Most of rare missense alleles in humans are deleterious: implications for evolution of complex disease and association studies

Gregory V. Kryukov, Len A. Pennacchio, Shamil R. Sunyaev

From Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard, Medical School, New Research Building, 77, Avenue Louis Pasteur, Boston, MA (G.V.K., S.R.S.) and

Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA (L.A.P.)

Address for correspondence and reprints: Shamil Sunyaev, Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School New Research Building, 77, Avenue Louis Pasteur, Boston, MA 02125. E-mail: ssunyaev@rics.bwh.harvard.edu

Tel: 617 525-4735, Fax: 617 525-4705

Short running title: Rare missense alleles

Abstract

The accumulation of mildly deleterious missense mutations in individual human genomes has been proposed to be a genetic basis for complex diseases. The plausibility of this hypothesis depends on quantitative estimates of the prevalence of mildly deleterious *de novo* mutations and polymorphic variants in humans and on the intensity of selective pressure against them. We combined analysis of mutations causing human Mendelian diseases, human-chimpanzee divergence and systematic data on human SNPs and found that about 20% of new missense mutations in humans result in a loss of function, while about 27% are effectively neutral. Thus, more than half of new missense mutations have mildly deleterious effects. These mutations give rise to many low frequency deleterious allelic variants in the human population as evident from a new dataset of 37 genes sequenced in over 1,500 individual human chromosomes. Surprisingly, up to 70% of low frequency missense alleles are mildly deleterious and associated with a heterozygous fitness loss in the range 0.001-0.003. Thus, the low allele frequency of an amino acid variant can by itself serve as a predictor of its functional significance. Several recent studies have reported a significant excess of rare missense variants in disease populations compared to controls in candidate genes or pathways. These studies would be unlikely to work if most rare variants were neutral or if rare variants were not a significant contributor to the genetic component of phenotypic inheritance. Our results provide a justification for these types of candidate gene (pathway) association studies and imply that mutation-selection balance may be a feasible mechanism for evolution of some common diseases.

Introduction

Many common human diseases have a strong heritable component. While modern human genetics has been incredibly successful in determining the genetic causes of rare Mendelian diseases, complex diseases have proven to be a more challenging problem[1,2]. Genetic variation that influences an individual's susceptibility to most such diseases is still largely unidentified. Of special interest are missense mutations, as many of them are believed to have non-marginal functional effects[1]. The effects of a missense mutation on molecular function, phenotype and organism fitness can be extremely diverse. A missense mutation can be lethal or cause severe Mendelian disease; alternatively, it can be mildly deleterious, effectively neutral or beneficial. Knowledge of relative frequencies of these types of mutations and their contributions to population genetic variation is important for understanding the evolutionary background of common disease and can inform design of human genetic studies.

The effect of a missense mutation on organism is always multifaceted and can be considered from multiple perspectives - biochemical, medical or evolutionary. The relationship between the effects of amino acid substitution on protein activity, human health and an individual's evolutionary fitness is not trivial. A mutation that damages protein structure does not necessarily lead to a detectable human disease phenotype, and a mutation that predisposes an individual towards a disease is not necessarily evolutionary deleterious. Fixation of completely deactivating mutations (pseudogenezation) was apparently a common event during recent human evolution[3], indicating that a mutation abolishing protein activity is not necessarily subject to purifying selection. Substitutions leading to abnormal hemoglobin function that cause sickle cell anemia

are apparently negative from both biochemical and medical points of view. Nevertheless, they can not be considered negative from an evolutionary point of view since balancing selection has brought them to high frequency in many parts of the world. To clearly distinguish different aspects of negative mutations we will use the term *damaging* to refer to a mutation that decreases protein activity, the term *detrimental* to refer to a mutation that predisposes an individual towards a disease, and the term *deleterious* to refer to a mutation that has been subject to purifying selection.

A high incidence rate for many complex diseases suggests that a surprisingly high cumulative frequency of medically detrimental variants should be present in the human population. It remains uncertain why such polymorphism may persist without being eliminated by purifying selection. Currently, two major lines of reasoning exist that explain this apparent paradox. The first considers various complex evolutionary scenarios and treats positive or balancing selection as a major force that can drive medically detrimental mutations to high frequencies. The second line of reasoning postulates a high mutation rate as a major factor that determines the cumulative frequency of detrimental polymorphism in population.

The first hypothesis states that the majority of polymorphism predisposing an individual to a complex disease, though *medically* detrimental at the present time, were not *evolutionarily* deleterious. There are a number of possible phenomena that might help a polymorphic variant which confers a susceptibility to a disease phenotype escape purifying selection. One of them is a late disease onset, when detrimental phenotypic consequences of a mutation strike after the reproductive age, and thus, do not affect the individual's number of offspring. A number of

earlier studies show, however, that this phenomenon is unlikely to be common. These studies suggest that most of the mutations that affect phenotypes at old age are likely to have a small pleiotropic effect earlier in life[4,5]. Changing direction of selection is another mechanism that can explain how presently detrimental mutations could have escaped purifying selection[6]. Human lifestyle, environment and nutrition have changed dramatically and some mutations that were neutral or even beneficial in hunter-gatherer societies tens of thousands of years ago might have become medically detrimental in modern human society. The best known example of this type of reasoning is a "thrifty genes" hypothesis[7] that postulates that polymorphisms that predispose modern humans to obesity and are presently medically detrimental were able to rise to high frequency in population due to associated selective advantages at the times of scarce food sources. Yet another mechanism, balancing selection, can also maintain deleterious mutations in a population if heterozygous individuals have a strong evolutionary advantage[8]. A classic example for this mechanism is the hemoglobin mutation that is protective against malaria in a heterozygote state and simultaneously leads to sickle cell anemia in a homozygous state[9]. The fourth phenomenon that can lead to a high frequency of medically detrimental mutation is antagonistic pleiotropy, when the negative effect of mutation on one trait is compensated by its positive effect on another[10]. None of the these evolutionary scenarious, however, have been shown to be frequent enough to account for a large number of human common complex diseases.

The second theory postulates that the majority of medically detrimental polymorphism are also mildly evolutionary deleterious and the observed frequency of disease-predisposing genetic variation is the result of mutation-selection balance. The key assumption here is that majority of disease-causing mutations are both medically detrimental *and* evolutionary deleterious, but the

pressure of purifying selection acting upon them is reasonably weak. A high rate of mildly deleterious mutations associated with disease risk counterbalances the action of purifying selection.

Mutation selection balance was, initially, theoretically treated by Kimura[11]. It was, then, both advocated[12] and argued against[13] as a mechanism capable of maintaining high level of genetic variance in natural populations. In the last decade, partially due to popularity of common disease/ common variant (CD/CV) hypothesis, mutation selection balance was frequently overlooked as a mechanism that can explain existence of many common, harmful, heritable disorders. Recently, however, mutation selection balance has been put forward, again, as a plausible mechanism of the maintenance of the deleterious genetic variation[14,15,16.]

Obviously, mutation-selection balance can only become a feasible evolutionary explanation for common disease if a sufficient fraction of *de novo* mutations in humans are mildly deleterious. The estimation of the fraction of mildly deleterious missense mutations and corresponding fraction among low frequency human polymorphism is the subject of this work.

The evolutionary origin of present-day detrimental polymorphism has an important implication for the spectrum of disease-predisposing alleles. If the majority of medically detrimental mutations were not evolutionarily deleterious, they might have risen to a high frequency in the population, and thus the common disease/common variant hypothesis is viable. On the contrary, if the majority of medically detrimental polymorphism are mildly deleterious and their cumulative high frequency in the population is being maintained by mutation/selection balance,

then the common disease/rare variant alternative is likely to hold true. Given the high incidence rate of many complex diseases, the mutation/selection balance hypothesis can be feasible only if a large fraction of *de novo* mutations is associated with moderate selection coefficients, so that, despite being deleterious, they can still achieve detectable frequencies in the human population.

Earlier theoretical studies in the framework of a mutation-selection model of common disease focused on the frequency spectrum of susceptibility alleles. Pritchard[17] and Pritchard & Cox[18] argued that low frequency alleles are major contributors to common disease, whereas Reich & Lander[19] advocated the common disease-common variant hypothesis. Pritchard[17] suggested that the rate of susceptibility mutations is high and that they are under pressure of weak purifying selection, leading to an abundance of rare variants. Reich & Lander[19] did not focus on the question of how an individual deleterious variant could reach high population frequency. Given the high frequency of the phenotype, they considered allelic identity in the disease class, which depended on the rate of mutations conferring disease susceptibility. Thus, both models depend on the rate of deleterious mutations involved in disease and on strength of selection against new mutations. Both studies used estimates of mutation rate based on examples from a few loci and mutations causing fully penetrant Mendelian phenotypes; mutations involved in complex disease, on the other hand, may have smaller effects. The difference in conclusions is mostly explained by the difference in numerical values of the rate of deleterious mutations and strength of selection acting on them. It should be stressed that the deleterious mutation rate depends not only on the raw per nucleotide substitution rate but also on fraction of *de novo* mutations that are deleterious. Reliable estimation of these parameters requires integration of several types of data.

We combined in our analysis data on human Mendelian disease-causing mutations, human-chimpanzee divergence and genetic variation in the modern human population. The availability of new data on human polymorphism detected in 37 obesity related genes sequenced in 756 individuals allowed us to investigate the important class of mildly deleterious mutations. Purifying selection acting on this type of mutation is strong enough to effectively prohibit their fixation, but they are present in the population at low frequencies and can be detected if a large number of individuals is sequenced.

We estimated that more than 50% of *de novo* missense mutations and 70% of missense SNPs detected only once among 1,500 chromosomes are mildly deleterious. Such mildly deleterious amino acid substitutions are associated with selection coefficients within a surprisingly narrow range of 0.001-0.003.

A high fraction of mildly deleterious mutations among missense mutations suggests that mutation-selection balance can be a possible explanation for the existence of common disease with complex inheritance. The observation that the majority of human rare non-synonymous variants are deleterious and thus, of significance to function and phenotype, strongly supports a resequencing strategy for candidate gene association studies: a disease population is expected to have a higher rate of rare amino acid variants than healthy controls in genes involved in disease.

Data and Methods

Mutations associated with Mendelian diseases

For information on strongly detrimental mutations we used Human Gene Mutation Database (HGMD)[20] that contains over 50,000 disease-causing mutations of various types, including missense, nonsense and splice-sites mutations. Disease-associated polymorphism comprise only very small fraction of HGMD and the majority of mutations included in the database are fully penetrant and cause simple Mendelian diseases[20]. HGMD lacks information on how many times each individual mutation has been identified. However, assuming that mutations follow the Poisson statistics, presence of mutations detected multiple times would not significantly affect our estimates unless majority of all possible nonsense mutations were detected, which is highly unlikely for most of genes.

HGMD may include genes with complete loss of function mutations being embryonically lethal. It contains some gain of function mutations and mutations with incomplete penetrance. Although fraction of these mutations in HGMD is probably small, their presence can cause bias of our estimates. Thus, we repeated our analysis on a smaller but well characterized set of autosomal dominant or X-linked disease genes originally collected by Kondrashov from locus specific databases for estimating mutation rate in humans[21]. We extracted information on missense and nonsense mutations in these genes from individual HGMD entries. To ensure that complete loss of function is not lethal we restricted our analysis to 26 genes that had at least 5 reported nonsense mutations. All missense mutations in these genes are believed to be loss-of-function mutations.

Human polymorphism data – new large resequencing dataset

For our analysis of very low frequency nonsynonymous SNPs we used a new large resequencing dataset described in the accompanying manuscript by Ahituv et al. Complete exonic sequences and their splice sites were sequenced in 58 genes with potential involvement in obesity in 379 obese and 378 lean individuals. Complete sequencing of more that 1,500 chromosomes provided us with an opportunity to study nonsynonymous SNPs at very low frequencies. All individuals included in the study are of Caucasian ancestry.

Since this dataset is not based on a random population sample and is phenotipically biased, we limited our analysis to 37 autosomal genes for which no evidence of their effect on obesity was detected. In these genes there was no statistically significant excess of variation in either obese or lean group. 71 and 79 missense variants were found within lean and obese cohorts respectively. Random resampling subsets matching the SeattleSNP dataset in size produced estimates highly similar to obtained from smaller systematic datasets, that further supported absence of bias in chosen 37 genes subset.

Human polymorphism data – publicly available data

In addition to our new resequencing dataset we used three publicly available datasets that contain sufficient information on rare nonsynonymous alleles: 1) dataset generated by Environmental Genome Project (NIEHS-EGP)[22] in which over 500 genes involved in DNA repair and cell cycle pathways were sequenced in at least 90 unrelated individuals; 2) dataset generated by SeattleSNPs project (SeattleSNPs) in which over 200 genes involved in the inflammatory

responses were sequenced in at least 46 individuals; and 3) dataset generated by Japanese Single Nucleotide Polymorphisms (JSNP)[23] project in which polymorphic sites in over 8,000 genes were discovered using panel of 12 individuals and later genotyped in 750 individuals.

Mutations fixed in the human lineage after divergence from chimpanzee

To obtain information on mutations fixed in the human lineage after divergence from chimpanzee we used human/chimpanzee and human/macaque whole genome pairwise alignments constructed with blastz program. These alignments were obtained from UCSC ftp site. We compared all nucleotide substitutions detected between human and chimp whole genome assemblies to dbSNP database[24] that contain information on most of known human genetic variation. Substitutions that were detected as present day polymorphisms were excluded from our analysis of human/chimpanzee divergence. Macaque sequence was used as an outgroup to distinguish mutations fixed in the human lineage versus mutations fixed in the chimpanzee lineage. Accordingly, we ignored sites at which macaque nucleotide was different from both human and chimpanzee nucleotides. Due to this procedure, we potentially excluded a small fraction of very rapidly evolving sites, thus slightly underestimating the fraction of "effectively neutral" amino acid substitutions.

Context-dependent mutation model

Accurate estimation of selective constraints in protein coding regions requires a context-dependent mutational model[25]. We constructed a context-dependent mutation rate matrix that described the probability of *de novo* mutation occurrence, before any selection has taken place. With such a mutability model we were able to compare the observed number of various

substitutions with the corresponding number expected under neutrality and, from such comparison, to draw a conclusion on the strength of purifying selection.

We restricted our analysis to single nucleotide substitutions, excluding other less common types of mutations. The real mutational spectrum is quite complex – CpG dinucleotides have mutation rate an order of magnitude higher than other dinucleotides, transitions are notably more frequent than transversions and some other, more subtle, context-dependent effects were reported to exist[26]. Various mutation models of different degrees of sophistication have been previously used[27]. The availability of human, chimpanzee and (incomplete) baboon genomic sequences allowed us to calculate an empirical "directed" 64x3 mutation matrix for triplets (with probabilities for all XY_1Z to XY_2Z single nucleotide substitutions). Such mutational model that takes into account nucleotide, its closest neighbors and direction of substitution should capture most of the known fine-scale mutation rate context dependencies. We used a second-order model of dependence on neighboring positions in order to capture all context-dependent effects on mutation rate. We also introduced statistical correction for back substitutions in the baboon lineage.

We made alignments of human, chimpanzee and baboon sequence taken from ENCODE regions[28] using the multiple sequence aligner TBA[29]. The frequencies of all nucleotide triples in the aligned sequences were counted (ignoring any triples containing gaps). Directionality of mutation (whether a substitution has occurred in human or in chimpanzee lineage) was determined using baboon sequence.

Calculated matrix was in good agreement with other recently published context-dependent models of neutral evolution[30-32]. For example, r^2 correlation coefficient between our matrix and matrix from Siepel and Haussler[30] is 96%)

Gene sets

Population and evolutionary dynamics of genes located on sex chromosomes differ from the rest of genome. To avoid unnecessary complications in our analysis we used only autosomal genes.

To calculate nonsynonymous to synonymous mutations ratio characteristic for the entire human genome we used 14095 reliably annotated autosomal genes from "Consensus CDS" project.

Confidence of the estimates

Analysis of human-chimpanzee divergence is based on the complete proteome and the analysis of the HGMD database incorporates the most comprehensive set of disease genes. However, human SNP data are represented by much smaller gene sets. Therefore, it is necessary to assign confidence to statistical estimates obtained from these sets. Even though these datasets have large numbers of SNPs, these mutations were found in different genes possessing different properties. Thus, sampling error of the estimates is expected to be mostly determined by sampling of genes rather than individual mutations.

We computed standard errors from standard deviation of estimates obtained from random non-overlapping subsets of the data. Although standard errors are mostly used for estimates of mean,

we tested in a series of simulations that standard errors computed using non-overlapping subsets satisfactory approximate the error of the estimates presented in this work.

Results

Strongly detrimental mutations

As a first step in our analysis of spectrum of potential effects of amino acid mutations we estimated the fraction of *de novo* missenses that are strongly detrimental. We define a missense mutation as strongly detrimental if it causes complete protein function loss often seen in Mendelian diseases.

Nonsense mutations that introduce premature stop codons can serve as a standard of "strong detrimentality". 26,305 missense and 6,764 nonsense mutations are listed in The Human Gene Mutation Database. Although HGMD has an inherent ascertainment bias, following[21,33], we believe that mutations causing the same phenotype are equally likely to be deposited in the database, regardless of their type – nonsense, missense or splice-site. If all missense mutations were as likely to result in complete loss of function, and subsequently, in strong disease phenotype as nonsense mutations, then the ratio of missense to nonsense mutations in HGMD database would be similar to the expected theoretical ratio for *de novo* mutations. In reality, it is significantly lower. Using our mutation rate model described above we estimated that approximately 19.7 missense substitutions occur per each nonsense substitution genome-wide (14,095 human genes from "Consensus CDS" project were used for calculation of genome-wide values). At the same time, only 3.9 missenses were observed per each nonsense substitution

among disease-causing mutations in HGMD database (Figure 1A). Such a difference indicates that a large fraction of missense mutations is not deleterious enough to be visible through such a "strong phenotype filter" as detection and inclusion in HGMD. Using the obtained numbers we can estimate that only 20% (3.9/19.7) of missense mutations are strongly detrimental.

Mutations that disrupt dinucleotides of the core splicing sites consensus (GT/AG) often lead to an almost complete loss of gene function. These splice site substitutions can be used as "reference" strongly detrimental mutations the same way as we used nonsense mutations. We calculated that genome-wide approximately 36.5 *de novo* missense mutations are expected to occur per each *de novo* splice site mutation. However, in HGMD only 7.6 missense mutations are listed per each mutation in the splice site (Figure 1B). Accordingly, we can estimate that the fraction of strongly deleterious mutations among missenses is approximately 21%. This number is in remarkable correspondence with the value obtained using nonsense mutations as a reference.

Although HGMD provides a comprehensive set of disease mutations, it is very heterogeneous and includes mutations of incomplete penetrance, gain of function mutations and possibly mutations in genes with embryonically lethal complete loss of function. Presence of mutations in two latter categories would bias our estimate upward, i.e. would lead to a conservative estimate for the purpose of this work. However, to test whether the upward bias can be significant, we analyzed a much smaller set of very well characterized genes involved in autosomal dominant or X-linked simple Mendelian diseases. All mutations in this set are believed to lead to loss of function and complete function loss is not lethal. After taking into account individual genes

lengths, 21% of new missense mutations in the set were estimated to be strongly detrimental.

Effectively neutral mutations

As the second step in our analysis of spectrum of potential effects of amino acid substitutions we estimated a fraction of *de novo* missense mutations that are effectively neutral. We define a missense mutation as effectively neutral if its probability to be fixed in the ancestral human population after divergence from chimpanzee would have been similar to that of synonymous substitutions. While synonymous substitutions were shown not to be completely selectively neutral[34,35,36] (reviewed in Chamary et al.[37]), the effect on fitness of the vast majority of them in the human population is believed to be relatively small.

We estimated (see Figure 1C) that 2.23 *de novo* missense mutations occur per each synonymous mutation. However, among substitutions fixed in the human lineage after divergence from chimpanzee (calculated with the use of macaque genomic sequence as an outgroup) only 0.60 missense mutations were present per a single synonymous substitution. The rest of missense mutations were apparently eliminated by purifying selection. From these values, we can estimate that approximately 27% (0.6/2.23) of missense mutations in human proteins are similar in effect to synonymous substitutions.

This estimate is sensitive to advantageous mutations driven to fixation by positive selection and to very slightly deleterious mutations fixed by drift. Presence of these mutations would bias our estimate of fraction of mildly deleterious mutations downward, i.e. our estimate is conservative for our purpose. It is also conservative with respect to purifying selection at synonymous sites.

Thus, this estimate can be viewed as an estimate of fraction of new missense mutations which are not associated with fitness loss larger than the reciprocal of effective population size.

Mildly deleterious polymorphisms

We analyzed two extremes of the potential effects of amino acid changes on fitness and estimated that, among *de novo* missense mutations in human proteins, approximately 20% are strongly detrimental while another 27% are effectively neutral. Simple arithmetic provides us with the conclusion that the majority – 53% - of all *de novo* missense mutations are, in fact, mildly deleterious. Mildly deleterious mutations can reach low, but detectable population frequencies. We next posed a question – if fraction of mildly deleterious mutations among *de novo* mutations is 53%, what is the fraction of mildly deleterious alleles among rare and common nonsynonymous SNPs in the human population?

First, we considered very rare missense polymorphisms. We analyzed only polymorphic sites at which at least 1400 chromosomes out of 1514 have been successfully sequenced. Over 60% of nonsynonymous SNPs discovered in our set of 37 autosomal genes sequenced in 757 individuals were detected as singletons (i.e. were found in heterozygote state in a single individual). Even though singletons represent very low frequency SNPs, probability that they are *de novo* mutations is diminishingly small. The number of nsSNPs detected at higher frequencies was significantly smaller and, thus, singletons represented the only group of rare missense alleles that had enough data for a statistically significant analysis.

We estimated the relative fraction of *de novo* missense mutations represented by singletons. This

estimate was based on a comparison of the observed number of nonsynonymous substitutions per one synonymous mutation (N_a/N_s ratio) with the corresponding theoretical number expected under neutral evolution (N_a^0/N_s^0 ratio). It should be noted that the assumption that natural selection is absent (neutral evolution) is equivalent to the assumption that natural selection has not yet acted (*de novo* mutations).

As presented in Figure 2 and Table 1, the N_a^0/N_s^0 ratio calculated using our neutral model of evolution for 37 genes of large resequencing dataset was equal 2.204. Experimental N_a/N_s ratio for singletons calculated for the same dataset was equal 1.49. Given these values we determined that only 32% (1-1.49/2.20) of missense mutations are deleterious to the extent that they have very low probability to be found even once in the sample of 1500 chromosomes. Standard error of this estimate is 2%.

The experimental N_a/N_s ratio for amino acid substitutions fixed in the set of 37 genes in the human lineage after divergence from chimpanzee is equal 0.44 This means that approximately 20% of all missense mutations in these genes are effectively neutral – a value close to the genome-wide average.

Human genetic variation detected in the set of 37 genes sequenced in 756 individuals provided an opportunity to get an insight into human genetic variation of very low allele frequency. However, it lacks a significant quantity of data on common genetic variation. To fill this gap we used three publicly available datasets: 1) dataset generated by Environmental Genome Project (NIEHS-EGP) in which over 500 genes involved in DNA repair and cell cycle pathways were

sequenced in at least 90 unrelated individuals; 2) dataset generated by SeattleSNPs project (SeattleSNPs) in which over 200 genes involved in the inflammatory responses were sequenced in at least 46 individuals; and 3) dataset generated by Japanese Single Nucleotide Polymorphisms (JSNP) project in which polymorphic sites in over 8,000 genes were discovered using panel of 12 individuals and later genotyped in 750 individuals. These datasets also contained information on rare genetic variation that we used, although a large fraction of rare nsSNPs with frequency below 1% might have been missed in the EGP and SeattleSNP datasets because less than 200 chromosomes were sequenced, and in JSNP, because only 12 individuals were used for SNP discovery.

Again, using our neutral model of evolution, we calculated predicted N_a^0/N_s^0 ratios, separately for EGP, SeattleSNPs and JSNP datasets. Then we calculated experimental N_a/N_s ratios for three types of substitutions. The first type is rare substitutions that a) were observed only once among all chromosomes sequenced (in EGP and SeattleSNPs) or b) had an observed frequency below 1%(JSNP). The second type is very common polymorphisms, with a frequency of the least common allele above 25%. The third type is substitutions fixed in the human lineage after divergence with chimpanzee.

We observed that N_a/N_s ratios for very common polymorphisms were only very slightly higher than N_a/N_s ratios for fixed substitutions (Figure 2, Table 1). This fact indicates that the fraction of deleterious substitutions among common SNPs is very low. It also supports the notion that the fraction of positively selected amino acid substitutions among all mutations in human proteins was low[38].

In sharp contrast with common SNPs, N_a/N_s ratio for very rare alleles significantly exceeds N_a/N_s ratios for fixed substitutions (that are presumably neutral). This fact indicates that a very large fraction of such rare substitutions are deleterious. The fraction of deleterious mutations among observed rare substitutions can be estimated as $(N_a^{rare}/N_s^{rare})/(N_a^{fixed}/N_s^{fixed})$. Simple calculations reveal that the majority (52%-71%, Table 2) of amino acid substitutions with the observed frequency below 1% are mildly deleterious in all datasets. This surprising finding indicates that a low frequency of missense mutation *per se* can serve as a strong predictor of deleterious effect of polymorphic variants.

Characteristic selection coefficients of mildly deleterious mutations - theory

We concluded that more than half of newly arising missense mutations are mildly deleterious. Such mutations are not present among common polymorphisms but are very common among substitutions detected only once in a hundred or more chromosomes. Based on these observations we estimated selection coefficients associated with such mutations.

First, we calculated theoretical expected number of mutations with selection coefficient *s* observed only once in a set of *m* randomly sampled chromosomes. Then, we divided it by the corresponding number of neutral mutations to obtain the ratio $R_{1/m}(s)$:

$$R_{1/m}(s)=\frac{[\int_0^1 \frac{e^{-2N_e s(1-x)}-1}{x(1-x)(e^{-2N_e s}-1)}(C_1^m x(1-x)^{m-1}+C_{m-1}^m x^{m-1}(1-x))dx]}{[\int_0^1 \frac{1}{x}(C_1^m x(1-x)^{m-1}+C_{m-1}^m x^{m-1}(1-x))dx]} \quad \{\text{Equation 1a}\}$$

where C^n_k is the binomial coefficient – the number of combinations of size *k* from a set with *n* elements $C_k^n = \frac{n!}{k!(n-k)!}$ {Equation 1b}

In the {Equation 1a} N_e represents an effective population size and *x* the frequency of an individual allele in the population. The number of detected mutations was obtained by multiplying the theoretical number of all alleles present in population with frequency *x* by the probability of a single allele with frequency *x* being detected as a singleton in a set of *m* chromosomes, and then by integrating over allele frequency *x* from 0 to 1. It should be noted that an allele is detected as a singleton in a set of *m* chromosomes if either 1 or *m*-1 copies are present.

Similarly we calculated the theoretical ratio $R_{MAF>0.25}(s)$ for mutations that have been detected as polymorphisms in a set of *m* chromosomes with minor allele frequency (MAF) higher than 25%.

$$R_{MAF>0.25}(s) = \frac{\left[\int_0^1 \left[\frac{e^{-2N_e s(1-x)}-1}{x(1-x)(e^{-2N_e s}-1)} \sum_{0.25m<j<0.75m} C_j^m x^j (1-x)^{m-j}\right]dx\right]}{\left[\int_0^1 \left[\frac{1}{x} \sum_{0.25m<j<0.75m} C_j^m x^j (1-x)^{m-j}\right]dx\right]} \quad \{\text{Equation 2}\}$$

These formulas result from an application of diffusion theory to the dynamics of polymorphism in populations[39]. Their two major underlying assumptions are constant effective population size and each novel mutation occurring at a new, previously monomorphic site (infinite number of

sites model). The dominance coefficient was assumed to be equal 0.5, so if *s* is the selection coefficient associated with a heterozygote, *2s* is the selection coefficient associated with a homozygote. The exact values of the dominance coefficient are not very important, as long as they are high enough to allow selection to operate primarily on heterozygotes.

The dependence of $R_{MAF>0.25}(s)$ and $R_{1/m}(s)$ (for various values of *m*) on selection coefficient *s* is shown in Figure 3A. We determined that roughly equal fractions of missense mutations can be observed as fixed substitutions and as common polymorphisms with minor allele frequency above 25% (Fig 2). This observation implies that mildly deleterious mutations are virtually absent among frequent SNPs and thus should have selection coefficient values at which $R_{MAF>0.25}(s)$ is very close to zero. On the other hand, the majority of mildly deleterious mutations have high relative chance to be detected as singletons. For each three resequencing datasets (obesity-related, EGP and SeattleSNP) we estimated fraction of *de novo* mildly deleterious missense mutations observed as singletons relative to neutral expectations:

$$R_{observed} = \frac{\frac{(N_a/N_s)^{singletons}}{(N_a^0/N_s^0)} - \frac{(N_a/N_s)^{hum.lin.subst.}}{(N_a^0/N_s^0)}}{1 - F_{strongly\,del.} - \frac{(N_a/N_s)^{hum.lin.subst.}}{(N_a^0/N_s^0)}} \quad \{\text{Equation 3}\}$$

Here the numerator reflects the observed fraction of mutations in the mildly deleterious class detectable as singletons, which is equal to fraction of all mutations detectable as singletons minus fraction of effectively neutral *de novo* mutations. Denumerator reflects the total fraction of mutations arising in the mildly deleterious class. It is given by substracting fraction of strongly deleterious *de novo* mutations (given by $F_{strongly\,del}$) and neutral mutations from 1.

Using theoretical dependence of R on selection coefficient (Fig. 3A) for corresponding number of sequenced chromosomes *m* we estimated characteristic selection coefficients for mildly deleterious *de novo* missense mutation for genes in each set. All three estimates fall into a relatively narrow range of selection coefficients 0.002-0.02.

Characteristic selection coefficients of mildly deleterious mutations – computer simulations

The major drawback of the above straightforward theoretical approach is the assumption that the effective size of the human population is constant and equals 10,000. This assumption was shown to be appropriate for common SNPs, most of which are ancient in origin. However, an order of magnitude larger effective population size might be more appropriate to describe the dynamics of rare polymorphisms. Larger effective population size leads to more efficient purifying selection and, as can be seen from {Equation 1a}, lower values of $R_{1/m}(s)$. By underestimating N_e we are overestimating $R_{1/m}(s)$, and in reality the $R_{1/m}(s)$ dependence-curve on *s* is shifted to the area of lower values of selection coefficients (Figure 3A).

To analyze the dependence of $R_{1/m}(s)$ and $R_{MAF>0.25}(s)$ under a more realistic human population history scenario we simulated evolution under the Wright-Fisher model (assuming constant mutation rate and infinite number of unlinked sites). Our model of human population history was comprised of the four principal epochs: 1) 100,000 generations of stable effective population size 15,000; 2) exponential reduction in 100 generations to the size of 7,000; 3) 500 generations of stable population size 7,000; 4) exponential growth in 3,000 generations to the effective population size 90,000. Such a scenario, where a long period of stable effective population size is followed by bottleneck and then by rapid expansion captures key features of human population

history. Exact values for the duration of each epoch and corresponding effective population sizes were chosen in general agreement with recent literature[40-42] and than slightly optimized to better describe alleles frequency distributions for non-coding SNPs detected by the Environmental Genome Project.

Usage of $R_{1/m}(s)$ and $R_{MAF>0.25}(s)$ dependencies obtained using direct simulation resulted in estimate that approximately 0.001–0.003 range of selection coefficients corresponds to mildly deleterious class of *de novo* missense mutations.

Cumulative equilibrium frequency of mildly deleterious polymorphisms in the human genome

We determined that mildly deleterious missense mutations that are not present among common polymorphisms are nevertheless numerous among substitutions with a detected population frequency below 1% and are associated with selection coefficients in the range of 0.001-0.003 Using this result, we estimated the equilibrium frequency of such mildly deleterious alleles in the human population. Under the assumption of strong selection and large population size ($Ns >> 1$) the *cumulative* equilibrium frequency of all mildly deleterious gene alleles in the population is:

$$Freq_{cumulative}=u/s \qquad \{Equation\ 4\}$$

where u is mutation rate per gene per generation and s is selection coefficient associated with heterozygotes. We used $2x10^{-8}$ value for mutation per nucleotide per generation rate estimated by Kondrashov[21]. Importantly, this estimate for cumulative frequency of deleterious alleles does not depend on population demographic history.

The mutation rate for mildly deleterious missense mutations per gene can be estimated as follows:

u = [Mutation rate per nucleotide]x[Average gene length]x[Fraction of missenses among all substitutions]x[Fraction of mildly deleterious among missenses] {Equation 5a}

$u = [2 \times 10^{-8}] \times [1500] \times [2.23/3.23] \times [0.53] = 1.1 \times 10^{-5}$ {Equation 5b}

Using equation 4, we calculate that for an average gene the combined frequency of alleles carrying mildly deleterious missense mutation is approximately 1%.

Taking into account that protein coding sequences comprise approximately 1.5% of the human genome, we estimated that approximately 600 missense mutations with selection coefficient 0.001-0.003 are present in the genome of an average individual. It should be noted that the number of mutations with significantly weaker effect on fitness (with associated selection coefficients below 10^{-4}) is likely to be significantly higher.

Discussion

It has been noted that, in some Mendelian diseases, less severe forms of the disease exist, associated with milder mutations that do not completely abrogate protein activity[1]. It was suggested that such mildly deleterious missense mutations may serve as a genetic basis of complex diseases. The plausibility of this hypothesis depends on how frequent such mildly deleterious substitutions are among *de novo* mutations and how strong the purifying selection that acts on them is. The presence of deleterious missense polymorphisms in the human genome has been reported in many previous studies[43-52]. In the present work we performed a quantitative analysis, calculating the fraction of mildly deleterious mutations among *de novo* amino acid changes, very rare and very common nsSNPs. Until recently, lack of large resequencing datasets that cover dozens of genes and hundreds of chromosomes precluded such analysis, although several earlier studies analyzed distribution of selection coefficients for new missense mutations in humans and flies using frequency spectrum in smaller samples[53,54], dependence of substitution rate on effective population size[50-52] and comparison of polymorphism and divergence[33,53].

We calculated the fraction of mildly deleterious missenses among all *de novo* substitutions as the remaining difference after subtraction of strongly detrimental mutations and effectively neutral mutations. We determined that among all *de novo* amino acid substitutions strongly detrimental mutations comprise approximately 20%. The genome-wide value that we obtained is similar to the fraction of amino acid replacements which destroy the protein function (25%) obtained by Yampolsky et al.[33]. Our estimate is also in general agreement with the results of Eyre-Walker et al.[53], who, by fitting a distribution of selection coefficients modeled as a gamma-function to the

data on human genetic variation, estimated that approximately 15% of missense substitutions are strongly deleterious with selection coefficients above 0.1. Earlier, Fay et al.[54] proposed that 54% of new missense mutations in humans are strongly deleterious. This estimate referred to the fraction of non-synonymous SNPs not observed in the sample of a hundred chromosomes, so this estimate is not directly comparable to the estimate presented here. Our value for the fraction of effectively neutral amino acid changes (27%) is only very slightly higher than a previous estimates[50] of 24% and the corresponding K_A/K_S estimate for human-chimpanzee comparison[38] (23%). The difference is most likely due to the application of the context-based mutation model. A smaller previous estimate 12%(ref. 33) possibly resulted from the use of the too distant mouse sequence as an outgroup.

On the basis of a very strong difference in N_a/N_s ratios for very rare and very common missense SNPs, we estimated that the majority of rare missense polymorphisms detected in the human population are associated with a surprisingly narrow range of selection coefficients: 0.001-0.003. Because of such relatively mild purifying selection acting on them, they can reach a high cumulative frequency in the human population, while still maintaining a highly heterogeneous spectrum of individual alleles.

An increase in the N_a/N_s ratio has been noted previously[43], however, the extent of such a difference was not always evident because of the low number of chromosome sequenced and crude binning by frequency. A strong, "threshold-like" increase in N_a/N_s ratio for SNPs with frequency below 6% was noted by Wong et al.[59] This work, however, has considered this observation only as a consequence of massive relaxation of selective constraints in recent human

evolution.

We estimated that for an average 500 amino acid protein the cumulative equilibrium frequency of alleles carrying deleterious missense SNPs is roughly equal to 1% in the human population. If we consider unusually long proteins or entire pathways as mutational targets, their mutation rate, and thus, equilibrium frequency of mildly deleterious alleles will be even significantly higher than 1%. Some traits, such as susceptibility to complex diseases, can be influenced by dozens or, perhaps, even hundreds of genes. The mutational target for such traits will be exceptionally large. A large mutational target size leads to a large mutation rate per generation and, as a consequence, to a large level of deleterious polymorphism, since at equilibrium it is directly proportional to the mutation rate. Even for a simple metabolic pathway that involves several multi-subunit enzymes, the fraction of individuals harboring at least one damaging mutation in the pathway can easily exceed 10%. This example shows that a combined frequency of rare mildly deleterious polymorphisms is high enough to serve as the basis for heritable susceptibility even to most common complex diseases such as hypertension or coronary heart disease. This suggests the possibility that mutation-selection balance can be a feasible evolutionary explanation at least for some common diseases.

Candidate genes association studies

These findings carry implications not only for our understanding of the role of rare nonsynonymous SNPs in susceptibility to complex diseases, but also to methods for detecting genes that harbor such detrimental genetic variation. Currently, there are two major approaches for identification of genes involved in complex diseases: linkage studies[60] and association studies

of population samples[60,61]. While the former approach is most effective for identification of Mendelian-like rare genetic variants with high penetrance, the latter approach is better suited for identification of common variants with relatively low penetrance. None of these approaches, however, is able to detect susceptibility loci that harbor numerous, but individually rare, mildly deleterious polymorphisms. However, our analysis of mildly deleterious mutations indicates that this situation is very plausible and might be true for many genes and common diseases.

In the light of our results, a recently proposed alternative approach looks more promising. This approach aims at the detection of enrichment in rare, potentially deleterious missense SNPs in a patient group versus a control subject group[62,63]. By considering the cumulative frequency of deleterious mutations rather than their individual frequencies, this method is suitable for investigation of common diseases that have very a heterogeneous spectrum of predisposing alleles. Such a missense accumulation approach to candidate gene association studies has been recently successfully utilized in studies of the MC4R gene[64], tyrosine phosphatome in colorectal cancers[65] and genes involved in lipid metabolism disorders[66].

However, two major potential disadvantages of this method have been pointed out[62]. First is the high cost of complete genomic sequencing of candidate genes in a large number of individuals. Second is the analytical challenge of selecting potentially deleterious SNPs from a large quantity of neutral genetic variation[62]. The first objection to the method will almost certainly be eliminated in the near future by continuing dramatic reduction in sequencing costs[67] and development of novel sequencing strategies[68]. The second objection is more fundamental, since the inability to distinguish between deleterious and neutral amino acid changes would lead to a

very low signal to noise ratio. To study enrichment in deleterious mutations one should be able to detect them among neutral genetic variation. The number of deleterious alleles in disease phenotypes was previously believed to be low relative to the large number of neutral missense SNPs present in both disease and control groups. Our analysis, however, reveals that the majority of missense substitutions with detected frequency below 1% are, in fact, deleterious. Thus, if only substitutions with detected frequency below 1% are counted, the enrichment in mutations number in disease group should be highly pronounced.

Conclusions

A high fraction of mildly deleterious mutations among missense mutations suggests that mutation-selection balance is a plausible explanation for the existence of common disease with complex inheritance, at least in some cases. Further, it is feasible that some common diseases may be caused by a multitude of rare allelic variants. The observation that the majority of human rare non-synonymous variants are deleterious, and thus of significance to function and phenotype, suggests a strategy for candidate gene association studies. Disease populations are expected to have a higher rate of rare amino acid variants than healthy controls in genes involved in disease. This difference can be easily detected in a deep re-sequencing study. Obviously, this strategy would be highly inefficient if the majority of coding variants at low frequency were neutral. Several recent reports demonstrated an excess of rare missense variants in individuals with phenotypes associated with disease risk. Our analysis provides an explanation for the success of these studies.

Web resources

The Human Gene Mutation Database (HGMD), http://www.hgmd.cf.ac.uk/ac/index.php

UCSC Sequence and Annotation Downloads, http://hgdownload.cse.ucsc.edu/downloads.html

Consensus CDS (CCDS) project, http://www.ncbi.nlm.nih.gov/CCDS/

SeattleSNPs. NHLBI Program for Genomic Applications, SeattleSNPs, Seattle, WA, http://pga.gs.washington.edu

NIEHS SNPs. NIEHS Environmental Genome Project, University of Washington, Seattle, WA, http://egp.gs.washington.edu

A database of Japanese Single Nucleotide Polymorphisms (JSNP), http://snp.ims.u-tokyo.ac.jp/

NCBI database of single nucleotide polymorphism (dbSNP), http://www.ncbi.nlm.nih.gov/projects/SNP/index.html

References

1.Glazier AM, Nadeau JH, Aitman TJ (2002) Finding genes that underlie complex traits. Science 298:2345-2349

2.Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33:228-237

3.Wang X, Grus WE, Zhang J (2006) Gene losses during human origins. PLoS Biol 4:e52

4.Charlesworth B (2001) Patterns of age-specific means and genetic variances of mortality rates predicted by the mutation-accumulation theory of ageing. J Theor Biol 210:47-65

5.Williams PD, Day T, Fletcher Q, Rowe L (2006) The shaping of senescence in the wild. Trends Ecol Evol 21:458-463

6.Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21:596-601

7.Neel JV (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14:353-362

8.Charlesworth D. (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64

9. Voet D, Voet JG (2004) Biochemistry, 3rd Edition, John Wiley & Sons, Inc. pp 183-185

10. Williams GC (1957) Pleiotropy, natural selection and the evolution of senescence. Evolution 11:398-411

11. Kimura M (1965) A stochastic model concerning the maintenance of genetic variability in quantitative characters. Proc Natl Acad Sci U S A 54:731-736

12. Lande R (1975) The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genet Res 26:221-235

13. Turelli M (1984) Heritable genetic variation via mutation-selection balance: Lerch's Zeta meets the abdominal bristle. Theor. Popul. Biol. 25:138–193

14. Zhang XS, Wang J, Hill WG (2004) Influence of dominance, leptokurtosis and pleiotropy of deleterious mutations on quantitative genetic variation at mutation-selection balance. Genetics 166:597-610

15. Zhang XS, Hill WG (2005) Genetic variability under mutation selection balance. Trends Ecol Evol 20:468-470

16. Keller MC, Miller G (2006) Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? Behav Brain Sci 29:385-404

17. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124-137

18.Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant or not? Hum Mol Genet 11:2417-2423

19.Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502-510

20.Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21:577-581

21.Kondrashov AS (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 21:12-27

22.Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821-1831

23.Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y (2002) JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30:158-162

24.Sherry ST, Ward, MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

25.Subramanian S, Kumar S (2006) Higher Intensity of Purifying Selection on >90% of the Human Genes Revealed by the Intrinsic Replacement Mutation Rates. Mol Biol Evol

23:2283-2287

26.Krawczak M, Ball EV, Cooper DN (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63:474-488

27.Li WH (1997) Molecular Evolution. Sinauer Associates, Inc. pp 80-84

28.ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636-640

29.Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708-715

30.Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21:468-488

31.Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A 101:13994-14001

32.Arndt PF, Hwa T (2005) Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics 21:2322-2328

33.Yampolsky LY, Kondrashov FA, Kondrashov AS (2005) Distribution of the strength of selection against amino acid replacements in human proteins. Hum Mol Genet 14:3191-3201

34.Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res 13:831-837

35.Chamary JV, Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 6(9):R75

36.Parmley JL, Chamary JV, Hurst LD (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol 23:301-309

37.Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98-108

38.Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69-87

39.Kimura M (1983) Infinite site model. In: The neutral theory of molecular evolution. Cambridge University Press, Cambridge, pp 236-240

40.Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166:351-372

41.Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A 102:7882-7887

42.Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15:1576-1583

43.Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22:239-247

44.Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231-238

45.Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R et al. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493

46.Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P (2001) Prediction of deleterious human alleles. Hum Mol Genet 10:591-597

47.Glatt CE, DeYoung JA, Delgado S, Service SK, Giacomini KM, Edwards RH, Risch N, Freimer NB (2001) Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. Nat Genet 27:435-438

48.Sunyaev S, Kondrashov FA, Bork P, Ramensky V (2003) Impact of selection, mutation rate and genetic drift on human genetic variation. Hum Mol Genet 12:3325-3330

49.Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U S A 100:15754-15757

50.Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. Nature 437:1153-1157

51.Crawford DC, Akey DT, Nickerson DA (2005) The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet 6:287-312

52.Gorlov IP, Kimmel M, Amos CI (2006) Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. Hum Mol Genet 15:1143-1150

53.Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173:891-900

54.Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158:1227-1234

55.Eyre-Walker A, Keightley PD, Smith NG, Gaffney D (2002) Quantifying the slightly deleterious mutation model of molecular evolution. Mol Biol Evol 19:2142-2149

56.Loewe L, Charlesworth B, Bartolome C, Noel V (2006) Estimating selection on

nonsynonymous mutations. Genetics 172:1079-1092

57.Kryukov GV, Schmidt S, Sunyaev S (2005) Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet 14:2221-2229

58.Bustamante CD, Nielsen R, Hartl DL (2003) Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. Theor Popul Biol 63:91-103

59.Wong GK, Yang Z, Passey DA, Kibukawa M, Paddock M, Liu CR, Bolund L, Yu J (2003) A population threshold for functional polymorphisms. Genome Res 13:1873-1879

60.Freimer N, Sabatti C (2004) The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nat Genet 36:1045-1051

61.Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517

62.Hirschhorn JN, Altshuler D (2002) Once and again-issues surrounding replication in genetic association studies. J Clin Endocrinol Metab 87:4438-4441

63.Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95-108

64.Vaisse C, Clement K, Durand E, Hercberg S, Guy-Grand B, Froguel P (2000) Melanocortin-4 receptor mutations are a frequent and heterogeneous cause of morbid obesity. J Clin Invest 106:253-262

65.Wang Z, Shen D, Parsons DW, Bardelli A, Sager J, Szabo S, Ptak J, Silliman N, Peters BA, van der Heijden MS, Parmigiani G, Yan H, Wang TL, Riggins G, Powell SM, Willson JK, Markowitz S, Kinzler KW, Vogelstein B, Velculescu VE (2004) Mutational analysis of the tyrosine phosphatome in colorectal cancers. Science 304:1164-1166

66.Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869-872

67.Service RF (2006) Gene sequencing. The race for the $1000 genome. Science 311:1544-1546

68.Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728-1732

Fig 1 Spectrum of effects of *de novo* missense mutations

A) Fraction of strongly detrimental mutations among *de novo* amino acid substitutions. Disease causing nonsense mutations in HGMD were used as a standard of "strong detrimentality"; B) fraction of strongly detrimental mutations among *de novo* amino acid substitutions. Disease causing splice-site mutations in HGMD were used as a standard of "strong detrimentality"; C) Fraction of effectively neutral mutations among *de novo* amino acid substitutions. Synonymous substitutions fixed in the human lineage after divergence from chimpanzee were used as a standard of "effective neutrality".

Fig 2 Fraction of *de novo* missense mutations represented at different levels of allele frequency

Normalized fraction of *de novo* amino acid substitutions detected in a given dataset was calculated from the difference of observed N_a/N_s ratio and theoretical N_a^0/N_s^0 ratio expected under neutrality. Data for rare polymorphisms are shown in orange, for common polymorphism in yellow, and for substitutions fixed in the human lineage after divergence from chimpanzee in green. Standard errors are shown by grey error bars.

Fig 3 A) $R_{1/m}(s)$ and $R_{MAF>0.25}(s)$ (see Equation 1a) calculated using equations derived from diffusion theory under the assumption of constant population size and an infinite number of sites. Expected shift of $R_{1/m}(s)$ curves due to recent population expansion is shown by red arrows. Black arrows illustrate estimation of characteristic selection coefficients for mildly deleterious class of missense mutations (see Results and Equation 3). B) $R_{1/m}(s)$ and $R_{MAF>0.25}(s)$ calculated by direct computer simulation of molecular evolution under the assumption of an infinite number of

sites and simple population history – stable population size epoch followed by bottleneck and then fast expansion.

Table 1 Ratios of missense to synonymous SNPs and substitutions.

Dataset	Description of dataset	Number of genes	Theoretical N_a^0/N_s^0	Singletons N_a/N_s	SNPs with minor allele freq. >25% N_a/N_s	Subst. in the human lineage N_a/N_s	Singletons $(N_a/N_s)/(N_a^0/N_s^0)$	SNPs with minor allele freq. >25% $(N_a/N_s)/(N_a^0/N_s^0)$	Subst. In the human lineage $(N_a/N_s)/(N_a^0/N_s^0)$
CCDS	"genome-wide"	14095	2.232	-	-	0.60	-	-	0.266+/-0.002
Obesity related genes	757 individuals sequenced	37	2.204	1.49	-	0.44	0.68+/-0.02	-	0.20+/-0.05
NIEHS-EGP	90-95 individuals sequenced	518	2.255	1.36	0.56	0.50	0.61+/-0.02	0.26+/-0.04	0.22+/-0.01
SeattleSNPs	46-47 individuals sequenced	236	2.203	1.43	0.66	0.66	0.65+/-0.04	0.32+/-0.05	0.30+/-0.02
JSNP	750 individuals genotyped	8786	2.244	1.37*	0.54	0.54	0.61	0.24	0.24

*) SNPs with observed frequency below 1% instead of singletons have been used in analysis of JSNP dataset.

Table 2 Fraction of deleterious substitutions among rare missense SNPs

Set	***Number of sequenced individuals***	***Percent of deleterious SNPs among missense "singletons"***
Resequencing dataset of obesity related genes	757	(71+/-8)%
NIEHS-EGP	90-95	(64+/-1)%
SeattleSNPs	46-47	(52+/-6)%