Trends in Genetics

CellPress
REVIEWS

Opinion

# Substitutions Are Boring: Some Arguments about Parallel Mutations and High Mutation Rates

Maximilian Oliver Press [1,*] Ashley N. Hall,[2,3] Elizabeth A. Morton,[2] and Christine Queitsch[2,*]

**Extant genomes are largely shaped by global transposition, copy-number fluctuation, and rearrangement of DNA sequences rather than by substitutions of single nucleotides. Although many of these large-scale mutations have low probabilities and are unlikely to repeat, others are recurrent or predictable in their effects, leading to stereotyped genome architectures and genetic variation in both eukaryotes and prokaryotes. Such recurrent, parallel mutation modes can profoundly shape the paths taken by evolution and undermine common models of evolutionary genetics. Similar patterns are also evident at the smaller scales of individual genes or short sequences. The scale and extent of this 'non-substitution' variation has recently come into focus through the advent of new genomic technologies; however, it is still not widely considered in genotype–phenotype association studies. In this review we identify common features of these disparate mutational phenomena and comment on the importance and interpretation of these mutational patterns.**

## Highlights

Single-nucleotide variants or mutations (e.g., point mutations) are less common than other variations and mutations, and cannot generate observed genomic diversity.

Genomic elements such as short tandem repeats, ribosomal RNA gene arrays, or transposable elements have extremely high mutation rates that likely contribute most mutations in eukaryotic genomes.

These high-rate elements are very diverse and their importance depends on their biological context. For example, in prokaryotes the more important such elements are plasmids and integrative and conjugative elements.

Functional elements with very high mutation rates behave very differently than functional elements with low mutation rates in evolution. Specifically, the same mutation can occur multiple times in different lineages, and evolution is no longer mutation-limited.

## The Dominance of Repetitive DNA in Mutation

**Substitution** (see Glossary) mutations, which lead to single-nucleotide variants (SNVs) do not substantially contribute to differences between species compared to other classes of mutations, and generally account for only a minority of new mutations (Table 1). As an example of the dominance of non-SNV variation, consider the difficulty of aligning whole genomes: most pairs of genomes are not sufficiently syntenic or similar in size that SNVs could play a substantial role in generating the observed diversity. Nevertheless, prominent reviews on mutation rate almost exclusively focus on substitution rates [1,2].

For example, the rate of spontaneous substitutions is lower than the rate of spontaneous short tandem repeat (STR) mutations in humans [3], and for decades **transposable elements** (TEs) have been thought to account for most spontaneous *Drosophila* mutations [4]. Such non-substitution mutational modes hold in common an idiosyncratic and high rate of per-locus mutation, and are sometimes referred to as 'repetitive' DNA mutations in that the affected DNA elements are usually present in high copy numbers in the genome. However, other mechanisms of high mutation rate are possible, as with plasmid acquisition and loss in prokaryotes. The importance of such mutational modes is illustrated by: (i) the above-cited numerical dominance of non-substitution mutations; (ii) The large genomic footprint of many classes of non-substitution mutations such as large indels, ploidy changes, and chromosomal rearrangements; (iii) the elaborate cellular machineries devoted to ameliorating or reducing the rate of devastating mutations {e.g., repeat-mediated deletion suppression in humans [5] and RNA-directed DNA methylation (RdDM) repression of TEs in plants [6]}; and (iv) the long-known

[1]Phase Genomics Inc., Seattle, WA 98195, USA
[2]Department of Genome Sciences, University of Washington, Seattle, WA 91895, USA
[3]Department of Molecular and Cellular Biology, University of Washington, Seattle, WA 91895, USA

*Correspondence:
maximilianpress5@gmail.com (M.O. Press) and queitsch@uw.edu (C. Queitsch).

overabundance in genomes of repetitive element families (particularly TEs), signifying past mutations [7].

In this review we take these points of importance as largely self-evident, given their longstanding and uncontested nature (although we touch on each as needed). We focus instead on the common characteristics of highly mutable genetic elements that meet two criteria. First, we require that these mutations are not substitutions because these are extremely well-studied and reviewed elsewhere [2]. Second, we require that the modes of mutation demonstrate **parallel mutation**; that is, their rate of mutation is sufficiently high to repeatedly give rise to recurrent or repeated mutations at the same locus. More specifically, we require that these mutational modes violate the **infinite-sites model** (in many interesting cases the infinite-alleles model also will be violated) [8]. The infinite-sites model assumes that the number of possible sites is very large compared to the mutation rate, and the infinite-alleles model assumes that the same allele never arises from mutation more than once; thus, both models assume no parallel evolution.

To illustrate some of the pertinent features of mutations fulfilling these two criteria, we begin by reviewing several important classes of genomic structural variation (including variation in copy number, **satellite DNA**, transposable elements, and others). We also discuss the example of STRs in some detail because they are relatively simple and easy to study. We continue by exploring some of the biological and evolutionary consequences of different mutational modes satisfying these criteria. We also discuss cases of particular interest, including **ribosomal DNA** (rDNA) copy-number variation, a fascinating and little-understood class of variation contributing to phenotypic variation.

## The Quantitative and Qualitative Preponderance of Non-SNV Variation

The vast majority of variation in DNA sequences between organisms is due to differences in ploidy and in TE content. This is best-described in plants [9], but is also marked in animal lineages [10]. Sister species/strains of maize [11,12], rice [13], and *Arabidopsis thaliana* [14–17] differ dramatically in their TE content. Moreover, it appears that these differences arise as a result of the preferential expansion and contraction of different TE families in closely related lineages [18]. Although qualitatively distinct from highly mutable, non-mobile elements in their mutational pattern and effects [18], TEs nonetheless indisputably evince parallel mutations of high rate. They also share other features, such as attenuated linkage with surrounding variation [15,19], limiting the power of SNP-based association approaches.

A further highly mutable class of variation is satellite DNA, one of the defining architectural features of eukaryotic genomes. Satellite DNA defines centromeres, telomeres, and other components of chromosomes. Such satellites consist of short motifs (usually less than 1000 bp) arranged tandemly in very high copy number. These crucially important elements, which participate in key genomic functions such as chromosome segregation and maintenance [20], evolve at remarkable speeds [21]. For example, *Drosophila melanogaster* centromeric repeats (which are generally 5–10 bp elements) are dramatically different from closely related *Drosophila simulans* and *Drosophila mauritania* centromeric repeats (mostly ~500 bp repeats) [22]. Non-centromeric satellite DNA follows similarly divergent patterns among *Drosophila* species [23,24]. In each case, as with TEs, it appears that different families of satellite repeats have expanded in different lineages of *Drosophila* by unknown mechanisms, leading to hotspots of diversification in the least-ascertainable portions of their genomes. Similar rapid evolutionary dynamics of satellite DNA have also been observed within and between primate lineages [25].

## Glossary

**Hard and soft selective sweeps:** when favorable mutations occur in populations, they tend to increase in frequency over time until they dominate the population as a result of positive selection. In hard sweeps, positive selection is very strong and the mutation goes to fixation very quickly. In soft sweeps, selection is weaker and multiple mutations are simultaneously under positive selection, leading to complex population dynamics which are more difficult to detect and interpret.

**Infinite-sites model:** a model of molecular evolution under which it is assumed that all mutations take place at different sites. Under this assumption, parallel (or recurrent) mutation does not occur. This condition is satisfied by simply assuming that the number of sites in the genome is infinite, while keeping the mutation rate constant, such that the probability of mutation at any specific site becomes infinitesimally small.

**Parallel mutation:** a mutation that occurs at the same locus as another previous mutation, but independently from the same starting allele, usually in different genetic lineages. Mutations at the same locus in the same lineage are called 'stepwise' mutations.

**Population mutation rate:** the total number of mutations arising across the entire population of an organism. Either a larger population or a higher rate of per-locus mutation can increase this measure. Sometimes written as $\theta$.

**Ribosomal DNA (rDNA):** regions of genomes consisting of many copies of ribosomal RNA genes, which vary dramatically in copy number across species and individuals while remaining conserved in the sequence of each gene.

**Satellite DNA:** regions of DNA consisting of tandemly repeated DNA sequences at high copy number. This copy number mutates rapidly. Genomic regions such as telomeres and centromeres tend to consist of satellite DNA. Short tandem repeats (STRs), also called microsatellites, consist of very short repeat units (<10 nt).

**Substitution:** a mutation that replaces a nucleotide at a single

There are multiple additional mechanisms for mutation that depend on specific aspects of genome architecture. For example, genomes with large families of closely related genes are amenable to gene-conversion mutations, which occur by recombination between highly similar loci. Specifically, trypanosomes and some other human pathogens rely on recombination between high-copy host interaction genes as a mechanism to generate diversity in response to host selection [26,27], with consequences for public health. In these cases, these recombination events are frequently facilitated by nearby TEs.

These adaptive mechanisms can easily be observed in the laboratory. When budding yeast is grown under nutrient-limited conditions, genes encoding fitness-limiting transporters are frequently amplified to high copy number [28,29]. These adaptations are highly replicable across parallel continuous culture systems because adjacent genomic features such as origins or inverted repeats, tandemly repeated homologous genes, and mobile elements all allow elevated rates of amplification and local copy-number expansion [30,28,29].

In summary, although these various sequence elements differ wildly in their mechanism of mutation, they hold in common the features of high rate, repeatability, and even predictability. These features are also well illustrated by STR variation.

## Lessons from Short Tandem Repeat Variation Concerning Genomic Elements with High Mutation Rates

STRs (also known as microsatellites) provide a useful model for understanding the dynamics of elements with elevated mutation rates. Specifically, they are abundant, highly mutagenic, contribute to phenotypic variation, but are more or less ignored in most population genomics. Thanks to technology advances coupled with longstanding theoretical work, we now have a basic understanding of this class of variation both in terms of its population variation and its molecular and phenotypic effects. Recent studies in humans [31–33] and *A. thaliana* [34] provide high-accuracy genotypes and evidence for selective and phenotypic consequences of STR variation. We use some examples from *A. thaliana* STRs to illustrate the previously identified features of elements with high mutation rates (Figure 1): (i) the expected number

position (A, C, G, T) with one of the other three nucleotides.

**Transposable element (TE):** DNA elements that reproduce themselves in genomes via 'cut-and-paste' or 'copy-and-paste' mechanisms, leading to large insertions and deletions of DNA. Sometimes called 'selfish DNA' or 'jumping genes'.

Table 1. Rate and Genomic Impact of Various Mutation Types across Eukaryotes

| Mutation type | Mutation rate (per element per generation) | Mutation rate (per genome copy per generation)[a] | Mutation rate (bp/generation)[b] | Refs |
|---|---|---|---|---|
| Substitution | $10^{-9}$ to $10^{-8}$ | ~30 (human)<br>~1 (weed)<br>~0.6 (fly)<br>~0.1 (yeast) | ~30 (human)<br>~1 (weed)<br>~0.6 (fly)<br>~0.1 (yeast) | [30,106–108] |
| Transposition | $10^{-6}$ to $10^{-4}$ | ~0.05 (human)<br>0.001–0.2 (fly) | ~60 (human)<br>2.9–581 (fly) | [109–111] |
| STR copy-number change | $10^{-5}$ to $10^{-3}$ | ~40 (human)<br>~0.24 to 2.4 (weed)<br>~0.014 to 0.14 (yeast) | >80 (human)<br>>0.5 to >5 (weed)<br>>0.03 to >0.3 (yeast) | [3,30,34,112,113,] |

[a]Where available, estimates are taken from the literature. For weed and yeast, rates are estimated as the product of the element-wise mutation rate and the number of relevant elements (taken from the references).

[b]Where available, estimates are taken from the literature. Estimates are made based on the product of element unit size and genome-wide mutation rate. Human transposition numbers are based on size and mutation rates of *Alu*, L1, and SVA elements reported in [109]; fly transposition numbers assume the size of the *P* element (2907 bp). As a lower bound on STR bp effects, we assume that all STR mutations are a one-unit change in a dinucleotide. Throughout, 'human' is *Homo sapiens,* 'fly' is *Drosophila melanogaster*, 'yeast' is *Saccharomyces cerevisiae*, and 'weed' is *Arabidopsis thaliana*.

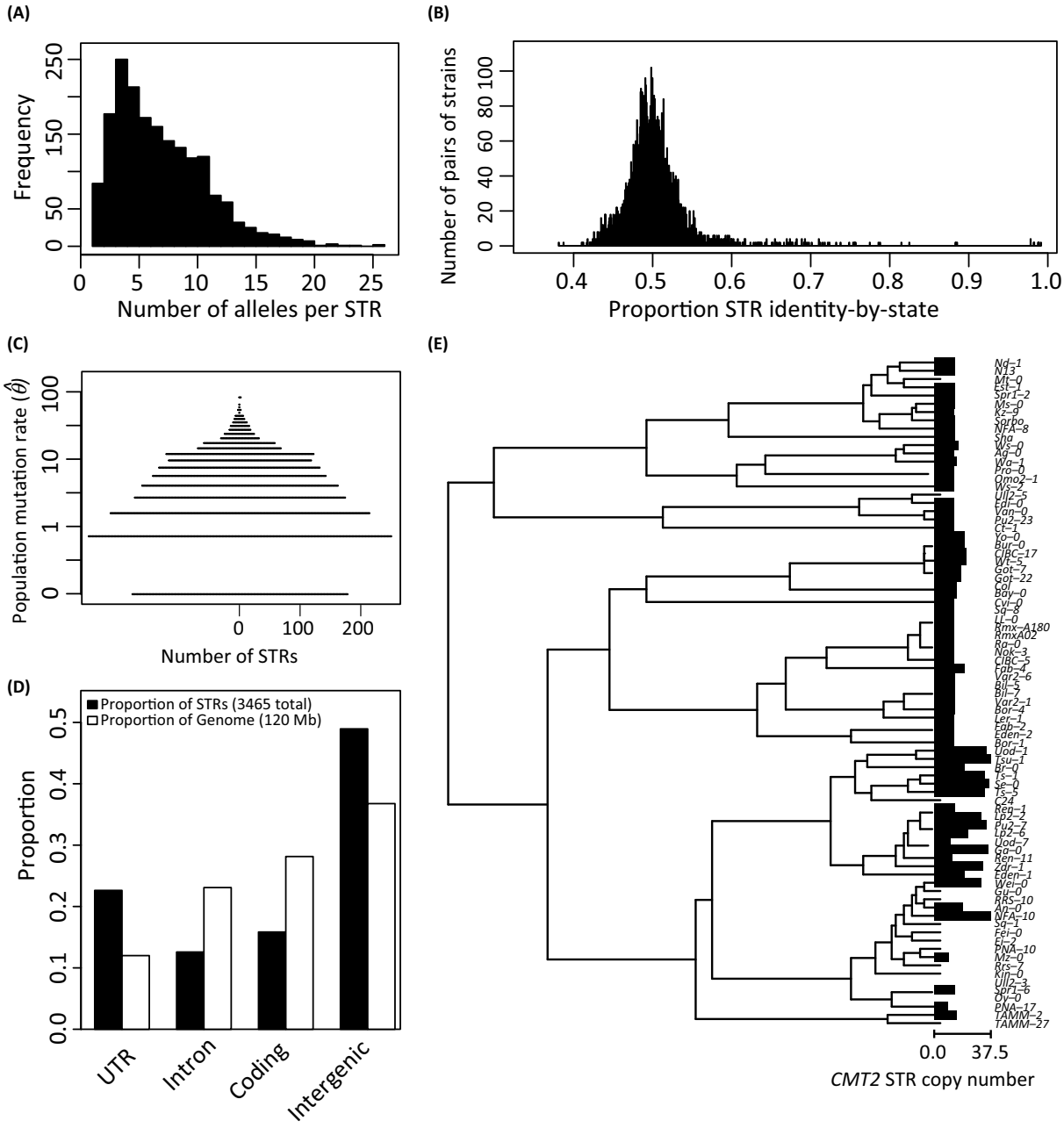


Figure 1. Short Tandem Repeats (STRs) Demonstrate Features of High Mutation-Rate Elements. STR loci demonstrate (A) multiallelism, (B) low allelic similarity between strains, (C) high mutation rate, (D) context-dependent mutation rate variation, and (E) parallelism. Data and figures adapted, with permission, from [34]. (A) Number of alleles at each STR locus. (B) All pairs of strains were compared at all positions where both strains had STR allele calls, and the proportion of alleles in common was computed. (C) Population mutation rate was computed according to [105]. Observed mutation rates of zero had a small nonzero value added such that they could be shown on the log scale. (D) Gross localization of STRs in the *A. thaliana* genome. Annotations from Araport11 were compared to STR calls from [34]; UTR, untranslated region. (E) Parallel expansions and contractions of the *CMT2* STR across *A. thaliana* strains (adapted, with permission, from [34]); the tree represents UPGMA (unweighted pair group method with arithmetic mean) clustering of strains according to full *CMT2* gene sequence, according to the Kimura 2-parameter model (which considers only transitions and transversions). Bars represent the relative copy number of the *CMT2* STR, bars are omitted in cases of missing data for a strain; observed values ranged from 8.5 to 37.5 repeat units.

of mutations and segregating alleles from high mutation-rate elements is very large (Figure 1A) and this variation has effects on phenotypic variation; (ii) the genomic context of an element strongly influences its mutation rate (Figure 1C); and (iii) several assumptions and qualitative expectations of classical evolutionary genetics are changed by high mutation rates (Figure 1E).

Most common population genomic methods and computer programs assume that loci are biallelic. This is true of less than 15% of 2046 typed STRs across 96 strains of *A. thaliana* (Figure 1A). Moreover, there is not even a 'major allele' for at least one half of STR loci because no single allele has a frequency above 50%. When comparing any two such *A. thaliana* strains, only half of STR loci will have the same allele (Figure 1B), whereas nucleotide positions will be identical at ~99% of ascertained sites in such comparisons. This demonstrates the massive population variation of these elements.

Substantial prior work has demonstrated the association of such STR variation with phenotypic variation in a variety of organisms [32,35–39]. Moreover, several studies have presented evidence that genic STRs are subject to substantial selective constraint [34,40], indicating that phenotypic effects of this STR variation contribute actively not only to evolutionary paths but also to the mutational load afflicting populations.

STR mutation rates are strongly influenced by the genomic context of the STR. For example, transcribed STRs have a substantially higher mutation rate than comparable untranscribed STRs [41]. Indeed, STRs disproportionately tend to be located in otherwise non-repetitive genic DNA [42], and specifically 5′ untranslated regions (5′-UTRs) (Figure 1D), even though selection should remove STRs from genic DNA to avoid gene disruptions. Presumably, STRs are maintained in genic DNA by a higher rate of expansion or birth in these regions. The mutagenic effect of transcription appears to increase the rate of STR unit insertions [41] that may lead to higher rates of STR 'birth' in genic sequences (although they may subsequently be removed by selection from coding sequences).

Finally, STRs show parallel mutation. For example, nonsense mutations in the *CMT2* gene in *A. thaliana* were previously described as being subject to positive selection [43], but more recently we showed that an intronic STR in this gene shows repeated dramatic changes in copy number, consistent with repeated mutation and selection (Figure 1E). Taking into account the local ancestry of this region in these *A. thaliana* strains, the most parsimonious explanation is multiple repeated mutations at this locus. The similarity of STR copy number between closely related strains suggests this may potentially (but not necessarily) occur via stepwise mutation of the STR.

## High Mutation Rates and Evolutionary Genetics

The genetic elements discussed in this paper all have high mutation rates. This is notable because the mutation rate is a key parameter in many evolutionary models. In fact, a simplifying assumption in population genetics is that the rate of evolution is equal to the mutation rate because evolution itself is often assumed to be mutation-limited [44]. This is sometimes called the strong-selection weak-mutation (SSWM) model. However, when mutation is not limiting relative to selection (e.g., 'strong mutation'), the dynamics of the evolutionary process change dramatically. For example, the seminal work on **soft selective sweeps** specifically noted recurrent mutation as a factor that would increase the frequency of soft sweeps from selected loci [45]. In distinction to **hard selective sweeps** – the rapid spread to fixation of a specific mutation on a distinctive haplotype – soft sweeps are characterized by the emergence of either multiple distinct adaptive mutations in a region, or the same adaptive mutation associated with

heterogeneous haplotypes. Soft sweeps will manifest with a **population mutation rate** ($\theta$) exceeding 0.01 even under very strong selection; the *A. thaliana* STRs discussed above have average population mutation rates in the order of 1000-fold higher than this threshold (Figure 1C). Experimental evolution experiments in microbes (which often have very high population mutation rates because of their large populations) frequently observe soft sweeps, with the emergence of multiple adaptive alleles leading to 'clonal interference' between lineages carrying different adaptive alleles, frequently at the same locus [46]. One such experiment in *Methylobacterium extorquens* yielded 17 distinct adaptive insertions into the same gene [47], a potent demonstration of the parallelism attainable with both large population sizes and high mutation rates.

High mutation-rate loci show qualitatively different behavior from low mutation-rate loci under selection (and their interaction with associated haplotypes), requiring different tools for detecting selection [45,48,49]. Therefore, vast differences in mutation rate within the genome and across mutational types can lead to dramatically different expectations for evolutionary outcomes. Simply, the SSWM model breaks down in the face of high mutation-rate genetic elements. This is because the rate of adaptation is no longer limited by the rate of mutation owing to the abundant supply of mutations at adaptive loci. Recent theoretical work suggests that this breakdown occurs with a rate $\theta > 0.1$ [50], leading to new dynamics such as population size-dependence of the rate of evolution. Again, the estimated average $\theta$ for STR mutations is $\sim$100-fold larger than this threshold (Figure 1C). This likely explains the observations of repeated mutations putatively contributing to adaptive variation at these loci, as observed for both STRs in *A. thaliana* (Figure 1E) and TEs in *Drosophila* [51]. Further theoretical work suggests that multimutation 'jumps' become possible when mutation rates are elevated relative to selection, changing the dynamics of evolution on rugged fitness landscapes [35,52].

Even in the absence of selection, there are consequences of very high mutation rates and multiallelism for the dynamics of molecular evolution [53], some of which we discuss in more detail in Box 1. Overall, the large number and apparent impact of high mutation-rate elements, combined with the proposition that adaptive evolution is mutation-limited, leads us to the natural conclusion that such elements contribute disproportionately to adaptation, even if presently available techniques are ill-suited to detecting this contribution.

## High Mutation Rates in the Human Genome

Mutation in the human genome is of inherent interest, and there is a large body of work on this subject, as reviewed elsewhere [54]. However, a few pertinent features of human mutation are worth noting here. First, much of the sequence difference between humans and great apes occurs in segmentally duplicated regions that are difficult to resolve because of high homology between duplicates [55,56]. Multiple human-specific genes with roles in neurodevelopmental processes appear to have arisen through such duplication events in the human lineage [56–58]. Second, copy-number variation is a major contributor to human genetic diversity [54], and tends to occur preferentially in repetitive regions such as pericentromeres and peritelomeres that are difficult to analyze with traditional short-read sequencing [59]. Some such regions consisting of low-copy repeats comprise 5% of the human genome, and show dramatic population variation consisting of rearrangements and large differences of copy number [60]. These variants are, moreover, nearly impossible to reconstruct without recently developed methods such as proximity ligation or optical mapping. These observations highlight again the effects of genomic context and the importance of low-complexity genomic regions in generating genetic diversity.

Box 1. Mutational Modes and Molecular Evolution

John Maynard Smith [114] proposed that molecular evolution might be understood with reference to a popular parlor game of his time, inferring the path of evolution by considering the most parsimonious number of single-letter substitutions to transform one word into another:

word→wore→gore→gone→gene

However, this set of rules (parsimony, single-letter substitution) does not necessarily describe the expected evolutionary path of a given DNA sequence. DNA sequences are altered according to rules that allow many more types of transition. For example, one might consider the following scenario instead, allowing also duplications, inversions, and rearrangements of letters:

word→drow→brow→brew→brewer→brewed→breed→breeder→breed→greed→green→greet→great →geat→gent→gene

Comparatively, this scenario is positively circuitous, and multiple steps are redundant, with no effect on the outcome. Many transitions involve addition, subtraction, or rearrangement of existing sequences. Some words ('drow', 'geat') may strain the dictionary. Nonetheless, we believe that many geneticists will (reluctantly) concede that it is a more familiar path than the simple one trod by Maynard Smith, while hastening to add that Maynard Smith's path has a higher tutelary value.

To defend this assertion, we can present some arguments which are based on the empirical failures of parsimony as a criterion in phylogenetic inference [115]. First, although the most parsimonious path may be the most likely single path, it may have a lower probability than the summation of other paths. Second, we cannot assume that all transitions have an equal probability [116], or even that transition probabilities are constant along the path [117]. These assumptions do not even hold for the single-letter substitutions in Maynard Smith's simplified model. Indeed, cursory reference to biological experience argues that transition probabilities must change based on the sequence state, and that there is very large variation in transition probabilities (i.e., mutation rates) between sites and types of transitions. Overall, we must confront the possibility that intuitively obvious paths in molecular evolution may not be the true ones, given the observed dynamics of genome architectures and sequence variation throughout evolution.

## rDNA Variation and Heritability

The ribosomal RNA genes, which are organized into high-copy regions known as rDNA, are notable for their high level of sequence conservation and are universally present throughout cellular life. The copy numbers of these genes vary enormously. rDNA copy-number variation in *A. thaliana* largely accounts for the size variation of the entire genome observed among strains [61]. Species estimates of rDNA copy number differ by orders of magnitude across eukaryotes [62], and natural isolates within a species may vary in rDNA copy number by as much as tenfold [63–68]. Moreover, rDNA copy number is highly labile as an off-target mutation in genetically manipulated yeast [69]. The expression and chromatin state of rDNA repeats are among the most tightly regulated features of the eukaryotic nucleus [70], and although only a subset of units are transcriptionally active, their gene products make up ~80% of total RNA in the cell [71]. Transcription from the rDNA (also termed nucleolar organizing regions) leads to formation of the nucleolus, the most obvious feature of gross nuclear morphology, as well as to genetic phenomena such as nucleolar dominance.

Perhaps as a result of such regulation, strong selection appears to maintain copy number, as observed in the large rDNA copy-number fluctuations observed upon disruption and subsequent complementation of yeast *orc2* mutants [54], and the return of yeast rDNA copy number back to the native ~150 copies after artificial reduction [72]. Reductions in germline rDNA copy number are heritable in *Drosophila*, but rDNA copy number also recovers rapidly in those progeny that inherited reduced rDNA arrays [73].

Many different mechanisms are proposed for the origin of rDNA copy-number variation. In yeast, transcription–replication conflicts may contribute to rDNA instability, in part because of the presence of an origin of replication in the rDNA intergenic spacer [74–76], another reminder of the importance of genomic context in determining mutation rates. Intrachromatid recombination has similarly been proposed to produce copy-number reduction [77], as well as unequal meiotic recombination leading to changes in rDNA copy number between generations; in humans there is a ~10% chance of a recombination event per meiosis per rDNA array that will result in a change in rDNA copy number [78].

The potential phenotypic consequences of rDNA variation are vast and largely unexplored. Beyond the documented fitness consequences of catastrophic reductions in rDNA copy number [79,80], no causal relationships have yet been demonstrated between phenotype and rDNA copy-number variation in the naturally occurring range, although a weak positive association has recently been found between rDNA copy number and flowering time in maize [81]. Extrachromosomal circular rDNA sequences accumulate with age in yeast [77], and both their accumulation as well as the instability of the rDNA locus itself have been proposed as causative agents of aging in yeast [82,83]. Recently, interest in the relationship between rDNA stability and cancer has arisen owing to observations that rDNA copy number modestly decreases in some cancers [84–86]. Whether rDNA may act on cell physiology through ribosome biogenesis, maintaining genome integrity [87] and the balance of heterochromatin [88], an influence on genome replication [75], or through some other mechanism remains to be resolved. One intriguing possibility is that, owing to its centrality in cellular physiology and the processing of genetic information, rDNA copy-number variation may modify the expressivity of other genetic variants [89]. rDNA copy-number alteration has been reported to have a genome-wide impact on gene expression in *Drosophila* [88], and to influence position effect variegation [90]. Human studies have further revealed an association between rDNA copy number and genome-wide gene expression, as well as an inverse relationship with mitochondrial DNA abundance [63]. The potential central role of rDNA copy-number variation in genomic structure and gene regulation places rDNA at a crucial position in research into human health and aging.

## High Mutation Rates and Parallel Evolution in Prokaryotes

Although we chiefly focus on eukaryotes, prokaryotic genomes also highlight diverse recurrent mutational modes. Indeed, large-scale reorganization and gene gain and loss are probably even more biologically significant in prokaryotes than in eukaryotes [91]. For example, pathogenic organisms carrying the genus name *Shigella* are not a genus, and indeed not even monophyletic [92]. Each lineage of *Shigella* in fact arose independently from *E. coli* ancestors by a concerted and localized process of massive gene loss and acquisition [92–94]. In this example, recurrent large mutations follow a predictable path, owing to the contextual influence of *E. coli* genome architecture, to yield the convergent outcome of the *Shigella* genome. A similar host-associated parallel mutation trajectory is known from the soil microbe *Mesorhizobium ciceri* in the form of a large 'symbiosis island' integration element that is broken up and integrated at three different genomic locations [95]. This element carries genes associated with diazotrophic symbiosis with plants, and its integration is repeatable, highly stereotyped, and can be recapitulated in the laboratory [96]. Moreover, it appears that this tripartite integration mechanism is conserved across at least the genus *Mesorhizobium* as a mechanism for facilitating the spread of beneficial mobile genetic elements [97]. More generally, adaptive horizontal-transfer events are repeatable as a result of epistasis [98], and are specifically facilitated by the genomic context of mobile elements and by associated cellular pathways and cellular features.

These well-trodden horizontal evolutionary pathways are superficially eye-catching, but in the context of microbial genomic evolution they are unremarkable. As seen in the above *M. extorquens* example [47], the large population sizes of microbes make them tractable systems for experimental evolution. Although the population sizes of experimentally evolved bacteria are sufficiently large that even substitutions are dominated by parallelism [99], these experiments emphasize the adaptive importance of non-SNV variation. Genomic optical mapping of parallel laboratory-evolved *E. coli* populations uncovered a dramatic diversity of rearrangements, generally mediated by recombination between distant insertion sequence elements [100]. Remarkably, these rearrangements were highly parallel in that most such events were observed in more than one among only 12 populations. In the same populations, the most dramatic fitness increase over decades of evolution consisted of a highly repeatable tandem gene amplification that depended on a predisposing genomic context [101].

These mechanisms for yielding repeated high-impact adaptive mutations in prokaryotes highlight the prevalence, diversity, and adaptive significance of recurrent high-rate mutation events in the dominant clades of cellular life. The phenotypic consequences of this form of variation are vast and largely unexplored.

## Concluding Remarks

We have discussed abundant cases where recurrently mutable DNA elements determine the architecture of genomes and variation in phenotypes. These highly abundant elements shape the direction of evolution through their large supply of ready genetic variation. In the past two decades the vastly improved ascertainment of SNVs and substitution mutations prompted genetic and genomic researchers to focus on this much more tractable subject of study. This focus was driven largely by the advent of automated DNA sequencers and efficient computer programs for sequence alignment, which in those early iterations experienced difficulties with other classes of genetic variation. These technological difficulties are, in some influential cases, the explicit reason for ignoring other forms of variation [1], likely biasing both results and discourse. This bias is potentially reinforced by the common assumption of quantitative genetics that genome-wide SNP genotyping is sufficient to ascertain neighboring mutations because of linkage [102] (several studies of STRs and TEs indicate that this is unlikely to be true for multiallelic loci with high mutation rates [34,31,51,103]). However, we are encouraged that recent methodological advances such as optical mapping, proximity ligation, multiplexed sequence capture, and long-read sequencing have vastly expanded the pool of accessible variants [104].

Genomic elements with high mutation rates are intrinsically difficult to analyze by molecular methods. It is possible that methodological artifacts have influenced our understanding of these elements, exactly as we argue that past methods have biased us. For this reason we must evaluate results regarding these elements with more caution than substitutional variation. Nonetheless, we believe that the balance of evidence argues for important roles of genomic elements with high mutation rates. In the future, we must investigate whether the dizzying array of molecular variation in these elements has a commensurate effect on phenotype, or whether this variation is merely a genomic extravagance (see Outstanding Questions).

### Outstanding Questions

What are the relative contributions of different mutation classes (substitutions, transpositions, copy-number changes) to heritable variation in different organisms?

The number of substitutions per generation is well ascertained across many organisms, but what is the total number of mutations – including other mutation types that are more difficult to observe?

Are there general rules for the emergence of new families of elements such as transposons or satellites with very high mutation rates?

Are there generalizable effects of different genomic contexts (e.g., pericentromeres, peritelomeres, transcribed regions, plasmids) on the rate of different mutational modes?

If the rate of evolution is not mutation-limited, does this undermine other assumptions or models in currency?

### Acknowledgments

We would like to thank members of the laboratory of C.Q. for helpful conversations. M.O.P. would like to thank attendees of the Population, Evolutionary, and Quantitative Genetics (PEQG) 2018 meeting, attendees of the Plasmid Biology 2018

meeting, and Emily Ebel for interesting scientific presentations and conversations. E.M., A.N.H., and C.Q. are supported by grant number R21AG052020 (NIA) and C.Q. is supported by grant number R01GM122088 (NIGMS). A.N.H. is supported by training grant T32GM007270 (NIGMS).

## Disclaimer Statement

M.O.P. is an employee of Phase Genomics Inc.

## References

1. Lynch, M. (2010) Evolution of the mutation rate. *Trends Genet.* 26, 345–352
2. Lynch, M. *et al.* (2016) Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* 17, 704–714
3. Willems, T. *et al.* (2016) Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *Am. J. Hum. Genet.* 98, 919–933
4. Finnegan, D.J. (1992) Transposable elements. In *The Genome of Drosophila Melanogaster* (Lindsley, D.L. and Zimm, G.G., eds), pp. 1096–1107, Academic Press
5. Mendez-Dorantes, C. *et al.* (2018) Repeat-mediated deletions can be induced by a chromosomal break far from a repeat, but multiple pathways suppress such rearrangements. *Genes Dev.* 32, 524–536
6. Bousios, A. and Gaut, B.S. (2016) Mechanistic and evolutionary questions about epigenetic conflicts between transposable elements and their plant hosts. *Curr. Opin. Plant Biol.* 30, 123–133
7. Britten, R.J. and Kohne, D.E. (1968) Repeated sequences in DNA. *Science* 161, 529–540
8. Tajima, F. (1996) Infinite-allele model and infinite-site model in population genetics. *J. Genet.* 75, 27
9. Wendel, J.F. *et al.* (2018) The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr. Opin. Genet. Dev.* 49, 1–7
10. Rodriguez, F. and Arkhipova, I.R. (2018) Transposable elements and polyploid evolution in animals. *Curr. Opin. Genet. Dev.* 49, 115–123
11. Vielle-Calzada, J.-P. *et al.* (2009) The Palomero genome suggests metal effects on domestication. *Science* 326, 1078
12. Jiao, Y. *et al.* (2017) Improved maize reference genome with single-molecule technologies. *Nature* 546, 524–527
13. Piegu, B. *et al.* (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 16, 1262–1269
14. Legrand, S. *et al.* (2018) Differential retention of transposable element-derived sequences in outcrossing *Arabidopsis* genomes. *bioRxiv* Published online July 13, 2018. http://dx.doi.org/10.1101/368985
15. Stuart, T. *et al.* (2016) Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife* 5, e20777
16. Li, Z.-W. *et al.* (2018) Transposable elements contribute to the adaptation of *Arabidopsis thaliana*. *Genome Biol. Evol.* 10, 2140–2150
17. Quadrana, L. *et al.* (2016) The *Arabidopsis thaliana* mobilome and its impact at the species level. *eLife* 5, e15716
18. Arkhipova, I.R. (2018) Neutral theory, transposable elements, and eukaryotic genome evolution. *Mol. Biol. Evol.* 35, 1332–1337
19. Kuhn, A. *et al.* (2014) Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 111, 8131–8136
20. Plohl, M. (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* 409, 72–82
21. Lower, S.S. (2018) Satellite DNA evolution: old ideas, new approaches. *Curr. Opin. Genet. Dev.* 49, 70–78
22. Talbert, P.B. *et al.* (2018) Simple and complex centromeric satellites in *Drosophila* sibling species. *Genetics* 208, 977–990
23. Jagannathan, M. *et al.* (2016) Comparative analysis of satellite DNA in the *Drosophila melanogaster* species complex. *G3 (Bethesda)* 7, 693–704
24. Wei, K.H.-C. *et al.* (2018) Variable rates of simple satellite gains across the *Drosophila* phylogeny. *Mol. Biol. Evol.* 35, 925–941
25. Cechova, M. *et al.* (2018) High inter- and intraspecific turnover of satellite repeats in great apes. *bioRxiv* Published online November 18, 2018. http://dx.doi.org/10.1101/470054
26. Talavera-Lopez, C. *et al.* (2018) Repeat-driven generation of antigenic diversity in a major human pathogen, *Trypanosoma cruzi*. *bioRxiv* Published online June 9, 2018. http://dx.doi.org/10.1101/283531
27. Faino, L. *et al.* (2016) Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res.* 26, 1091–1100
28. Brown, C.J. *et al.* (1998) Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol. Biol. Evol.* 15, 931–942
29. Gresham, D. *et al.* (2008) The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet.* 4, e1000303
30. Lynch, M. *et al.* (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 105, 9272–9277
31. Willems, T.F. *et al.* (2014) The landscape of human STR variation. *Genome Res.* 24, 1894–1904
32. Gymrek, M. *et al.* (2015) Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* 48, 22–29
33. Quilez, J. *et al.* (2016) Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* 44, 3750–3762
34. Press, M.O. *et al.* (2018) Massive variation of short tandem repeats with functional consequences across strains of *Arabidopsis thaliana*. *Genome Res.* 28, 1169–1178
35. Press, M.O. *et al.* (2014) The overdue promise of short tandem repeat variation for heritability. *Trends Genet.* 30, 504–512
36. Tabib, A. *et al.* (2016) A polynucleotide repeat expansion causing temperature-sensitivity persists in wild Irish accessions of *Arabidopsis thaliana*. *Front. Plant Sci.* 7, 1311
37. Hannan, A.J. (2018) Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* 19, 286–298
38. Bilgin Sonay, T. *et al.* (2015) Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Res.* 25, 1591–1599
39. Mackay, T.F.C. *et al.* (2012) The *Drosophila melanogaster* genetic reference panel. *Nature* 482, 173–178
40. Gymrek, M. *et al.* (2017) Interpreting short tandem repeat variations in humans using mutational constraint. *Nat. Genet.* 49, 1495–1501
41. Zavodna, M. *et al.* (2018) The effects of transcription and recombination on mutational dynamics of short tandem repeats. *Nucleic Acids Res.* 46, 1321–1330
42. Morgante, M. *et al.* (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30, 194–200

43. Shen, X. *et al.* (2014) Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. *PLoS Genet.* 10, e1004842
44. Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* 217, 624–626
45. Pennings, P.S. and Hermisson, J. (2006) Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.* 2, e186
46. Lang, G.I. *et al.* (2013) Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* 500, 571–574
47. Lee, M.-C. and Marx, C.J. (2013) Synchronous waves of failed soft sweeps in the laboratory: remarkably rampant clonal interference of alleles at a single locus. *Genetics* 193, 943–952
48. Haasl, R.J. *et al.* (2014) The Effects of microsatellite selection on linked sequence diversity. *Genome Biol. Evol.* 6, 1843–1861
49. Haasl, R.J. and Payseur, B.A. (2016) Fifteen years of genome-wide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol. Ecol.* 25, 5–23
50. de Koning, A.P.J. and De Sanctis, B.D.D. (2018) The rate of molecular evolution when mutation may not be weak. *bioRxiv* Published online August 28, 2018. http://dx.doi.org/10.1101/259507
51. Rech, G.E. *et al.* (2018) Stress response, behavior, and development are shaped by transposable element-induced mutations in *Drosophila*. *bioRxiv* Published online July 30, 2018. http://dx.doi.org/10.1101/380618
52. Katsnelson, M.I. *et al.* (2018) On the feasibility of saltational evolution. *bioRxiv* Published online August 23, 2018. http://dx.doi.org/10.1101/399022
53. Kimura, M. and Ohta, T. (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. U. S. A.* 75, 2868–2872
54. Campbell, C.D. and Eichler, E.E. (2013) Properties and rates of germline mutations in humans. *Trends Genet.* 29, 575–584
55. Dennis, M.Y. and Eichler, E.E. (2016) Human adaptation and evolution by segmental duplication. *Curr. Opin. Genet. Dev.* 41, 44–52
56. Dennis, M.Y. *et al.* (2017) The evolution and population diversity of human-specific segmental duplications. *Nat. Ecol. Evol.* 1, 69
57. Dennis, M.Y. *et al.* (2012) Human-specific evolution of novel SRGAP2 genes by incomplete segmental duplication. *Cell* 149, 912–922
58. Dougherty, M.L. *et al.* (2017) The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biol.* 18, 49
59. Monlong, J. *et al.* (2017) Human copy number variants are enriched in regions of low mappability. *Nucleic Acids Res.* 46, 7236–7249
60. Demaerel, W. *et al.* (2018) The 22q11 low copy repeats are characterized by unprecedented size and structure variability. *bioRxiv* Published online September 12, 2018. http://dx.doi.org/10.1101/403873
61. Long, Q. *et al.* (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.* 45, 884–890
62. Prokopowich, C.D. *et al.* (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome* 46, 48–50
63. Gibbons, J.G. *et al.* (2014) Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nat. Commun.* 5, 4850
64. Thompson, O. *et al.* (2013) The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. *Genome Res.* 23, 1749–1762
65. Parks, M.M. *et al.* (2018) Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Sci. Adv.* 4, eaao0665
66. Ritossa, F.M. and Scala, G. (1969) Equilibrium variations in the redundancy of rDNA in *Drosophila melanogaster*. *Genetics* 61, 305–317
67. Mohan, J. and Ritossa, F.M. (1970) Regulation of ribosomal RNA synthesis and its bearing on the bobbed phenotype in *Drosophila melanogaster*. *Dev. Biol.* 22, 495–512
68. James, S.A. *et al.* (2009) Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing. *Genome Res.* 19, 626–635
69. Kwan, E.X. *et al.* (2016) rDNA copy number variants are frequent passenger mutations in *Saccharomyces cerevisiae* deletion collections and de novo transformants. *G3* 6, 2829–2838
70. Preuss, S. and Pikaard, C.S. (2007) rRNA gene silencing and nucleolar dominance: insights into a chromosome-scale epigenetic on/off switch. *Biochim. Biophys. Acta Gene Struct. Expr.* 1769, 383–392
71. Warner, J.R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* 24, 437–440
72. Kobayashi, T. *et al.* (1998) Expansion and contraction of ribosomal DNA repeats in *Saccharomyces cerevisiae*: requirement of replication fork blocking (Fob1) protein and the role of RNA polymerase I. *Genes Dev.* 12, 3821–3830
73. Lu, K.L. *et al.* (2018) Transgenerational dynamics of rDNA copy number in *Drosophila* male germline stem cells. *eLife* 7, e32421
74. Sanchez, J.C. *et al.* (2017) Defective replication initiation results in locus specific chromosome breakage and a ribosomal RNA deficiency in yeast. *PLoS Genet.* 13, e1007041
75. Kwan, E.X. *et al.* (2013) A natural polymorphism in rDNA replication origins links origin activation with calorie restriction and lifespan. *PLoS Genet.* 9, e1003329
76. Takeuchi, Y. *et al.* (2003) Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes Dev.* 17, 1497–1506
77. Sinclair, D.A. and Guarente, L. (1997) Extrachromosomal rDNA circles – a cause of aging in yeast. *Cell* 91, 1033–1042
78. Stults, D.M. *et al.* (2008) Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res.* 18, 13–18
79. Ritossa, F.M. *et al.* (1966) A molecular explanation of the bobbed mutants of *Drosophila* as partial deficiencies of 'ribosomal' DNA. *Genetics* 54, 819–834
80. French, S.L. *et al.* (2003) In exponentially growing *Saccharomyces cerevisiae* cells, rRNA synthesis is determined by the summed RNA polymerase I loading rate rather than by the number of active genes. *Mol. Cell. Biol.* 23, 1558–1568
81. Li, B. *et al.* (2018) Co-regulation of ribosomal RNA with hundreds of genes contributes to phenotypic variations. *Genome Res.* 28, 1555–1565
82. Ganley, A.R.D. *et al.* (2009) The effect of replication initiation on gene amplification in the rDNA and its relationship to aging. *Mol. Cell* 35, 683–693
83. Saka, K. *et al.* (2013) Cellular senescence in yeast is regulated by rDNA noncoding transcription. *Curr. Biol.* 23, 1794–1798
84. Wang, M. and Lemos, B. (2017) Ribosomal DNA copy number amplification and loss in human cancers is linked to tumor genetic context, nucleolus activity, and proliferation. *PLoS Genet.* 13, e1006994
85. Xu, B. *et al.* (2017) Ribosomal DNA copy number loss and sequence variation in cancer. *PLoS Genet.* 13, e1006771
86. Udugama, M. *et al.* (2018) Ribosomal DNA copy loss and repeat instability in ATRX-mutated cancers. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4737–4742
87. Kobayashi, T. (2008) A new role of the rDNA and nucleolus in the nucleus – rDNA instability maintains genome integrity. *Bioessays* 30, 267–272
88. Paredes, S. *et al.* (2011) Ribosomal DNA deletions modulate genome-wide gene expression: 'rDNA-sensitive' genes and natural variation. *PLoS Genet.* 7, e1001376

89. Waddington, C.H. (1942) Canalization of development and the inheritance of acquired characters. *Nature* 150, 563–565
90. Paredes, S. and Maggert, K.A. (2009) Ribosomal DNA contributes to global chromatin regulation. *Proc. Natl. Acad. Sci. U. S. A.* 106, 17829–17834
91. Puigbò, P. *et al.* (2014) Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* 12, 66
92. Pupo, G.M. *et al.* (2000) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10567–10572
93. Zhang, Y. and Lin, K. (2012) A phylogenomic analysis of *Escherichia coli/Shigella* group: implications of genomic features associated with pathogenicity and ecological adaptation. *BMC Evol. Biol.* 12, 174
94. Touchon, M. *et al.* (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5, e1000344
95. Haskett, T.L. *et al.* (2016) Assembly and transfer of tripartite integrative and conjugative genetic elements. *Proc. Natl. Acad. Sci. U. S. A.* 113, 12268–12273
96. Haskett, T.L. *et al.* (2018) Sequential induction of three recombination directionality factors directs assembly of tripartite integrative and conjugative elements. *PLoS Genet.* 14, e1007292
97. Haskett, T.L. (2017) Evolutionary persistence of tripartite integrative and conjugative elements. *Plasmid* 92, 30–36
98. Press, M.O. *et al.* (2016) Evolutionary assembly patterns of prokaryotic genomes. *Genome Res.* 26, 826–833
99. Woods, R. (2006) Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 103, 9107–9112
100. Raeside, C. *et al.* (2014) Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *mBio* 5, e01377-14
101. Blount, Z.D. (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 489, 513–518
102. Bush, W.S. and Moore, J.H. (2012) Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* 8, e1002822
103. Sawaya, S. *et al.* (2016) Linkage disequilibrium between single nucleotide polymorphisms and hypermutable loci. *bioRxiv* Published online August 10, 2016. http://dx.doi.org/10.1101/020909
104. Dijk, E.L. van (2018) The third revolution in sequencing technology. *Trends Genet.* 34, 666–668
105. Haasl, R.J. and Payseur, B.A. (2010) The number of alleles at a microsatellite defines the allele frequency spectrum and facilitates fast accurate estimation of θ. *Mol. Biol. Evol.* 27, 2702–2715
106. Ossowski, S. *et al.* (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327, 92–94
107. Keightley, P.D. *et al.* (2014) Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics* 196, 313–320
108. Campbell, C.D. *et al.* (2012) Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* 44, 1277–1281
109. Burns, K.H. and Boeke, J.D. (2012) Human transposon tectonics. *Cell* 149, 740–752
110. Nuzhdin, S.V. and Mackay, T.F. (1995) The genomic rate of transposable element movement in *Drosophila melanogaster*. *Mol. Biol. Evol.* 12, 180–181
111. Bergman, C.M. and Bensasson, D. (2007) Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* 104, 11340–11345
112. Verstrepen, K.J. *et al.* (2005) Intragenic tandem repeats generate functional variability. *Nat. Genet.* 37, 986–990
113. Vinces, M.D. *et al.* (2009) Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* 324, 1213–1216
114. Maynard Smith, J. (1970) Natural selection and the concept of a protein space. *Nature* 225, 563–564
115. Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410
116. Yang, Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11, 367–372
117. Fitch, W.M. and Markowitz, E. (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* 4, 579–593