REVIEW

The origins, determinants, and consequences of human mutations

Jay Shendure and Joshua M. Akey

Germline mutations are the principal cause of heritable disease and the ultimate source of evolutionary change. Similarly, somatic mutations are the primary cause of cancer and may contribute to the burden of human disease more broadly than previously appreciated. Here, we review recent insights into the rates, spectrum, and determinants of genomic mutations and how these parameters inform our understanding of both Mendelian and complex human diseases. We also consider models for conceptualizing mutational consequences and outline several key areas for future research, including the development of new technologies to access and quantify the full spectrum of mutations, as well as to better interpret the consequences of mutations with respect to molecular functionality, evolutionary fitness, and disease pathogenicity.

Despite the exquisite molecular mechanisms that have evolved to replicate and repair DNA with high fidelity, mutations happen. Each human is estimated to carry on average ~60 de novo point mutations (with considerable variability among individuals) that arose in the germline of their parents (*1*–*4*). Consequently, across all seven billion humans, about 10^{11} germline mutations—well in excess of the number of nucleotides in the human genome—occurred in just the last generation (*5*). Furthermore, the number of somatic mutations that arise during development and throughout the lifetime of each individual human is potentially staggering, with proliferative tissues such as the intestinal epithelium expected to harbor a mutation at nearly every genomic site in at least one cell by the time an individual reaches the age of 60 (*6*).

Advances in DNA sequencing (*7*) have enabled the identification of human germline and somatic mutations at a genome-wide scale. These studies have confirmed, refined, and extended our understanding on the origins, mechanistic basis, and empirical characteristics of human mutations, including both replicative and nonreplicative errors (*8*), heterogeneity in the rates and spectrum of mutations within and between the genomes of individuals (*1*–*3*, *9*–*11*), the influence of sex and parental age on mutation rates (*2*, *4*, *12*), and the similarities and differences between patterns and characteristics of germline and somatic mutations (*13*–*15*). Yet, gaps in interpreting the functional, phenotypic, and fitness effects of mutations remain. These gaps must be filled if we are to effectively identify de novo disease-causing mutations (*16*), distinguish between causal and noncausal mutations in cancer (*13*), and interpret the genetic architecture of human diseases (*17*).

The germline human mutation rate

A number of distinct approaches have been used to estimate the germline mutation rate of base substitutions (Fig. 1), which we focus on here unless otherwise noted. Historically (*18*), and even more recently (*6*), estimates of mutation rates have been derived from the incidence of highly penetrant Mendelian diseases. The largest such study aggregated data across ~60 loci, estimating an average germline mutation rate of 1.28×10^{-8} per base pair (bp) per generation (*6*). However, disease-based estimates make a number of assumptions, and because inferences are confined to a small number of loci, they may not be representative of mutation rates at large. Phylogenetic methods have also been used to estimate mutation rates at putatively neutral loci on the basis of the amount of sequence divergence between humans and nonhuman primates, yielding a higher genome-wide average germline mutation rate of 2.2×10^{-8} per bp per generation (*19*). Phylogenetic methods also make assumptions such as the time to most recent common ancestor between humans and nonhuman primates, generation time, and that the loci studied do not have fitness consequences. In addition, phylogenetic estimates may be influenced by evolutionary processes other than mutation and selection, such as biased gene conversion, which influences substitution rates in mammals (*20*).

New sequencing technologies have enabled more direct estimates of mutation rates by identifying de novo mutations in pedigrees (i.e., those observed in a child but not their parents). Whole-genome sequencing studies (*1*–*3*, *9*–*11*) of pedigrees estimate the germline mutation rate to be ~1.0×10^{-8} per bp per generation [extensively reviewed in (*8*)], which is less than half that of phylogenetic methods but in better agreement with disease-based estimates. An important caveat to pedigree-based sequencing is that heavy data filtering is necessary and analysis choices may influence both false-positive and false-negative rates (*8*). Nonetheless, complementary approaches for estimating mutation rates on the basis of the number of "missing mutations" that would be expected to have occurred in the time between when an archaic hominin individual (such as Neanderthal) died and the present (*21*) and the accumulation of heterozygous variants within autozygous regions of founder populations (*1*) are broadly consistent with pedigree-based approaches (~1.1 and 1.2×10^{-8} per bp per generation, respectively).

Although a twofold range for the estimated germline mutation rate might appear inconsequential, it has important implications for our understanding of human evolution and disease (e.g., influencing estimates of effective population sizes as well as the inferred timing of when modern humans separated from other archaic hominin groups, when modern humans dispersed out of Africa, and the time of divergence between modern human populations more generally) (*22*). Moreover, there is accumulating evidence that de novo mutations and rare variants in coding sequences contribute not only to rare Mendelian diseases (*16*) but also to common but genetically heterogeneous diseases such as neurodevelopmental disorders (*23*) and early-onset breast cancer (*24*). The observation that de novo mutations and rare variants play a major role in some common diseases suggests that these phenotypes have a large mutational target size (*16*), such that mutations of large effect that occur in any one of many distinct genes can influence disease risk (*25*). Thus, an accurate estimate of the germline mutation rate is also critical for interpreting the patterns, prevalence, and architecture of human disease (*26*).

The estimates discussed above are specifically for the germline mutation rate of single-nucleotide substitutions. There have also been some attempts to estimate the de novo mutation rate for small insertions and deletions (indels) as well as copy number variants (CNVs). For example, whole-genome sequencing in families estimated a rate of 2.94 small indels (1 to 20 bp) and 0.16 structural variants (>20 bp) per generation (*27*). Of note, structural mutations affect many more nucleotides of the genome, on average, than substitutions. It is important to recognize that whole-genome sequencing with short reads has likely underascertained structural events, particularly insertions or deletions of modest size (*28*). More refined estimates of the germline mutation rate for structural variants will likely emerge as sequencing technologies continue to improve. De novo structural events can lead both to well-defined Mendelian syndromes (*29*) and to the same neurodevelopmental diseases that de novo point mutations contribute to [reviewed in (*16*)]. Indeed, family-based discovery of such de novo CNVs helped to motivate investigations of the role of de novo point mutations in these diseases (*30*).

Germline mutations exhibit remarkable heterogeneity in rates and patterns across the genome at both fine and broad scales (*31*), with sequence composition (*32*) and functional context influencing local mutation rates. The largest effect occurs at CpG dinucleotides, where the mutation rate of cytosine is higher by a factor of ~10 than other dinucleotides, consequent to spontaneous deamination of methylated cytosine to thymine. The higher CpG content of coding sequences, along with other differences in sequence composition, may contribute to the higher germline mutation

Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA.
E-mail: shendure@uw.edu (J.S.); akeyj@uw.edu (J.M.A.)

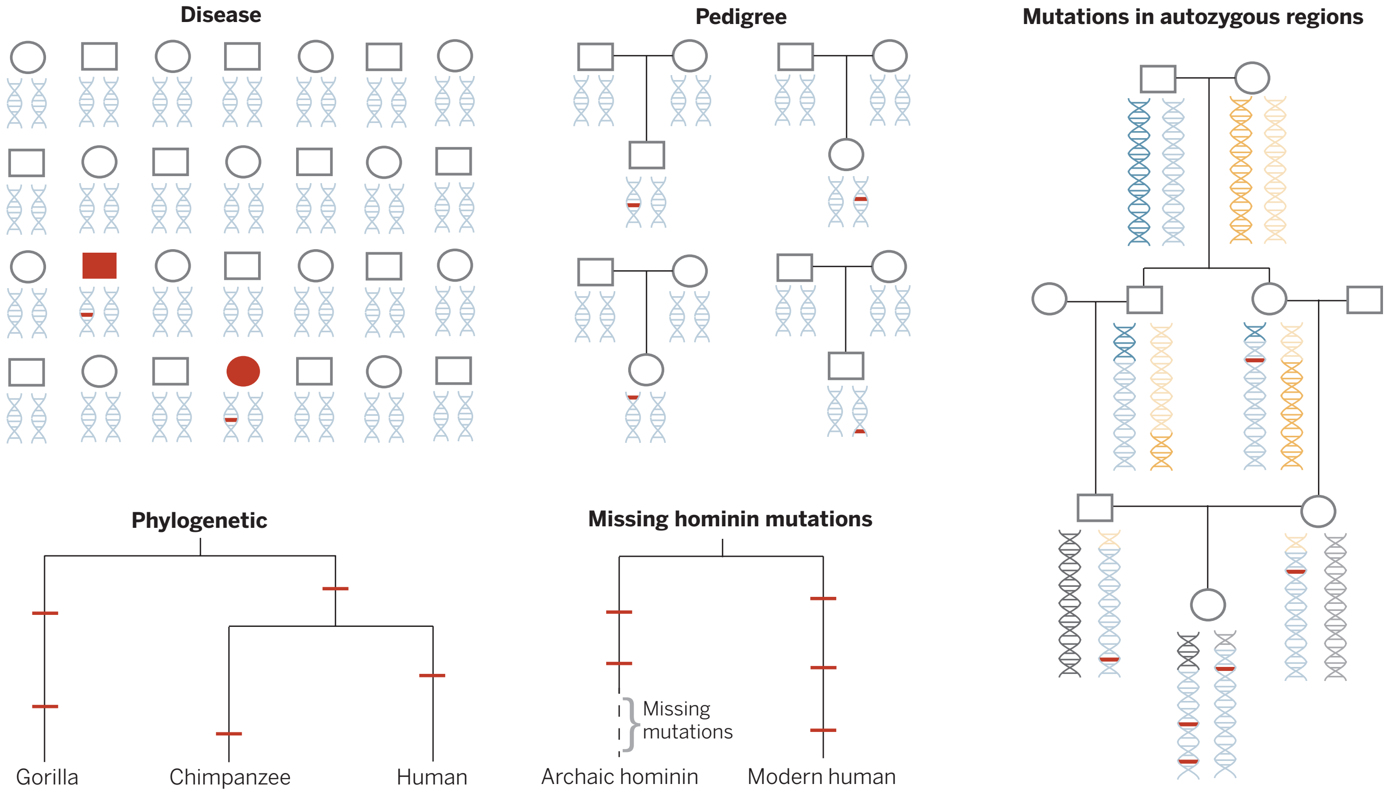

Fig. 1. Approaches to infer the human germline mutation rate. (**Top**) Methods based on the prevalence of individuals with highly penetrant Mendelian disease (denoted as filled circles and squares), identification of de novo mutations in pedigrees (mutations found in offspring but not their parents), and by finding mutations that arise in autozygous regions in pedigrees from founder populations (mutations that appear as heterozygous sites in regions of long stretches of homozygosity). (**Bottom**) Comparative genomics approaches include phylogenetic estimates (based on the number of mutations that have occurred between human and nonhuman primates) and by inferring how many mutations are missing in an archaic hominin sequence (how many mutations would have accumulated if the archaic group did not go extinct). In all panels, red lines indicate mutations.

rate observed in the exome than the genome in pedigree-based studies (~1.5 × 10^{-8} versus ~1.0 × 10^{-8} per bp per generation) (*8*). Heterogeneity in rates and patterns can also result from nonuniform repair—e.g., transcription-coupled repair of genes expressed in the germline (*33*). Fluctuations in mutation rates at larger scales, reflected in patterns of single-nucleotide polymorphism density and human-chimpanzee nucleotide divergence, likely relate in part to factors such as chromatin structure (*14*) and replication timing (*2*, *15*).

Sex and parental age effects

Consequent to sex-specific differences in germ cell biology (*34*), the majority of germline mutations resulting from errors in DNA replication are inherited from fathers. Furthermore, pedigree-based sequencing has recently yielded quantitative insights into the relationship between paternal age and the rate of de novo point mutations. Specifically, paternal age is estimated to explain 95% of the variation in the number of de novo mutations among offspring; following puberty, an additional ~1 to 2 mutations are observed per paternal year (*4*); the proportion of de novo mutations in genic regions increases by 0.26% per paternal year, such that offspring born to 40-year-old fathers carry twice as many genic mutations compared with children of 20-year-old fathers (~19.1 versus ~9.6, respectively) (*2*). These results are qualitatively similar to exome sequencing studies of sporadic autism cases, which found that de novo point mutations showed a 4:1 parental bias (*12*). Maternal age effects are generally not seen for point mutations but instead are well documented for chromosomal nondisjunction errors (*35*). Beyond point mutations, rates of nonrecurrent de novo CNVs also show a strong paternal bias and age effect (*36*), implicating replication-based mechanisms of CNV formation such as fork stalling and template switching (*37*). More broadly, differences in mutational rates, spectrum, and age effects in males and females reflect the underlying mechanisms by which various classes of mutations originate.

Somatic mutations and disease

From zygote to adult, a human undergoes trillions of cell divisions, with somatic mutations accumulating at each division. Tissues such as epithelial cells divide throughout life, and even terminally differentiated tissues continue to acquire somatic mutations through nonreplicative processes. It has been estimated that mutation rates in somatic cells are 4 to 25 times as high as in germline cells (*6*), and the acquisition of somatic mutations is intimately related to cancer. Because mutations need only be compatible with the life of a cell, rather than that of the full organism, the spectrum of mutations observed in cancer is much more diverse than typically observed in germline mutations (e.g., aneuploidy, chromothripsis, etc.). Furthermore, each individual cancer exhibits a characteristic burden and spectrum of mutation, although commonalities are present within cancer types (*13*, *38*).

Variation in fine-scale somatic mutation patterns reflect the contributions of environmental mutagens and/or intrinsic dysfunction in DNA replication or repair (*39*). At broader scales, variation in mutation rates across cancer types correlate with histone marks defining repressive versus open chromatin, replication timing, and transcription-coupled repair (*13*, *14*). It remains unclear whether mutational spectra inherently contribute to specific types of cancer, but understanding these patterns has nonetheless been key for the identification of genes that are significantly mutated over cancer-specific background levels (*13*).

It is increasingly recognized that somatic mutations underlie a much broader spectrum of human disease beyond cancer. For example, somatic mutations occurring early in development underlie a surprising fraction (6 to 20%) of studied

Mendelian disorders (*40*). Some single-gene disorders are exclusively caused by somatic mutations (*41*), presumably because germline mutations are lethal during embryonic development. Somatic mutations affecting gonadal tissues in an unaffected parent can result in multiple children with the same de novo mutation (germline mosaicism). Finally, somatic mutations also may result in reversion of constitutional disease in subsets of cells (*42*). Overall, it is likely that the role of somatic mutation in diseases other than cancer is greater than documented (*43*).

Conceptual models of mutational effects

Although the rates of germline and somatic mutations are coming into focus, interpreting their consequences remains challenging. A straightforward way to conceptualize the consequences of mutations is as having a distribution of effects. For example, in population genetics, the distribution of fitness effects (DFE) of new mutations is a well-established concept. Although estimating or measuring the DFE is challenging (discussed below), it is clear that the DFE is a complex distribution that differs between organisms as well as across genomes (*44*). Deleterious mutations have a distribution-of-fitness effects that is likely multimodal and varies by functional class, e.g., protein coding versus noncoding (*44*, *45*). Evolutionarily advantageous mutations are rare, and the distribution of their fitness effect sizes remains largely unexplored.

In considering the consequences of new mutations, we can distinguish between the concepts of fitness, pathogenicity, and molecular function (Fig. 2). All of these relate to "function" but in different ways. Fitness is a continuous property that can be defined in terms of the reproductive success of genotypes carrying the mutant allele relative to genotypes carrying the wild-type allele. Fitness effects and their distribution (i.e., the DFE) are relevant over a wide range of time scales. For example, new mutations with extremely large deleterious fitness effects may fail to transmit for even a single generation, whereas weakly deleterious mutations will exhibit allele frequency trajectories that vary over many generations as a function of population history and chance.

Pathogenicity refers to the propensity of a mutation to cause clinically manifest disease within a single individual. Historically, pathogenic mutations have been defined for Mendelian diseases, in discrete terms wherein a given mutation is classified as either disease-causing or not. However, it may be more useful to think of pathogenicity as a continuous property by which mutations or variants confer risk for a particular disease or diseases, perhaps quantified by an odds ratio. Analogous to the DFE, we can conceptualize a distribution of pathogenic effects for new mutations, in relation to a specific disease or to disease in the broader sense [referred to below as a distribution of pathogenic effects (DPE)]. At one extreme are Mendelian disease-causing variants with extremely high odds ratios (although it is worth noting that

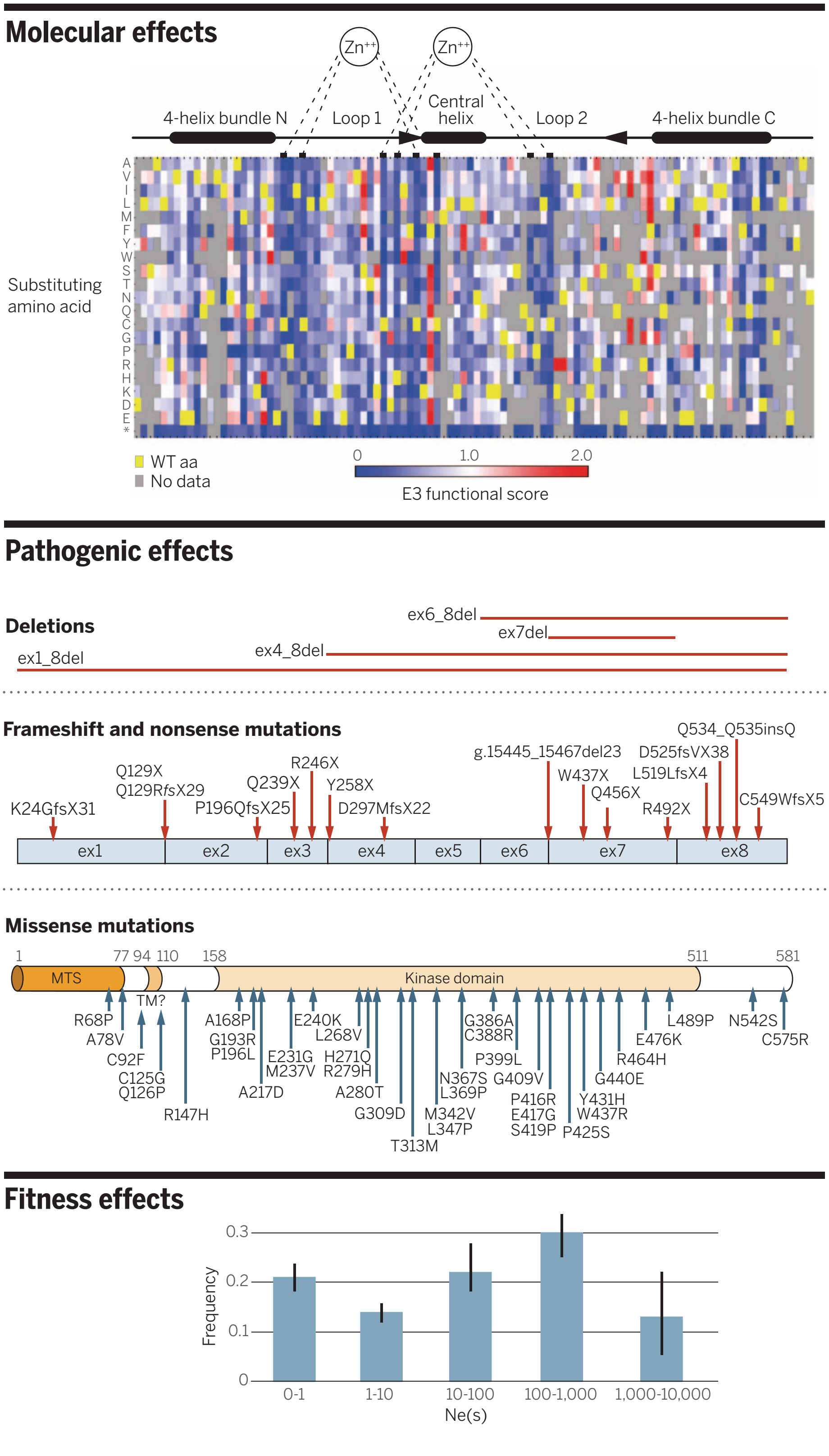

Fig. 2. Conceptual models for mutational effects. (**Top**) Molecular effects of mutation on protein structure and function, illustrated by deep mutational scanning of the RING domain of BRCA1. Figure reproduced from (*51*) with permission. (**Middle**) Pathogenic effects, illustrated by a recessive form of Parkinson's disease caused by mutations in the Parkin protein. Figure reproduced from (*60*) with permission. (**Bottom**) Fitness effects, illustrated by a histogram of inferred values for N_e*s, the product of effective population size (N_e), and the strength of selection (s). [Figure reproduced from (*44*) with permission]

because Mendelian disease-causing mutations typically are only ascertained in diseased individuals, we often lack formal measurements of their penetrance). In contrast, the vast majority of variants or haplotypes implicated by genome-wide association studies (GWAS) have very small odds ratios and may be neither necessary nor sufficient to cause disease.

Finally, mutations have evolutionary (e.g., fitness) or organismal (e.g., pathogenic) consequences by virtue of disruptive effects that they have at the molecular level. The molecular consequences of a given mutation—whether in a protein-coding or regulatory sequence—are of course highly dependent on the function of the sequence in which it resides and, moreover, likely to be highly context dependent (i.e., relevant in only certain cell types and developmental stages). Nonetheless, all of these contingencies can be conceptualized within a distribution of molecular effects (DME) for mutations in particular sequences or across the genome.

The DFE, DPE, and DME for a given sequence are undoubtedly correlated. For instance, variants with large molecular effects are more likely to be deleterious and/or pathogenic, and pathogenicity can be predicted on the basis of deleteriousness (*46*). However, the relationship between these distributions is not straightforward. For example, a variant might have no measurable contribution to disease status but might nonetheless affect reproductive fitness over short or long time scales. A highly penetrant mutation for disease that affects later life may not affect reproductive fitness. A variant might have a large molecular effect, but this might have only environmentally dependent consequences. At present, there is no single measure of mutational effects that is useful in all circumstances, and the choice of what type of mutational effect makes the most sense to measure or predict depends heavily on one's goals.

The DFE has been estimated experimentally by mutation accumulation or mutagenesis, followed by fitness measurements. These approaches are limited to model organisms, often rely on mutagens, and generally only identify mutations with large fitness effects. An alternative approach is to compare genomes for patterns of fixed or standing variation (*47*). The DFE clearly differs for coding and regulatory sequences in the human genome, and although a higher proportion of coding mutations have large known fitness effects, the absolute amount of noncoding sequence under purifying selection is greater than that of coding sequence [although precisely how much greater is debated (*48*)].

In principle, the DPE can be measured by examining the odds ratios for individual mutations in a gene associated with a particular disease. However, in practice, this is challenging. First, mutations are usually ascertained on the basis of phenotype, such that we lack quantification of how often pathogenic mutations occur in disease-free individuals. Second, although in principle every mutation that is compatible with life occurs in several hundred present-day humans (*5*), we lack whole genomes for billions of humans such that most mutations are—and will, for the near future, remain—unobserved. Finally, we can only quantify odds ratios for those variants in GWAS that have risen to high allele frequencies. For rare alleles or new mutations, even sequencing every human on the planet might be insufficient to detect low or modest effect sizes.

In contrast to the DFE and DPE, distributions of molecular effects for particular sequences of interest are amenable to empirical measurement. For example, saturation mutagenesis studies characterized the distribution of effects on transcriptional activation (for cis-regulatory elements) or on enzymatic or signaling activities (for protein-coding sequences) for a diversity of sequences. The modern equivalents of these methods, termed massively parallel reporter assays and deep mutational scans, enable multiplex mutagenesis and functional characterization of thousands of mutations (e.g., all possible point mutations or amino acid swaps in a sequence of interest) (*49*, *50*). Of course, how the DME informs the DFE and DPE for any given sequence of interest will remain unknown. However, provided that the experimental assay used to test the sequence of interest is appropriate for the physiologically relevant function of a sequence, it may be reasonable to infer the DFE or DPE from the DME. For example, one could define subsets of mutations in a gene of interest with similar molecular effects and treat these as a group for estimating the odds ratio with which such mutations confer disease risk (*51*).

The interplay of mutation rates, mutation effects, and human disease

Any given disease has its own set of potential mutations that contribute to risk, with a particular distribution with respect to both the odds of occurring as well as odds of causing disease. For most Mendelian diseases, the mutational target consists primarily of protein-altering mutations in a specific gene. For genetically heterogeneous disorders, there may be as many as hundreds of target genes. However, this picture is even more complex if one considers that there likely exists a broader range of mutations that much more modestly affect expressivity and severity. However, such modifiers remain largely undiscoverable at present because we lack the sample sizes to robustly identify them.

Mendelian diseases are typically early onset and severe, such that the consequences of mutations underlying them with respect to pathogenicity and deleteriousness are tightly linked. Genetically complex, common disorders with more variable and often much later onset, such as type 2 diabetes and cardiovascular disease, may have substantially different genetic architectures, perhaps reflected in the fact that de novo mutations and rare variants of large effect have not, at least to date, been shown to be major contributors to the burden of these diseases. Instead, although enriched within coding sequences, the majority of the GWAS signal appears to lie within regulatory sequences defined by deoxyribonuclease I hypersensitivity (*52*). However, variants implicated in common disease tend to have very modest effects with respect to pathogenicity. Moreover, many variants underlying GWAS are unlikely to have been strongly deleterious, allowing allele frequencies to rise within the population and be detected through association studies.

To illustrate the diversity of relationships between patterns of mutation and disease, we consider three examples. First, the mutational spectrum of disease caused by nonsynonymous mutations is highly heterogeneous, with changes at arginine and glycine residues accounting for ~30% of such mutations (Fig. 3A) (*26*, *53*). This is largely due to the higher mutation rates at arginine and glycine codons, most of which begin with a CpG dinucleotide. CpG mutations at arginine codons result in amino acid changes that may be particularly disrupting to protein structure. Thus, the intrinsic mutability of particular codons combined with their biochemical effects can shape the mutational spectrum of disease-causing nonsynonymous mutations.

Second, as discussed above, there is a substantial increase in mutation rate as a function of paternal age. Whole-genome sequencing of ~250 parent-child trios identifying >11,000 de novo mutations showed that replication timing bias in de novo mutations was present in younger, but not older, fathers (*2*) (Fig. 3B). Because early replicating regions of the genome are associated with higher levels of gene density and transcriptional activity, this bias reduces the proportion of de novo mutations in coding regions. In older fathers, in addition to the overall increased rate of mutations, the mitigation or absence of this bias results in an increased proportion of mutations in genic regions (Fig. 3B).

Finally, it has been noted that certain somatic mutations in cancer appear more likely to occur on specific germline haplotypes. Replication timing quantitative trait loci (rtQTLs) have been identified that influence variation in replication timing between individuals. One such rtQTL influences replication timing of *JAK2* (Fig. 3C), a gene in which haplotype-dependent variation in mutation likelihood had previously been noted (*54*, *55*). Thus, this germline variant locally influences replication timing, which in turn may affect the rate of somatic mutations in a gene that underlies specific types of cancer.

Future challenges

The recent progress in cataloging and characterizing human mutations has raised myriad new questions, challenges, and opportunities. Studies of de novo mutations highlight the large gaps in knowledge that remain about the basic biological mechanisms that shape the mutational spectrum. Empirical patterns of mutations requiring more study include clustering of mutations (*2*), context-dependent effects (*32*), and recurrent mutation (*56*), among others. The drive to understand the underlying mechanisms and selective forces behind these phenomena should stimulate experimental studies providing a more comprehensive understanding into how and how often mutations occur.

We are also on the precipice of major opportunities, driven in large part by ongoing advances with respect to the quality and cost of DNA sequencing. We anticipate that the number of de novo germline mutations ascertained by whole-genome sequencing will exponentially grow in the next few years, identifying hundreds of thousands or millions of de novo mutations of all types across studies. Additionally, the ability to accurately sequence whole genomes from single cells, directly or expanded ex vivo (*57*), will further our understanding of rates and patterns of somatic mutation. The increasing depth of mutational catalogs—which we predict may approach saturation of the ~10^{10} possible substitution mutations to the human genome, will further our understanding not only of mutational processes but also of their consequences with respect to molecular function, deleteriousness, and pathogenicity.

It is also of considerable interest to better delimit the extent of heritable variation in germline and somatic mutation rates among individuals and how such heritable variation contributes to disease burden, and to integrate individualized mutation rate estimates into inferences of lifetime risk for particular diseases. Heritable variation of mutation rates seems plausible given the fact that genetic variation occurs in genes encoding components of DNA repair pathways, influences susceptibility to cancers (*58*), and is apparent from studies of de novo mutations (*10*). For example, there is evidence that European populations have a 50% higher germline mutation rate of TCC → TTC transitions (*59*), the most common somatic mutation found in melanoma cancers. Although it is not clear if this observation is related to ultraviolet exposure, the results are striking and suggest that mutation rates may evolve over much shorter time scales than previously thought, influencing individual and population-specific disease risks.

Last, we anticipate that massively parallel experimental approaches, including some based on genome editing, will facilitate measurements of the DME and DFE for diverse sequences, both protein-coding and regulatory. Although of great interest in their own right, such dense empirical catalogs will also inform the DPE for the same sequences, ideally enabling a solution to the long-standing challenge of variants of uncertain significance in clinical genetics.

The study of human mutations has entered an exciting new era. Although considerable technological, computational, and conceptual challenges remain, the ensuing discoveries will have profound implications for interpreting human evolutionary history, patterns and prevalence of human disease, and the mechanisms underlying one of life's most fundamental biological processes.

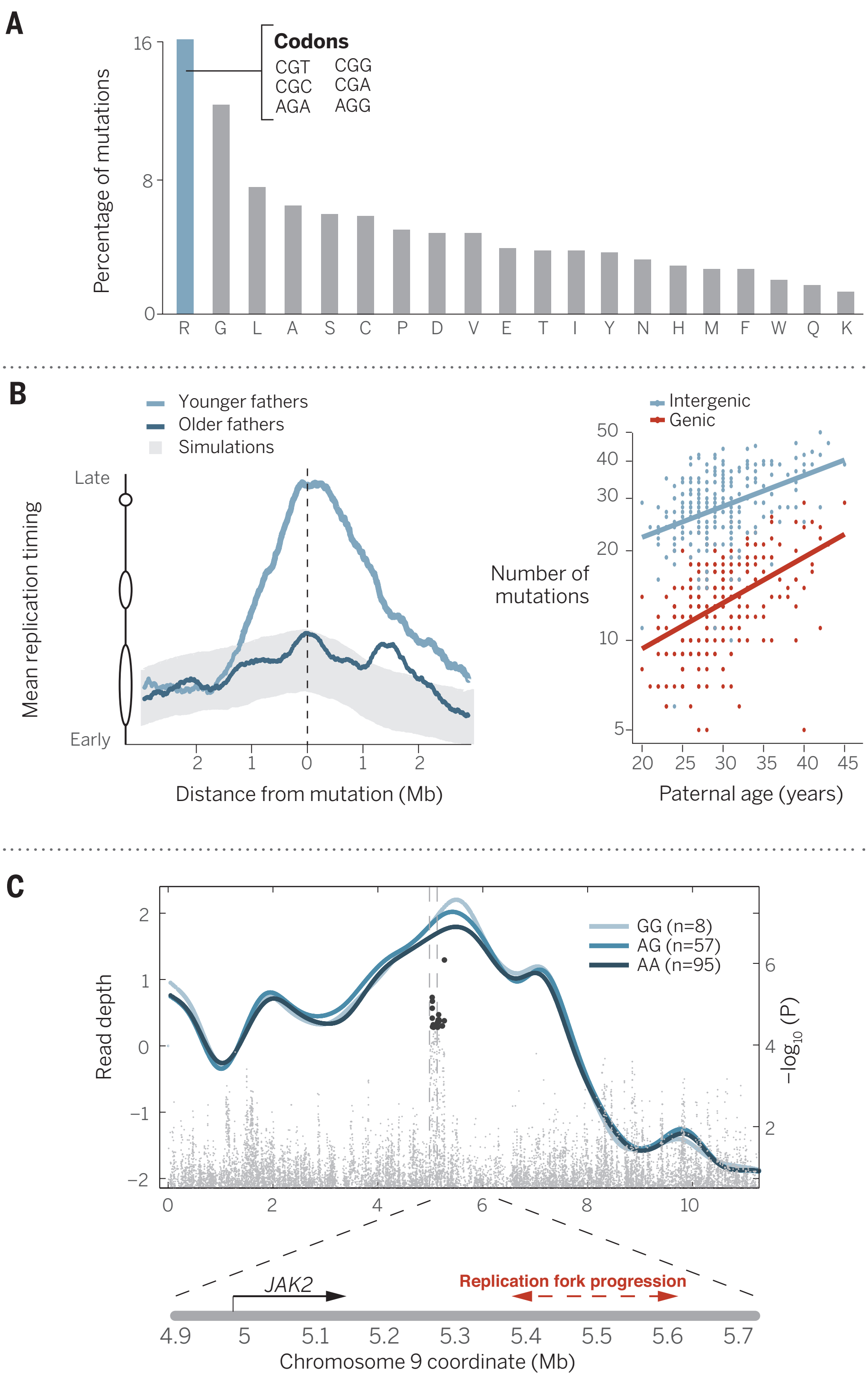

Fig. 3. Relationship between patterns of mutation and disease. (**A**) The frequency of mutations that occur at each amino acid among a panel of ~4000 putative disease-causing nonsynonymous mutations [data from (*53*)]. The high percentage of mutations at arginine (R) residues is due, in part, to its CpG-rich codons. (**B**) Mean replication timing as a function of distance from de novo mutations shows that mutations in children of younger fathers (<28 years old) are biased toward late-replicating DNA sequences (left). No significant replication timing bias is observed in the children of older fathers. The replication timing bias difference results in the accumulation of mutations in protein-coding regions at a higher rate in older compared with younger fathers (right). Figure reproduced with permission from (*2*). (**C**) Identification of a replication timing QTL (higher read depth is correlated with earlier replicating sequences) in the *JAK2* gene, which may influence rates of somatic mutation. Note the significant difference in mean replication timing as a function of genotype as measured in 160 individuals. [Figure reproduced with permission from (*54*)]

REFERENCES AND NOTES

1. C. D. Campbell *et al.*, *Nat. Genet.* **44**, 1277–1281 (2012).
2. L. C. Francioli *et al.*, *Nat. Genet.* **47**, 822–826 (2015).
3. J. C. Roach *et al.*, *Science* **328**, 636–639 (2010).

4. A. Kong *et al.*, *Nature* **488**, 471–475 (2012).
5. W. Fu, J. M. Akey, *Annu. Rev. Genomics Hum. Genet.* **14**, 467–489 (2013).
6. M. Lynch, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 961–968 (2010).
7. J. Shendure, H. Ji, *Nat. Biotechnol.* **26**, 1135–1145 (2008).
8. L. Ségurel, M. J. Wyman, M. Przeworski, *Annu. Rev. Genomics Hum. Genet.* **15**, 47–70 (2014).
9. Y. H. Jiang *et al.*, *Am. J. Hum. Genet.* **93**, 249–263 (2013).
10. D. F. Conrad *et al.*, *Nat. Genet.* **43**, 712–714 (2011).
11. J. J. Michaelson *et al.*, *Cell* **151**, 1431–1442 (2012).
12. B. J. O'Roak *et al.*, *Nature* **485**, 246–250 (2012).
13. M. S. Lawrence *et al.*, *Nature* **499**, 214–218 (2013).
14. B. Schuster-Böckler, B. Lehner, *Nature* **488**, 504–507 (2012).
15. J. A. Stamatoyannopoulos *et al.*, *Nat. Genet.* **41**, 393–395 (2009).
16. J. A. Veltman, H. G. Brunner, *Nat. Rev. Genet.* **13**, 565–575 (2012).
17. J. K. Pritchard, N. J. Cox, *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
18. J. B. Haldane, *J. Genet.* **83**, 235–244 (2004).
19. Chimpanzee Sequencing and Analysis Consortium, *Nature* **437**, 69–87 (2005).
20. L. Duret, P. F. Arndt, *PLOS Genet.* **4**, e1000071 (2008).
21. Q. Fu *et al.*, *Nature* **514**, 445–449 (2014).
22. A. Scally, R. Durbin, *Nat. Rev. Genet.* **13**, 745–753 (2012).
23. D. H. Geschwind, J. Flint, *Science* **349**, 1489–1494 (2015).
24. T. Walsh *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12629–12633 (2010).
25. J. McClellan, M. C. King, *Cell* **141**, 210–217 (2010).
26. D. N. Cooper, M. Krawczak, *Hum. Genet.* **85**, 55–74 (1990).
27. W. P. Kloosterman *et al.*, *Genome Res.* **25**, 792–801 (2015).
28. M. J. Chaisson *et al.*, *Nature* **517**, 608–611 (2015).
29. J. R. Lupski, *Nat. Genet.* **39** (suppl.), S43–S47 (2007).
30. J. Sebat *et al.*, *Science* **316**, 445–449 (2007).
31. A. Hodgkinson, A. Eyre-Walker, *Nat. Rev. Genet.* **12**, 756–766 (2011).
32. D. G. Hwang, P. Green, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13994–14001 (2004).
33. P. Green, B. Ewing, W. Miller, P. J. Thomas, E. D. Green, *Nat. Genet.* **33**, 514–517 (2003).
34. J. F. Crow, *Nat. Rev. Genet.* **1**, 40–47 (2000).
35. S. L. Sherman *et al.*, *Hum. Mol. Genet.* **3**, 1529–1535 (1994).
36. J. Y. Hehir-Kwa *et al.*, *J. Med. Genet.* **48**, 776–778 (2011).
37. J. A. Lee, C. M. Carvalho, J. R. Lupski, *Cell* **131**, 1235–1247 (2007).
38. L. B. Alexandrov *et al.*, *Nature* **500**, 415–421 (2013).
39. I. Martincorena, P. J. Campbell, *Science* **349**, 1483–1489 (2015).
40. R. P. Erickson, *Mutat. Res.* **705**, 96–106 (2010).
41. M. J. Lindhurst *et al.*, *N. Engl. J. Med.* **365**, 611–619 (2011).
42. R. Hirschhorn, *J. Med. Genet.* **40**, 721–728 (2003).
43. A. Poduri, G. D. Evrony, X. Cai, C. A. Walsh, *Science* **341**, 1237758 (2013).
44. A. Eyre-Walker, P. D. Keightley, *Nat. Rev. Genet.* **8**, 610–618 (2007).
45. M. Soskine, D. S. Tawfik, *Nat. Rev. Genet.* **11**, 572–582 (2010).
46. M. Kircher *et al.*, *Nat. Genet.* **46**, 310–315 (2014).
47. F. Racimo, J. G. Schraiber, *PLOS Genet.* **10**, e1004697 (2014).
48. P. Green, B. Ewing, *Science* **340**, 682 (2013).
49. D. M. Fowler, S. Fields, *Nat. Methods* **11**, 801–807 (2014).
50. R. P. Patwardhan *et al.*, *Nat. Biotechnol.* **27**, 1173–1175 (2009).
51. L. M. Starita *et al.*, *Genetics* **200**, 413–422 (2015).
52. A. Gusev *et al.*, *Am. J. Hum. Genet.* **95**, 535–552 (2014).
53. D. Vitkup, C. Sander, G. M. Church, *Genome Biol.* **4**, R72 (2003).
54. A. Koren *et al.*, *Cell* **159**, 1015–1026 (2014).
55. P. J. Campbell, *Nat. Genet.* **41**, 385–386 (2009).
56. A. Hodgkinson, E. Ladoukakis, A. Eyre-Walker, *PLOS Biol.* **7**, e1000027 (2009).
57. S. Behjati *et al.*, *Nature* **513**, 422–425 (2014).
58. F. Altieri, C. Grillo, M. Maceroni, S. Chichiarelli, *Antioxid. Redox Signal.* **10**, 891–938 (2008).
59. K. Harris, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3439–3444 (2015).
60. O. Corti, S. Lesage, A. Brice, *Physiol. Rev.* **91**, 1161–1218 (2011).

ACKNOWLEDGMENTS

This work was supported by NIH grants R01GM110068 and U01HG007591 to J.M.A. and DP1HG007811 to J.S. Due to space constraints, all relevant literature could not be cited, and we apologize for omissions.

10.1126/science.aaa9119

REVIEW

Somatic mutation in cancer and normal cells

Iñigo Martincorena[1] and Peter J. Campbell[1,2]*

Spontaneously occurring mutations accumulate in somatic cells throughout a person's lifetime. The majority of these mutations do not have a noticeable effect, but some can alter key cellular functions. Early somatic mutations can cause developmental disorders, whereas the progressive accumulation of mutations throughout life can lead to cancer and contribute to aging. Genome sequencing has revolutionized our understanding of somatic mutation in cancer, providing a detailed view of the mutational processes and genes that drive cancer. Yet, fundamental gaps remain in our knowledge of how normal cells evolve into cancer cells. We briefly summarize a number of the lessons learned over 5 years of cancer genome sequencing and discuss their implications for our understanding of cancer progression and aging.

Although most of the somatic mutations that steadily accumulate in our cells are harmless, occasionally a mutation affects a gene or regulatory element and leads to a phenotypic consequence. A fraction of these mutations can confer a selective advantage to the cell, leading to preferential growth or survival of a clone. We use the term "driver mutation" to denote mutations under positive selection within a population of cells, and we use "passenger mutation" for variants that have either no phenotypic consequences or biological effects that are not selectively advantageous to the clone (*1*). One end product of somatic cell evolution is cancer, a disease in which an autonomous clone of cells escapes from both the in-built programs of normal somatic cell behavior and the exogenous restraints on cell proliferation.

A very brief history of somatic mutation and cancer

Cancer results from the clonal expansion of a single abnormal cell. In 1914, the observation of chromosomal abnormalities in cancer cells was one of the first links between mutation and cancer (*1*). The causal role of somatic mutations in cancer was later supported by the discovery that many carcinogenic chemicals are also mutagenic (*2*). Conclusive evidence came from studies showing that the introduction of DNA fragments from cancer cells into normal cells led to malignant transformation and also from the identification of the responsible mutations in the transforming DNA (*1*). This work led to the discovery of the first oncogenes, whose mutation can bring about a gain of function that drives transformation into cancer. In parallel, studies on hereditary cancers led to the discovery of tumor suppressor genes (*3*), which are typically inactivated by mutations, either germline or somatic.

As the link between somatic mutation and cancer was established, cancer was described as an example of Darwinian evolution, in which cells acquire the hallmarks of cancer through somatic mutation and selection (*4*, *5*). This remains a widely accepted framework for understanding the progression of cancer, but we still lack quantitative information about the role of different factors in the evolution of normal cells into cancer cells.

In the past decade, high-throughput DNA sequencing has enabled the systematic sequencing of more than 10,000 cancer exomes and 2500 whole cancer genomes. This has revolutionized our understanding of the genetics of cancer, leading to the discovery of previously unrecognized cancer genes, new mutational signatures, and fresh insights into cancer evolution.

Mutational processes in cancer

Mutations arise from replication errors or from DNA damage that is either repaired incorrectly or left unrepaired. DNA damage can be caused by exogenous factors, including chemicals, ultraviolet (UV) light, and ionizing radiation; by endogenous factors, such as reactive oxygen species, aldehydes, or mitotic errors; or by enzymes involved in DNA repair or genome editing, among others (*6*). Additionally, viruses and endogenous retrotransposons can cause insertions of DNA sequence.

The rates of different mutational processes vary among tumors and cancer types (Fig. 1A). Though numbers vary widely, most cancers carry 1000 to 20,000 somatic point mutations and a few to hundreds of insertions, deletions, and rearrangements (*7*–*10*). Pediatric brain tumors and leukemias typically have the lowest numbers of mutations, whereas tumors induced by exposure to mutagens, such as lung cancers (tobacco) or skin cancers (UV rays), present the highest rates (*8*–*10*). Although these are common figures, some cancers acquire dramatically increased mutation rates due to the loss of repair pathways or chromosome integrity checkpoints (*6*, *8*). Depending on which process is affected, this can manifest as a very high rate of point mutations, microsatellite instability, or chromosome instability.

[1] Wellcome Trust Sanger Institute, Hinxton CB10 1SA, Cambridgeshire, UK. [2] Department of Haematology, University of Cambridge, Cambridge, UK.
*Corresponding author. E-mail: pc8@sanger.ac.uk

The origins, determinants, and consequences of human mutations

Jay Shendure and Joshua M. Akey

Science **349**, 1478 (2015);
DOI: 10.1126/science.aaa9119

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.