular Component	cytoskeleton	18.4-30.0	148	889	2.14E-02
GO Cellular Component	cytoskeleton	18.4-30.0	148	889	2.14E-02
GO Cellular Component	cell	14.5-37.8	2491	10056	4.52E-11

GO Cellular Component	cytoskeleton	14.5-37.8	272	889	3.04E-03
GO Cellular Component	transcription factor complex	14.5-37.8	197	628	1.96E-02
GO Cellular Component	membrane	10.6-22.3	629	4338	3.15E-07
GO Cellular Component	cell	10.6-22.3	1261	10056	1.35E-04
GO Cellular Component	integral to membrane	10.6-22.3	431	2911	2.13E-04
GO Cellular Component	cytoskeleton	10.6-22.3	157	889	5.24E-04
GO Cellular Component	cell	6.7-30.0	2486	10056	1.05E-07
GO Cellular Component	cytoskeleton	6.7-30.0	295	889	1.27E-07
GO Cellular Component	membrane	6.7-30.0	1175	4338	5.59E-07
GO Cellular Component	integral to membrane	6.7-30.0	812	2911	1.03E-05
GO Cellular Component	integrin complex	6.7-30.0	25	36	3.39E-05
GO Cellular Component	plasma membrane	6.7-30.0	524	1812	1.37E-04
GO Cellular Component	integral to plasma membrane	6.7-30.0	371	1269	6.65E-03
GO Cellular Component	pore complex	6.7-30.0	28	52	1.11E-02
GO Cellular Component	nuclear pore	6.7-30.0	28	52	1.11E-02
GO Cellular Component	extracellular matrix	2.8-14.5	76	302	4.35E-07
GO Cellular Component	collagen	2.8-14.5	18	33	1.18E-05
GO Cellular Component	cytoskeleton	2.8-14.5	162	889	4.71E-05
GO Cellular Component	membrane	2.8-14.5	612	4338	2.68E-04
GO Cellular Component	voltage-gated sodium channel complex	2.8-14.5	10	13	3.29E-04
GO Cellular Component	fibrillar collagen	2.8-14.5	8	9	9.08E-04
GO Cellular Component	integrin complex	2.8-14.5	15	36	1.87E-02
GO Cellular Component	integral to membrane	2.8-14.5	418	2911	2.16E-02
GO Cellular Component	striated muscle thick filament	2.8-14.5	9	15	3.43E-02
GO Molecular Function	structural constituent of ribosome	88.3-100	89	210	7.25E-31
GO Molecular Function	G-protein-coupled receptor binding	88.3-100	40	50	1.58E-27
GO Molecular Function	receptor binding	88.3-100	144	522	8.59E-27
GO Molecular Function	chemokine activity	88.3-100	38	48	1.00E-25
GO Molecular Function	chemokine receptor binding	88.3-100	38	48	1.00E-25
GO Molecular Function	chemoattractant activity	88.3-100	38	50	1.52E-24
GO Molecular Function	hydrogen ion transporter activity	88.3-100	56	113	9.00E-23
GO Molecular Function	hormone activity	88.3-100	51	99	1.47E-21
GO Molecular Function	monovalent inorganic cation transporter activity	88.3-100	56	123	1.95E-20
GO Molecular Function	cytokine activity	88.3-100	73	203	1.20E-19
GO Molecular Function	cation transporter activity	88.3-100	58	200	2.33E-10
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors\, oxygen as acceptor	88.3-100	15	20	3.08E-08
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors	88.3-100	15	20	3.08E-08
GO Molecular Function	cytochrome-c oxidase activity	88.3-100	15	20	3.08E-08
GO Molecular Function	heme-copper terminal oxidase activity	88.3-100	15	20	3.08E-08
GO Molecular Function	NADH dehydrogenase (ubiquinone) activity	88.3-100	20	36	5.18E-08
GO Molecular Function	NADH dehydrogenase activity	88.3-100	20	36	5.18E-08
GO Molecular Function	primary active transporter activity	88.3-100	54	202	5.62E-08
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, quinone or similar compound as acceptor	88.3-100	21	40	6.95E-08
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, other acceptor	88.3-100	20	38	1.94E-07
GO Molecular Function	neuropeptide hormone activity	88.3-100	15	22	2.77E-07

GO Molecular Function	antifungal peptide activity	88.3-100	10	10	2.94E-07
GO Molecular Function	sodium ion transporter activity	88.3-100	20	40	6.51E-07
GO Molecular Function	ion transporter activity	88.3-100	60	261	3.26E-06
GO Molecular Function	enzyme inhibitor activity	88.3-100	50	200	3.89E-06
GO Molecular Function	antimicrobial peptide activity	88.3-100	17	35	2.52E-05
GO Molecular Function	cysteine protease inhibitor activity	88.3-100	13	21	2.89E-05
GO Molecular Function	protein translocase activity	88.3-100	12	18	3.15E-05
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH	88.3-100	23	63	6.01E-05
GO Molecular Function	RNA binding	88.3-100	88	484	1.40E-04
GO Molecular Function	molecular_function unknown	88.3-100	101	584	1.97E-04
GO Molecular Function	small monomeric GTPase activity	88.3-100	34	132	7.25E-04
GO Molecular Function	structural molecule activity	88.3-100	112	711	5.14E-03
GO Molecular Function	hematopoietin/interferon-class (D200-domain) cytokine receptor binding	88.3-100	16	43	5.90E-03
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	88.3-100	41	190	7.36E-03
GO Molecular Function	isoprenoid binding	88.3-100	7	10	2.46E-02
GO Molecular Function	retinoid binding	88.3-100	7	10	2.46E-02
GO Molecular Function	acute-phase response protein activity	88.3-100	5	5	2.58E-02
GO Molecular Function	GTPase activity	88.3-100	43	217	4.16E-02
GO Molecular Function	metal ion transporter activity	88.3-100	22	82	4.26E-02
GO Molecular Function	small monomeric GTPase activity	80.6-92.2	62	132	5.93E-22
GO Molecular Function	structural constituent of ribosome	80.6-92.2	72	210	5.01E-16
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	80.6-92.2	63	190	6.17E-13
GO Molecular Function	GTPase activity	80.6-92.2	66	217	1.49E-11
GO Molecular Function	RAB small monomeric GTPase activity	80.6-92.2	29	53	2.21E-11
GO Molecular Function	GTP binding	80.6-92.2	65	244	2.04E-08
GO Molecular Function	guanyl nucleotide binding	80.6-92.2	65	251	7.87E-08
GO Molecular Function	growth factor activity	80.6-92.2	46	151	1.91E-07
GO Molecular Function	receptor binding	80.6-92.2	105	522	1.47E-06
GO Molecular Function	cytokine activity	80.6-92.2	50	203	8.97E-05
GO Molecular Function	Rho small monomeric GTPase activity	80.6-92.2	14	27	5.15E-04
GO Molecular Function	structural constituent of eye lens	80.6-92.2	12	24	6.46E-03
GO Molecular Function	RAS small monomeric GTPase activity	80.6-92.2	13	29	1.16E-02
GO Molecular Function	structural molecule activity	80.6-92.2	117	711	1.50E-02
GO Molecular Function	protein transporter activity	80.6-92.2	53	262	2.48E-02
GO Molecular Function	hydrolase activity\, acting on acid anhydrides	80.6-92.2	86	492	2.76E-02
GO Molecular Function	structural constituent of ribosome	76.7-100	145	210	2.02E-46
GO Molecular Function	receptor binding	76.7-100	246	522	1.10E-36
GO Molecular Function	small monomeric GTPase activity	76.7-100	94	132	6.64E-31
GO Molecular Function	cytokine activity	76.7-100	119	203	2.90E-27
GO Molecular Function	hormone activity	76.7-100	67	99	2.35E-19
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	76.7-100	102	190	6.02E-19
GO Molecular Function	G-protein-coupled receptor binding	76.7-100	42	50	2.34E-17
GO Molecular Function	chemokine activity	76.7-100	40	48	3.62E-16

GO Molecular Function	chemokine receptor binding	76.7-100	40	48	3.62E-16
GO Molecular Function	RAB small monomeric GTPase activity	76.7-100	42	53	1.65E-15
GO Molecular Function	hydrogen ion transporter activity	76.7-100	68	113	1.84E-15
GO Molecular Function	GTPase activity	76.7-100	105	217	3.87E-15
GO Molecular Function	chemoattractant activity	76.7-100	40	50	6.17E-15
GO Molecular Function	monovalent inorganic cation transporter activity	76.7-100	68	123	9.08E-13
GO Molecular Function	growth factor activity	76.7-100	75	151	6.05E-11
GO Molecular Function	GTP binding	76.7-100	100	244	1.41E-08
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, other acceptor	76.7-100	28	38	3.16E-08
GO Molecular Function	glutathione transferase activity	76.7-100	19	21	9.32E-08
GO Molecular Function	guanyl nucleotide binding	76.7-100	100	251	1.02E-07
GO Molecular Function	NADH dehydrogenase (ubiquinone) activity	76.7-100	26	36	3.72E-07
GO Molecular Function	NADH dehydrogenase activity	76.7-100	26	36	3.72E-07
GO Molecular Function	neuropeptide hormone activity	76.7-100	19	22	5.44E-07
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH\, quinone or similar compound as acceptor	76.7-100	27	40	1.88E-06
GO Molecular Function	structural molecule activity	76.7-100	221	711	2.53E-06
GO Molecular Function	hematopoietin/interferon-class (D200-domain) cytokine receptor binding	76.7-100	28	43	3.19E-06
GO Molecular Function	protein transporter activity	76.7-100	99	262	4.04E-06
GO Molecular Function	oxidoreductase activity\, acting on NADH or NADPH	76.7-100	35	63	1.31E-05
GO Molecular Function	sodium ion transporter activity	76.7-100	26	40	1.35E-05
GO Molecular Function	Rho small monomeric GTPase activity	76.7-100	20	27	2.70E-05
GO Molecular Function	cytochrome-c oxidase activity	76.7-100	16	20	1.38E-04
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors	76.7-100	16	20	1.38E-04
GO Molecular Function	oxidoreductase activity\, acting on heme group of donors\, oxygen as acceptor	76.7-100	16	20	1.38E-04
GO Molecular Function	heme-copper terminal oxidase activity	76.7-100	16	20	1.38E-04
GO Molecular Function	cation transporter activity	76.7-100	76	200	2.56E-04
GO Molecular Function	enzyme inhibitor activity	76.7-100	76	200	2.56E-04
GO Molecular Function	molecular_function unknown	76.7-100	178	584	5.71E-04
GO Molecular Function	antifungal peptide activity	76.7-100	10	10	7.07E-04
GO Molecular Function	structural constituent of eye lens	76.7-100	17	24	1.07E-03
GO Molecular Function	small protein conjugating enzyme activity	76.7-100	30	58	1.38E-03
GO Molecular Function	RAS small monomeric GTPase activity	76.7-100	19	29	1.44E-03
GO Molecular Function	protein translocase activity	76.7-100	14	18	1.90E-03
GO Molecular Function	ubiquitin conjugating enzyme activity	76.7-100	29	56	2.08E-03
GO Molecular Function	cysteine protease inhibitor activity	76.7-100	15	21	4.61E-03
GO Molecular Function	RNA binding	76.7-100	147	484	8.61E-03
GO Molecular Function	electron transporter activity	76.7-100	99	305	1.84E-02
GO Molecular Function	primary active transporter activity	76.7-100	71	202	1.90E-02
GO Molecular Function	protein kinase inhibitor activity	76.7-100	15	23	2.67E-02
GO Molecular Function	small monomeric GTPase activity	72.8-84.4	39	132	3.88E-05
GO Molecular Function	receptor binding	72.8-84.4	99	522	6.10E-04
GO Molecular Function	growth factor activity	72.8-84.4	40	151	6.90E-04
GO Molecular Function	tumor necrosis factor receptor binding	72.8-84.4	11	19	3.91E-03
GO Molecular Function	nucleobase\, nucleoside\, nucleotide kinase activity	72.8-84.4	17	44	8.81E-03
GO Molecular Function	nucleoside kinase activity	72.8-84.4	8	11	1.00E-02

GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	72.8-84.4	43	190	2.34E-02
GO Molecular Function	small monomeric GTPase activity	68.9-92.2	92	132	3.76E-27
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	68.9-92.2	96	190	1.15E-13
GO Molecular Function	GTP binding	68.9-92.2	109	244	5.62E-11
GO Molecular Function	GTPase activity	68.9-92.2	100	217	6.11E-11
GO Molecular Function	RAB small monomeric GTPase activity	68.9-92.2	38	53	1.61E-10
GO Molecular Function	guanyl nucleotide binding	68.9-92.2	109	251	5.71E-10
GO Molecular Function	structural constituent of ribosome	68.9-92.2	95	210	1.11E-09
GO Molecular Function	growth factor activity	68.9-92.2	73	151	1.22E-08
GO Molecular Function	receptor binding	68.9-92.2	185	522	3.70E-08
GO Molecular Function	Rho small monomeric GTPase activity	68.9-92.2	21	27	7.22E-06
GO Molecular Function	structural constituent of eye lens	68.9-92.2	19	24	2.60E-05
GO Molecular Function	RAS small monomeric GTPase activity	68.9-92.2	20	29	5.26E-04
GO Molecular Function	cytokine activity	68.9-92.2	78	203	1.03E-03
GO Molecular Function	glutathione transferase activity	68.9-92.2	15	21	1.02E-02
GO Molecular Function	apoptosis regulator activity	68.9-92.2	46	113	4.89E-02
GO Molecular Function	rhodopsin-like receptor activity	65.0-76.7	70	331	1.51E-03
GO Molecular Function	rhodopsin-like receptor activity	61.1-84.4	126	331	2.66E-06
GO Molecular Function	nucleotide receptor activity\, G-protein coupled	61.1-84.4	21	29	1.08E-04
GO Molecular Function	purinergic nucleotide receptor activity	61.1-84.4	21	29	1.08E-04
GO Molecular Function	nucleotide receptor activity	61.1-84.4	21	29	1.08E-04
GO Molecular Function	purinergic nucleotide receptor activity\, G-protein coupled	61.1-84.4	21	29	1.08E-04
GO Molecular Function	tumor necrosis factor receptor binding	61.1-84.4	15	19	1.74E-03
GO Molecular Function	MHC class II receptor activity	61.1-84.4	13	16	6.03E-03
GO Molecular Function	growth factor activity	61.1-84.4	60	151	1.61E-02
GO Molecular Function	small monomeric GTPase activity	61.1-84.4	53	132	4.04E-02
GO Molecular Function	rhodopsin-like receptor activity	57.2-68.9	90	331	3.42E-11
GO Molecular Function	G-protein coupled receptor activity	57.2-68.9	96	410	7.10E-08
GO Molecular Function	purinergic nucleotide receptor activity	57.2-68.9	15	29	5.73E-04
GO Molecular Function	nucleotide receptor activity	57.2-68.9	15	29	5.73E-04
GO Molecular Function	nucleotide receptor activity\, G-protein coupled	57.2-68.9	15	29	5.73E-04
GO Molecular Function	purinergic nucleotide receptor activity\, G-protein coupled	57.2-68.9	15	29	5.73E-04
GO Molecular Function	peptide receptor activity	57.2-68.9	33	109	6.50E-04
GO Molecular Function	peptide receptor activity\, G-protein coupled	57.2-68.9	33	109	6.50E-04
GO Molecular Function	rhodopsin-like receptor activity	53.4-76.7	163	331	5.73E-21
GO Molecular Function	G-protein coupled receptor activity	53.4-76.7	173	410	1.52E-13
GO Molecular Function	purinergic nucleotide receptor activity\, G-protein coupled	53.4-76.7	21	29	1.63E-04
GO Molecular Function	nucleotide receptor activity	53.4-76.7	21	29	1.63E-04
GO Molecular Function	purinergic nucleotide receptor activity	53.4-76.7	21	29	1.63E-04
GO Molecular Function	nucleotide receptor activity\, G-protein coupled	53.4-76.7	21	29	1.63E-04
GO Molecular Function	peptide receptor activity	53.4-76.7	51	109	4.48E-04
GO Molecular Function	peptide receptor activity\, G-protein coupled	53.4-76.7	51	109	4.48E-04
GO Molecular Function	transmembrane receptor activity	53.4-76.7	266	852	5.95E-04
GO Molecular Function	oxidoreductase activity\, acting on CH-OH group of	53.4-76.7	46	96	7.73E-04

	donors				
GO Molecular Function	oxidoreductase activity\, acting on the CH-OH group of donors\, NAD or NADP as acceptor	53.4-76.7	44	92	1.52E-03
GO Molecular Function	S-adenosylmethionine-dependent methyltransferase activity	53.4-76.7	38	78	5.22E-03
GO Molecular Function	peptide binding	53.4-76.7	58	143	2.20E-02
GO Molecular Function	rhodopsin-like receptor activity	49.5-61.1	98	331	9.76E-15
GO Molecular Function	G-protein coupled receptor activity	49.5-61.1	112	410	2.92E-14
GO Molecular Function	transmembrane receptor activity	49.5-61.1	171	852	6.92E-09
GO Molecular Function	peptide receptor activity\, G-protein coupled	49.5-61.1	39	109	3.22E-07
GO Molecular Function	peptide receptor activity	49.5-61.1	39	109	3.22E-07
GO Molecular Function	receptor activity	49.5-61.1	228	1306	2.36E-06
GO Molecular Function	peptide binding	49.5-61.1	42	143	5.90E-05
GO Molecular Function	oxidoreductase activity\, acting on CH-OH group of donors	49.5-61.1	28	96	1.55E-02
GO Molecular Function	rhodopsin-like receptor activity	45.6-68.9	182	331	3.11E-31
GO Molecular Function	G-protein coupled receptor activity	45.6-68.9	199	410	1.05E-24
GO Molecular Function	peptide receptor activity\, G-protein coupled	45.6-68.9	72	109	4.00E-17
GO Molecular Function	peptide receptor activity	45.6-68.9	72	109	4.00E-17
GO Molecular Function	transmembrane receptor activity	45.6-68.9	308	852	2.67E-13
GO Molecular Function	peptide binding	45.6-68.9	77	143	3.89E-11
GO Molecular Function	receptor activity	45.6-68.9	422	1306	7.22E-10
GO Molecular Function	MHC class I receptor activity	45.6-68.9	17	21	2.30E-04
GO Molecular Function	transcription factor activity	45.6-68.9	255	809	8.96E-04
GO Molecular Function	amine receptor activity	45.6-68.9	24	40	4.66E-03
GO Molecular Function	transcription regulator activity	45.6-68.9	328	1117	2.87E-02
GO Molecular Function	transferase activity	45.6-68.9	382	1327	3.75E-02
GO Molecular Function	monooxygenase activity	41.7-53.4	29	72	2.47E-06
GO Molecular Function	peptide receptor activity	41.7-53.4	34	109	1.97E-04
GO Molecular Function	peptide receptor activity\, G-protein coupled	41.7-53.4	34	109	1.97E-04
GO Molecular Function	rhodopsin-like receptor activity	41.7-53.4	73	331	2.72E-04
GO Molecular Function	receptor activity	41.7-53.4	212	1306	1.48E-03
GO Molecular Function	oxygen binding	41.7-53.4	13	24	1.97E-03
GO Molecular Function	G-protein coupled receptor activity	41.7-53.4	82	410	3.52E-03
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	41.7-53.4	27	87	4.92E-03
GO Molecular Function	transmembrane receptor activity	41.7-53.4	146	852	5.77E-03
GO Molecular Function	peptide binding	41.7-53.4	37	143	9.32E-03
GO Molecular Function	neurotransmitter receptor activity	41.7-53.4	18	50	2.69E-02
GO Molecular Function	neurotransmitter binding	41.7-53.4	18	51	3.70E-02
GO Molecular Function	rhodopsin-like receptor activity	37.8-61.1	152	331	2.19E-15
GO Molecular Function	G-protein coupled receptor activity	37.8-61.1	168	410	2.10E-11
GO Molecular Function	monooxygenase activity	37.8-61.1	48	72	6.30E-11
GO Molecular Function	peptide receptor activity	37.8-61.1	61	109	2.25E-09
GO Molecular Function	peptide receptor activity\, G-protein coupled	37.8-61.1	61	109	2.25E-09
GO Molecular Function	neurotransmitter receptor activity	37.8-61.1	35	50	2.69E-08
GO Molecular Function	neurotransmitter binding	37.8-61.1	35	51	6.59E-08
GO Molecular Function	transmembrane receptor activity	37.8-61.1	286	852	8.60E-08
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	37.8-61.1	49	87	3.15E-07

GO Molecular Function	GABA-A receptor activity	37.8-61.1	20	23	1.15E-06
GO Molecular Function	peptide binding	37.8-61.1	68	143	2.00E-06
GO Molecular Function	receptor activity	37.8-61.1	402	1306	3.96E-06
GO Molecular Function	GABA receptor activity	37.8-61.1	20	25	2.07E-05
GO Molecular Function	oxygen binding	37.8-61.1	19	24	7.00E-05
GO Molecular Function	amine receptor activity	37.8-61.1	26	40	1.57E-04
GO Molecular Function	catalytic activity	37.8-61.1	1123	4200	1.83E-04
GO Molecular Function	steroid hormone receptor activity	37.8-61.1	30	51	3.92E-04
GO Molecular Function	transferase activity	37.8-61.1	394	1327	6.52E-04
GO Molecular Function	transcription factor activity	37.8-61.1	255	809	7.71E-04
GO Molecular Function	neuropeptide binding	37.8-61.1	16	20	7.71E-04
GO Molecular Function	neuropeptide receptor activity	37.8-61.1	16	20	7.71E-04
GO Molecular Function	ligand-dependent nuclear receptor activity	37.8-61.1	31	57	2.70E-03
GO Molecular Function	extracellular ligand-gated ion channel activity	37.8-61.1	32	60	3.21E-03
GO Molecular Function	steroid hydroxylase activity	37.8-61.1	11	12	5.35E-03
GO Molecular Function	unspecific monooxygenase activity	37.8-61.1	17	24	6.11E-03
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen\, reduced flavin or flavoprotein as one donor\, and incorporation of one atom of oxygen	37.8-61.1	17	24	6.11E-03
GO Molecular Function	neuropeptide Y receptor activity	37.8-61.1	11	13	2.72E-02
GO Molecular Function	monooxygenase activity	33.9-45.6	36	72	4.65E-12
GO Molecular Function	oxygen binding	33.9-45.6	18	24	3.54E-09
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	33.9-45.6	36	87	6.47E-09
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen\, reduced flavin or flavoprotein as one donor\, and incorporation of one atom of oxygen	33.9-45.6	15	24	1.54E-05
GO Molecular Function	unspecific monooxygenase activity	33.9-45.6	15	24	1.54E-05
GO Molecular Function	catalytic activity	33.9-45.6	569	4200	1.96E-02
GO Molecular Function	steroid hydroxylase activity	33.9-45.6	8	12	3.52E-02
GO Molecular Function	monooxygenase activity	30.0-53.4	56	72	2.26E-18
GO Molecular Function	oxygen binding	30.0-53.4	24	24	4.26E-12
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	30.0-53.4	54	87	1.21E-10
GO Molecular Function	unspecific monooxygenase activity	30.0-53.4	22	24	1.25E-08
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen\, reduced flavin or flavoprotein as one donor\, and incorporation of one atom of oxygen	30.0-53.4	22	24	1.25E-08
GO Molecular Function	transcription factor activity	30.0-53.4	259	809	1.08E-04
GO Molecular Function	neurotransmitter binding	30.0-53.4	30	51	3.72E-04
GO Molecular Function	acetylcholine binding	30.0-53.4	16	20	7.47E-04
GO Molecular Function	neurotransmitter receptor activity	30.0-53.4	29	50	9.31E-04
GO Molecular Function	steroid hormone receptor activity	30.0-53.4	29	51	1.67E-03
GO Molecular Function	acetylcholine receptor activity	30.0-53.4	15	19	2.53E-03
GO Molecular Function	amine receptor activity	30.0-53.4	24	40	4.37E-03
GO Molecular Function	steroid hydroxylase activity	30.0-53.4	11	12	5.23E-03
GO Molecular Function	ligand-dependent nuclear receptor activity	30.0-53.4	30	57	9.71E-03
GO Molecular Function	transcription regulator activity	30.0-53.4	326	1117	3.82E-02
GO Molecular Function	monooxygenase activity	22.3-45.6	44	72	6.84E-08

GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	22.3-45.6	49	87	3.16E-07
GO Molecular Function	porter activity	22.3-45.6	79	184	3.24E-05
GO Molecular Function	electrochemical potential-driven transporter activity	22.3-45.6	79	185	4.34E-05
GO Molecular Function	oxygen binding	22.3-45.6	19	24	7.02E-05
GO Molecular Function	organic cation transporter activity	22.3-45.6	17	22	7.74E-04
GO Molecular Function	catalytic activity	22.3-45.6	1108	4200	6.13E-03
GO Molecular Function	unspecific monooxygenase activity	22.3-45.6	17	24	6.13E-03
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen\, reduced flavin or flavoprotein as one donor\, and incorporation of one atom of oxygen	22.3-45.6	17	24	6.13E-03
GO Molecular Function	ATP binding	22.3-45.6	315	1059	1.15E-02
GO Molecular Function	adenyl nucleotide binding	22.3-45.6	318	1071	1.22E-02
GO Molecular Function	adenyl nucleotide binding	18.4-30.0	194	1071	3.36E-06
GO Molecular Function	porter activity	18.4-30.0	51	184	1.81E-05
GO Molecular Function	ATP binding	18.4-30.0	189	1059	1.92E-05
GO Molecular Function	electrochemical potential-driven transporter activity	18.4-30.0	51	185	2.21E-05
GO Molecular Function	symporter activity	18.4-30.0	29	78	3.07E-05
GO Molecular Function	solute\:cation symporter activity	18.4-30.0	19	43	4.41E-04
GO Molecular Function	neurotransmitter transporter activity	18.4-30.0	11	18	3.26E-03
GO Molecular Function	solute\:sodium symporter activity	18.4-30.0	15	32	3.61E-03
GO Molecular Function	neurotransmitter\:sodium symporter activity	18.4-30.0	10	16	7.71E-03
GO Molecular Function	purine nucleotide binding	18.4-30.0	210	1303	8.89E-03
GO Molecular Function	nucleotide binding	18.4-30.0	210	1317	1.88E-02
GO Molecular Function	amine/polyamine transporter activity	18.4-30.0	20	60	4.04E-02
GO Molecular Function	adenyl nucleotide binding	18.4-30.0	194	1071	3.36E-06
GO Molecular Function	porter activity	18.4-30.0	51	184	1.81E-05
GO Molecular Function	ATP binding	18.4-30.0	189	1059	1.92E-05
GO Molecular Function	electrochemical potential-driven transporter activity	18.4-30.0	51	185	2.21E-05
GO Molecular Function	symporter activity	18.4-30.0	29	78	3.07E-05
GO Molecular Function	solute\:cation symporter activity	18.4-30.0	19	43	4.41E-04
GO Molecular Function	neurotransmitter transporter activity	18.4-30.0	11	18	3.26E-03
GO Molecular Function	solute\:sodium symporter activity	18.4-30.0	15	32	3.61E-03
GO Molecular Function	neurotransmitter\:sodium symporter activity	18.4-30.0	10	16	7.71E-03
GO Molecular Function	purine nucleotide binding	18.4-30.0	210	1303	8.89E-03
GO Molecular Function	nucleotide binding	18.4-30.0	210	1317	1.88E-02
GO Molecular Function	amine/polyamine transporter activity	18.4-30.0	20	60	4.04E-02
GO Molecular Function	porter activity	14.5-37.8	95	184	1.50E-12
GO Molecular Function	electrochemical potential-driven transporter activity	14.5-37.8	95	185	2.37E-12
GO Molecular Function	adenyl nucleotide binding	14.5-37.8	364	1071	3.28E-11
GO Molecular Function	ATP binding	14.5-37.8	358	1059	1.36E-10
GO Molecular Function	solute\:cation symporter activity	14.5-37.8	31	43	1.50E-07
GO Molecular Function	solute\:sodium symporter activity	14.5-37.8	25	32	6.69E-07
GO Molecular Function	symporter activity	14.5-37.8	45	78	7.69E-07
GO Molecular Function	metallopeptidase activity	14.5-37.8	76	169	6.67E-06
GO Molecular Function	neurotransmitter\:sodium symporter activity	14.5-37.8	15	16	2.48E-05
GO Molecular Function	neurotransmitter transporter activity	14.5-37.8	16	18	4.40E-05
GO Molecular Function	purine nucleotide binding	14.5-37.8	395	1303	9.77E-05

GO Molecular Function	amine/polyamine transporter activity	14.5-37.8	34	60	2.34E-04
GO Molecular Function	nucleotide binding	14.5-37.8	396	1317	2.64E-04
GO Molecular Function	metalloendopeptidase activity	14.5-37.8	45	93	1.06E-03
GO Molecular Function	peptidase activity	14.5-37.8	171	508	1.46E-03
GO Molecular Function	protein serine/threonine kinase activity	14.5-37.8	126	355	2.37E-03
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	14.5-37.8	195	598	2.81E-03
GO Molecular Function	protein kinase activity	14.5-37.8	168	508	6.48E-03
GO Molecular Function	organic acid transporter activity	14.5-37.8	34	67	7.66E-03
GO Molecular Function	carboxylic acid transporter activity	14.5-37.8	34	67	7.66E-03
GO Molecular Function	metal ion binding	14.5-37.8	337	1135	1.08E-02
GO Molecular Function	ATPase activity\, coupled	14.5-37.8	88	236	1.24E-02
GO Molecular Function	DNA binding	14.5-37.8	569	2039	2.18E-02
GO Molecular Function	amino acid transporter activity	14.5-37.8	26	48	2.63E-02
GO Molecular Function	ATP dependent helicase activity	14.5-37.8	45	103	3.46E-02
GO Molecular Function	endopeptidase activity	14.5-37.8	91	252	3.89E-02
GO Molecular Function	cell adhesion molecule activity	10.6-22.3	118	374	4.35E-21
GO Molecular Function	calcium-dependent cell adhesion molecule activity	10.6-22.3	51	94	8.43E-20
GO Molecular Function	adenyl nucleotide binding	10.6-22.3	216	1071	1.53E-12
GO Molecular Function	ATP binding	10.6-22.3	213	1059	3.48E-12
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	10.6-22.3	31	64	1.90E-09
GO Molecular Function	transmembrane receptor protein kinase activity	10.6-22.3	33	76	1.58E-08
GO Molecular Function	metal ion binding	10.6-22.3	210	1135	3.65E-08
GO Molecular Function	purine nucleotide binding	10.6-22.3	233	1303	7.82E-08
GO Molecular Function	nucleotide binding	10.6-22.3	234	1317	1.35E-07
GO Molecular Function	calcium ion binding	10.6-22.3	118	576	4.78E-06
GO Molecular Function	protein kinase activity	10.6-22.3	107	508	5.79E-06
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	10.6-22.3	119	598	2.44E-05
GO Molecular Function	metallopeptidase activity	10.6-22.3	47	169	4.59E-05
GO Molecular Function	protein-tyrosine kinase activity	10.6-22.3	60	241	4.97E-05
GO Molecular Function	metalloendopeptidase activity	10.6-22.3	31	93	1.45E-04
GO Molecular Function	glutamate receptor activity	10.6-22.3	17	34	1.72E-04
GO Molecular Function	small GTPase regulatory/interacting protein activity	10.6-22.3	44	163	3.38E-04
GO Molecular Function	peptidase activity	10.6-22.3	100	508	6.85E-04
GO Molecular Function	ATPase activity\, coupled	10.6-22.3	55	236	1.98E-03
GO Molecular Function	GTPase regulator activity	10.6-22.3	54	231	2.25E-03
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	10.6-22.3	33	119	6.53E-03
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	10.6-22.3	33	119	6.53E-03
GO Molecular Function	aminophospholipid transporter activity	10.6-22.3	7	8	6.86E-03
GO Molecular Function	kainate selective glutamate receptor activity	10.6-22.3	7	8	6.86E-03
GO Molecular Function	phospholipid-translocating ATPase activity	10.6-22.3	7	8	6.86E-03
GO Molecular Function	kinase activity	10.6-22.3	128	725	7.59E-03
GO Molecular Function	ATPase activity	10.6-22.3	55	249	1.15E-02
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	10.6-22.3	58	275	2.87E-02
GO Molecular Function	ephrin receptor activity	10.6-22.3	9	15	3.47E-02

GO Molecular Function	calcium-dependent cell adhesion molecule activity	6.7-30.0	85	94	3.67E-39
GO Molecular Function	cell adhesion molecule activity	6.7-30.0	208	374	7.54E-37
GO Molecular Function	adenyl nucleotide binding	6.7-30.0	434	1071	1.04E-32
GO Molecular Function	ATP binding	6.7-30.0	427	1059	1.76E-31
GO Molecular Function	purine nucleotide binding	6.7-30.0	462	1303	2.09E-19
GO Molecular Function	nucleotide binding	6.7-30.0	465	1317	3.98E-19
GO Molecular Function	metal ion binding	6.7-30.0	408	1135	7.01E-18
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	6.7-30.0	49	64	4.03E-15
GO Molecular Function	calcium ion binding	6.7-30.0	227	576	1.37E-13
GO Molecular Function	ATPase activity\, coupled	6.7-30.0	115	236	3.29E-13
GO Molecular Function	protein kinase activity	6.7-30.0	202	508	2.84E-12
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	6.7-30.0	229	598	4.72E-12
GO Molecular Function	transmembrane receptor protein kinase activity	6.7-30.0	51	76	7.90E-12
GO Molecular Function	ATPase activity	6.7-30.0	116	249	1.55E-11
GO Molecular Function	small GTPase regulatory/interacting protein activity	6.7-30.0	82	163	1.00E-09
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	6.7-30.0	120	275	1.71E-09
GO Molecular Function	GTPase regulator activity	6.7-30.0	105	231	2.26E-09
GO Molecular Function	protein-tyrosine kinase activity	6.7-30.0	107	241	8.59E-09
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	6.7-30.0	64	119	9.06E-09
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	6.7-30.0	64	119	9.06E-09
GO Molecular Function	kinase activity	6.7-30.0	253	725	3.05E-08
GO Molecular Function	electrochemical potential-driven transporter activity	6.7-30.0	84	185	5.03E-07
GO Molecular Function	porter activity	6.7-30.0	83	184	9.69E-07
GO Molecular Function	metallopeptidase activity	6.7-30.0	77	169	2.42E-06
GO Molecular Function	guanyl-nucleotide exchange factor activity	6.7-30.0	50	95	6.11E-06
GO Molecular Function	symporter activity	6.7-30.0	43	78	1.31E-05
GO Molecular Function	P-P-bond-hydrolysis-driven transporter activity	6.7-30.0	64	137	1.97E-05
GO Molecular Function	metalloendopeptidase activity	6.7-30.0	48	93	2.99E-05
GO Molecular Function	transferase activity\, transferring phosphorus-containing groups	6.7-30.0	247	750	3.11E-05
GO Molecular Function	peptidase activity	6.7-30.0	178	508	3.26E-05
GO Molecular Function	carrier activity	6.7-30.0	152	421	4.44E-05
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of ions\, phosphorylative mechanism	6.7-30.0	28	44	1.03E-04
GO Molecular Function	cell adhesion receptor activity	6.7-30.0	27	42	1.34E-04
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of ions	6.7-30.0	28	46	4.16E-04
GO Molecular Function	hydrolase activity	6.7-30.0	495	1699	4.42E-04
GO Molecular Function	neurotransmitter\:sodium symporter activity	6.7-30.0	14	16	5.82E-04
GO Molecular Function	neurotransmitter transporter activity	6.7-30.0	15	18	7.34E-04
GO Molecular Function	amine/polyamine transporter activity	6.7-30.0	33	60	9.30E-04
GO Molecular Function	ATP-binding cassette (ABC) transporter activity	6.7-30.0	38	74	1.27E-03
GO Molecular Function	magnesium ion binding	6.7-30.0	58	133	1.94E-03
GO Molecular Function	glutamate receptor activity	6.7-30.0	22	34	2.14E-03
GO Molecular Function	binding	6.7-30.0	1708	6657	5.75E-03
GO Molecular Function	inorganic anion transporter activity	6.7-30.0	16	22	7.47E-03

GO Molecular Function	calcium-transporting ATPase activity	6.7-30.0	9	9	1.01E-02
GO Molecular Function	protein serine/threonine kinase activity	6.7-30.0	123	355	1.28E-02
GO Molecular Function	calcium ion transporter activity	6.7-30.0	10	11	2.09E-02
GO Molecular Function	ubiquitin thiolesterase activity	6.7-30.0	25	45	2.09E-02
GO Molecular Function	amino acid transporter activity	6.7-30.0	26	48	2.54E-02
GO Molecular Function	solute\:cation symporter activity	6.7-30.0	24	43	2.88E-02
GO Molecular Function	guanylate cyclase activity	6.7-30.0	13	17	2.92E-02
GO Molecular Function	ephrin receptor activity	6.7-30.0	12	15	3.00E-02
GO Molecular Function	ATP dependent helicase activity	6.7-30.0	45	103	3.36E-02
GO Molecular Function	kainate selective glutamate receptor activity	6.7-30.0	8	8	4.20E-02
GO Molecular Function	cell adhesion molecule activity	2.8-14.5	142	374	5.95E-36
GO Molecular Function	adenyl nucleotide binding	2.8-14.5	254	1071	5.52E-26
GO Molecular Function	ATP binding	2.8-14.5	252	1059	5.54E-26
GO Molecular Function	calcium-dependent cell adhesion molecule activity	2.8-14.5	57	94	8.19E-26
GO Molecular Function	nucleotide binding	2.8-14.5	266	1317	3.07E-16
GO Molecular Function	purine nucleotide binding	2.8-14.5	263	1303	5.76E-16
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	2.8-14.5	36	64	5.18E-14
GO Molecular Function	metal ion binding	2.8-14.5	230	1135	1.13E-13
GO Molecular Function	ATPase activity	2.8-14.5	79	249	2.14E-13
GO Molecular Function	GTPase regulator activity	2.8-14.5	75	231	3.17E-13
GO Molecular Function	ATPase activity\, coupled	2.8-14.5	76	236	3.30E-13
GO Molecular Function	calcium ion binding	2.8-14.5	138	576	6.69E-13
GO Molecular Function	transmembrane receptor protein kinase activity	2.8-14.5	36	76	6.69E-11
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	2.8-14.5	79	275	1.02E-10
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	2.8-14.5	46	119	2.18E-10
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	2.8-14.5	46	119	2.18E-10
GO Molecular Function	small GTPase regulatory/interacting protein activity	2.8-14.5	54	163	2.53E-09
GO Molecular Function	P-P-bond-hydrolysis-driven transporter activity	2.8-14.5	46	137	7.26E-08
GO Molecular Function	protein kinase activity	2.8-14.5	112	508	1.56E-07
GO Molecular Function	guanyl-nucleotide exchange factor activity	2.8-14.5	35	95	9.69E-07
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	2.8-14.5	123	598	1.92E-06
GO Molecular Function	binding	2.8-14.5	905	6657	2.38E-06
GO Molecular Function	guanylate cyclase activity	2.8-14.5	13	17	4.07E-06
GO Molecular Function	magnesium ion binding	2.8-14.5	42	133	4.48E-06
GO Molecular Function	phosphorus-oxygen lyase activity	2.8-14.5	14	20	6.24E-06
GO Molecular Function	motor activity	2.8-14.5	37	113	1.37E-05
GO Molecular Function	GTPase activator activity	2.8-14.5	35	108	4.78E-05
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of ions\, phosphorylative mechanism	2.8-14.5	20	44	9.30E-05
GO Molecular Function	adenylate cyclase activity	2.8-14.5	9	10	1.19E-04
GO Molecular Function	ATP-binding cassette (ABC) transporter activity	2.8-14.5	27	74	1.25E-04
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of ions	2.8-14.5	20	46	2.33E-04
GO Molecular Function	transmembrane receptor protein tyrosine phosphatase activity	2.8-14.5	12	19	5.09E-04
GO Molecular Function	transmembrane receptor protein phosphatase activity	2.8-14.5	12	19	5.09E-04

GO Molecular Function	protein-tyrosine kinase activity	2.8-14.5	57	241	7.87E-04
GO Molecular Function	enzyme regulator activity	2.8-14.5	106	549	9.10E-04
GO Molecular Function	epidermal growth factor receptor activity	2.8-14.5	7	7	9.26E-04
GO Molecular Function	extracellular matrix structural constituent	2.8-14.5	28	89	2.46E-03
GO Molecular Function	transferase activity\, transferring phosphorus-containing groups	2.8-14.5	134	750	2.58E-03
GO Molecular Function	calmodulin binding	2.8-14.5	31	105	3.10E-03
GO Molecular Function	kinase activity	2.8-14.5	130	725	3.11E-03
GO Molecular Function	chloride transporter activity	2.8-14.5	8	10	3.99E-03
GO Molecular Function	ATP dependent helicase activity	2.8-14.5	30	103	6.26E-03
GO Molecular Function	anion exchanger activity	2.8-14.5	7	8	6.63E-03
GO Molecular Function	inorganic anion exchanger activity	2.8-14.5	7	8	6.63E-03
GO Molecular Function	bicarbonate transporter activity	2.8-14.5	7	8	6.63E-03
GO Molecular Function	cation\:chloride symporter activity	2.8-14.5	7	8	6.63E-03
GO Molecular Function	helicase activity	2.8-14.5	34	125	7.60E-03
GO Molecular Function	primary active transporter activity	2.8-14.5	47	202	1.36E-02
GO Molecular Function	voltage-gated sodium channel activity	2.8-14.5	10	17	1.39E-02
GO Molecular Function	glucosidase activity	2.8-14.5	7	9	2.67E-02
GO Molecular Function	anion\:anion antiporter activity	2.8-14.5	7	9	2.67E-02
GO Molecular Function	calcium-transporting ATPase activity	2.8-14.5	7	9	2.67E-02
GO Molecular Function	enzyme activator activity	2.8-14.5	43	184	3.01E-02
GO Molecular Function	ephrin receptor activity	2.8-14.5	9	15	3.37E-02
GO Molecular Function	cell adhesion receptor activity	2.8-14.5	16	42	3.57E-02
SwissProt keyword	Ribosomal protein	88.3-100	46	88	2.98E-19
SwissProt keyword	Nucleosome core	88.3-100	26	30	1.23E-18
SwissProt keyword	Cytokine	88.3-100	59	146	4.99E-18
SwissProt keyword	Hormone	88.3-100	36	59	1.13E-17
SwissProt keyword	Chemotaxis	88.3-100	33	52	8.51E-17
SwissProt keyword	Cleavage on pair of basic residues	88.3-100	32	56	2.82E-14
SwissProt keyword	Amidation	88.3-100	23	32	6.80E-13
SwissProt keyword	Chromosomal protein	88.3-100	30	54	8.65E-13
SwissProt keyword	Acetylation	88.3-100	56	178	2.90E-11
SwissProt keyword	Mitochondrion	88.3-100	91	392	2.21E-10
SwissProt keyword	Inflammatory response	88.3-100	26	52	2.24E-09
SwissProt keyword	Neuropeptide	88.3-100	13	15	3.64E-08
SwissProt keyword	Ubiquinone	88.3-100	20	36	7.77E-08
SwissProt keyword	Fungicide	88.3-100	10	10	3.71E-07
SwissProt keyword	CF(0)	88.3-100	9	10	3.21E-05
SwissProt keyword	Thiol protease inhibitor	88.3-100	11	15	3.55E-05
SwissProt keyword	Antibiotic	88.3-100	12	19	1.01E-04
SwissProt keyword	Pyrrolidone carboxylic acid	88.3-100	20	52	2.56E-04
SwissProt keyword	Hydrogen ion transport	88.3-100	17	39	2.65E-04
SwissProt keyword	Inner membrane	88.3-100	19	54	2.66E-03
SwissProt keyword	3D-structure	88.3-100	159	1067	3.23E-03
SwissProt keyword	Vitamin A	88.3-100	6	7	1.95E-02
SwissProt keyword	Lipid-binding	88.3-100	11	24	2.63E-02
SwissProt keyword	Defensin	88.3-100	5	5	2.91E-02
SwissProt keyword	Prenylation	80.6-92.2	37	91	2.64E-10

SwissProt keyword	Ribosomal protein	80.6-92.2	35	88	2.69E-09
SwissProt keyword	GTP-binding	80.6-92.2	43	158	1.22E-05
SwissProt keyword	Growth factor	80.6-92.2	33	109	4.53E-05
SwissProt keyword	Eye lens protein	80.6-92.2	11	17	3.62E-04
SwissProt keyword	Lipoprotein	80.6-92.2	65	317	3.90E-04
SwissProt keyword	Protein transport	80.6-92.2	43	189	3.07E-03
SwissProt keyword	3D-structure	80.6-92.2	161	1067	4.30E-03
SwissProt keyword	Cytokine	80.6-92.2	35	146	8.65E-03
SwissProt keyword	Nucleosome core	80.6-92.2	13	30	1.26E-02
SwissProt keyword	Acetylation	80.6-92.2	39	178	2.50E-02
SwissProt keyword	Ribosomal protein	76.7-100	69	88	2.58E-26
SwissProt keyword	Cytokine	76.7-100	89	146	2.16E-21
SwissProt keyword	Hormone	76.7-100	47	59	1.06E-17
SwissProt keyword	Cleavage on pair of basic residues	76.7-100	44	56	5.41E-16
SwissProt keyword	Acetylation	76.7-100	93	178	6.72E-16
SwissProt keyword	Prenylation	76.7-100	59	91	2.79E-15
SwissProt keyword	Amidation	76.7-100	29	32	4.27E-13
SwissProt keyword	Nucleosome core	76.7-100	27	30	7.56E-12
SwissProt keyword	Mitochondrion	76.7-100	145	392	4.62E-09
SwissProt keyword	Chemotaxis	76.7-100	35	52	7.04E-09
SwissProt keyword	Mitogen	76.7-100	24	29	1.39E-08
SwissProt keyword	Growth factor	76.7-100	56	109	2.72E-08
SwissProt keyword	Chromosomal protein	76.7-100	35	54	3.70E-08
SwissProt keyword	3D-structure	76.7-100	318	1067	8.71E-08
SwissProt keyword	Ubiquinone	76.7-100	26	36	4.33E-07
SwissProt keyword	GTP-binding	76.7-100	68	158	4.14E-06
SwissProt keyword	Inflammatory response	76.7-100	30	52	6.02E-05
SwissProt keyword	Neuropeptide	76.7-100	13	15	5.20E-04
SwissProt keyword	Eye lens protein	76.7-100	14	17	5.81E-04
SwissProt keyword	Protein transport	76.7-100	72	189	6.22E-04
SwissProt keyword	CF(0)	76.7-100	10	10	7.61E-04
SwissProt keyword	Fungicide	76.7-100	10	10	7.61E-04
SwissProt keyword	Lipid-binding	76.7-100	17	24	1.20E-03
SwissProt keyword	Insulin family	76.7-100	9	9	3.51E-03
SwissProt keyword	Lipoprotein	76.7-100	105	317	4.29E-03
SwissProt keyword	Pyrrolidone carboxylic acid	76.7-100	27	52	5.43E-03
SwissProt keyword	Antibiotic	76.7-100	14	19	6.34E-03
SwissProt keyword	Pharmaceutical	76.7-100	17	26	6.86E-03
SwissProt keyword	Thiol protease inhibitor	76.7-100	12	15	8.35E-03
SwissProt keyword	Threonine protease	76.7-100	12	16	2.67E-02
SwissProt keyword	Hydrogen ion transport	76.7-100	21	39	3.72E-02
SwissProt keyword	Mitogen	72.8-84.4	15	29	4.47E-04
SwissProt keyword	Growth factor	72.8-84.4	32	109	1.42E-03
SwissProt keyword	Proteasome	72.8-84.4	16	37	3.84E-03
SwissProt keyword	Threonine protease	72.8-84.4	10	16	5.57E-03
SwissProt keyword	Prenylation	68.9-92.2	58	91	1.08E-13
SwissProt keyword	GTP-binding	68.9-92.2	76	158	3.07E-09
SwissProt keyword	Lipoprotein	68.9-92.2	119	317	1.82E-06

SwissProt keyword	Ribosomal protein	68.9-92.2	46	88	3.36E-06
SwissProt keyword	Threonine protease	68.9-92.2	15	16	8.47E-06
SwissProt keyword	Growth factor	68.9-92.2	52	109	1.84E-05
SwissProt keyword	Eye lens protein	68.9-92.2	15	17	5.69E-05
SwissProt keyword	Mitogen	68.9-92.2	19	29	3.11E-03
SwissProt keyword	Proteasome	68.9-92.2	21	37	2.37E-02
SwissProt keyword	G-protein coupled receptor	61.1-84.4	112	316	3.48E-03
SwissProt keyword	Homeobox	61.1-84.4	61	152	1.40E-02
SwissProt keyword	MHC II	61.1-84.4	11	13	2.31E-02
SwissProt keyword	G-protein coupled receptor	57.2-68.9	87	316	1.50E-10
SwissProt keyword	G-protein coupled receptor	53.4-76.7	150	316	5.46E-17
SwissProt keyword	Developmental protein	53.4-76.7	99	261	8.88E-04
SwissProt keyword	G-protein coupled receptor	49.5-61.1	93	316	9.65E-14
SwissProt keyword	G-protein coupled receptor	45.6-68.9	170	316	4.54E-27
SwissProt keyword	Palmitate	45.6-68.9	62	130	1.71E-05
SwissProt keyword	Homeobox	45.6-68.9	66	152	5.43E-04
SwissProt keyword	Developmental protein	45.6-68.9	98	261	3.35E-03
SwissProt keyword	Glycosyltransferase	45.6-68.9	50	115	1.68E-02
SwissProt keyword	Monooxygenase	41.7-53.4	25	56	5.37E-06
SwissProt keyword	Intermediate filament	41.7-53.4	19	44	9.36E-04
SwissProt keyword	Keratin	41.7-53.4	15	33	7.62E-03
SwissProt keyword	Microsome	41.7-53.4	26	82	8.95E-03
SwissProt keyword	G-protein coupled receptor	41.7-53.4	65	316	4.77E-02
SwissProt keyword	G-protein coupled receptor	37.8-61.1	140	316	9.13E-12
SwissProt keyword	Monooxygenase	37.8-61.1	41	56	6.55E-11
SwissProt keyword	Microsome	37.8-61.1	51	82	1.30E-09
SwissProt keyword	Postsynaptic membrane	37.8-61.1	35	62	2.43E-04
SwissProt keyword	Palmitate	37.8-61.1	57	130	3.10E-03
SwissProt keyword	Heme	37.8-61.1	39	82	1.58E-02
SwissProt keyword	Intermediate filament	37.8-61.1	25	44	1.62E-02
SwissProt keyword	Transferase	37.8-61.1	237	765	3.24E-02
SwissProt keyword	Keratin	37.8-61.1	20	33	3.97E-02
SwissProt keyword	Monooxygenase	33.9-45.6	33	56	2.41E-13
SwissProt keyword	Microsome	33.9-45.6	33	82	2.89E-07
SwissProt keyword	Heme	33.9-45.6	33	82	2.89E-07
SwissProt keyword	Electron transport	33.9-45.6	27	70	4.42E-05
SwissProt keyword	Zinc-finger	33.9-45.6	110	581	2.25E-03
SwissProt keyword	Oxidoreductase	33.9-45.6	73	351	5.96E-03
SwissProt keyword	Monooxygenase	30.0-53.4	49	56	6.46E-20
SwissProt keyword	Microsome	30.0-53.4	52	82	1.86E-10
SwissProt keyword	Heme	30.0-53.4	49	82	2.53E-08
SwissProt keyword	Electron transport	30.0-53.4	42	70	6.86E-07
SwissProt keyword	DNA-binding	30.0-53.4	302	974	7.22E-04
SwissProt keyword	Transcription regulation	30.0-53.4	277	884	9.17E-04
SwissProt keyword	Serine/threonine-protein kinase	30.0-53.4	87	226	3.92E-03
SwissProt keyword	Zinc-finger	30.0-53.4	188	581	1.06E-02
SwissProt keyword	Oxidoreductase	30.0-53.4	120	351	4.68E-02
SwissProt keyword	Zinc-finger	22.3-45.6	214	581	1.26E-09

SwissProt keyword	Monooxygenase	22.3-45.6	38	56	1.47E-08
SwissProt keyword	DNA-binding	22.3-45.6	310	974	5.08E-06
SwissProt keyword	Transcription regulation	22.3-45.6	285	884	6.29E-06
SwissProt keyword	Metal-binding	22.3-45.6	163	479	6.75E-04
SwissProt keyword	Heme	22.3-45.6	41	82	9.60E-04
SwissProt keyword	Microsome	22.3-45.6	40	82	3.09E-03
SwissProt keyword	Serine/threonine-protein kinase	22.3-45.6	85	226	8.85E-03
SwissProt keyword	Nuclear protein	22.3-45.6	481	1707	1.10E-02
SwissProt keyword	Oxidoreductase	22.3-45.6	121	351	1.36E-02
SwissProt keyword	Activator	22.3-45.6	93	260	3.51E-02
SwissProt keyword	Kelch repeat	22.3-45.6	16	24	4.28E-02
SwissProt keyword	Symport	18.4-30.0	24	60	7.43E-05
SwissProt keyword	Repeat	18.4-30.0	307	1993	8.52E-05
SwissProt keyword	Motor protein	18.4-30.0	18	42	1.23E-03
SwissProt keyword	Alternative splicing	18.4-30.0	281	1867	4.31E-03
SwissProt keyword	ATP-binding	18.4-30.0	118	670	1.05E-02
SwissProt keyword	Symport	18.4-30.0	24	60	7.43E-05
SwissProt keyword	Repeat	18.4-30.0	307	1993	8.52E-05
SwissProt keyword	Motor protein	18.4-30.0	18	42	1.23E-03
SwissProt keyword	Alternative splicing	18.4-30.0	281	1867	4.31E-03
SwissProt keyword	ATP-binding	18.4-30.0	118	670	1.05E-02
SwissProt keyword	Repeat	14.5-37.8	605	1993	9.77E-13
SwissProt keyword	ATP-binding	14.5-37.8	234	670	7.86E-09
SwissProt keyword	Alternative splicing	14.5-37.8	551	1867	1.84E-08
SwissProt keyword	Symport	14.5-37.8	37	60	1.09E-06
SwissProt keyword	DNA-binding	14.5-37.8	304	974	1.05E-05
SwissProt keyword	Zinc-finger	14.5-37.8	196	581	1.66E-05
SwissProt keyword	Transport	14.5-37.8	172	498	2.36E-05
SwissProt keyword	Metalloprotease	14.5-37.8	47	92	2.90E-05
SwissProt keyword	Nuclear protein	14.5-37.8	490	1707	5.03E-05
SwissProt keyword	Transcription regulation	14.5-37.8	275	884	1.01E-04
SwissProt keyword	Phosphorylation	14.5-37.8	296	997	3.77E-03
SwissProt keyword	Serine/threonine-protein kinase	14.5-37.8	85	226	4.08E-03
SwissProt keyword	Metal-binding	14.5-37.8	155	479	1.34E-02
SwissProt keyword	Hydrolase	14.5-37.8	235	779	1.69E-02
SwissProt keyword	Neurotransmitter transport	14.5-37.8	15	21	1.73E-02
SwissProt keyword	Trans-acting factor	14.5-37.8	13	17	2.29E-02
SwissProt keyword	Zinc	14.5-37.8	102	294	2.64E-02
SwissProt keyword	Sodium transport	14.5-37.8	21	36	2.94E-02
SwissProt keyword	Phorbol-ester binding	14.5-37.8	18	29	4.00E-02
SwissProt keyword	Repeat	10.6-22.3	394	1993	1.59E-29
SwissProt keyword	Cell adhesion	10.6-22.3	88	270	1.24E-16
SwissProt keyword	ATP-binding	10.6-22.3	153	670	1.42E-13
SwissProt keyword	Tyrosine-protein kinase	10.6-22.3	36	98	3.99E-07
SwissProt keyword	Alternative splicing	10.6-22.3	302	1867	4.43E-07
SwissProt keyword	Transmembrane	10.6-22.3	315	2026	1.74E-05
SwissProt keyword	Phosphorylation	10.6-22.3	174	997	5.98E-05
SwissProt keyword	Glycoprotein	10.6-22.3	311	2023	8.31E-05

SwissProt keyword	Receptor	10.6-22.3	84	420	1.79E-03
SwissProt keyword	Integrin	10.6-22.3	13	25	3.14E-03
SwissProt keyword	EGF-like domain	10.6-22.3	31	119	4.12E-02
SwissProt keyword	Metalloprotease	10.6-22.3	26	92	4.24E-02
SwissProt keyword	Repeat	6.7-30.0	737	1993	9.43E-53
SwissProt keyword	Cell adhesion	6.7-30.0	156	270	3.90E-31
SwissProt keyword	ATP-binding	6.7-30.0	286	670	4.76E-27
SwissProt keyword	Alternative splicing	6.7-30.0	608	1867	3.34E-21
SwissProt keyword	Tyrosine-protein kinase	6.7-30.0	61	98	8.65E-13
SwissProt keyword	Phosphorylation	6.7-30.0	338	997	4.39E-12
SwissProt keyword	Integrin	6.7-30.0	22	25	5.55E-08
SwissProt keyword	Glycoprotein	6.7-30.0	580	2023	1.23E-06
SwissProt keyword	Transport	6.7-30.0	176	498	2.05E-06
SwissProt keyword	Receptor	6.7-30.0	152	420	5.13E-06
SwissProt keyword	Transmembrane	6.7-30.0	576	2026	6.98E-06
SwissProt keyword	Calcium-binding	6.7-30.0	111	286	1.15E-05
SwissProt keyword	EGF-like domain	6.7-30.0	57	119	1.66E-05
SwissProt keyword	Coiled coil	6.7-30.0	149	416	1.71E-05
SwissProt keyword	Calcium transport	6.7-30.0	14	15	6.78E-05
SwissProt keyword	Metalloprotease	6.7-30.0	46	92	9.75E-05
SwissProt keyword	Symport	6.7-30.0	34	60	1.20E-04
SwissProt keyword	Magnesium	6.7-30.0	55	121	3.02E-04
SwissProt keyword	Metal-binding	6.7-30.0	159	479	1.73E-03
SwissProt keyword	Hydrolase	6.7-30.0	240	779	2.18E-03
SwissProt keyword	Sodium transport	6.7-30.0	22	36	5.77E-03
SwissProt keyword	Motor protein	6.7-30.0	24	42	1.05E-02
SwissProt keyword	Zinc-finger	6.7-30.0	183	581	1.20E-02
SwissProt keyword	Calcium	6.7-30.0	41	91	1.76E-02
SwissProt keyword	Nuclear protein	6.7-30.0	472	1707	1.77E-02
SwissProt keyword	Microtubule	6.7-30.0	35	75	3.33E-02
SwissProt keyword	GTPase activation	6.7-30.0	28	55	3.37E-02
SwissProt keyword	Repeat	2.8-14.5	466	1993	1.73E-69
SwissProt keyword	Cell adhesion	2.8-14.5	104	270	1.80E-28
SwissProt keyword	ATP-binding	2.8-14.5	175	670	3.69E-25
SwissProt keyword	Alternative splicing	2.8-14.5	317	1867	3.78E-13
SwissProt keyword	Coiled coil	2.8-14.5	101	416	8.83E-11
SwissProt keyword	Tyrosine-protein kinase	2.8-14.5	39	98	1.02E-09
SwissProt keyword	Receptor	2.8-14.5	96	420	1.87E-08
SwissProt keyword	Magnesium	2.8-14.5	41	121	1.15E-07
SwissProt keyword	Connective tissue	2.8-14.5	19	32	2.31E-07
SwissProt keyword	Extracellular matrix	2.8-14.5	41	124	2.75E-07
SwissProt keyword	Guanine-nucleotide releasing factor	2.8-14.5	24	52	9.13E-07
SwissProt keyword	Helicase	2.8-14.5	26	62	2.26E-06
SwissProt keyword	Collagen	2.8-14.5	21	46	1.50E-05
SwissProt keyword	Integrin	2.8-14.5	15	25	1.90E-05
SwissProt keyword	Basement membrane	2.8-14.5	14	22	2.07E-05
SwissProt keyword	Calmodulin-binding	2.8-14.5	30	86	2.18E-05
SwissProt keyword	Glycoprotein	2.8-14.5	304	2023	2.22E-05

SwissProt keyword	Phosphorylation	2.8-14.5	170	997	2.99E-05
SwissProt keyword	Bromodomain	2.8-14.5	15	27	8.10E-05
SwissProt keyword	Calcium	2.8-14.5	29	91	3.76E-04
SwissProt keyword	cAMP biosynthesis	2.8-14.5	7	7	6.48E-04
SwissProt keyword	Calcium-binding	2.8-14.5	61	286	2.03E-03
SwissProt keyword	Signal	2.8-14.5	264	1791	2.26E-03
SwissProt keyword	Hydroxylation	2.8-14.5	19	52	5.35E-03
SwissProt keyword	EGF-like domain	2.8-14.5	31	119	1.82E-02
SwissProt keyword	Thick filament	2.8-14.5	9	15	2.20E-02

Table C.4: Correlations to M_{dbSNP}-ranked gene list

System	Gene Category	Percentile (%)	# genes	# in category	p-value
GO Biological Process	cell-cell signaling	88.3-100	97	555	4.56E-03
GO Biological Process	chromatin assembly/disassembly	88.3-100	27	93	3.65E-03
GO Biological Process	defense response	88.3-100	125	787	2.41E-02
GO Biological Process	gas transport	88.3-100	8	11	6.53E-03
GO Biological Process	immune response	88.3-100	116	713	1.66E-02
GO Biological Process	nucleosome assembly	88.3-100	27	61	9.70E-08
GO Biological Process	oxygen transport	88.3-100	8	11	6.53E-03
GO Biological Process	response to biotic stimulus	88.3-100	136	849	5.92E-03
GO Biological Process	small GTPase mediated signal transduction	88.3-100	45	212	2.45E-02
GO Biological Process	mitochondrial transport	80.6-92.2	11	18	1.97E-03
GO Biological Process	small GTPase mediated signal transduction	80.6-92.2	51	212	5.39E-04
GO Biological Process	cell-cell signaling	76.7-100	176	555	9.45E-04
GO Biological Process	defense response	76.7-100	253	787	4.02E-07
GO Biological Process	immune response	76.7-100	232	713	7.30E-07
GO Biological Process	nucleosome assembly	76.7-100	36	61	3.33E-06
GO Biological Process	organismal physiological process	76.7-100	373	1318	5.93E-04
GO Biological Process	response to biotic stimulus	76.7-100	267	849	1.47E-06
GO Biological Process	response to external stimulus	76.7-100	320	1060	4.07E-06
GO Biological Process	response to stimulus	76.7-100	369	1320	2.81E-03
GO Biological Process	small GTPase mediated signal transduction	76.7-100	90	212	2.47E-07
GO Biological Process	cell surface receptor linked signal transduction	72.8-84.4	174	974	7.59E-06
GO Biological Process	cyclic-nucleotide-mediated signaling	72.8-84.4	38	109	5.94E-07
GO Biological Process	G-protein coupled receptor protein signaling pathway	72.8-84.4	127	602	3.09E-08
GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	72.8-84.4	36	103	1.64E-06
GO Biological Process	second-messenger-mediated signaling	72.8-84.4	40	123	2.18E-06
GO Biological Process	cell surface receptor linked signal transduction	68.9-92.2	306	974	5.84E-06
GO Biological Process	cyclic-nucleotide-mediated signaling	68.9-92.2	47	109	1.58E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	68.9-92.2	218	602	1.06E-09
GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	68.9-92.2	45	103	1.65E-02
GO Biological Process	organismal physiological process	68.9-92.2	378	1318	6.29E-03
GO Biological Process	signal transduction	68.9-92.2	607	2214	2.55E-03
GO Biological Process	small GTPase mediated signal transduction	68.9-92.2	79	212	1.40E-02
GO Biological Process	cell surface receptor linked signal transduction	65.0-76.7	166	974	1.31E-03
GO Biological Process	G-protein coupled receptor protein signaling pathway	65.0-76.7	122	602	2.90E-06
GO Biological Process	cAMP-mediated signaling	61.1-84.4	33	69	3.60E-02
GO Biological Process	cell surface receptor linked signal transduction	61.1-84.4	315	974	1.82E-07
GO Biological Process	cyclic-nucleotide-mediated signaling	61.1-84.4	55	109	3.49E-06
GO Biological Process	G-protein coupled receptor protein signaling pathway	61.1-84.4	223	602	7.37E-11
GO Biological Process	G-protein signaling\, coupled to cAMP nucleotide second messenger	61.1-84.4	33	67	1.61E-02

GO Biological Process	G-protein signaling\, coupled to cyclic nucleotide second messenger	61.1-84.4	53	103	2.82E-06
GO Biological Process	G-protein signaling\, coupled to IP3 second messenger (phospholipase C activating)	61.1-84.4	37	79	1.95E-02
GO Biological Process	neurophysiological process	61.1-84.4	98	276	1.69E-02
GO Biological Process	organismal physiological process	61.1-84.4	391	1318	1.52E-04
GO Biological Process	second-messenger-mediated signaling	61.1-84.4	60	123	3.40E-06
GO Biological Process	carbohydrate metabolism	53.4-76.7	130	390	4.22E-02
GO Biological Process	gamma-aminobutyric acid signaling pathway	53.4-76.7	13	17	2.84E-02
GO Biological Process	G-protein coupled receptor protein signaling pathway	53.4-76.7	194	602	4.11E-03
GO Biological Process	regulation of synapse	49.5-61.1	11	22	4.03E-02
GO Biological Process	metabolism	33.9-45.6	815	6240	4.89E-02
GO Biological Process	homophilic cell adhesion	30.0-53.4	53	116	6.42E-04
GO Biological Process	metabolism	30.0-53.4	1578	6240	3.55E-02
GO Biological Process	protein modification	30.0-53.4	289	975	2.19E-02
GO Biological Process	homophilic cell adhesion	26.2-37.8	32	116	9.95E-03
GO Biological Process	phosphate metabolism	26.2-37.8	120	681	1.40E-02
GO Biological Process	phosphorus metabolism	26.2-37.8	120	681	1.40E-02
GO Biological Process	protein amino acid phosphorylation	26.2-37.8	90	491	4.83E-02
GO Biological Process	protein metabolism	26.2-37.8	338	2296	1.09E-02
GO Biological Process	protein modification	26.2-37.8	170	975	2.51E-04
GO Biological Process	cell adhesion	22.3-45.6	185	589	3.50E-02
GO Biological Process	cell-cell adhesion	22.3-45.6	85	223	3.91E-03
GO Biological Process	homophilic cell adhesion	22.3-45.6	61	116	6.34E-08
GO Biological Process	metabolism	22.3-45.6	1588	6240	8.81E-03
GO Biological Process	phosphate metabolism	22.3-45.6	232	681	9.98E-07
GO Biological Process	phosphorus metabolism	22.3-45.6	232	681	9.98E-07
GO Biological Process	phosphorylation	22.3-45.6	181	537	2.48E-04
GO Biological Process	protein amino acid phosphorylation	22.3-45.6	174	491	7.48E-06
GO Biological Process	protein metabolism	22.3-45.6	645	2296	1.98E-04
GO Biological Process	protein modification	22.3-45.6	310	975	6.01E-06
GO Biological Process	cell adhesion	18.4-30.0	105	589	1.71E-02
GO Biological Process	enzyme linked receptor protein signaling pathway	18.4-30.0	37	119	3.85E-05
GO Biological Process	homophilic cell adhesion	18.4-30.0	31	116	2.08E-02
GO Biological Process	phosphate metabolism	18.4-30.0	127	681	1.26E-04
GO Biological Process	phosphorus metabolism	18.4-30.0	127	681	1.26E-04
GO Biological Process	phosphorylation	18.4-30.0	97	537	2.23E-02
GO Biological Process	protein amino acid phosphorylation	18.4-30.0	95	491	1.27E-03
GO Biological Process	transmembrane receptor protein tyrosine kinase signaling pathway	18.4-30.0	27	86	2.80E-03
GO Biological Process	cell adhesion	14.5-37.8	191	589	1.83E-03
GO Biological Process	enzyme linked receptor protein signaling pathway	14.5-37.8	51	119	1.08E-02
GO Biological Process	homophilic cell adhesion	14.5-37.8	53	116	5.89E-04
GO Biological Process	integrin-mediated signaling pathway	14.5-37.8	25	45	1.57E-02
GO Biological Process	phosphate metabolism	14.5-37.8	234	681	2.26E-07
GO Biological Process	phosphorus metabolism	14.5-37.8	234	681	2.26E-07
GO Biological Process	phosphorylation	14.5-37.8	180	537	3.19E-04

GO Biological Process	protein amino acid phosphorylation	14.5-37.8	176	491	1.61E-06
GO Biological Process	protein metabolism	14.5-37.8	637	2296	1.09E-03
GO Biological Process	protein modification	14.5-37.8	308	975	9.81E-06
GO Biological Process	transmembrane receptor protein tyrosine kinase signaling pathway	14.5-37.8	39	86	3.12E-02
GO Biological Process	cell adhesion	6.7-30.0	212	589	9.12E-09
GO Biological Process	enzyme linked receptor protein signaling pathway	6.7-30.0	53	119	1.35E-03
GO Biological Process	muscle development	6.7-30.0	56	132	4.43E-03
GO Biological Process	phosphate metabolism	6.7-30.0	221	681	1.36E-04
GO Biological Process	phosphorus metabolism	6.7-30.0	221	681	1.36E-04
GO Biological Process	phosphorylation	6.7-30.0	174	537	4.57E-03
GO Biological Process	protein amino acid phosphorylation	6.7-30.0	171	491	2.29E-05
GO Biological Process	striated muscle contraction	6.7-30.0	18	29	4.01E-02
GO Biological Process	cytoskeletal anchoring	2.8-14.5	10	15	1.97E-03
GO Biological Process	striated muscle contraction	2.8-14.5	14	29	2.81E-03
GO Cellular Component	cytosolic large ribosomal subunit (sensu Eukarya)	88.3-100	15	39	1.96E-02
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	88.3-100	27	69	3.04E-06
GO Cellular Component	extracellular	88.3-100	203	1230	1.07E-06
GO Cellular Component	extracellular space	88.3-100	86	410	4.63E-06
GO Cellular Component	hemoglobin complex	88.3-100	9	13	2.84E-03
GO Cellular Component	large ribosomal subunit	88.3-100	20	55	1.75E-03
GO Cellular Component	nucleosome	88.3-100	27	79	9.25E-05
GO Cellular Component	ribonucleoprotein complex	88.3-100	84	441	6.89E-04
GO Cellular Component	ribosome	88.3-100	67	282	1.50E-06
GO Cellular Component	small ribosomal subunit	88.3-100	17	45	6.51E-03
GO Cellular Component	cytosol	76.7-100	128	388	5.17E-03
GO Cellular Component	cytosolic ribosome (sensu Eukarya)	76.7-100	39	69	4.25E-06
GO Cellular Component	cytosolic small ribosomal subunit (sensu Eukarya)	76.7-100	19	29	3.62E-03
GO Cellular Component	eukaryotic 48S initiation complex	76.7-100	19	29	3.62E-03
GO Cellular Component	extracellular	76.7-100	362	1230	1.14E-05
GO Cellular Component	extracellular space	76.7-100	141	410	8.83E-05
GO Cellular Component	hemoglobin complex	76.7-100	12	13	6.88E-04
GO Cellular Component	mitochondrion	76.7-100	206	690	1.41E-02
GO Cellular Component	nucleosome	76.7-100	38	79	2.08E-03
GO Cellular Component	ribosome	76.7-100	109	282	2.67E-06
GO Cellular Component	small ribosomal subunit	76.7-100	27	45	2.81E-04
GO Cellular Component	integral to membrane	68.9-92.2	796	2911	1.38E-05
GO Cellular Component	integral to membrane	65.0-76.7	414	2911	4.35E-03
GO Cellular Component	integral to plasma membrane	65.0-76.7	200	1269	1.09E-02
GO Cellular Component	integral to membrane	61.1-84.4	808	2911	1.97E-06
GO Cellular Component	integral to plasma membrane	61.1-84.4	388	1269	2.63E-06
GO Cellular Component	membrane	61.1-84.4	1119	4338	2.14E-02
GO Cellular Component	plasma membrane	61.1-84.4	526	1812	9.55E-06
GO Cellular Component	integral to membrane	53.4-76.7	783	2911	7.17E-03
GO Cellular Component	integral to plasma membrane	53.4-76.7	382	1269	6.74E-05

GO Cellular Component plasma membrane	53.4-76.7	510	1812	4.16E-03
GO Cellular Component cell	30.0-53.4	2444	10056	4.62E-02
GO Cellular Component cytoskeleton	22.3-45.6	266	889	2.67E-02
GO Cellular Component cytoskeleton	18.4-30.0	160	889	4.31E-05
GO Cellular Component extracellular matrix	18.4-30.0	63	302	1.21E-02
GO Cellular Component actin cytoskeleton	14.5-37.8	105	294	8.47E-03
GO Cellular Component cytoskeleton	14.5-37.8	297	889	3.26E-08
GO Cellular Component extracellular matrix	14.5-37.8	110	302	1.61E-03
GO Cellular Component collagen	10.6-22.3	16	33	6.64E-04
GO Cellular Component cytoskeleton	10.6-22.3	154	889	1.23E-03
GO Cellular Component extracellular matrix	10.6-22.3	64	302	6.35E-03
GO Cellular Component fibrillar collagen	10.6-22.3	7	9	2.71E-02
GO Cellular Component voltage-gated calcium channel complex	10.6-22.3	11	22	3.68E-02
GO Cellular Component voltage-gated sodium channel complex	10.6-22.3	11	13	1.13E-05
GO Cellular Component actin cytoskeleton	6.7-30.0	123	294	1.18E-08
GO Cellular Component basement membrane	6.7-30.0	27	46	1.50E-03
GO Cellular Component collagen	6.7-30.0	28	33	7.39E-10
GO Cellular Component cytoskeleton	6.7-30.0	341	889	1.10E-20
GO Cellular Component extracellular matrix	6.7-30.0	136	302	6.80E-13
GO Cellular Component fibrillar collagen	6.7-30.0	9	9	8.87E-03
GO Cellular Component microtubule associated complex	6.7-30.0	46	103	8.91E-03
GO Cellular Component myosin	6.7-30.0	54	95	1.52E-08
GO Cellular Component striated muscle thick filament	6.7-30.0	12	15	2.57E-02
GO Cellular Component voltage-gated sodium channel complex	6.7-30.0	12	13	1.20E-03
GO Cellular Component basement membrane	2.8-14.5	18	46	3.75E-03
GO Cellular Component collagen	2.8-14.5	19	33	6.02E-07
GO Cellular Component cytoskeleton	2.8-14.5	152	889	4.64E-04
GO Cellular Component extracellular matrix	2.8-14.5	61	302	1.96E-02
GO Cellular Component fibrillar collagen	2.8-14.5	8	9	6.85E-04
GO Cellular Component muscle myosin	2.8-14.5	12	26	2.97E-02
GO Cellular Component myosin	2.8-14.5	30	95	3.75E-04
GO Cellular Component myosin II	2.8-14.5	17	45	1.30E-02
GO Cellular Component striated muscle thick filament	2.8-14.5	9	15	2.53E-02
GO Cellular Component voltage-gated sodium channel complex	2.8-14.5	9	13	4.49E-03
GO Molecular Function cytochrome-c oxidase activity	88.3-100	10	20	3.69E-02
GO Molecular Function cytokine activity	88.3-100	56	203	4.97E-08
GO Molecular Function heme-copper terminal oxidase activity	88.3-100	10	20	3.69E-02
GO Molecular Function hormone activity	88.3-100	29	99	9.08E-04
GO Molecular Function hydrogen ion transporter activity	88.3-100	33	113	1.56E-04
GO Molecular Function hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	88.3-100	41	190	2.85E-02
GO Molecular Function monovalent inorganic cation transporter activity	88.3-100	33	123	1.39E-03
GO Molecular Function oxidoreductase activity\, acting on heme group of donors	88.3-100	10	20	3.69E-02
GO Molecular Function oxidoreductase activity\, acting on heme group of donors\, oxygen as acceptor	88.3-100	10	20	3.69E-02
GO Molecular Function oxygen transporter activity	88.3-100	8	11	5.86E-03

GO Molecular Function receptor binding	88.3-100	114	522	1.71E-10
GO Molecular Function small monomeric GTPase activity	88.3-100	39	132	6.87E-06
GO Molecular Function structural constituent of ribosome	88.3-100	65	210	2.72E-12
GO Molecular Function GTP binding	80.6-92.2	56	244	6.03E-04
GO Molecular Function GTPase activity	80.6-92.2	51	217	9.66E-04
GO Molecular Function guanyl nucleotide binding	80.6-92.2	57	251	6.86E-04
GO Molecular Function hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	80.6-92.2	49	190	7.51E-05
GO Molecular Function RAB small monomeric GTPase activity	80.6-92.2	21	53	3.65E-04
GO Molecular Function Rho small monomeric GTPase activity	80.6-92.2	12	27	4.73E-02
GO Molecular Function rhodopsin-like receptor activity	80.6-92.2	69	331	1.25E-03
GO Molecular Function small monomeric GTPase activity	80.6-92.2	45	132	1.31E-08
GO Molecular Function cytochrome-c oxidase activity	76.7-100	14	20	2.63E-02
GO Molecular Function cytokine activity	76.7-100	82	203	1.54E-05
GO Molecular Function G-protein coupled receptor activity	76.7-100	131	410	1.07E-02
GO Molecular Function GTP binding	76.7-100	101	244	4.50E-08
GO Molecular Function GTPase activity	76.7-100	96	217	1.34E-09
GO Molecular Function guanyl nucleotide binding	76.7-100	102	251	1.31E-07
GO Molecular Function heme-copper terminal oxidase activity	76.7-100	14	20	2.63E-02
GO Molecular Function hydrogen ion transporter activity	76.7-100	49	113	1.66E-03
GO Molecular Function hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	76.7-100	90	190	5.29E-11
GO Molecular Function monovalent inorganic cation transporter activity	76.7-100	50	123	1.28E-02
GO Molecular Function oxidoreductase activity\, acting on heme group of donors	76.7-100	14	20	2.63E-02
GO Molecular Function oxidoreductase activity\, acting on heme group of donors\, oxygen as acceptor	76.7-100	14	20	2.63E-02
GO Molecular Function RAB small monomeric GTPase activity	76.7-100	31	53	4.77E-05
GO Molecular Function RAS small monomeric GTPase activity	76.7-100	19	29	2.78E-03
GO Molecular Function receptor binding	76.7-100	183	522	2.55E-08
GO Molecular Function Rho small monomeric GTPase activity	76.7-100	19	27	4.93E-04
GO Molecular Function rhodopsin-like receptor activity	76.7-100	123	331	1.38E-06
GO Molecular Function small monomeric GTPase activity	76.7-100	77	132	1.01E-15
GO Molecular Function structural constituent of ribosome	76.7-100	101	210	3.44E-13
GO Molecular Function amine receptor activity	72.8-84.4	20	40	7.32E-06
GO Molecular Function G-protein coupled receptor activity	72.8-84.4	110	410	1.51E-14
GO Molecular Function peptide binding	72.8-84.4	38	143	1.80E-03
GO Molecular Function peptide receptor activity	72.8-84.4	36	109	7.49E-06
GO Molecular Function peptide receptor activity\, G-protein coupled	72.8-84.4	36	109	7.49E-06
GO Molecular Function receptor activity	72.8-84.4	209	1306	8.18E-04
GO Molecular Function rhodopsin-like receptor activity	72.8-84.4	103	331	1.27E-18
GO Molecular Function signal transducer activity	72.8-84.4	302	2102	3.59E-02
GO Molecular Function transmembrane receptor activity	72.8-84.4	160	852	4.34E-07
GO Molecular Function amine receptor activity	68.9-92.2	25	40	4.75E-04
GO Molecular Function cytokine binding	68.9-92.2	29	57	1.90E-02
GO Molecular Function galactosyltransferase activity	68.9-92.2	14	20	4.56E-02
GO Molecular Function G-protein coupled receptor activity	68.9-92.2	184	410	2.38E-19
GO Molecular Function GTP binding	68.9-92.2	99	244	3.12E-06

GO Molecular Function	GTPase activity	68.9-92.2	87	217	6.47E-05
GO Molecular Function	guanyl nucleotide binding	68.9-92.2	100	251	8.27E-06
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, acting on GTP\, involved in cellular and subcellular movement	68.9-92.2	83	190	1.05E-06
GO Molecular Function	oxidoreductase activity\, acting on CH-OH group of donors	68.9-92.2	44	96	3.30E-03
GO Molecular Function	oxidoreductase activity\, acting on the CH-OH group of donors\, NAD or NADP as acceptor	68.9-92.2	42	92	6.63E-03
GO Molecular Function	peptide binding	68.9-92.2	68	143	5.00E-07
GO Molecular Function	peptide receptor activity	68.9-92.2	57	109	1.41E-07
GO Molecular Function	peptide receptor activity\, G-protein coupled	68.9-92.2	57	109	1.41E-07
GO Molecular Function	RAB small monomeric GTPase activity	68.9-92.2	28	53	1.06E-02
GO Molecular Function	receptor activity	68.9-92.2	380	1306	3.96E-04
GO Molecular Function	Rho small monomeric GTPase activity	68.9-92.2	17	27	4.52E-02
GO Molecular Function	rhodopsin-like receptor activity	68.9-92.2	172	331	3.84E-27
GO Molecular Function	small monomeric GTPase activity	68.9-92.2	66	132	5.72E-08
GO Molecular Function	transmembrane receptor activity	68.9-92.2	276	852	4.81E-07
GO Molecular Function	extracellular ligand-gated ion channel activity	65.0-76.7	20	60	2.82E-02
GO Molecular Function	G-protein coupled receptor activity	65.0-76.7	103	410	6.15E-11
GO Molecular Function	neurotransmitter binding	65.0-76.7	19	51	7.18E-03
GO Molecular Function	neurotransmitter receptor activity	65.0-76.7	19	50	5.06E-03
GO Molecular Function	peptide receptor activity	65.0-76.7	30	109	1.70E-02
GO Molecular Function	peptide receptor activity\, G-protein coupled	65.0-76.7	30	109	1.70E-02
GO Molecular Function	receptor activity	65.0-76.7	220	1306	1.67E-05
GO Molecular Function	rhodopsin-like receptor activity	65.0-76.7	89	331	5.71E-11
GO Molecular Function	transmembrane receptor activity	65.0-76.7	159	852	2.52E-06
GO Molecular Function	alcohol dehydrogenase activity	61.1-84.4	15	22	4.23E-02
GO Molecular Function	alcohol dehydrogenase activity\, zinc-dependent	61.1-84.4	13	17	2.28E-02
GO Molecular Function	alpha-type channel activity	61.1-84.4	114	335	2.21E-02
GO Molecular Function	amine receptor activity	61.1-84.4	27	40	1.60E-05
GO Molecular Function	channel/pore class transporter activity	61.1-84.4	118	353	3.96E-02
GO Molecular Function	extracellular ligand-gated ion channel activity	61.1-84.4	32	60	2.04E-03
GO Molecular Function	G-protein coupled receptor activity	61.1-84.4	186	410	1.05E-19
GO Molecular Function	neurotransmitter binding	61.1-84.4	32	51	9.35E-06
GO Molecular Function	neurotransmitter receptor activity	61.1-84.4	32	50	4.49E-06
GO Molecular Function	peptide binding	61.1-84.4	70	143	8.34E-08
GO Molecular Function	peptide receptor activity	61.1-84.4	65	109	2.05E-12
GO Molecular Function	peptide receptor activity\, G-protein coupled	61.1-84.4	65	109	2.05E-12
GO Molecular Function	receptor activity	61.1-84.4	404	1306	1.05E-07
GO Molecular Function	rhodopsin-like receptor activity	61.1-84.4	168	331	2.52E-24
GO Molecular Function	signal transducer activity	61.1-84.4	573	2102	1.54E-02
GO Molecular Function	transmembrane receptor activity	61.1-84.4	300	852	1.94E-12
GO Molecular Function	extracellular ligand-gated ion channel activity	57.2-68.9	20	60	3.22E-02
GO Molecular Function	ligand-gated ion channel activity	57.2-68.9	27	97	4.59E-02
GO Molecular Function	neurotransmitter binding	57.2-68.9	18	51	3.61E-02
GO Molecular Function	neurotransmitter receptor activity	57.2-68.9	18	50	2.62E-02
GO Molecular Function	acetylcholine receptor activity	53.4-76.7	14	19	2.46E-02

GO Molecular Function	alpha-type channel activity	53.4-76.7	120	335	1.38E-03
GO Molecular Function	amine receptor activity	53.4-76.7	23	40	1.96E-02
GO Molecular Function	channel/pore class transporter activity	53.4-76.7	126	353	8.81E-04
GO Molecular Function	extracellular ligand-gated ion channel activity	53.4-76.7	35	60	4.03E-05
GO Molecular Function	GABA receptor activity	53.4-76.7	17	25	1.40E-02
GO Molecular Function	GABA-A receptor activity	53.4-76.7	17	23	2.16E-03
GO Molecular Function	G-protein coupled receptor activity	53.4-76.7	152	410	2.40E-06
GO Molecular Function	ion channel activity	53.4-76.7	112	310	2.08E-03
GO Molecular Function	ligand-gated ion channel activity	53.4-76.7	45	97	3.47E-03
GO Molecular Function	neurotransmitter binding	53.4-76.7	35	51	5.72E-08
GO Molecular Function	neurotransmitter receptor activity	53.4-76.7	35	50	2.33E-08
GO Molecular Function	nicotinic acetylcholine-activated cation-selective channel activity	53.4-76.7	13	16	8.24E-03
GO Molecular Function	peptide receptor activity	53.4-76.7	51	109	4.70E-04
GO Molecular Function	peptide receptor activity\, G-protein coupled	53.4-76.7	51	109	4.70E-04
GO Molecular Function	receptor activity	53.4-76.7	379	1306	8.60E-03
GO Molecular Function	rhodopsin-like receptor activity	53.4-76.7	135	331	1.09E-08
GO Molecular Function	transmembrane receptor activity	53.4-76.7	267	852	4.17E-04
GO Molecular Function	electrochemical potential-driven transporter activity	45.6-68.9	71	185	2.11E-02
GO Molecular Function	monooxygenase activity	45.6-68.9	38	72	3.57E-04
GO Molecular Function	neurotransmitter binding	45.6-68.9	28	51	6.23E-03
GO Molecular Function	neurotransmitter receptor activity	45.6-68.9	28	50	3.63E-03
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen	45.6-68.9	41	87	5.90E-03
GO Molecular Function	oxidoreductase activity\, acting on paired donors\, with incorporation or reduction of molecular oxygen\, reduced flavin or flavoprotein as one donor\, and incorporation of one atom of oxygen	45.6-68.9	16	24	3.91E-02
GO Molecular Function	oxygen binding	45.6-68.9	17	24	5.53E-03
GO Molecular Function	porter activity	45.6-68.9	70	184	3.49E-02
GO Molecular Function	unspecific monooxygenase activity	45.6-68.9	16	24	3.91E-02
GO Molecular Function	calcium ion binding	33.9-45.6	105	576	1.48E-02
GO Molecular Function	metal ion binding	33.9-45.6	190	1135	1.15E-03
GO Molecular Function	calcium-dependent cell adhesion molecule activity	30.0-53.4	48	94	3.82E-05
GO Molecular Function	catalytic activity	30.0-53.4	1102	4200	1.74E-02
GO Molecular Function	metal ion binding	30.0-53.4	345	1135	2.94E-04
GO Molecular Function	metalloendopeptidase activity	30.0-53.4	47	93	8.31E-05
GO Molecular Function	metallopeptidase activity	30.0-53.4	67	169	1.33E-02
GO Molecular Function	adenyl nucleotide binding	26.2-37.8	184	1071	3.61E-04
GO Molecular Function	ATP binding	26.2-37.8	179	1059	1.71E-03
GO Molecular Function	calcium-dependent cell adhesion molecule activity	26.2-37.8	29	94	2.74E-03
GO Molecular Function	catalytic activity	26.2-37.8	579	4200	1.98E-02
GO Molecular Function	magnesium ion binding	26.2-37.8	34	133	3.63E-02
GO Molecular Function	metalloendopeptidase activity	26.2-37.8	29	93	2.14E-03
GO Molecular Function	metallopeptidase activity	26.2-37.8	45	169	4.26E-04
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	26.2-37.8	111	598	3.77E-03
GO Molecular Function	protein kinase activity	26.2-37.8	94	508	2.86E-02

GO Molecular Function protein-tyrosine kinase activity	26.2-37.8	59	241	1.43E-04
GO Molecular Function adenyl nucleotide binding	22.3-45.6	364	1071	1.72E-11
GO Molecular Function ATP binding	22.3-45.6	356	1059	2.14E-10
GO Molecular Function ATPase activity	22.3-45.6	94	249	2.53E-03
GO Molecular Function ATPase activity\, coupled	22.3-45.6	93	236	2.91E-04
GO Molecular Function calcium ion binding	22.3-45.6	188	576	3.11E-03
GO Molecular Function calcium-dependent cell adhesion molecule activity	22.3-45.6	56	94	5.99E-10
GO Molecular Function catalytic activity	22.3-45.6	1107	4200	1.22E-02
GO Molecular Function cell adhesion molecule activity	22.3-45.6	141	374	4.60E-06
GO Molecular Function glutamate receptor activity	22.3-45.6	21	34	1.13E-02
GO Molecular Function hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	22.3-45.6	99	275	1.55E-02
GO Molecular Function kinase activity	22.3-45.6	237	725	9.08E-05
GO Molecular Function magnesium ion binding	22.3-45.6	60	133	2.51E-04
GO Molecular Function metal ion binding	22.3-45.6	366	1135	5.70E-08
GO Molecular Function metalloendopeptidase activity	22.3-45.6	52	93	1.29E-07
GO Molecular Function metallopeptidase activity	22.3-45.6	75	169	1.44E-05
GO Molecular Function nucleotide binding	22.3-45.6	398	1317	7.66E-05
GO Molecular Function phosphotransferase activity\, alcohol group as acceptor	22.3-45.6	211	598	4.23E-07
GO Molecular Function protein kinase activity	22.3-45.6	177	508	4.23E-05
GO Molecular Function protein-tyrosine kinase activity	22.3-45.6	99	241	8.64E-06
GO Molecular Function purine nucleotide binding	22.3-45.6	396	1303	4.00E-05
GO Molecular Function transferase activity	22.3-45.6	382	1327	3.86E-02
GO Molecular Function transferase activity\, transferring phosphorus-containing groups	22.3-45.6	241	750	3.53E-04
GO Molecular Function adenyl nucleotide binding	18.4-30.0	219	1071	1.48E-13
GO Molecular Function ATP binding	18.4-30.0	216	1059	3.42E-13
GO Molecular Function ATP dependent helicase activity	18.4-30.0	36	103	2.86E-06
GO Molecular Function ATPase activity	18.4-30.0	63	249	1.09E-05
GO Molecular Function ATPase activity\, coupled	18.4-30.0	62	236	2.94E-06
GO Molecular Function calcium-dependent cell adhesion molecule activity	18.4-30.0	29	94	2.51E-03
GO Molecular Function cell adhesion molecule activity	18.4-30.0	81	374	1.54E-04
GO Molecular Function helicase activity	18.4-30.0	37	125	2.68E-04
GO Molecular Function hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	18.4-30.0	66	275	4.52E-05
GO Molecular Function kinase activity	18.4-30.0	130	725	2.57E-03
GO Molecular Function magnesium ion binding	18.4-30.0	35	133	1.23E-02
GO Molecular Function nucleotide binding	18.4-30.0	229	1317	1.64E-06
GO Molecular Function phosphotransferase activity\, alcohol group as acceptor	18.4-30.0	115	598	2.92E-04
GO Molecular Function protein kinase activity	18.4-30.0	101	508	3.45E-04
GO Molecular Function protein-tyrosine kinase activity	18.4-30.0	59	241	1.23E-04
GO Molecular Function purine nucleotide binding	18.4-30.0	227	1303	1.65E-06
GO Molecular Function transferase activity\, transferring phosphorus-containing groups	18.4-30.0	131	750	9.77E-03
GO Molecular Function transmembrane receptor protein kinase activity	18.4-30.0	26	76	9.75E-04
GO Molecular Function transmembrane receptor protein tyrosine kinase activity	18.4-30.0	25	64	8.09E-05
GO Molecular Function adenyl nucleotide binding	14.5-37.8	420	1071	7.18E-28

GO Molecular Function	ATP binding	14.5-37.8	412	1059	2.32E-26
GO Molecular Function	ATP dependent helicase activity	14.5-37.8	49	103	5.61E-04
GO Molecular Function	ATPase activity	14.5-37.8	110	249	5.43E-09
GO Molecular Function	ATPase activity\, coupled	14.5-37.8	107	236	1.44E-09
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of ions\, phosphorylative mechanism	14.5-37.8	24	44	4.75E-02
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	14.5-37.8	54	119	9.39E-04
GO Molecular Function	binding	14.5-37.8	1715	6657	5.24E-04
GO Molecular Function	calcium-dependent cell adhesion molecule activity	14.5-37.8	48	94	4.40E-05
GO Molecular Function	cell adhesion molecule activity	14.5-37.8	141	374	5.25E-06
GO Molecular Function	GTPase regulator activity	14.5-37.8	100	231	2.39E-07
GO Molecular Function	guanyl-nucleotide exchange factor activity	14.5-37.8	49	95	2.01E-05
GO Molecular Function	helicase activity	14.5-37.8	53	125	1.56E-02
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	14.5-37.8	54	119	9.39E-04
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	14.5-37.8	116	275	5.43E-08
GO Molecular Function	kinase activity	14.5-37.8	242	725	9.34E-06
GO Molecular Function	magnesium ion binding	14.5-37.8	65	133	1.32E-06
GO Molecular Function	metal ion binding	14.5-37.8	353	1135	2.25E-05
GO Molecular Function	metalloendopeptidase activity	14.5-37.8	49	93	7.90E-06
GO Molecular Function	metallopeptidase activity	14.5-37.8	75	169	1.56E-05
GO Molecular Function	nucleotide binding	14.5-37.8	449	1317	3.73E-15
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	14.5-37.8	220	598	1.85E-09
GO Molecular Function	protein kinase activity	14.5-37.8	186	508	2.10E-07
GO Molecular Function	protein-tyrosine kinase activity	14.5-37.8	108	241	2.84E-09
GO Molecular Function	purine nucleotide binding	14.5-37.8	446	1303	2.17E-15
GO Molecular Function	small GTPase regulatory/interacting protein activity	14.5-37.8	72	163	4.05E-05
GO Molecular Function	transferase activity\, transferring phosphorus-containing groups	14.5-37.8	245	750	6.64E-05
GO Molecular Function	transmembrane receptor protein kinase activity	14.5-37.8	39	76	8.64E-04
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	14.5-37.8	37	64	2.67E-05
GO Molecular Function	adenyl nucleotide binding	10.6-22.3	237	1071	9.11E-20
GO Molecular Function	ATP binding	10.6-22.3	235	1059	9.75E-20
GO Molecular Function	ATPase activity	10.6-22.3	59	249	4.05E-04
GO Molecular Function	ATPase activity\, coupled	10.6-22.3	57	236	3.29E-04
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	10.6-22.3	32	119	1.86E-02
GO Molecular Function	ATP-binding cassette (ABC) transporter activity	10.6-22.3	25	74	2.08E-03
GO Molecular Function	binding	10.6-22.3	876	6657	5.55E-03
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	10.6-22.3	32	119	1.86E-02
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	10.6-22.3	62	275	1.29E-03
GO Molecular Function	nucleotide binding	10.6-22.3	254	1317	5.79E-13
GO Molecular Function	protein kinase activity	10.6-22.3	93	508	4.04E-02
GO Molecular Function	purine nucleotide binding	10.6-22.3	249	1303	3.97E-12
GO Molecular Function	sodium channel activity	10.6-22.3	15	29	5.89E-04

GO Molecular Function	voltage-gated sodium channel activity	10.6-22.3	12	17	8.15E-05
GO Molecular Function	actin binding	6.7-30.0	93	220	4.36E-06
GO Molecular Function	adenyl nucleotide binding	6.7-30.0	452	1071	2.50E-40
GO Molecular Function	ATP binding	6.7-30.0	449	1059	1.18E-40
GO Molecular Function	ATP dependent helicase activity	6.7-30.0	58	103	6.01E-09
GO Molecular Function	ATPase activity	6.7-30.0	120	249	1.41E-13
GO Molecular Function	ATPase activity\, coupled	6.7-30.0	115	236	2.28E-13
GO Molecular Function	ATPase activity\, coupled to transmembrane movement of substances	6.7-30.0	55	119	3.06E-04
GO Molecular Function	ATP-binding cassette (ABC) transporter activity	6.7-30.0	38	74	1.10E-03
GO Molecular Function	binding	6.7-30.0	1768	6657	3.04E-11
GO Molecular Function	calcium channel activity	6.7-30.0	27	52	4.34E-02
GO Molecular Function	calcium ion binding	6.7-30.0	196	576	5.32E-05
GO Molecular Function	calmodulin binding	6.7-30.0	50	105	3.65E-04
GO Molecular Function	cell adhesion molecule activity	6.7-30.0	149	374	1.20E-08
GO Molecular Function	cytoskeletal protein binding	6.7-30.0	114	293	2.23E-05
GO Molecular Function	extracellular matrix structural constituent	6.7-30.0	47	89	1.37E-05
GO Molecular Function	extracellular matrix structural constituent conferring tensile strength	6.7-30.0	13	15	1.98E-03
GO Molecular Function	GTPase regulator activity	6.7-30.0	88	231	3.47E-03
GO Molecular Function	guanyl-nucleotide exchange factor activity	6.7-30.0	44	95	5.49E-03
GO Molecular Function	helicase activity	6.7-30.0	61	125	4.37E-06
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances	6.7-30.0	55	119	3.06E-04
GO Molecular Function	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides	6.7-30.0	124	275	2.55E-11
GO Molecular Function	kinase activity	6.7-30.0	238	725	5.33E-05
GO Molecular Function	metal ion binding	6.7-30.0	343	1135	8.24E-04
GO Molecular Function	microtubule motor activity	6.7-30.0	22	37	1.53E-02
GO Molecular Function	motor activity	6.7-30.0	64	113	2.93E-10
GO Molecular Function	nucleotide binding	6.7-30.0	478	1317	2.68E-23
GO Molecular Function	phosphotransferase activity\, alcohol group as acceptor	6.7-30.0	214	598	6.55E-08
GO Molecular Function	P-P-bond-hydrolysis-driven transporter activity	6.7-30.0	56	137	2.92E-02
GO Molecular Function	protein binding	6.7-30.0	468	1548	2.75E-06
GO Molecular Function	protein kinase activity	6.7-30.0	186	508	1.67E-07
GO Molecular Function	protein-tyrosine kinase activity	6.7-30.0	93	241	8.67E-04
GO Molecular Function	purine nucleotide binding	6.7-30.0	471	1303	1.99E-22
GO Molecular Function	small GTPase regulatory/interacting protein activity	6.7-30.0	67	163	3.07E-03
GO Molecular Function	structural molecule activity	6.7-30.0	235	711	3.41E-05
GO Molecular Function	transferase activity\, transferring phosphorus-containing groups	6.7-30.0	247	750	2.00E-05
GO Molecular Function	transmembrane receptor protein kinase activity	6.7-30.0	42	76	1.54E-05
GO Molecular Function	transmembrane receptor protein phosphatase activity	6.7-30.0	16	19	1.91E-04
GO Molecular Function	transmembrane receptor protein tyrosine kinase activity	6.7-30.0	40	64	2.09E-07
GO Molecular Function	transmembrane receptor protein tyrosine phosphatase activity	6.7-30.0	16	19	1.91E-04
GO Molecular Function	voltage-gated sodium channel activity	6.7-30.0	13	17	2.72E-02
GO Molecular Function	actin binding	2.8-14.5	51	220	2.04E-03

GO Molecular Function	binding	2.8-14.5	841	6657	4.56E-02
GO Molecular Function	calcium-release channel activity	2.8-14.5	6	7	4.19E-02
GO Molecular Function	cytoskeletal protein binding	2.8-14.5	59	293	3.59E-02
GO Molecular Function	extracellular matrix structural constituent	2.8-14.5	28	89	1.19E-03
GO Molecular Function	extracellular matrix structural constituent conferring tensile strength	2.8-14.5	10	15	2.02E-03
GO Molecular Function	motor activity	2.8-14.5	36	113	1.89E-05
GO Molecular Function	structural molecule activity	2.8-14.5	144	711	1.04E-08
SwissProt keyword	Chromosomal protein	88.3-100	22	54	8.46E-05
SwissProt keyword	Cytokine	88.3-100	39	146	5.93E-04
SwissProt keyword	Hormone	88.3-100	24	59	2.08E-05
SwissProt keyword	Myelin	88.3-100	6	7	3.62E-02
SwissProt keyword	Nucleosome core	88.3-100	19	30	4.52E-08
SwissProt keyword	Ribosomal protein	88.3-100	34	88	8.67E-08
SwissProt keyword	GTP-binding	80.6-92.2	42	158	7.24E-04
SwissProt keyword	Lipoprotein	80.6-92.2	68	317	1.56E-03
SwissProt keyword	Prenylation	80.6-92.2	27	91	1.09E-02
SwissProt keyword	Chromosomal protein	76.7-100	31	54	3.31E-04
SwissProt keyword	Cytokine	76.7-100	58	146	3.15E-02
SwissProt keyword	G-protein coupled receptor	76.7-100	109	316	1.81E-02
SwissProt keyword	GTP-binding	76.7-100	72	158	2.78E-06
SwissProt keyword	Hormone	76.7-100	32	59	1.24E-03
SwissProt keyword	Lipoprotein	76.7-100	122	317	3.75E-06
SwissProt keyword	Mitochondrion	76.7-100	144	392	4.95E-06
SwissProt keyword	Nucleosome core	76.7-100	23	30	4.25E-06
SwissProt keyword	Prenylation	76.7-100	46	91	6.54E-05
SwissProt keyword	Ribosomal protein	76.7-100	50	88	6.33E-08
SwissProt keyword	G-protein coupled receptor	72.8-84.4	97	316	2.45E-15
SwissProt keyword	G-protein coupled receptor	68.9-92.2	158	316	2.12E-20
SwissProt keyword	GTP-binding	68.9-92.2	70	158	8.44E-05
SwissProt keyword	Lipoprotein	68.9-92.2	129	317	1.36E-07
SwissProt keyword	Palmitate	68.9-92.2	59	130	3.83E-04
SwissProt keyword	Transmembrane	68.9-92.2	597	2026	1.38E-06
SwissProt keyword	G-protein coupled receptor	65.0-76.7	85	316	2.03E-10
SwissProt keyword	Transmembrane	65.0-76.7	299	2026	8.76E-03
SwissProt keyword	Glycoprotein	61.1-84.4	565	2023	7.24E-03
SwissProt keyword	G-protein coupled receptor	61.1-84.4	158	316	7.13E-21
SwissProt keyword	Transmembrane	61.1-84.4	593	2026	9.32E-07
SwissProt keyword	G-protein coupled receptor	53.4-76.7	126	316	3.99E-07
SwissProt keyword	Ionic channel	53.4-76.7	88	219	1.88E-04
SwissProt keyword	Postsynaptic membrane	53.4-76.7	34	62	5.95E-04
SwissProt keyword	Microsome	45.6-68.9	40	82	3.01E-03
SwissProt keyword	Monooxygenase	45.6-68.9	33	56	8.11E-05
SwissProt keyword	Monooxygenase	37.8-61.1	29	56	2.52E-02
SwissProt keyword	Transcription regulation	37.8-61.1	265	884	4.22E-02

SwissProt keyword	Repeat	33.9-45.6	304	1993	4.72E-04
SwissProt keyword	ATP-binding	30.0-53.4	208	670	4.24E-02
SwissProt keyword	Metalloprotease	30.0-53.4	43	92	6.28E-03
SwissProt keyword	Repeat	30.0-53.4	575	1993	3.39E-05
SwissProt keyword	Serine/threonine-protein kinase	30.0-53.4	84	226	2.27E-02
SwissProt keyword	ATP-binding	26.2-37.8	125	670	2.50E-04
SwissProt keyword	Hydrolase	26.2-37.8	138	779	1.23E-03
SwissProt keyword	Magnesium	26.2-37.8	32	121	2.49E-02
SwissProt keyword	Repeat	26.2-37.8	308	1993	8.83E-05
SwissProt keyword	Tyrosine-protein kinase	26.2-37.8	31	98	4.94E-04
SwissProt keyword	Alternative splicing	22.3-45.6	508	1867	2.70E-02
SwissProt keyword	ATP-binding	22.3-45.6	235	670	2.41E-09
SwissProt keyword	Cell adhesion	22.3-45.6	102	270	2.01E-04
SwissProt keyword	Hydrolase	22.3-45.6	241	779	9.20E-04
SwissProt keyword	Magnesium	22.3-45.6	57	121	3.01E-05
SwissProt keyword	Metalloprotease	22.3-45.6	44	92	8.92E-04
SwissProt keyword	Phorbol-ester binding	22.3-45.6	18	29	3.69E-02
SwissProt keyword	Phosphorylation	22.3-45.6	299	997	7.23E-04
SwissProt keyword	Repeat	22.3-45.6	614	1993	3.06E-15
SwissProt keyword	Tyrosine-protein kinase	22.3-45.6	48	98	9.78E-05
SwissProt keyword	Alternative splicing	18.4-30.0	270	1867	8.49E-04
SwissProt keyword	ATP-binding	18.4-30.0	138	670	2.19E-10
SwissProt keyword	Cell adhesion	18.4-30.0	59	270	6.01E-04
SwissProt keyword	Magnesium	18.4-30.0	34	121	5.92E-04
SwissProt keyword	Repeat	18.4-30.0	346	1993	5.51E-19
SwissProt keyword	Tyrosine-protein kinase	18.4-30.0	33	98	6.34E-06
SwissProt keyword	Alternative splicing	14.5-37.8	541	1867	2.95E-08
SwissProt keyword	ATP-binding	14.5-37.8	271	670	2.37E-22
SwissProt keyword	Cell adhesion	14.5-37.8	106	270	4.05E-06
SwissProt keyword	Coiled coil	14.5-37.8	136	416	1.03E-02
SwissProt keyword	Helicase	14.5-37.8	31	62	1.20E-02
SwissProt keyword	Hydrolase	14.5-37.8	234	779	5.49E-03
SwissProt keyword	Integrin	14.5-37.8	17	25	8.32E-03
SwissProt keyword	Magnesium	14.5-37.8	62	121	4.09E-08
SwissProt keyword	Metalloprotease	14.5-37.8	43	92	1.77E-03
SwissProt keyword	Phosphorylation	14.5-37.8	307	997	4.74E-06
SwissProt keyword	Repeat	14.5-37.8	670	1993	1.70E-32
SwissProt keyword	Tyrosine-protein kinase	14.5-37.8	55	98	5.05E-09
SwissProt keyword	Zinc-finger	14.5-37.8	180	581	1.28E-02
SwissProt keyword	Alternative splicing	10.6-22.3	296	1867	2.12E-09
SwissProt keyword	ATP-binding	10.6-22.3	149	670	1.20E-14
SwissProt keyword	Bromodomain	10.6-22.3	14	27	5.86E-04
SwissProt keyword	Cell adhesion	10.6-22.3	54	270	3.17E-02
SwissProt keyword	Coiled coil	10.6-22.3	88	416	2.74E-06
SwissProt keyword	Connective tissue	10.6-22.3	15	32	1.17E-03

SwissProt keyword	Guanine-nucleotide releasing factor	10.6-22.3	19	52	3.92E-03
SwissProt keyword	Helicase	10.6-22.3	21	62	4.64E-03
SwissProt keyword	Phosphorylation	10.6-22.3	170	997	3.93E-06
SwissProt keyword	Repeat	10.6-22.3	394	1993	2.50E-37
SwissProt keyword	Zinc-finger	10.6-22.3	102	581	3.20E-03
SwissProt keyword	Actin-binding	6.7-30.0	61	136	1.82E-05
SwissProt keyword	Alkylation	6.7-30.0	11	11	2.71E-04
SwissProt keyword	Alternative splicing	6.7-30.0	559	1867	2.43E-14
SwissProt keyword	ATP-binding	6.7-30.0	279	670	1.23E-27
SwissProt keyword	Basement membrane	6.7-30.0	19	22	1.28E-06
SwissProt keyword	Bromodomain	6.7-30.0	22	27	4.61E-07
SwissProt keyword	Calcium channel	6.7-30.0	22	37	4.91E-03
SwissProt keyword	Calmodulin-binding	6.7-30.0	48	86	6.75E-08
SwissProt keyword	Cell adhesion	6.7-30.0	110	270	2.53E-08
SwissProt keyword	Coiled coil	6.7-30.0	189	416	6.10E-23
SwissProt keyword	Collagen	6.7-30.0	32	46	4.56E-08
SwissProt keyword	Connective tissue	6.7-30.0	27	32	6.26E-10
SwissProt keyword	Cytoskeleton	6.7-30.0	72	165	3.39E-06
SwissProt keyword	EGF-like domain	6.7-30.0	55	119	2.81E-05
SwissProt keyword	Extracellular matrix	6.7-30.0	63	124	1.37E-08
SwissProt keyword	Guanine-nucleotide releasing factor	6.7-30.0	29	52	7.06E-04
SwissProt keyword	Helicase	6.7-30.0	35	62	2.48E-05
SwissProt keyword	Hydroxylation	6.7-30.0	32	52	5.62E-06
SwissProt keyword	Laminin EGF-like domain	6.7-30.0	15	17	5.79E-05
SwissProt keyword	Myosin	6.7-30.0	23	32	1.37E-05
SwissProt keyword	Nuclear protein	6.7-30.0	451	1707	3.66E-02
SwissProt keyword	Phosphorylation	6.7-30.0	314	997	5.53E-09
SwissProt keyword	Receptor	6.7-30.0	135	420	7.23E-03
SwissProt keyword	Repeat	6.7-30.0	765	1993	1.59E-75
SwissProt keyword	Structural protein	6.7-30.0	30	52	1.52E-04
SwissProt keyword	Thick filament	6.7-30.0	12	15	1.38E-02
SwissProt keyword	Tyrosine-protein kinase	6.7-30.0	44	98	2.43E-03
SwissProt keyword	Zinc-finger	6.7-30.0	193	581	1.89E-06
SwissProt keyword	Actin-binding	2.8-14.5	36	136	7.95E-04
SwissProt keyword	Alkylation	2.8-14.5	10	11	6.55E-06
SwissProt keyword	Basement membrane	2.8-14.5	14	22	1.24E-05
SwissProt keyword	Bromodomain	2.8-14.5	12	27	2.80E-02
SwissProt keyword	Calcium channel	2.8-14.5	17	37	2.13E-04
SwissProt keyword	Calmodulin-binding	2.8-14.5	24	86	2.88E-02
SwissProt keyword	Coiled coil	2.8-14.5	106	416	2.05E-14
SwissProt keyword	Collagen	2.8-14.5	19	46	3.02E-04
SwissProt keyword	Connective tissue	2.8-14.5	18	32	1.29E-06
SwissProt keyword	Epidermolysis bullosa	2.8-14.5	6	7	3.00E-02
SwissProt keyword	Extracellular matrix	2.8-14.5	36	124	6.05E-05
SwissProt keyword	Hydroxylation	2.8-14.5	20	52	5.72E-04

SwissProt keyword	Laminin EGF-like domain	2.8-14.5	13	17	1.31E-06
SwissProt keyword	Myosin	2.8-14.5	17	32	1.29E-05
SwissProt keyword	Repeat	2.8-14.5	340	1993	9.84E-19
SwissProt keyword	Thick filament	2.8-14.5	9	15	1.65E-02

REFERENCES

Adams, W. T. and T. R. Skopek (1987). "Statistical test for the comparison of samples from mutational spectra." J Mol Biol **194**(3): 391-6.

Ahmadian, A., J. Lundeberg, et al. (2000). "Analysis of the p53 tumor suppressor gene by pyrosequencing." Biotechniques **28**(1): 140-4, 146-7.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Altshuler, D., V. J. Pollara, et al. (2000). "An SNP map of the human genome generated by reduced representation shotgun sequencing." Nature **407**(6803): 513-6.

Alvarez-Valin, F., O. Clay, et al. (2004). "Inaccurate reconstruction of ancestral GC levels creates a "vanishing isochores" effect." Mol Phylogenet Evol **31**(2): 788-93.

Archetti, M. (2004). "Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code." J Mol Evol **59**(2): 258-66.

Association, A. H. (2003). Heart Disease and Stroke Statistics -- 2004 Update. A. H. Association. Dallas, TX, American Heart Association.

Bazykin, G. A., F. A. Kondrashov, et al. (2004). "Positive selection at sites of multiple amino acid replacements since rat-mouse divergence." Nature **429**(6991): 558-62.

Bell, P. A., S. Chaturvedi, et al. (2002). "SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery." Biotechniques **Suppl**: 70-2, 74, 76-7.

Beltrami, C. A., C. Di Loreto, et al. (1997). "DNA Content in End-Stage Heart Failure." Adv Clin Path **1**(1): 59-73.

Bernardi, G. (2001). "Misunderstandings about isochores. Part 1." Gene **276**(1-2): 3-13.

Bobadilla, J. L., M. Macek, Jr., et al. (2002). "Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening." Hum Mutat **19**(6): 575-606.

Braastad, C. D., H. Hovhannisyan, et al. (2004). "Functional characterization of a human histone gene cluster duplication." Gene **342**(1): 35-40.

Brookes, A. J. (1999). "The essence of SNPs." GENE **234**: 177-186.

Buetow, K. H., M. Edmonson, et al. (2001). "High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry." PNAs **98**: 581-584.

Buetow, K. H., M. N. Edmonson, et al. (1999). "Reliable identification of large numbers of candidate SNPs from public EST data." Nat Genet **21**(3): 323-5.

Cargill, M., D. Altschuler, et al. (1999). "Characterization of single-nucleotide polymorphisms in coding regions of human genes." Nat. Genet. **22**: 231-238.

Castresana, J. (2002). "Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content." Nucleic Acids Res **30**(8): 1751-6.

Charney, D. S. and H. K. Manji (2004). "Life stress, genes, and depression: multiple pathways lead to increased risk and new opportunities for intervention." Sci STKE **2004**(225): re5.

Chuang, J. H. and H. Li (2004). "Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome." PLoS Biol **2**(2): E29.

Clark, A. G., S. Glanowski, et al. (2003). "Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios." Science **302**(5652): 1960-3.

Claustres, M., O. Horaitis, et al. (2002). "Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases." Genome Res **12**(5): 680-8.

Clifford, R., M. Edmonson, et al. (2000). "Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project." Genome Res **10**(8): 1259-65.

Cooper, D. N. and M. Krawczak (1990). "The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions." Hum Genet **85**(1): 55-74.

Cooper, D. N. and M. Krawczak (1993). Human Gene Mutation. Oxford, BIOS Scientific Publishers Ltd.

Desai, S. and M. Jessup (2004). "Practice guidelines: role of internists and primary care physicians." Med Clin North Am **88**(5): 1369-80, xiii.

Dickinson, W. J. and J. Seger (1999). "Cause and effect in evolution." Nature **399**(6731): 30.

Ding, Y. C., H. C. Chi, et al. (2002). "Evidence of positive selection acting at the human dopamine receptor D4 gene locus." Proc Natl Acad Sci U S A **99**(1): 309-14.

Duret, L., M. Semon, et al. (2002). "Vanishing GC-rich isochores in mammalian genomes." Genetics **162**(4): 1837-47.

Emanueli, C., R. Maestri, et al. (1999). "Dilated and failing cardiomyopathy in bradykinin B(2) receptor knockout mice." Circulation **100**(23): 2359-65.

Epstein, R. J., K. Lin, et al. (2000). "A functional significance for codon third bases." Gene **245**(2): 291-8.

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-94.

Ewing, B., L. Hillier, et al. (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-85.

Eyre-Walker, A. and L. D. Hurst (2001). "The evolution of isochores." Nat Rev Genet **2**(7): 549-55.

Eyre-Walker, A. and P. D. Keightley (1999). "High genomic deleterious mutation rates in hominids." Nature **397**(6717): 344-7.

Fay, J. C., G. J. Wyckoff, et al. (2001). "Positive and negative selection on the human genome." Genetics **158**(3): 1227-34.

Fay, J. C., G. J. Wyckoff, et al. (2002). "Testing the neutral theory of molecular evolution with genomic data from Drosophila." Nature **415**(6875): 1024-6.

Fentzke, R. C., C. E. Korcarz, et al. (1998). "Dilated cardiomyopathy in transgenic mice expressing a dominant-negative CREB transcription factor in the heart." J Clin Invest **101**(11): 2415-26.

Flood, E. M., F. Tang, et al. (2002). "SNPCEQer: detecting SNPs in sequences generated by the Beckman CEQ2000 DNA Analysis System." Biotechniques **33**(4): 814, 816, 818-20 passim.

Force, T., K. Kuida, et al. (2004). "Inhibitors of protein kinase signaling pathways: emerging therapies for cardiovascular disease." Circulation **109**(10): 1196-205.

Fraser, H. B., A. E. Hirsh, et al. (2002). "Evolutionary rate in the protein interaction network." Science **296**(5568): 750-2.

Freeland, S. J. (2002). "The Darwinian Genetic Code: An Adaptation for Adapting?" Genetic Programming and Evolvable Machines **3**: 113-127.

Galindo, B. E., V. D. Vacquier, et al. (2003). "Positive selection in the egg receptor for abalone sperm lysin." Proc Natl Acad Sci U S A **100**(8): 4639-43.

Gasteiger, E., E. Jung, et al. (2001). "SWISS-PROT: connecting biomolecular knowledge via a protein database." Curr Issues Mol Biol **3**(3): 47-55.

Geer, R. C. and E. W. Sayers (2003). "Entrez: making use of its power." Brief Bioinform **4**(2): 179-84.

Glusman, G., I. Yanai, et al. (2001). "The complete human olfactory subgenome." Genome Res **11**(5): 685-702.

Gordon, D., C. Abajian, et al. (1998). "Consed: a graphical tool for sequence finishing." Genome Res **8**(3): 195-202.

Griffin, T. J. and L. M. Smith (2000). "Single-nucleotide polymorphism analysis by MALDI-TOF mass spectrometry." Trends Biotechnol **18**(2): 77-84.

Gu, Y. H., H. Kodama, et al. (2001). "ATP7A gene mutations in 16 patients with Menkes disease and a patient with occipital horn syndrome." Am J Med Genet **99**(3): 217-22.

Guo, B. (1999). "Mass Spectrometry in DNA Analysis." Anal. Chem. **71**: 333R-337R.

Hagmann, M. (1999). "A Good SNP May Be Hard to Find." Science **285**: 21-22.

Halapi, E. and H. Hakonarson (2004). "Recent development in genomic and proteomic research for asthma." Curr Opin Pulm Med **10**(1): 22-30.

Halushka, M. K., J. B. Fan, et al. (1999). "Patterns of single-nucleotide polymorphisms in candidate genes for blood pressure homeostatsis." Nat. Genet. **22**: 239-247.

Harris, M. A., J. Clark, et al. (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Res **32 Database issue**: D258-61.

Helgadottir, A., A. Manolescu, et al. (2004). "The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke." Nat Genet **36**(3): 233-9.

Henikoff, S. and J. G. Henikoff (1993). "Performance evaluation of amino acid substitution matrices." Proteins **17**(1): 49-61.

Hermjakob, H., F. Lang, et al. (1998). SPTR - A comprehensive, non-redundant and up-to-date view of the protein sequence world. CCP11 newsletter. Oxon, UK. **2.3**.

Herrmann, S., K. Schmidt-Petersen, et al. (2001). "A polymorphism in the endothelin-A receptor gene predicts survival in patients with idiopathic dilated cardiomyopathy." Eur Heart J **22**(20): 1948-53.

Hirakawa, M., T. Tanaka, et al. (2002). "JSNP: a database of common gene variations in the Japanese population." Nucleic Acids Res **30**(1): 158-62.

Horvath, M. M., J. W. Fondon, 3rd, et al. (2003). "Low hanging fruit: a subset of human cSNPs is both highly non-uniform and predictable." Gene **312**: 197-206.

Hosack, D. A., G. Dennis, Jr., et al. (2003). "Identifying biological themes within lists of genes with EASE." Genome Biol **4**(10): R70.

Jackson, P. E., P. F. Scholl, et al. (2000). "Mass spectrometry for genotyping: an emerging tool for molecular medicine." Molec. Medicine Today **6**: 271-276.

Johnson, J. A. and J. J. Lima (2003). "Drug receptor/effector polymorphisms and pharmacogenetics: current status and challenges." Pharmacogenetics **13**(9): 525-34.

Katzmarzyk, P. T., L. Perusse, et al. (2000). "Familial resemblance for coronary heart disease risk: the HERITAGE Family Study." Ethn Dis **10**(2): 138-47.

Kimura, M. (1987). "Molecular evolutionary clock and the neutral theory." J Mol Evol **26**(1-2): 24-33.

Krawczak, M., E. V. Ball, et al. (1998). "Neighboring-Nucleotide Effects on the Rates of Germ-Line Single Base Pair Substitution in Human Genes." Am. J. Hum. Genet. **63**: 474-488.

Krawczak, M., E. V. Ball, et al. (2000). "Human Gene Mutation Database--A Biomedical Information and Research Resource." Human Mutation **15**: 45-51.

Krawczak, M. and D. N. Cooper (1991). "Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment." Hum Genet **86**(5): 425-41.

Ku, L., J. Feiger, et al. (2003). "Cardiology patient page. Familial dilated cardiomyopathy." Circulation **108**(17): e118-21.

Kumar, S. and S. Subramanian (2002). "Mutation rates in mammalian genomes." Proc Natl Acad Sci U S A **99**(2): 803-8.

Lee, S., D. C. Russo, et al. (2001). "Molecular defects underlying the Kell null phenotype." J Biol Chem **276**(29): 27281-9.

Lenski, R. E. and J. E. Mittler (1993). "The directed mutation controversy and neo-Darwinism." Science **259**(5092): 188-94.

Lercher, M. J. and L. D. Hurst (2002). "Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)?" Gene **300**(1-2): 53-8.

Lercher, M. J., A. O. Urrutia, et al. (2003). "A unification of mosaic structures in the human genome." Hum Mol Genet **12**(19): 2411-5.

Lunter, G. and J. Hein (2004). "A nucleotide substitution model with nearest-neighbour interactions." Bioinformatics **20 Suppl 1**: I216-I223.

Majewski, J. and J. Ott (2003). "Amino acid substitutions in the human genome: evolutionary implications of single nucleotide polymorphisms." Gene **305**(2): 167-73.

Marth, G., G. Schuler, et al. (2003). "Sequence variations in the public human genome data reflect a bottlenecked population history." Proc Natl Acad Sci U S A **100**(1): 376-81.

McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature **351**(6328): 652-4.

Morimura, H., G. A. Fishman, et al. (1998). "Mutations in the RPE65 gene in patients with autosomal recessive retinitis pigmentosa or leber congenital amaurosis." Proc Natl Acad Sci U S A **95**(6): 3088-93.

Mullikin, J. C., S. E. Hunt, et al. (2000). "An SNP map of human chromosome 22." Nature **407**(6803): 516-20.

Nakamura, Y., T. Gojobori, et al. (2000). "Codon usage tabulated from international DNA sequence databases: status for the year 2000." Nucleic Acids Res **28**(1): 292.

Nelson, M. R., G. Marnellos, et al. (2004). "Large-scale validation of single nucleotide polymorphisms in gene regions." Genome Res **14**(8): 1664-8.

Ng, P. C. and S. Henikoff (2002). "Accounting for human polymorphisms predicted to affect protein function." Genome Res **12**(3): 436-46.

Nickerson, D. A., V. O. Tobe, et al. (1997). "PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing." Nucleic Acids Res **25**(14): 2745-51.

Ogawa, T., K. Nakashima, et al. (1996). "Accelerated evolution of snake venom phospholipase A2 isozymes for acquisition of diverse physiological functions." Toxicon **34**(11-12): 1229-36.

Ohnishi, Y., T. Tanaka, et al. (2000). "Identification of 187 single nucleotide polymorphisms (SNPs) among 41 candidate genes for ischemic heart disease in the Japanese population." Hum Genet **106**(3): 288-92.

Pruitt, K. D. and D. R. Maglott (2001). "RefSeq and LocusLink: NCBI gene-centered resources." Nucleic Acids Res **29**(1): 137-40.

Rajamannan, N. M., M. Subramaniam, et al. (2003). "Human aortic valve calcification is associated with an osteoblast phenotype." Circulation **107**(17): 2181-4.

Repka-Ramirez, M. S. (2003). "New concepts of histamine receptors and actions." Curr Allergy Asthma Rep **3**(3): 227-31.

Riggins, G. J. and R. L. Strausberg (2001). "Genome and genetic resources from the Cancer Genome Anatomy Project." Hum Mol Genet **10**(7): 663-7.

Rogozin, I., F. Kondrashov, et al. (2001). "Use of mutation spectra analysis software." Hum Mutat **17**(2): 83-102.

Rogozin, I. B., Y. I. Pavlov, et al. (2001). "Somatic mutation hotspots correlate with DNA polymerase eta error spectrum." Nat Immunol **2**(6): 530-6.

Ross, P., L. Hall, et al. (1998). "High level multiplex genotyping by MALDI-TOF mass spectrometry." Nat. Biotech. **16**: 1347-1351.

Rothermel, B. A., T. A. McKinsey, et al. (2001). "Myocyte-enriched calcineurin-interacting protein, MCIP1, inhibits cardiac hypertrophy in vivo." Proc Natl Acad Sci U S A **98**(6): 3328-33.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-86.

Rutter, M. T. and R. A. Zufall (2004). "Pathway length and evolutionary constraint in amino acid biosynthesis." J Mol Evol **58**(2): 218-24.

Sachidanandam, R., D. Weissman, et al. (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature **409**(6822): 928-33.

Shapiro, G. S., K. Aviszus, et al. (1999). "Predicting regional mutability in antibody V genes based solely on di- and trinucleotide sequence composition." J Immunol **163**(1): 259-68.

Shapiro, G. S., K. Aviszus, et al. (2002). "Evolution of Ig DNA sequence to target specific base positions within codons for somatic hypermutation." J Immunol **168**(5): 2302-6.

Sherry, S. T., M. H. Ward, et al. (2001). "dbSNP: the NCBI database of genetic variation." Nucleic Acids Res **29**(1): 308-11.

Singer, G. A. and D. A. Hickey (2000). "Nucleotide bias causes a genomewide bias in the amino acid composition of proteins." Mol Biol Evol **17**(11): 1581-8.

Smigielski, E. M., K. Sirotkin, et al. (2000). "dbSNP: A database of single nucleotide polymorphisms." Nucleic Acids Res **28**(1): 352-5.

Stefansson, H., V. Steinthorsdottir, et al. (2004). "Neuregulin 1 and schizophrenia." Ann Med **36**(1): 62-71.

Sunyaev, S., J. Hanke, et al. (1999). "Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes." J. Mol. Med. **77**: 754-760.

Tang, K., D.-J. Fu, et al. (1999). "Chip-based genotyping by mass spectrometry." Proc. Natl. Acad. Sci. USA **96**: 10016-10020.

TheHapMapConsortium (2003). "The International HapMap Project." Nature **426**(6968): 789-96.

Todorova, A. and G. A. Danieli (1997). "Large majority of single-nucleotide mutations along the dystrophin gene can be explained by more than one mechanism of mutagenesis." Hum Mutat **9**(6): 537-47.

Visser, R., O. Shimokawa, et al. (2004). "Identification of a 3.0-kb Major Recombination Hotspot in Patients with Sotos Syndrome Who Carry a Common 1.9-Mb Microdeletion." Am J Hum Genet **76**(1).

Wang, R., S. J. Shattil, et al. (1997). "Truncation of the cytoplasmic domain of beta3 in a variant form of Glanzmann thrombasthenia abrogates signaling through the integrin alpha(IIb)beta3 complex." J Clin Invest **100**(9): 2393-403.

Weil, M. R., P. Widlak, et al. (2004). "Global survey of chromatin accessibility using DNA microarrays." Genome Res **14**(7): 1374-81.

Wilbur, W. J. (1985). "On the PAM matrix model of protein evolution." Mol Biol Evol **2**(5): 434-47.

Wright, A. F., A. D. Caraothers, et al. (1999). "Population choice in mapping genes for complex diseases." Nat. Genet. **23**: 397-404.

Yamada, T., K. I. Mizuno, et al. (2004). "Roles of histone acetylation and chromatin remodeling factor in a meiotic recombination hotspot." Embo J **23**(8): 1792-1803.

Yang, J., B. Rothermel, et al. (2000). "Independent signals control expression of the calcineurin inhibitory proteins MCIP1 and MCIP2 in striated muscles." Circ Res **87**(12): E61-8.

Yeadon, P. J., F. J. Bowring, et al. (2004). "Alleles of the hotspot cog are codominant in effect on recombination in the his-3 region of Neurospora." Genetics **167**(3): 1143-53.

You, Y. H., C. Li, et al. (1999). "Involvement of 5-methylcytosine in sunlight-induced mutagenesis." J Mol Biol **293**(3): 493-503.

Ziche, M., S. Donnini, et al. (2004). "Development of new drugs in angiogenesis." Curr Drug Targets **5**(5): 485-93.

VITAE

Monica Marie Horvath was born to Grace Patton-Horvath and Thomas Horvath on September 11th, 1977 in Jackson, Michigan. Raised in Elizabeth, Pennsylvania, she matriculated into the University of Pittsburgh and discovered a passion for biology through her work in macromolecular X-ray crystallography with Dr. John Rosenberg. She completed a B.S. in Chemistry in May 1999 (summa cum laude), married computer game designer Joshua Jay on July 31st, 1999, and moved to Dallas, Texas to attend UT Southwestern to pursue a doctoral degree in Molecular Biophysics. She currently resides in Apex, North Carolina with her husband, four dogs: Xerxes Barkimedes (Xerx), William Waddlebottom (Bill), Zootyfruit (Zoot), and Merry Ploppins (Plop), and a naughty cat named Chairman Meow (aka Gretyl).

Permanent address: 1306 Eastham Drive
Apex, NC 27502