Review

Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations

Sudhir Kumar[1,2], Joel T. Dudley[3,4], Alan Filipski[1,2] and Li Liu[2]

[1] School of Life Sciences, Arizona State University, Tempe, AZ 85287-4501, USA
[2] Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, AZ 85287-5301, USA
[3] Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305, USA
[4] Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA 94305, USA

Modern technologies have made the sequencing of personal genomes routine. They have revealed thousands of nonsynonymous (amino acid altering) single nucleotide variants (nSNVs) of protein-coding DNA per genome. What do these variants foretell about an individual's predisposition to diseases? The experimental technologies required to carry out such evaluations at a genomic scale are not yet available. Fortunately, the process of natural selection has lent us an almost infinite set of tests in nature. During long-term evolution, new mutations and existing variations have been evaluated for their biological consequences in countless species, and outcomes are readily revealed by multispecies genome comparisons. We review studies that have investigated evolutionary characteristics and *in silico* functional diagnoses of nSNVs found in thousands of disease-associated genes. We conclude that the patterns of long-term evolutionary conservation and permissible sequence divergence are essential and instructive modalities for functional assessment of human genetic variations.

Evolutionary genomic medicine

Thousands of individuals in the general public have begun to gain access to their genetic variation profiles by using direct-to-consumer DNA tests available from commercial vendors; these tests profile hundreds of thousands of genomic markers at a cost of a few hundred dollars (Figure 1a). Through this genetic profiling individuals hope to learn about not only their ancestry but also about genetic variations underlying their physical characteristics and predispositions to diseases. In biomedicine, scientists have been profiling genome-wide variations in healthy and diseased individuals in a variety of disease contexts and populations. This has led to the discovery of thousands of disease-associated genes and DNA variants [1–6]. Meanwhile, following sharp declines in the per-base cost of sequencing, complete genomic sequencing of individuals and cohorts is underway and expanding {[7–11]; 1000 Genomes Project (www.1000genomes.org); Personal Genomes Project (www.personalgenomes.org)}. Taken together, these efforts have begun to paint a more robust picture of the amount and types of variations found within and between human individuals and populations. Any one personal genome contains more than a million variants, the majority of which are single nucleotide variants (SNVs) (Figure 1b). With the complete sequencing of each new genome, the number of novel variants discovered is decreasing, but the total number of known variants is growing quickly (Figure 2a). Our knowledge of the number of disease genes and the total number of known disease-associated SNVs has grown with these advances [12].

Today, the vast majority of the known disease-associated variants are found within protein-coding genes (Figure 1c) although genome-wide association studies beginning to reveal thousands of non-coding variants. Proteins are encoded in genomic DNA by exon regions, and these comprise only ~1% of the genomic sequence (exome; Glossary) [11,13]. It is this part of our genome for which we have the best understanding of how DNA sequence relates to function, and is arguably the best chance to connect genetic variations with disease pathophysiology. The exome of an individual carries about 6000–10 000

Glossary

Complex disease: refers to any disease having some genetic component of etiology that is characterized as involving the effects of many genes. Complex diseases are typically common in the population, exhibit complex patterns of inheritance, and often involve the interaction of genetic and environmental factors.
Driver mutation: somatic mutations implicated as having a causal role in the pathogenesis of cancer.
Evolutionary retention: a position-specific measure of conservation taking into account the number of times a human amino acid position is missing a homolog in the multiple sequence alignment with other species.
Exome: the complete collection of (known) exons that ultimately constitute proteins expressed by an individual.
Genetic drift: the change in the population frequency of alleles due to random sampling of neutral or effectively neutral alleles.
Mendelian disease: a genetic disease trait exhibiting a Mendelian inheritance pattern for an underlying mutation at a single genetic locus.
Passenger mutation: somatic mutations observed in cancer cell genomes that do not contribute to cancer pathogenesis. Can be seen in high frequencies in tumors if they occur in the same lineage as driver mutations that contribute to the clonal expansion of the cancer cell lineage.
Purifying selection: a type of directional evolutionary selection that acts to remove deleterious alleles from a population.
Somatic mutation: a change in the genetic structure that is neither inherited nor passed to offspring.

Corresponding author: Kumar, S. (s.kumar@asu.edu)

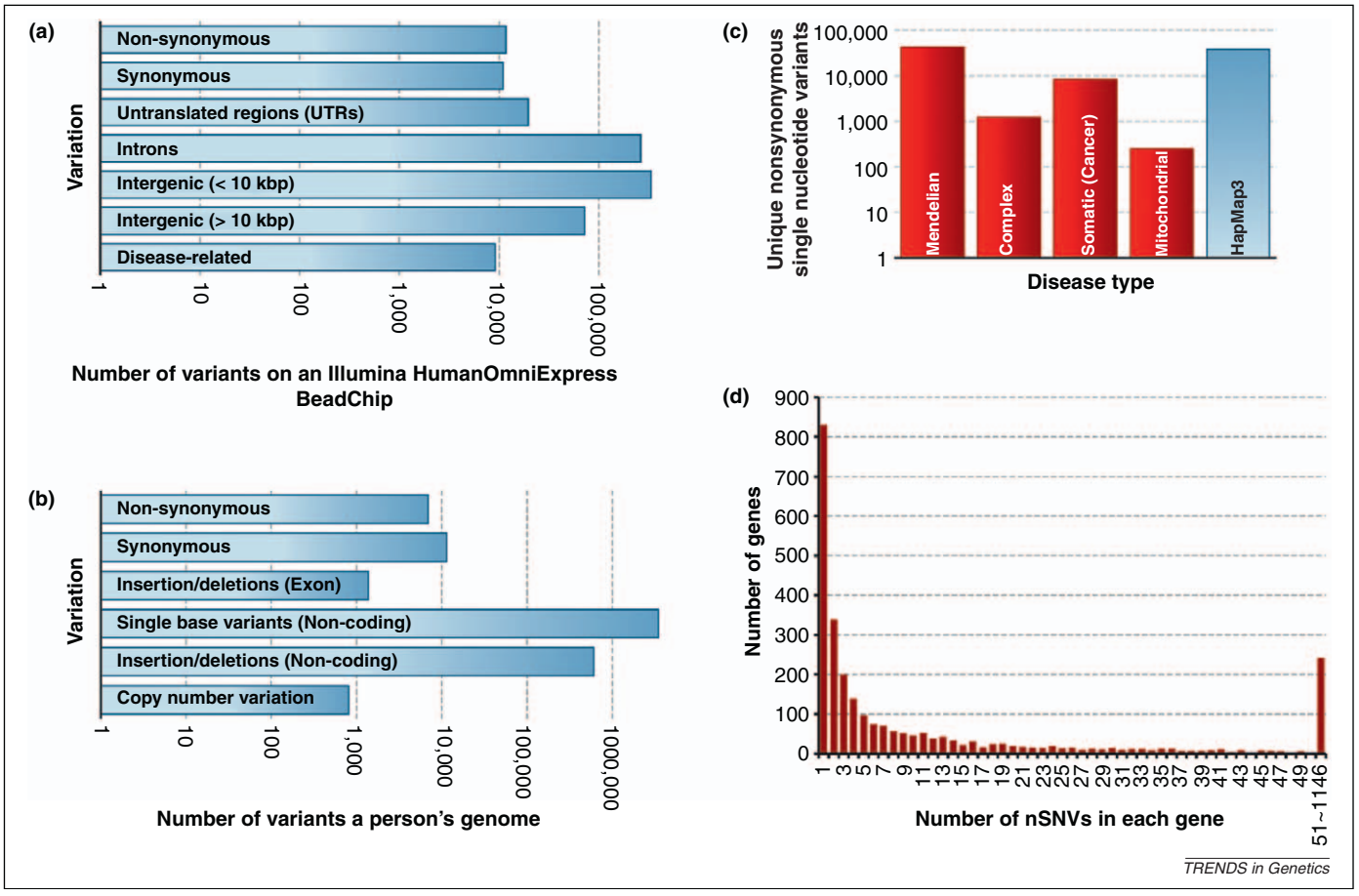

Figure 1. Profiles of personal and population variations. **(a)** Counts of different types of genetic variants profiled by 23andMe using the Illumina HumanOmniExpress BeadChip. 733 202 SNP identifiers (rsIDs) were retrieved from the Illumina website and mapped to the dbSNP database. Cross-referenced by rsIDs, disease-related variants were determined using data from the HGMD [12] and VARIMED [96] datasets. **(b)** The numbers of different types of variants found per human genome [97]. **(c)** The numbers of known non-synonymous single nucleotide variants (nSNVs) in the human nuclear and mitochondrial genomes that are associated with Mendelian diseases, complex diseases, and somatic cancers. Compared to complex diseases and somatic cancers, nSNVs related to Mendelian diseases account for the most variants discovered to date. Data were retrieved from HGMD [12], VARIMED [96], COSMIC [98], MITOMAP [41], and HapMap3 [99] resources. **(d)** The number of nSNVs in each gene related to Mendelian diseases. The majority of genes have only one or a few mutations, whereas some genes host hundreds or even more than 1000 mutations. Data were retrieved from HGMD. The numbers of variants in panels **(a–c)** are on a $\log_{10}$ scale. Information for disease-associated variants is shown in red and the personal and population variations are shown in blue.

amino-acid-altering nSNVs [2,7,9,10,14]. These protein alleles are already known to be associated with more than 1000 major diseases [12]. A large number of exome projects are poised to reveal protein mutations of tens of thousands of individuals from disease cohorts and healthy populations for disorders of various complexities [11,15–17]. With the sequencing of each new exome we are currently discovering hundreds of new nSNVs, and this points to the existence of a large number of different protein alleles in the genomes of humans (Figure 2b). In addition to the variations arising in the germline, protein-coding regions of somatic cancer cells contain tens of thousands of non-synonymous mutations of somatic and germline mutational origins (Figure 1c). Adding to the variation in the nuclear genetic material are mutations in the mitochondrial genome, many of which are also implicated in diseases (Figure 1c).

Translating a personal variation profile into useful phenotypic information (e.g. relating to predisposition to disease, differential drug response, or other health concerns) is a grand challenge in the field of genomic medicine. Genomic medicine is concerned with enabling healthcare that is tailored to the individual based on genomic information [18]. This is a daunting task because common variants derived from large population-based studies typically describe relatively small proportions of disease risk. In addition, each individual genome carries many private variants that are not typically seen in a limited sampling of the human populations. Although only a small fraction of all personal variations are likely to modulate health, the sheer volume of genomic and exomic variants is far too large to apply traditional laboratory or experimental techniques to aid in their diagnosis. Higher-throughput techniques are now becoming available to evaluate the functional consequences of hundreds of specified mutant proteins, or much greater numbers of random mutants. However, these methods are still inadequate to handle the volume of variation information arising from modern sequencing methods in a scalable or economical manner [19–23].

Fortunately, results from the great natural experiment of molecular evolution are recorded in the genomes of humans and other living species. All new mutations and pre-existing variations are subjected to the process of natural selection which eliminates mutants with negative effects on phenotype. Variants escaping the sieve of

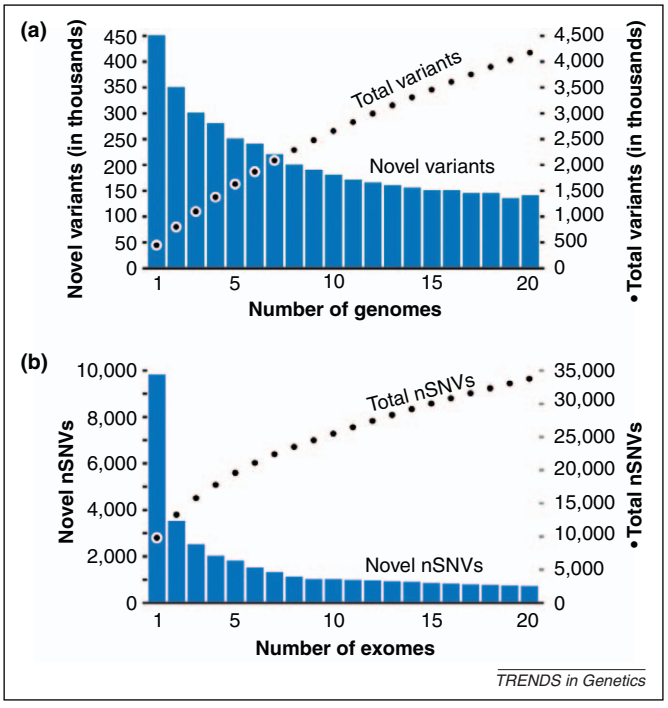

Figure 2. Novel SNV discovery with genome and exome sequencing. **(a)** The number of novel SNVs discovered by sequencing one or more genomes [97]. With increasing numbers of genomes sequenced, the number of novel SNVs decreases (bars), whereas the cumulative count of SNVs increases (filled circles). **(b)** The number of nSNVs discovered by sequencing one or more exomes [14]. With more exomes sequenced, the number of novel SNVs discovered decreases (bars) and the cumulative count of nSNVs increases (filled circles). Panels **(a)** and **(b)** are redrawn with permission from [97] and [14], respectively.

natural selection appear in the form of differences among the genomes of humans, great apes, and other species. Through multispecies comparisons of these data, using the models and methods of molecular evolution, it is possible to mine this information and evaluate the severity of each variant computationally (*in silico*). With the availability of a large number of genomes from the tree of life it is becoming clear that evolution can serve as a type of telescope for exploring the universe of genetic variation. In this evolutionary telescope, the degree of historical conservation of individual position (and regions) and the sets of substitutions permitted among species at individual positions serve as two lenses. This tool has the ability to provide first glimpses into the functional and health consequences of variations that are being discovered by high-throughput sequencing efforts. Consequently, phylomedicine will emerge as an important discipline at the intersection of molecular evolution and genomic medicine with a focus on understanding of human disease and health through the application of long-term molecular evolutionary history. Phylomedicine expands the purview of contemporary evolutionary medicine [24–28] to use evolutionary patterns beyond the short-term history (e.g. populations) by means of multispecies genomics [29,30].

In the following we review scientific investigations that have analyzed the evolutionary properties of disease-associated nSNVs and predicted function-altering propensities of individual variants *in silico* using multispecies data. We have primarily focused on variants of exomes because the function of proteins is currently best understood independently of comparative genomics. Furthermore, protein point mutations are associated with more than 1000 major diseases, and generally with a statistically significant association beyond chance alone. Furthermore, the cost of exome sequencing is declining to the point that a legion of small scientific laboratories are now able to economically profile complete exomes [17,31,32]. Therefore, the chosen emphasis on exome variations reflects current directions in clinical and research applications of genomic sequencing.

Mendelian (monogenic) diseases

For centuries it has been known that particular diseases run in families, notably in some royal families where there was a degree of inbreeding. Once Mendel's principles of inheritance became widely known in the early 1900s it became evident from family genealogies that specific heritable diseases fit Mendelian predictions. These are termed Mendelian diseases (reviewed in [33]). Such diseases can have substantial impact on the affected individual but tend to be rare, on the order of one case per several thousand or several tens of thousands of individuals.

Over the last three decades mutations in single (candidate) genes in many families have been linked to individual Mendelian diseases (e.g. Box 1). Sometimes more than a hundred SNVs in the same gene have been implicated in a particular disease (Figure 1d). For example, by the turn of this century, individual patient and family studies revealed that over 500 different nSNVs in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene can cause cystic fibrosis (CF). This enabled first efforts to examine evolutionary properties of the positions harboring *CFTR* nSNVs [29]. The disease-associated nSNVs were found to be overabundant at positions that had permitted only a very small amount of change over evolutionary time [29] (Figure 3a, b). Soon after, this trend was confirmed at the proteome scale in analyses of thousands of nSNVs from hundreds of genes (Figure 3c) [34–37]. These patterns were in sharp contrast to the variations seen in non-patients, which are enriched in the fast-evolving positions (Figure 4a) [29,35]. In population polymorphism data, faster-evolving positions also show higher minor-allele frequencies than those at slow-evolving positions [29,35], which translates into an enrichment of rare alleles in slowly evolving and functionally important genomic positions [38].

Looking at patterns of evolutionary retention at positions, another type of evolutionary conservation, a similar pattern was found: positions preferentially retained over the history of vertebrates were more likely to be involved in Mendelian diseases as compared to the patterns of natural variation (Figure 4b) [35]. Somatic mutations in a variety of cancers have also been found to occur disproportionately at conserved positions [39,40]. A similar pattern has emerged for mitochondrial disease-associated nSNVs [41].

The relationship between evolutionary conservation and disease association has been explained by the effect of natural selection [29,34–37]. There is a high degree of purifying selection on variation at highly conserved positions because of their potential effect on inclusive fitness (fecundity, reproductive success) due to the functional importance of the position [29,34,35,37,38]. At the

Box 1. Variation in the dihydroorotate dehydrogenase 1 (*DHODH*) protein found in individuals suffering from the Miller syndrome

Miller syndrome is a rare genetic disorder characterized by distinctive craniofacial malformations that occur in association with limb abnormalities (Figure I on the left; reproduced with permission from [102]). It is a typical Mendelian disease that is inherited as an autosomal recessive genetic trait. By sequencing the exomes of four affected individuals in three independent kindreds, ten mutations in a single candidate gene, *DHODH*, were found to be associated with this disease [102]. In the figure on the right, the ten mutations are shown in the context of the DHODH orthologs from six primates (including human) and the timing of their evolutionary relationships (timetree from [57]). They are in slow-evolving sites that are highly conserved not only in primates, but also among distantly related vertebrates. Specifically, 50% of these mutations are found at completely conserved positions among 46 vertebrates, including human. The average evolutionary rate, estimated using methods in [57], for sites containing these disease-related mutations is 0.50 substitutions per billion years, which is ~40% slower than those sites hosting four non-disease-related population polymorphisms of DHODH available in the public databases. Biochemically, the average severity of these ten mutations is more than twice that of the four population polymorphisms, as measured using the Grantham's [54] index (112 and 55, respectively). PolyPhen-2 [103], a computational program used to predict the propensity of individual amino acid changes to damage protein function, diagnosed all ten mutations to be potentially damaging and the four population polymorphisms to be benign. This case study demonstrated clear patterns of long-term evolutionary conservation for Mendelian disease-linked variations, and the promising applications of *in silico* tools in assisting functional diagnosis.

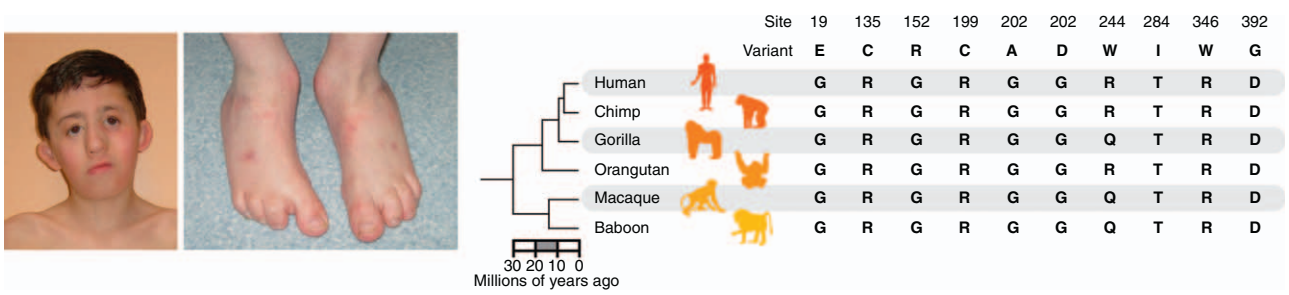

Figure I. Disease-associated genetic variants identified in patients with Miller syndrome.

faster-evolving positions, many substitutions have been tolerated over evolutionary time in different species. This points to the 'neutrality' of some mutations that spread through the population primarily by the process of random genetic drift and appear as fixed differences between species. Therefore, fewer mutations are culled at fast-evolving positions, producing a relative under-abundance of disease mutations at such positions. Of course, the above arguments hold true only when the functional importance of a position has remained unchanged over evolutionary time, an assumption that is expected to be fulfilled for a large fraction of positions in orthologous proteins.

Multigenic (complex) diseases

Despite successes in identifying and mapping genes causing Mendelian diseases, it is now clear that most common diseases with significant genetic components, although they are often seen to cluster in families, do not approximate to the simple paradigm of high penetrance based on a dominant/recessive genotype. Instead, common diseases appear to result from a more complex pattern where many genes, and probably other non-genetic factors, contribute in non-additive ways, and individual monogenic factors have a low and inconsistent correlation with the disease phenotype [42–44]. Examples of such diseases include heart disease, asthma, rheumatoid arthritis, and type 2 diabetes [45–49]. These diseases often appear relatively later in life, and the associated nSNVs are often present in one or more human populations at substantial frequencies.

An early examination of the evolutionary patterns of the occurrence of a small set of 37 nSNVs associated with complex diseases did not find any tendency for these variations to occur at sites with high conservation (Figure 4c) [37]. These trends were confirmed with larger datasets containing alleles associated with seven complex diseases [50]. These patterns stand in stark contrast to those seen for Mendelian diseases. At the level of overall rate of protein evolution, the genes associated with complex diseases are not under strong purifying selection as compared to the proteins implicated in the Mendelian diseases [51]. The rate of nonsynonymous substitutions in complex-disease genes is more than twice that of the Mendelian disease genes [52]. One reason for the lack of evolutionary conservation of positions associated with complex diseases is that their effects appear later in life, which means that these variants are frequently inherited without being acted upon by natural selection and without any impact on fecundity. For this reason, molecular evolutionary analyses are sometimes not deemed to be useful for complex diseases [53].

Evolutionary and biochemical constraints on disease-associated nSNVs

In addition to the evolutionary conservation of the positions in the protein, the biochemical properties of the amino acid change can also provide rich information. Not all changes at a position have an equal effect because one set of amino acid alternatives could be optimal, another set tolerable, and a third crippling to protein structure and function. Although the actual effect of a mutation is expected to be a complex function of the protein structure and its cellular *milieu*, many biologists have used a simple measure of biochemical difference (Grantham distance [54]) to quantify the severity of amino acid changes. In

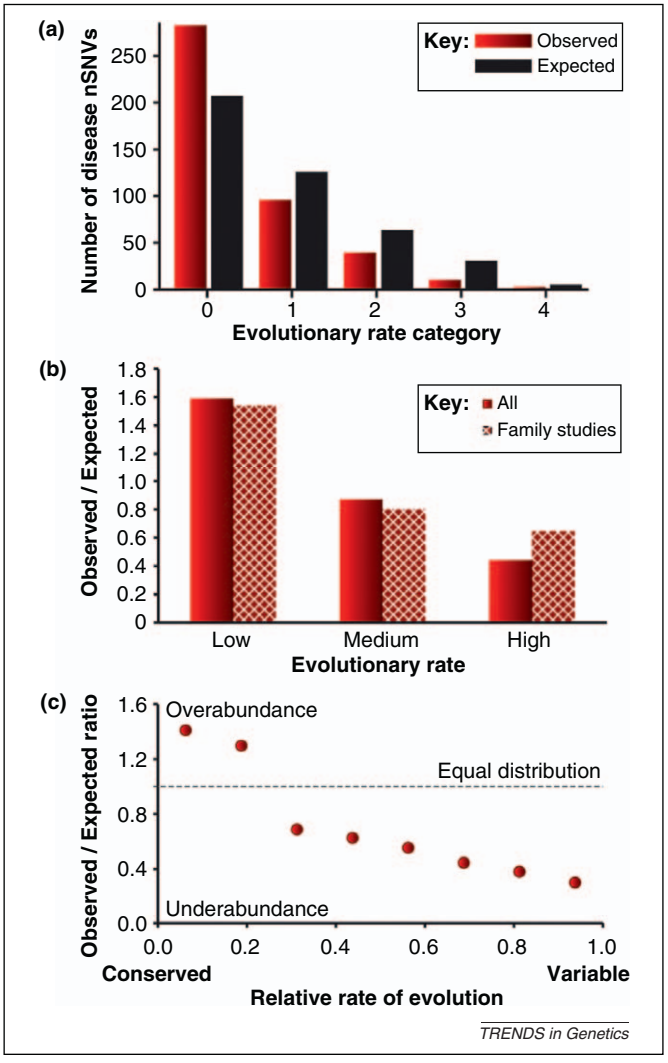

Figure 3. Evolutionary properties of positions affected by disease-associated nonsynonymous single nucleotide variants (nSNVs). **(a)** The observed and expected numbers of disease-associated nSNVs in positions that have evolved with different evolutionary rates in the CFTR protein [29]. The disease-associated nSNVs are enriched in positions evolving with the lowest rates, which belong to the rate category 0. **(b)** The ratio of observed to expected numbers of nSNVs in different rate categories for all CFTR variants (solid pattern; 431 variants) and those reported in publications profiling one or more families (hatched pattern; 59 variants). Data and publications were obtained from HGMD for all variants with a deposition date up to and including the year 2000. This comparison shows that the initial practice of the use of all available variants, including those reported by clinicians from individual patients (>80% of the variants), did not bias the observed trends. **(c)** The proteome-scale relationship of the observed/expected ratios of Mendelian disease-associated nSNVs in positions that have evolved with different evolutionary rates. The results are from an analysis of disease-associated nSNVs from 2717 genes (public release of HGMD). Just as for individual diseases, nSNVs are enriched in positions evolving with the lowest rates. Panel **(a)** is redrawn with permission from [29].

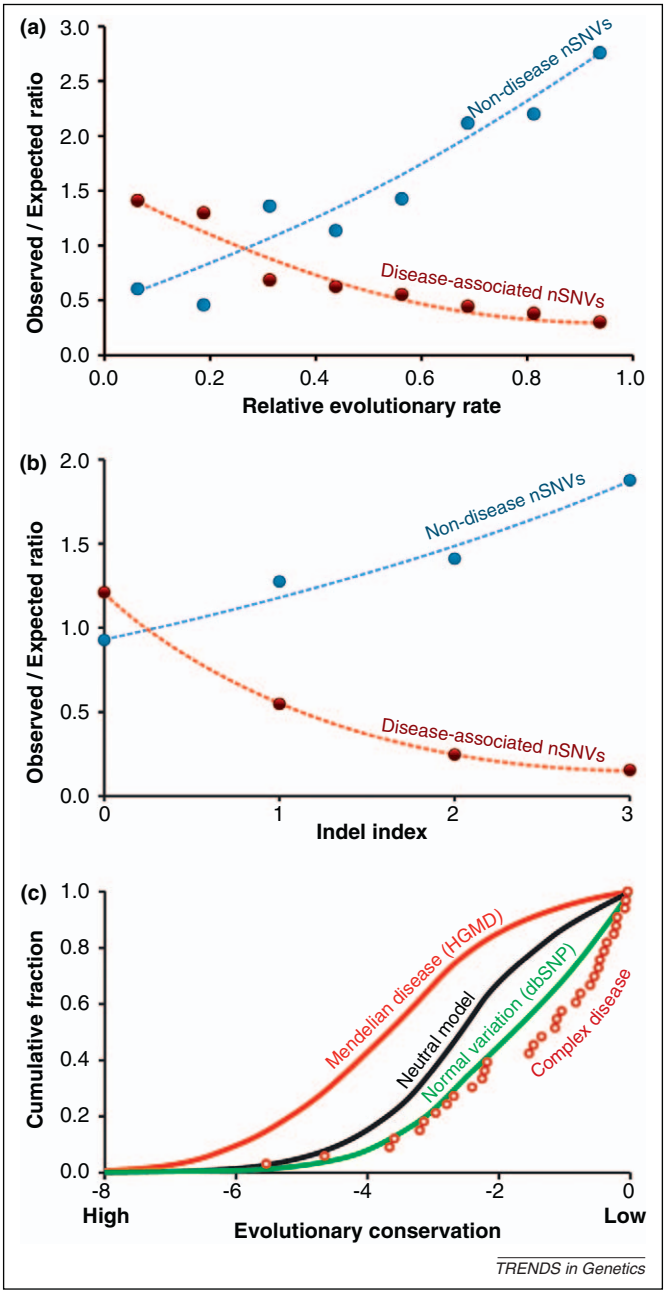

Figure 4. The enrichment of disease-associated nSNVs (red) and the deficit of population polymorphisms (blue) in human amino acid positions **(a)** evolving with different rates and **(b)** with differing degrees of insertions/deletions [35]. In both cases, smaller numbers on the *x* axis correspond to more conserved positions. There is an enrichment of disease-associated nSNVs and a deficit of population nSNPs in conserved positions. This trend is reversed for the fastest-evolving positions. **(c)** The cumulative distributions of the evolutionary conservation scores for nSNVs associated with Mendelian diseases (solid red line), complex diseases (open red circles), and population polymorphisms (green line). The shift towards the left in Mendelian nSNVs indicates higher position-specific evolutionary conservation. Conversely, a shift towards the right in complex disease nSNVs indicates lower evolutionary conservation, which overlaps with normal variations observed in the population. Data for the neutral model (black line) were generated by simulation [37]. Panels **(a–c)** are redrawn with permission from [35], [35], and [37], respectively.

an analysis of seven genes it was noted early on that amino acid changes of Mendelian disease-associated nSNVs were, on average, 67% more severe than substitutions observed between species in the same proteins [29]. The generality of this trend was confirmed in subsequent analyses of a larger number of Mendelian disease genes [34,35]. Interestingly, the timing of the onset of a disease also shows a correlation with the biochemical severity of an amino acid change: late-onset diseases involve amino acids with smaller biochemical differences [35]. Similarly, the severity of the phenotype also shows a relationship with the biochemical dissimilarity of the variation (e.g. [55]). In addition, the severity of Mendelian nSNVs has been quantified by using the substitution probability of one amino acid into another. These analyses show that disease-associated nSNVs are amino acid changes that are unlike those observed between species proteome-wide (e.g. [29,34]).

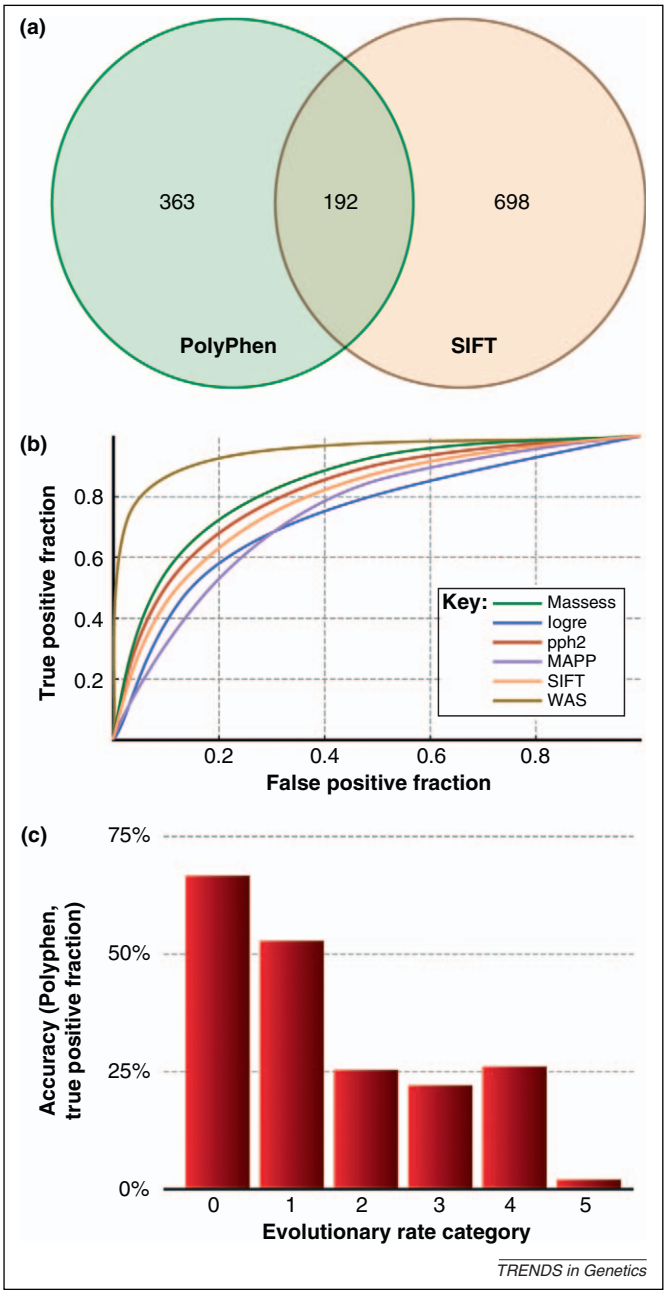

Figure 5. Some applications of evolutionary *in silico* tools in diagnosing pathogenic variants. **(a)** The comparison of PolyPhen [100] and SIFT [69] predictions for 7534 high-quality variants present within the Venter genome [2]. The numbers of variants diagnosed as probably damaging (PolyPhen) and intolerant (SIFT) are shown. The *in silico* diagnosis of personal variants by different tools produces highly discordant results. **(b)** ROC (receiver operating characteristic [101]) curves produced by PolyPhen-2 (pph2), SIFT, MAPP, Mutation Assessor (masses), Log R Pfam E-value (logre), and Condel (WAS). Condel used a weighted average of the normalized scores of the other five methods and outperformed each of them [85]. The ROC curve for Condel rises much more quickly, which means that it has a much greater rate of diagnosing damaging variants (true positives) at the expense of much smaller rate of incorrect diagnosis (false positives). **(c)** The relationship of the accuracy of the PolyPhen prediction for disease-associated nSNVs at positions evolving with different long-term rates (0–5 are categories of slowest to fastest-evolving sites) [57]. This shows that the accuracy of the PolyPhen prediction is the highest for the most slow-evolving positions for disease-associated nSNVs. Panels **(a–c)** are redrawn with permission from [2], [57], and [85], respectively.

A large number of Mendelian disease-associated variations occur at positions that show evolutionary substitutions among species. For example, more than a hundred variants of the CFTR protein in CF patients occur at positions that have undergone at least one change (Figure 3a). In any position, evolutionary differences (substitutions) between species are expected to be neutral in nature, in other words they are unlikely to have negative fitness effects provided that the protein function has not changed. They constitute a set of evolutionarily permissible alleles (EPAs) at a given position, which are expected to not be involved in diseases at those positions. Indeed, an overwhelming fraction of Mendelian nSNVs (~90%) are not evolutionarily permissible [35,55,56]. This is in sharp contrast to population polymorphisms that frequently (59%) appear in the set of EPAs in individual positions [57]. Disease-associated nSNVs in mitochondrion-encoded proteins also show similar patterns [58].

Nevertheless, scientists have been interested in investigating why some nSNVs are associated with diseases in humans, but appear as natural alleles in other species [35,56,58,59]. One possibility is that the function of the affected amino acid position has changed either in humans or in other species. In this case, evolutionary differences among species cannot be used to determine permissible amino acids at the affected positions. Another reason for the overlap between the disease nSNVs and evolutionarily permissible alleles is that the amino acid position has undergone compensatory changes. In this case the negative effects of the mutation(s) at one position of the same or different proteins compensates for the negative effects of the other mutation [35,56,59–61]. Such compensation could occur, for example, by antagonistic pleiotropy [62,63] or for protein functional reasons (e.g. [64,65]). Whatever the reason, the initial mutation needs to escape natural selection for a period of time before it is compensated by another mutation in the same or another protein. This is likely to be possible only for mutations that have very small negative fitness effects, resulting in such mutations occurring at faster-evolving positions that are biochemically less radical (e.g. [35]).

Evolutionary diagnosis of function-altering mutations *in silico*

Over a decade ago, first methods were proposed to predict computationally whether a mutation will negatively affect the structure and function of a human protein [30,66–68]. These methods, now part of the PolyPhen software package, employed physical properties of the mutational change along with a multispecies alignment as a basis to evaluate mutations. This method showed promise: 69% of mutations associated with human disorders could be correctly diagnosed to be damaging to protein function (true positives) and 66% of known population polymorphisms were correctly diagnosed as non-damaging (true negatives) [67]. Most recently, a true-positive rate of 92% was achieved by PolyPhen-2 when only damaging alleles with known effects on the molecular function causing Mendelian diseases were tested [63], which reduced to 73% when all human disease-associated mutations were analyzed. The false-positive rate was close to ~20% for PolyPhen-2.

Another early method [sorting intolerant from tolerant (SIFT)] employed multispecies alignments to distinguish between functionally neutral and deleterious amino acid changes [69]. Applications of SIFT and PolyPhen/PolyPhen-2 to predict well-characterized variants in selected

sets of genes revealed similar true-positive rates for the two programs [70,71], but these investigations revealed much higher false-positive rates (up to 68%). Comparative analyses have also revealed that the prediction accuracy of *in silico* tools depends on both the algorithm and the sequence alignment employed [71–73], with predictions from the PolyPhen-2 showing the least dependence on the alignment employed.

Over the years these *in silico* prediction tools have frequently been employed to predict the proportion of benign mutations in newly sequenced human genomes and to prioritize polymorphisms for further experimental research in humans and other species [74–81]. In all of these investigations the focus has been on diagnosing monogenic disease mutations because *in silico* tools based on evolutionary considerations are not expected to be effective for identifying nSNVs associated with complex diseases. The patterns of evolutionary conservation of known complex disease nSNVs are no different from those of natural polymorphisms found among populations (Figure 4c).

Even for Mendelian disease mutations, *in silico* diagnosis has been challenging because the diagnoses from different programs are not the same for the same variant. For example, PolyPhen and SIFT diagnoses for protein-altering mutations in the Venter genome disagreed more often than they agreed [2] (Figure 5a). Because of such problems, efforts have gone into the development of composite and ensemble methods that: (i) incorporate increasingly larger numbers of clinical and biological attributes in the decision-making process, and (ii) combine the results from existing tools by using logistic regression, Bayesian neutral networks, decision trees, support vector machines, random forests, and multiple selection rule voting [82–85]. These efforts are beginning to improve prediction accuracy significantly, and one recent method combining many less successful methods into a new composite approach was found to outperform each method used separately (Figure 5b) [85].

Many evolutionary features used by classical and advanced versions of SIFT and PolyPhen (among others) for diagnosing Mendelian disease variants are also discriminatory for differentiating between driver and passenger mutations [39,86]. This prompted the development of a hybrid method, CanPredict [86], that integrated gene function information (e.g. gene ontology) to screen somatic mutations (also [87]). This tool diagnoses mutations found in samples of more than ten patients to be damaging 50% more often than mutations that were seen in only one patient [86]. Driver mutations contribute to cancer progression and have a tendency to be found in many independent samples as compared to passenger mutations that, as the name suggests, hitchhike causing the cells with driver mutations to increase in number by the processes of natural selection and adaptation [39,40,88–90]. For mitochondrial DNA (mtDNA), four different tools (including PolyPhen and SIFT) have been combined along with the biochemical features and frequency of variants to evaluate mitochondrial nSNVs [91]. This approach was adopted because only 5% of disease-associated nSNVs in mtDNA were found to be harmful by all four *in silico* methods, even though each of these SNVs was predicted to be damaging by at least one method [91].

Efforts have been made to identify *a priori* determinants of the protein position where *in silico* tools will most probably succeed [57]. This knowledge will empower biologists to quantify the reliability of inference and use the *in silico* predictions only when they are expected to be reliable. Initial research has revealed a clear-cut relationship between the sensitivity (true-positive diagnosis) and specificity (true-negative diagnosis) of predictions with the rate at which the given position has evolved over species as diverse as fish and lamprey. The disease-associated nSNVs at slow-evolving positions were more likely to be diagnosed correctly than those at fast-evolving positions (Figure 5c). This is consistent with earlier findings that the evolutionary rate is overwhelmingly the most important determinant of the accuracy of *in silico* prediction methods [92,93]. It is also clear that the accuracy of *in silico* tools is severely degraded when the observed disease-associated variant is found in other species at the same position [57]. Therefore, the *in silico* diagnosis failures are systematic and are probably predictable.

By using evolutionary rates derived from multispecies analyses *a priori*, it should be possible to develop adaptive classifiers that have potential to generate more reliable predictions based on the evolutionary context of specific positions. Because high-quality genomic alignments between human and many closely and distantly related species are publicly available, it is possible to enumerate each multispecies aligned position in the human genome to compute position-specific features such as evolutionary rate of change. These pre-computed evolutionary features could be incorporated into prediction methods to adaptively adjust the classifier thresholds to optimize for the type of nSNVs that are likely to be observed. For example, fast-evolving positions are expected to harbor a higher proportion of neutral nSNVs, and thresholds could therefore be fine-tuned to improve overall accuracy.

Concluding remarks

The cosmic analogy used in the title of this review is intended to convey the enormity of the challenge that researchers in genomic medicine face as they attempt to decipher the functional consequences of the constellation of genomic changes carried in each personal genome. In tackling this challenge the evolutionary telescope is among a set of initial tools to generate functional predictions. Clearly, the progress made to date prompts enthusiasm, but there is an urgent need to develop better *in silico* approaches to aid and complement the array of experimental, clinical, and physical tools that must be combined to assay accurately the diversity of the functional effects of the variants present in the human population and of the *de novo* mutations that continually arise in the natural processes of cell division and population propagation.

Many limits to the use of the evolutionary approaches in genomic medicine are already evident. As mentioned earlier, *in silico* analysis of nSNVs underlying complex diseases remains a major challenge. Furthermore, there are few cases when disease categorization can be seen as a black and white decision: diseases represent a continuum

from predominately monogenic to highly polygenic [94]. Some classical monogenic diseases will surely be caused by mutations in multiple genes, whereas some classic polygenic diseases will have a few major effect alleles. This complicates the choice of when to apply evolutionary knowledge in diagnosing the function-altering potential of variants. The distinction between the neutrality and non-neutrality of function alteration is also not straightforward because it depends on both environmental and genomic contexts (e.g. compensatory mutations) and could well involve fitness trade-offs (e.g. between rapid maturation and risk of disease). Moreover, the extent to which personal variations manifest themselves as health concerns in individuals remains unknown. With an enhanced quantification of health and disease, and an improved understanding of genome and disease biology, we will have a better idea of the powers and pitfalls of evolutionary analysis in genomic medicine. At the same time there is a need to profile exome variants experimentally and connect them with individual health via predictive frameworks. Some cell-based and *in vitro* assays are already showing promise in deciphering the pathogenic roles of variants in cancers [23,95], an important step forward towards satisfying the urgent need for the development of higher throughput biological and functional approaches.

Nonetheless, the rapid emergence of clinical genome sequencing has established a pressing need to incorporate evolutionary information into clinical diagnostics. An individual genome contains hundreds of thousands of variants of different antiquities, and the long-term evolutionary history of genomic positions provides an immediate means to derive and apply predictive and quantitative assessement of the potential functional effect of any given variant observed. Using the evolutionary anatomies of positions, clinicians can be provided with ready access to evolutionary-guided *in silico* diagnostic tools to identify and diagnose the observed variants that are most likely to have consequences for the health or clinical course of treatment for a patient.

Acknowledgments

We thank Vanessa Gray and Alicia Varma for literature searching, Maxwell Sanderford for mapping 23andMe information to the dbSNP database, and Carol Williams for edits. This work is supported by research grants from National Institutes of Health to S.K.

References

1 Li, Y. and Agarwal, P. (2009) A pathway-based view of human diseases and disease relationships. *PLoS ONE* 4, e4346
2 Chun, S. and Fay, J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.* 19, 1553–1561
3 Hindorff, L.A. *et al.* (2011) A catalog of published genome-wide association studies, www.genome.gov/gwastudies
4 Hindorff, L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367
5 McCarthy, M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369
6 Roberts, R. *et al.* (2010) The genome-wide association study-a new era for common polygenic disorders. *J. Cardiovasc. Translational Res.* 3, 173–182
7 Levy, S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.* 5, e254
8 Bentley, D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59
9 Wang, J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature* 456, 60–65
10 Wheeler, D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–876
11 Ng, S.B. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272–276
12 Stenson, P. *et al.* (2009) The Human Gene Mutation Database: 2008 update. *Genome Med.* 1, 1–6
13 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
14 Ng, S.B. *et al.* (2010) Massively parallel sequencing and rare disease. *Hum. Mol. Genet.* 19, R119–R124
15 Ku, C-S. *et al.* (2011) Revisiting Mendelian disorders through exome sequencing. *Hum. Genet.* 1–20
16 Choi, M. *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19096–19101
17 Teer, J.K. and Mullikin, J.C. (2010) Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* 19, R145–R151
18 Green, E.D. and Guyer, M.S. (2011) Charting a course for genomic medicine from base pairs to bedside. *Nature* 470, 204–213
19 Carapito, R. *et al.* (2006) Automated high-throughput process for site-directed mutagenesis, production, purification, and kinetic characterization of enzymes. *Anal. Biochem.* 355, 110–116
20 Saboulard, D. *et al.* (2005) High-throughput site-directed mutagenesis using oligonucleotides synthesized on DNA chips. *Biotechniques* 39, 363–368
21 Sylvestre, J. (2010) Massive Mutagenesis: high-throughput combinatorial site-directed mutagenesis. *Methods Mol. Biol.* 634, 233–238
22 van Boxtel, R. *et al.* (2010) Systematic generation of *in vivo* G protein-coupled receptor mutants in the rat. *Pharmacogenomics J.* DOI: 10.1038/tpj.2010.44
23 Davis, E.E. *et al.* (2011) TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum. *Nat. Genet.* 43, 189–196
24 Williams, G.C. and Nesse, R.M. (1991) The dawn of Darwinian medicine. *Q. Rev. Biol.* 66, 1–22
25 Gluckman, P.D. *et al.* (2009) *Principles of Evolutionary Medicine*, Oxford University Press
26 Nesse, R.M. and Williams, G.C. (1994) *Why We Get Sick: The New Science of Darwinian Medicine*, Times Books
27 Harper, R.M.J. (1975) *Evolutionary Origins of Disease*, G. Mosdell
28 Stearns, S.C. and Koella, J.C. (2008) *Evolution in Health and Disease*, Oxford University Press
29 Miller, M.P. and Kumar, S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* 10, 2319–2328
30 Sunyaev, S. *et al.* (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.* 10, 591–597
31 Bonetta, L. (2010) Whole-genome sequencing breaks the cost barrier. *Cell* 141, 917–919
32 Coffey, A.J. *et al.* (2011) The GENCODE exome: sequencing the complete human exome. *Eur. J. Hum. Genet.* 19, 827–831
33 Harper, P.S. (2008) *A Short History of Medical Genetics*, Oxford University Press
34 Vitkup, D. *et al.* (2003) The amino-acid mutational spectrum of human genetic disease. *Genome Biol.* 4, R72
35 Subramanian, S. and Kumar, S. (2006) Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics* 7, 306
36 Kulkarni, V. *et al.* (2008) Exhaustive prediction of disease susceptibility to coding base changes in the human genome. *BMC Bioinform.* 9, S3
37 Thomas, P.D. and Kejariwal, A. (2004) Coding single-nucleotide polymorphisms associated with complex vs.Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15398–15403
38 Zhu, Q. *et al.* (2011) A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *Am. J. Hum. Genet.* 88, 458–468

39 Kaminker, J.S. *et al.* (2007) Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res.* 67, 465–473
40 Forbes, S. *et al.* (2006) COSMIC 2005. *Br. J. Cancer* 94, 318–322
41 Montoya, J. *et al.* (2009) 20 years of human mtDNA pathologic point mutations: Carefully reading the pathogenicity criteria. *Biochim. Biophys. Acta: Bioenerg.* 1787, 476–483
42 Lander, E. and Schork, N. (1994) Genetic dissection of complex traits. *Science* 265, 2037–2048
43 Thomson, G. and Esposito, M.S. (1999) The genetics of complex diseases. *Trends Genet.* 15, M17–M20
44 Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* 33, 228–237
45 Fujimura, J.H. (1996) *Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer*, Harvard University Press
46 Marenberg, M.E. *et al.* (1994) Genetic susceptibility to death from coronary heart disease in a study of twins. *N. Eng. J. Med.* 330, 1041–1046
47 Sarafino, E.P. and Goldfedder, J. (1995) Genetic factors in the presence, severity, and triggers of asthma. *Arch. Dis. Child.* 73, 112–116
48 MacGregor, A.J. *et al.* (2000) Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.* 43, 30–37
49 O'Rahilly, S. *et al.* (2005) Genetic factors in type 2 diabetes: the end of the beginning? *Science* 307, 370–373
50 Corona, E. *et al.* (2010) Extreme evolutionary disparities seen in positive selection across seven complex diseases. *PLoS ONE* 5, e12236
51 Blekhman, R. *et al.* (2008) Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* 18, 883–889
52 Podder, S. and Ghosh, T.C. (2010) Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human. *Mol. Biol. Evol.* 27, 934–941
53 Nathan, D.G. and Orkin, S.H. (2009) Musings on genome medicine: genome wide association studies. *Genome Med.* 1, 3
54 Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science* 185, 862–864
55 Briscoe, A.D. *et al.* (2004) The spectrum of human rhodopsin disease mutations through the lens of interspecific variation. *Gene* 332, 107–118
56 Kondrashov, A.S. *et al.* (2002) Dobzhansky–Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14878–14883
57 Kumar, S. *et al.* (2009) Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Res.* 19, 1562–1569
58 Magalhães, J. (2005) Human disease-associated mitochondrial mutations fixed in nonhuman primates. *J. Mol. Evol.* 61, 491–497
59 Gao, L. and Zhang, J. (2003) Why are some human disease-associated mutations fixed in mice? *Trends Genet.* 19, 678–681
60 Kulathinal, R.J. *et al.* (2004) Compensated deleterious mutations in insect genomes. *Science* 306, 1553–1554
61 Liao, B.Y. and Zhang, J. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet.* 23, 378–381
62 Williams, G.C. (1957) Pleiotropy, natural selection, and the evolution of senescence. *Evolution* 11, 398–411
63 He, X. and Zhang, J. (2006) Toward a molecular understanding of pleiotropy. *Genetics* 173, 1885–1891
64 Lee, B.C. *et al.* (2008) Analysis of the residue-residue coevolution network and the functionally important residues in proteins. *Proteins* 72, 863–872
65 Ferrer-Costa, C. *et al.* (2007) Characterization of compensated mutations in terms of structural and physico-chemical properties. *J. Mol. Biol.* 365, 249–256
66 Sunyaev, S. *et al.* (1999) Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. *J. Mol. Med.* 77, 754–760
67 Sunyaev, S. *et al.* (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.* 16, 198–200
68 Sunyaev, S.R. *et al.* (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* 12, 387–394
69 Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814
70 Flanagan, S.E. *et al.* (2010) Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet. Test. Mol. Biomarkers* 14, 533–537
71 Hicks, S. *et al.* (2011) Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum. Mutat.* 32, 661–668
72 Reva, B. *et al.* (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* 8, R232
73 Mathe, E. *et al.* (2006) Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res.* 34, 1317–1325
74 Arnett, J. *et al.* (2011) Autosomal dominant progressive sensorineural hearing loss due to a novel mutation in the *KCNQ4* gene. *Arch. Otolaryngol. Head Neck Surg.* 137, 54–59
75 Čalşkan, M. *et al.* (2011) Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the *TECR* gene on chromosome 19p13. *Hum. Mol. Genet.* 20, 1285–1289
76 Hoefele, J. *et al.* (2011) Novel *PKD1* and *PKD2* mutations in autosomal dominant polycystic kidney disease (ADPKD). *Nephrol. Dial. Transplant.* 26, 2181–2188
77 Saccone, S.F. *et al.* (2010) SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Res.* 38, W201–W209
78 McGee, T.L. *et al.* (2010) Novel mutations in the long isoform of the *USH2A* gene in patients with Usher syndrome type II or non-syndromic retinitis pigmentosa. *J. Med. Genet.* 47, 499–506
79 Doherty, D. *et al.* (2010) Mutations in 3 genes (*MKS3*, *CC2D2A* and *RPGRIP1L*) cause COACH syndrome (Joubert syndrome with congenital hepatic fibrosis). *J. Med. Genet.* 47, 8–21
80 Lee, P.H. and Shatkay, H. (2009) An integrative scoring system for ranking SNPs by their potential deleterious effects. *Bioinformatics* 25, 1048–1055
81 Kantaputra, P.N. *et al.* (2011) Cleft lip with cleft palate, ankyloglossia, and hypodontia are associated with *TBX22* mutations. *J. Dent. Res.* 90, 450–455
82 Huang, T. *et al.* (2010) Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. *PLoS ONE* 5, e11900
83 Ng, P.C. and Henikoff, S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* 7, 61–80
84 Mort, M. *et al.* (2010) *In silico* functional profiling of human disease-associated and polymorphic amino acid substitutions. *Hum. Mutat.* 31, 335–346
85 Gonzalez-Perez, A. and Lopez-Bigas, N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* 88, 440–449
86 Kaminker, J.S. *et al.* (2007) CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res.* 35, W595–W598
87 Carter, H. *et al.* (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* 69, 6660–6667
88 Greenman, C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153–158
89 Hon, L.S. *et al.* (2008) Computational approaches for predicting causal missense mutations in cancer genome projects. *Curr. Bioinform.* 3, 46–55
90 Bignell, G.R. *et al.* (2010) Signatures of mutation and selection in the cancer genome. *Nature* 463, 893–898
91 Bhardwaj, A. *et al.* (2009) MtSNPscore: a combined evidence approach for assessing cumulative impact of mitochondrial variations in disease. *BMC Bioinform.* 10, S7
92 Tian, J. *et al.* (2007) Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinform.* 8, 450
93 Jiang, R. *et al.* (2007) Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations. *Am. J. Hum. Genet.* 81, 346–360
94 McKusick, V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* 80, 588–604

95 Molatore, S. *et al.* (2010) Characterization of a naturally-occurring p27 mutation predisposing to multiple endocrine tumors. *Mol. Cancer* 9, 116
96 Chen, R. *et al.* (2010) Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS ONE* 5, e13574
97 Pelak, K. *et al.* (2010) The characterization of twenty sequenced human genomes. *PLoS Genet.* 6, e1001111
98 Forbes, S.A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 39, D945–D950
99 Altshuler, D.M. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58
100 Ramensky, V. *et al.* (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30, 3894–3900
101 Lasko, T.A. *et al.* (2005) The use of receiver operating characteristic curves in biomedical informatics. *J. Biomed. Inform.* 38, 404–415
102 Ng, S.B. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* 42, 30–35
103 Adzhubei, I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249