Substitutions are boring: some arguments about parallel mutations and repetitive DNA

Maximilian Oliver Press^{1*}, Ashley N Hall^{2,3}, Elizabeth A Morton², Christine Queitsch^{2*}

*: to whom correspondence should be addressed: maximilianpress5@gmail.com,
queitsch@uw.edu

1: Phase Genomics Inc. 4000 Mason Rd, Seattle, WA 98195

2: University of Washington Department of Genome Sciences, Seattle, WA 91895

3: University of Washington Department of Molecular and Cellular Biology, Seattle, WA 91895

Keywords: mutation, short tandem repeats, transposons, repetitive DNA, rDNA, parallel mutation

Abstract

Extant genomes are largely shaped by the global transposition, copy number fluctuation, and rearrangement of DNA sequences, rather than by the 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 directly undermine the Wright-Fisher model of evolutionary genetics. Similar patterns are also evident at the smaller scales of individual genes or short genomic sequences. The scale and extent of this ‘non-substitution’ variation has only recently come into focus through the advent of new genomic technologies; however, it still not widely included in genotype-phenotype association studies. In this review, we identify the common features of these disparate mutational phenomena and comment on the importance and interpretation of these mutational patterns.

The dominance of repetitive DNA in mutation

Substitution mutations do not substantially contribute to evolutionary novelty compared to other classes of mutations, in terms of either genome fraction or number of new mutation events. Consider the task of aligning whole genomes: most pairs of genomes are not syntenic

1 enough or similar enough in size for a substantial role of substitutions in generating the
2 observed diversity. Further, the rate of spontaneous substitutions is substantially lower than
3 the rate of spontaneous short tandem repeat mutations in humans [1]; similarly, it has been
4 estimated for decades that transposable elements (TEs) account for most spontaneous
5 *Drosophila* mutations [2]. Such non-substitution mutational modes hold in common an
6 idiosyncratic and high rate of per-locus mutation, and are sometimes referred to as “repetitive”
7 DNA mutations, in that the affected DNA elements usually exist in high copy numbers in the
8 genome. However, other mechanisms of high mutation rate are possible, as with plasmid
9 acquisition and loss in prokaryotes. The importance of such mutational modes is illustrated by:

- 11 1) The above-cited numerical dominance of non-substitution mutations;
- 12 2) The large effect size of many classes of non-substitution mutations such as large indels,
13 ploidy changes, and chromosomal rearrangements;
- 14 3) The elaborate cellular machineries devoted to ameliorating or reducing the rate of
15 devastating mutations (*e.g.* repeat-mediated deletion suppression in humans [3] and
16 RNA-directed DNA methylation (RdDM) repression of TEs in plants [4])
- 17 4) The long-known high prevalence of repetitive elements signifying past mutations in
18 genomes [5].

19
20 In this Review, we take these points of importance as largely self-evident, given their long-
21 standing and uncontested nature (though we do touch on each as needed). We focus
22 principally on the common features of the classes of mutation that they describe, along with
23 their relevance to evolutionary and complex trait genetics.

24
25 For the purposes of this review, specific classes of mutations must fulfill only two requirements.
26 First, we require that these mutations *not be substitutions*, as this mode of mutation is
27 extremely well-studied and reviewed elsewhere [6] (and also arguably less important,
28 considering the above points). Second, we require that the modes of mutation demonstrate
29 *parallel mutation*; that is, that their rate of mutation is sufficiently high to repeatedly give rise

1 to recurrent or repeated mutations at the same locus. More specifically, we require that these
2 mutational modes violate the infinite-sites model (in many interesting cases the infinite-alleles
3 model also will be violated) [7].

4 To illustrate some of the pertinent features of mutations fulfilling these two criteria, we
5 will begin by discussing short tandem repeats (STRs), which are (relatively) simple and easy to
6 study. We will continue by exploring some of the biological and evolutionary consequences of
7 different mutational modes satisfying these criteria. We will additionally discuss cases of
8 particular interest, including ribosomal DNA (rDNA) copy number variation, a fascinating and
9 little-understood class of variation contributing to phenotypic variation.

11 **Lessons from short tandem repeat (STR) variation concerning genomic elements with high** 12 **mutation rates.**

13 STRs (also known as microsatellites) provide a useful model for understanding the dynamics of
14 elements with elevated mutation rates. Specifically, they are abundant, highly mutagenic,
15 contribute to phenotypic variation, and more or less ignored in most population genomics.
16 Thanks to technology advances coupled with long-standing theoretical work, we now have a
17 fairly complete understanding of this class of variation, in terms of both its population variation
18 and its molecular and phenotypic effects. Recent studies in humans [8–10] and *Arabidopsis*
19 *thaliana* [11] provide high-accuracy genotypes and evidence for selective and phenotypic
20 consequences of STR variation. We use some examples from our previous study of *A. thaliana*
21 STRs to illustrate compelling features of elements with high mutation rates (Figure 1):

- 23 1. The expected number of mutations and segregating alleles from high-mutation-rate
24 elements is very large (Figure 1A; and this variation has effects on phenotypic variation).
- 25 2. The genomic context of an element strongly influences its mutation rate (Figure 1C).
- 26 3. Several assumptions and qualitative expectations of classical evolutionary genetics are
27 changed by high mutation rates (Figure 1E).

1 Most common population genomic methods and computer programs assume that loci are
2 biallelic. This is true of less than 15% of 2,046 typed STRs across 96 strains of *A. thaliana* (Figure
3 1A). Moreover, there is not even a “major allele” for at least half of STR loci, because no single
4 allele has a frequency above 50%. When comparing any two such *A. thaliana* strains, only half
5 of STR loci will have the same allele (Figure 1B), whereas nucleotide positions will be identical at
6 ~99% of sites in such comparisons. This demonstrates the massive population variation of these
7 elements.

8 This variation has phenotypic consequences. Substantial prior work has demonstrated
9 the importance of such STR variation for phenotypic variation in a variety of organisms [9,12–
10 16]. Moreover, several studies have presented evidence that genic STRs are subject to
11 substantial selective constraint [11,17], indicating that phenotypic effects of this STR variation
12 contribute actively not only to evolutionary paths, but also to the mutational load afflicting
13 populations. Additionally, several groups have noted the empirical tendency of STR variation to
14 demonstrate attenuated linkage disequilibrium (LD) with neighboring single nucleotide
15 polymorphisms (SNPs) [8,11,18,19]. This is due both to the high mutation rate of STRs and also
16 the formal definition of LD, which will generally decrease with many alleles. Consequently, we
17 cannot rely on SNPs as markers to tag nearby STRs, as these SNPs will show only weak linkage
18 with any given, potentially causal STR allele. Attenuated linkage disequilibrium is likely true for
19 any highly mutagenic element.

20 STR mutation rates are strongly influenced by the genomic location of the STR. For
21 example, transcribed STRs have a substantially higher mutation rate than comparable
22 untranscribed STRs [20]. This observation may resolve the long-standing but puzzling
23 observation that STRs disproportionately tend to be located in otherwise non-repetitive genic
24 DNA [21], and specifically 5' UTRs (Figure 1D), even though selection should remove STRs from
25 genic DNA due to their propensity to lead to gene disruptions. Presumably, STRs are maintained
26 in genic DNA by a higher rate of expansion or birth from low-complexity sequence in these
27 regions. This likely leads to the surprisingly large number of STRs in untranslated regions,
28 because selection would remove most STR mutations in coding or intronic regions, which are
29 more sensitive to such variation. Specifically, the mutagenic effect of transcription appears to

1 increase the rate of STR unit insertions [20], which may lead to higher rates of STR “birth” in
2 genic sequences (though they may be removed at some high rate from coding sequence).

3 Finally, STRs show extensive parallel mutation. For example, the nonsense mutations in
4 the *CMT2* gene in *A. thaliana* were previously described as subject to positive selection [22],
5 but more recently we showed that an intronic STR in this gene shows repeated dramatic
6 changes in copy number consistent with repeated mutation and selection (Figure 1E). Taking
7 into account the local ancestry of this region in these *A. thaliana* strains, the most parsimonious
8 explanation is multiple repeated mutations at this locus. The phylogenetic non-independence
9 of STR copy number between closely related strains suggests this may potentially (but not
10 necessarily) occur via stepwise mutation of the STR.

11

Figure 1. 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 from [11]. (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 number of alleles in common was computed. (C): Population mutation rate was computed according to [23]. (D): Gross localization of STRs in the *A. thaliana* genome. Annotations from Araport11 were compared to STR calls from [11]; UTR: untranslated regions. (E) Parallel expansions and contractions of the *CMT2* STR across *A. thaliana* strains, adapted from [11]; tree represents UPGMA clustering of strains according to

1 full *CMT2* gene sequence, according to the Kimura 2-parameter model (which considers only
2 transitions and transversions). Bars represent the relative copy number of the *CMT2* STR, bars
3 are omitted in cases of missing data for a strain, ranging from 8.5 to 37.5 repeat units.

5 **The quantitative and qualitative preponderance of non-SNV variation.**

6 The vast majority of variation in DNA sequences between organisms is due to differences in
7 ploidy and in transposable element (TE) content. This is most well-known in plants [24], but is
8 also marked in animal lineages [25]. Sister species/strains of *Arabidopsis* [26–29], maize [30,31],
9 or rice [32] differ dramatically in their TE content. Moreover, it appears that these differences
10 arise due to the preferential expansion and contraction of different TE families in closely related
11 lineages [33]. Although qualitatively distinct from highly mutable, non-mobile elements in their
12 mutational pattern and effects [33], TEs nonetheless indisputably evince parallel mutations of
13 high rate. They additionally share other features, such as attenuated linkage with surrounding
14 variation [27,34], limiting the power of SNP-based association approaches.

15
16 Even ignoring variation from TEs and STRs, a recent study comparing human and three great
17 ape genomes estimates that more than twice as many positions are affected by indels as by
18 substitutions, even when restricting analysis to the positions tractable to alignment [35]. We
19 can expect that a complete accounting of all sequence content will only exaggerate this ratio.

20 For example, one of the defining architectural features of eukaryotic genomes is large
21 regions of satellite DNA, which define centromeres, telomeres, and other features. Such
22 satellites consist of short motifs (usually less than 1000bp) arranged tandemly in very high copy
23 number. These critically important elements, which participate in crucial genomic functions
24 such as chromosome segregation and maintenance, evolve at remarkable speeds. For example,
25 *Drosophila melanogaster* centromeric repeats (which are generally 5-10bp elements) are
26 dramatically different from closely related *Drosophila simulans* and *Drosophila mauritania*
27 centromeric repeats (mostly ~500bp repeats) [36]. Non-centromeric satellite DNA follows
28 similarly divergent patterns among *Drosophila* species [37,38]. In each case, as with TEs, it
29 appears that different families of satellite repeats have expanded in different lineages of

1 *Drosophila* by unknown mechanisms, leading to hotspots of diversification in the least-
2 ascertainable portions of their genomes.

3
4 There are a large number of additional mechanisms for mutation that depend on specific
5 aspects of genome architecture. For example, genomes with large families of closely related
6 genes are amenable to gene conversion mutations, which occur by recombination between
7 highly similar loci. Specifically, trypanosomes and some other human pathogens rely on
8 recombination between high-copy host interaction genes as a mechanism to generate diversity
9 in response to host selection [39,40], with consequences for public health. In these cases, these
10 recombination events are frequently facilitated by nearby TEs.

11
12 Lastly, when budding yeast is grown under sulfur-limited conditions, within 50-100 generations
13 the *SUL1* sulfate transport gene will be amplified to high copy number [41], leading to a fitness
14 increase of ~40%. This adaptation is highly replicable across parallel continuous culture systems
15 due to an adjacent origin of replication and inverted repeats flanking *SUL1*, which allow for a
16 highly predictable mechanism of amplification and local copy number expansion [42].

17 In summary, although these various sequence elements differ wildly in their mechanism of
18 mutation, they hold in common the features of high rate, repeatability and even predictability.
19 These features have consequences when considered in the context of evolution, as we discuss
20 next.

21 **High mutation rates and evolutionary genetics.**

22 Mutation rate is a key parameter in many evolutionary models. It is in fact a simplifying
23 assumption in population genetics that the rate of evolution is equal to the mutation rate, as
24 evolution itself is often assumed to be mutation-limited [43]. This is sometimes called the
25 strong-selection weak-mutation (SSWM) model. However, when mutation is not limiting
26 relative to selection (*e.g.* “strong mutation”), the dynamics of the evolutionary process change
27 dramatically. For example, the seminal work on “soft” selective sweeps specifically noted
28 recurrent mutation as a factor that would increase the frequency of soft sweeps from selected
29

1 loci [44]. In distinction to “hard” sweeps, the rapid spread to fixation of a specific mutation on a
2 distinctive haplotype, soft sweeps are characterized by the emergence of either multiple
3 distinct adaptive mutations in a region, or the same adaptive mutation associated with
4 heterogeneous haplotypes. Soft sweeps will manifest with population mutation rates (θ)
5 greater than 0.01 even under very strong selection; the *A. thaliana* STRs discussed above have
6 average population mutation rates on the order of one thousand times higher than this
7 threshold (Figure 1C).

8 This difference leads to qualitatively different behavior of high-mutation-rate loci under
9 selection (and their interaction with associated haplotypes), requiring different tools for
10 detecting selection [44–46]. Therefore, vast differences in mutation rate within the genome
11 and across mutational types can lead to dramatically different expectations for evolutionary
12 outcomes. Simply, the SSWM model breaks down in the face of high-mutation-rate genetic
13 elements. Other recent theoretical work notes that this breakdown occurs when $\theta > 0.1$ [47],
14 leading to new dynamics such as population-size-dependence of the rate of evolution. Again,
15 estimated average θ for STR mutations is approximately 100 times as large as this threshold
16 (Figure 1C). This likely explains the observations of repeated mutations putatively contributing
17 to adaptive variation for these loci, as observed for both microsatellites in *A. thaliana* (Figure
18 1E) and TEs in *Drosophila* [48]. Further theoretical work has suggested that multi-mutation
19 “jumps” become possible with elevated mutation rates relative to selection, changing dynamics
20 of evolution on rugged fitness landscapes [12,49].

21 Even in the absence of selection, there are consequences of very high mutation rates
22 and multiallelism for the dynamics of molecular evolution [50], some of which we discuss in
23 more detail in Box 1. Overall, the large number and apparent impact of high-mutation-rate
24 elements, combined with the proposition that adaptive evolution is mutation-limited, leads us
25 to the natural conclusion that such elements contribute disproportionately to adaptation, even
26 if presently available techniques are ill-suited to detecting this contribution.

27
28 **High mutation rates in the human genome.**

Mutation of the human genome specifically is of inherent interest, and there exists a large body of work on this subject, reviewed elsewhere [51]. 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 due to high homology between duplicates [52,53]. More specifically, multiple human-specific genes with roles in neurodevelopmental processes appear to have arisen through such duplication events in the human lineage [53–55]. Second, copy number variation is major contributor to human genetic diversity [51], and tends to occur preferentially in repetitive regions such as the pericentromeres and peritelomeres that are difficult to analyze with traditional short read sequencing [56]. These observations highlight again the effects of genomic context and the importance of low-complexity genomic regions in generating genetic diversity.

rDNA variation and heritability.

The ribosomal RNA genes, which are organized into high copy regions known as the rDNA, are notable in the genome for their high level of sequence conservation. The copy numbers of these genes, however, are a genetic feature of enormous variability. rDNA copy number variation in *A. thaliana* largely accounts for the size variation of the entire genome observed among strains [57]. Species estimates of rDNA copy number differ by orders of magnitude across eukaryotes [58], and natural isolates within a species may vary in rDNA copy number by as much as 10-fold [59–64]. Moreover, rDNA copy number is highly labile as an off-target mutation in yeast [65]. The expression and chromatin state of rDNA repeats are among the most tightly regulated features of the eukaryotic nucleus [66], and while only a subset of units are transcriptionally active, the gene products from them make up ~80% of total RNA in the cell [67]. Transcription from the rDNA (also termed nucleolar organizing regions) leads to formation of the nucleolus, the most obvious feature of gross nuclear morphology, and to such long-appreciated genetic phenomena as nucleolar dominance.

Perhaps due to 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 [63] and the return of yeast rDNA copy

1 number back to the native ~150 copies after artificial reduction [69]. Reductions in germline
2 rDNA copy number are heritable in *Drosophila*, yet rDNA copy number also recovers rapidly in
3 those progeny that inherited reduced rDNA arrays [70].

4 Of relevance to understanding the consequences of rDNA copy number variation are the
5 many different mechanisms proposed for its generation. In yeast, transcription-replication
6 conflicts may contribute to rDNA instability, due in part to the presence of an origin of
7 replication in the rDNA intergenic spacer [68,71,72], another reminder of the importance of
8 genomic context in determining mutation rates. Intrachromatid recombination has similarly
9 been proposed to produce copy number reduction [73], along with unequal meiotic
10 recombination leading to changes in rDNA copy number between generations; in humans there
11 is ~10% chance of a recombination in an rDNA array that will result in a change in rDNA copy
12 number [74].

13 The potential phenotypic consequences of rDNA variation are vast and largely
14 unexplored. Beyond the documented fitness consequences of catastrophic reductions in rDNA
15 copy number [75,76], no causal relationships have yet been demonstrated between phenotype
16 and rDNA copy number variation in the naturally-occurring range, although a weak positive
17 association has recently been found between rDNA copy number and flowering time in maize
18 [77]. Extrachromosomal circular rDNA sequences accumulate with age in yeast [73], and both
19 their accumulation as well as instability of the rDNA locus itself have been proposed causative
20 agents of aging in yeast [78,79]. Recently, interest in the relationship between rDNA stability
21 and cancer has arisen, due to observations that rDNA copy number modestly decreases in some
22 cancers [80–82]. Whether rDNA may act on cell physiology through ribosome biogenesis,
23 maintaining genome integrity [83], balance of heterochromatin [84], influence on genome
24 replication [71], or through some other mechanism remains to be resolved. One intriguing
25 possibility is that, due to its centrality in cellular physiology and the processing of genetic
26 information, rDNA copy number variation may affect not specific traits but the expressivity of
27 other genetic variants [85]. rDNA copy number alteration has been reported to have genome-
28 wide impact on gene expression in *Drosophila* [84], and to influence position effect variegation
29 [86]. Human studies have further revealed an association between rDNA copy number and

1 genome-wide gene expression, as well as an inverse relationship with mitochondrial DNA
2 abundance [59]. The potential central role of rDNA copy number variation in genomic structure
3 and gene regulation puts rDNA at a critical point of research into human health and aging.

5 **High mutation rates in prokaryotes**

6 While we chiefly focus on eukaryotes in this Review, prokaryotic genomes also highlight diverse
7 recurrent mutational modes. Indeed, large-scale reorganization and gene gain and loss is
8 probably even more biologically significant in prokaryotes than it is in eukaryotes [87]. For
9 example, pathogenic organisms carrying the genus name *Shigella* are not a genus, and indeed
10 not even monophyletic [88]. Each lineage of *Shigella* in fact arose independently from *E. coli*
11 ancestors by a concerted and specifically localized process of massive gene loss and acquisition
12 [88–90]. In this example, recurrent large mutations follow a predictable path, due to the
13 contextual influence of *E. coli* genome architecture, to yield the convergent outcome of the
14 *Shigella* genome.

15 Genomic context can also take the form of a global rather than a local feature. For
16 example, we have previously shown that certain gene orthology groups in prokaryotic genomes
17 actively predict the acquisition of specific new genes via horizontal transfer, by means of
18 epistatic associations [91]. In many cases, such gene-gene associations are previously known
19 and in fact mechanically required for propagation. For example, the presence of secretion
20 systems and conjugative pili in a microbial population’s pangenome will directly lead to a higher
21 incidence of gene mobilization and acquisition in that community, an obvious but somewhat
22 counterintuitive example of genomic context controlling the mutation rate.

23 We will conclude with a final startling example of how genomic context controls massive
24 recurrent mutation in prokaryotes. In most cases in bacteria, chromosomal integration of
25 mobile elements occurs at defined sequences with homology to integrating elements known as
26 attachment (*att*) sites [92]. However, in the soil microbe *Mesorhizobium ciceri*, a large
27 “symbiosis island” integration element is instead broken up and rearranged into three pieces to
28 be integrated at three different genomic *att* sites [93]. This integration is repeatable, highly
29 stereotyped, and reversible, as all steps can be directly observed in the laboratory [94]. The

tripartite architecture apparently arises following recombination between three different integration elements with distinct attachment sites [95]. Moreover, it appears that these *att* sites, and this tripartite mode of integration of mobile elements, are conserved across at least the genus *Mesorhizobium* as a mechanism for facilitating the spread of beneficial mobile genetic elements [95].

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 potential phenotypic consequences of this form of variation are likely vast and largely unexplored.

Concluding Remarks

We have discussed abundant cases of recurrently mutable DNA elements determining 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 last two decades, the vastly improved ascertainment of single nucleotide variants and substitutions altered the focus of genetic and genomic researchers to 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. However, newer technologies are resurrecting old ideas and restoring the balance by elucidating and acknowledging the full range of genomic variation and its likely phenotypic impact.

Acknowledgments

We would like to thank members of the Queitsch lab for helpful conversations. MOP would like to thank attendees of the Population, Evolutionary, and Quantitative Genetics (PEQG) 2018 meeting and attendees of the Plasmid Biology 2018 meeting for stimulating scientific presentations and conversations. EM, ANH, and CQ are supported by R21AG052020 and R01GM122088 to CQ. ANH is supported by training grant T32GM007270.

1 **Conflicts of Interest**

2 MOP is an employee of Phase Genomics Inc.

3

- 4 1. Willems T, Gymrek M, Poznik GD, Tyler-Smith C, 1000 Genomes Project Chromosome Y
5 Group, Erlich Y. Population-Scale Sequencing Data Enable Precise Estimates of Y-STR
6 Mutation Rates. *Am J Hum Genet.* 2016;98: 919–933. doi:10.1016/j.ajhg.2016.04.001
- 7 2. Finnegan DJ. TRANSPOSABLE ELEMENTS. In: Lindsley DL, Zimm GG, editors. *The Genome of*
8 *Drosophila Melanogaster*. San Diego: Academic Press; 1992. pp. 1096–1107.
9 doi:10.1016/B978-0-12-450990-0.50010-1
- 10 3. Mendez-Dorantes C, Bhargava R, Stark JM. Repeat-mediated deletions can be induced by a
11 chromosomal break far from a repeat, but multiple pathways suppress such
12 rearrangements. *Genes Dev.* 2018; doi:10.1101/gad.311084.117
- 13 4. Bousios A, Gaut BS. Mechanistic and evolutionary questions about epigenetic conflicts
14 between transposable elements and their plant hosts. *Current Opinion in Plant Biology.*
15 2016;30: 123–133. doi:10.1016/j.pbi.2016.02.009
- 16 5. Britten RJ, Kohne DE. Repeated Sequences in DNA. *Science.* 1968;161: 529–540.
17 doi:10.1126/science.161.3841.529
- 18 6. Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, et al. Genetic drift,
19 selection and the evolution of the mutation rate. *Nat Rev Genet.* 2016;17: 704–714.
20 doi:10.1038/nrg.2016.104
- 21 7. Tajima F. Infinite-allele model and infinite-site model in population genetics. *J Genet.*
22 1996;75: 27. doi:10.1007/BF02931749
- 23 8. Willems TF, Gymrek M, Highnam G, Mittelman D, Erlich Y. The landscape of human STR
24 variation. *Genome Research.* 2014; gr.177774.114–. doi:10.1101/gr.177774.114
- 25 9. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant
26 contribution of short tandem repeats to gene expression variation in humans. *Nature*
27 *Genetics.* 2015;48: 22–29. doi:10.1038/ng.3461
- 28 10. Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, et al. Polymorphic tandem
29 repeats within gene promoters act as modifiers of gene expression and DNA methylation in
30 humans. *Nucleic Acids Res.* 2016;44: 3750–3762. doi:10.1093/nar/gkw219
- 31 11. Press MO, McCoy RC, Hall AN, Akey JM, Queitsch C. Massive variation of short tandem
32 repeats with functional consequences across strains of *Arabidopsis thaliana*. *Genome Res.*
33 2018; doi:10.1101/gr.231753.117

12. Press MO, Carlson KD, Queitsch C. The overdue promise of short tandem repeat variation for heritability. *Trends in Genetics*. 2014;30: 504–512. doi:10.1016/j.tig.2014.07.008
13. Press MO, Queitsch C. Variability in a Short Tandem Repeat Mediates Complex Epistatic Interactions in *Arabidopsis thaliana*. *Genetics*. 2016; genetics.116.193359. doi:10.1534/genetics.116.193359
14. Rival P, Press MO, Bale J, Grancharova T, Undurraga SF, Queitsch C. The conserved PFT1 tandem repeat is crucial for proper flowering in *Arabidopsis thaliana*. *Genetics*. 2014;198: 747–754. doi:10.1534/genetics.114.167866
15. Tabib A, Vishwanathan S, Seleznev A, McKeown PC, Downing T, Dent C, et al. A Polynucleotide Repeat Expansion Causing Temperature-Sensitivity Persists in Wild Irish Accessions of *Arabidopsis thaliana*. *Arabidopsis thaliana*. 2016; 1311. doi:10.3389/fpls.2016.01311
16. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics*. 2018; doi:10.1038/nrg.2017.115
17. Gymrek M, Willems T, Erlich Y, Reich DE. A framework to interpret short tandem repeat variation in humans. *bioRxiv*. 2016; 092734. doi:10.1101/092734
18. Payseur BA, Place M, Weber JL. Linkage disequilibrium between STRPs and SNPs across the human genome. *American journal of human genetics*. 2008;82: 1039–50. doi:10.1016/j.ajhg.2008.02.018
19. Sawaya S, Jones M, Keller M. Linkage disequilibrium between single nucleotide polymorphisms and hypermutable loci [Internet]. *Cold Spring Harbor Labs Journals*; 2015 Jun p. 020909. Available: <http://biorxiv.org/content/early/2015/06/15/020909.abstract>
20. Zavodna M, Bagshaw A, Brauning R, Gemmell NJ. The effects of transcription and recombination on mutational dynamics of short tandem repeats. *Nucleic Acids Res*. 2018;46: 1321–1330. doi:10.1093/nar/gkx1253
21. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet*. 2002;30: 194–200. doi:10.1038/ng822
22. Shen X, Jonge JD, Forsberg SKG, Pettersson ME, Sheng Z, Hennig L, et al. Natural CMT2 Variation Is Associated With Genome-Wide Methylation Changes and Temperature Seasonality. *PLOS Genetics*. 2014;10: e1004842. doi:10.1371/journal.pgen.1004842
23. Haasl RJ, Payseur BA. The Number of Alleles at a Microsatellite Defines the Allele Frequency Spectrum and Facilitates Fast Accurate Estimation of θ . *Mol Biol Evol*. 2010;27: 2702–2715. doi:10.1093/molbev/msq164

24. Wendel JF, Lisch D, Hu G, Mason AS. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Current Opinion in Genetics & Development*. 2018;49: 1–7. doi:10.1016/j.gde.2018.01.004
25. Rodriguez F, Arkhipova IR. Transposable elements and polyploid evolution in animals. *Current Opinion in Genetics & Development*. 2018;49: 115–123. doi:10.1016/j.gde.2018.04.003
26. Legrand S, Caron T, Maumus F, Schvartzman S, Quadrana L, Durand E, et al. Differential retention of transposable element-derived sequences in outcrossing *Arabidopsis* genomes. *bioRxiv*. 2018; 368985. doi:10.1101/368985
27. Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. In: *eLife* [Internet]. 2 Dec 2016 [cited 4 Aug 2018]. doi:10.7554/eLife.20777
28. Li Z-W, Hou X-H, Chen J-F, Xu Y-C, Wu Q, González J, et al. Transposable elements contribute to the adaptation of *Arabidopsis thaliana*. *Genome Biol Evol*. doi:10.1093/gbe/evy171
29. Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddloh JA, et al. The *Arabidopsis thaliana* mobilome and its impact at the species level. *eLife*. 2016;5: e15716. doi:10.7554/eLife.15716
30. Vielle-Calzada J-P, Martínez de la Vega O, Hernández-Guzmán G, Ibarra-Laclette E, Alvarez-Mejía C, Vega-Arreguín JC, et al. The Palomero genome suggests metal effects on domestication. *Science*. 2009;326: 1078. doi:10.1126/science.1178437
31. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. *Nature*. 2017;546: 524–527. doi:10.1038/nature22971
32. Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Saniyal A, et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res*. 2006;16: 1262–1269. doi:10.1101/gr.5290206
33. Arkhipova IR. Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution. *Mol Biol Evol*. 2018;35: 1332–1337. doi:10.1093/molbev/msy083
34. Kuhn A, Ong YM, Cheng C-Y, Wong TY, Quake SR, Burkholder WF. Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome. *PNAS*. 2014;111: 8131–8136. doi:10.1073/pnas.1401532111
35. Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, et al. High-resolution comparative analysis of great ape genomes. *Science*. 2018;360: eaar6343. doi:10.1126/science.aar6343

36. Talbert PB, Kasinathan S, Henikoff S. Simple and Complex Centromeric Satellites in *Drosophila* Sibling Species. *Genetics*. 2018;208: 977–990. doi:10.1534/genetics.117.300620
37. Jagannathan M, Warsinger-Pepe N, Watase GJ, Yamashita YM. Comparative Analysis of Satellite DNA in the *Drosophila melanogaster* Species Complex. *G3 (Bethesda)*. 2016;7: 693–704. doi:10.1534/g3.116.035352
38. Wei KH-C, Lower SE, Caldas IV, Sless TJS, Barbash DA, Clark AG. Variable Rates of Simple Satellite Gains across the *Drosophila* Phylogeny. *Mol Biol Evol*. 2018;35: 925–941. doi:10.1093/molbev/msy005
39. Talavera-Lopez C, Messenger LA, Lewis MD, Yeo M, Reis-Cunha JL, Bartholomeu DC, et al. Repeat-driven generation of antigenic diversity in a major human pathogen, *Trypanosoma cruzi*. *bioRxiv*. 2018; 283531. doi:10.1101/283531
40. Faino L, Seidl MF, Shi-Kunne X, Pauper M, van den Berg GCM, Wittenberg AHJ, et al. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res*. 2016;26: 1091–1100. doi:10.1101/gr.204974.116
41. Gresham D, Desai MM, Tucker CM, Jenq HT, Pai D a, Ward A, et al. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genetics*. 2008;4: e1000303. doi:10.1371/journal.pgen.1000303
42. Brewer BJ, Payen C, Rienzi SCD, Higgins MM, Ong G, Dunham MJ, et al. Origin-Dependent Inverted-Repeat Amplification: Tests of a Model for Inverted DNA Amplification. *PLOS Genetics*. 2015;11: e1005699. doi:10.1371/journal.pgen.1005699
43. Kimura M. Evolutionary Rate at the Molecular Level. *Nature*. 1968;217: 624–626.
44. Pennings PS, Hermisson J. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. *PLOS Genetics*. 2006;2: e186. doi:10.1371/journal.pgen.0020186
45. Haasl RJ, Johnson RC, Payseur BA. The Effects of Microsatellite Selection on Linked Sequence Diversity. *Genome Biol Evol*. 2014;6: 1843–1861. doi:10.1093/gbe/evu134
46. Haasl RJ, Payseur BA. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol Ecol*. 2016;25: 5–23. doi:10.1111/mec.13339
47. Koning APJ de, Sanctis BDD. The Rate of Observable Molecular Evolution When Mutation May Not Be Weak. *bioRxiv*. 2018; 259507. doi:10.1101/259507
48. Rech GE, Bogaerts-Marquez M, Barron MG, Merenciano M, Villanueva-Canas JL, Horvath V, et al. Stress response, behavior, and development are shaped by transposable element-induced mutations in *Drosophila*. *bioRxiv*. 2018; 380618. doi:10.1101/380618

- 1 49. Katsnelson MI, Wolf YI, Koonin EV. On the feasibility of saltational evolution. *bioRxiv*. 2018;
2 399022. doi:10.1101/399022
- 3 50. Kimura M, Ohta T. Stepwise mutation model and distribution of allelic frequencies in a
4 finite population. *Proc Natl Acad Sci U S A*. 1978;75: 2868–2872.
- 5 51. Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. *Trends in*
6 *Genetics*. 2013;29: 575–584. doi:10.1016/j.tig.2013.04.005
- 7 52. Dennis MY, Eichler EE. Human adaptation and evolution by segmental duplication. *Curr*
8 *Opin Genet Dev*. 2016;41: 44–52. doi:10.1016/j.gde.2016.08.001
- 9 53. Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, et al. The evolution
10 and population diversity of human-specific segmental duplications. *Nat Ecol Evol*. 2017;1.
11 doi:10.1038/s41559-016-0069
- 12 54. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, et al. Human-specific
13 evolution of novel SRGAP2 genes by incomplete segmental duplication. *Cell*. 2012;149:
14 912–922. doi:10.1016/j.cell.2012.03.033
- 15 55. Dougherty ML, Nuttle X, Penn O, Nelson BJ, Huddleston J, Baker C, et al. The birth of a
16 human-specific neural gene by incomplete duplication and gene fusion. *Genome Biology*.
17 2017;18: 49. doi:10.1186/s13059-017-1163-9
- 18 56. Monlong J, Cossette P, Meloche C, Rouleau G, Girard SL, Bourque G. Human copy number
19 variants are enriched in regions of low mappability. *Nucleic Acids Res*.
20 doi:10.1093/nar/gky538
- 21 57. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic
22 variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat Genet*.
23 2013;45: 884–890. doi:10.1038/ng.2678
- 24 58. Prokopowich CD, Gregory TR, Crease TJ. The correlation between rDNA copy number and
25 genome size in eukaryotes. *Genome*. 2003;46: 48–50. doi:10.1139/g02-103
- 26 59. Gibbons JG, Branco AT, Yu S, Lemos B. Ribosomal DNA copy number is coupled with gene
27 expression variation and mitochondrial abundance in humans. *Nat Commun*. 2014;5: 4850.
28 doi:10.1038/ncomms5850
- 29 60. Thompson O, Edgley M, Strasbourger P, Flibotte S, Ewing B, Adair R, et al. The million
30 mutation project: a new approach to genetics in *Caenorhabditis elegans*. *Genome Res*.
31 2013;23: 1749–1762. doi:10.1101/gr.157651.113
- 32 61. Parks MM, Kurylo CM, Dass RA, Bojmar L, Lyden D, Vincent CT, et al. Variant ribosomal RNA
33 alleles are conserved and exhibit tissue-specific expression. *Sci Adv*. 2018;4: eaao0665.
34 doi:10.1126/sciadv.aao0665

62. Ritossa FM, Scala G. Equilibrium variations in the redundancy of rDNA in *Drosophila melanogaster*. *Genetics*. 1969; Available: http://agris.fao.org/agris-search/search.do;jsessionid=1E08D6134C19B3ACC4F456D8CE1F5E86?request_locale=ru&recordID=US201301233618&query=&sourceQuery=&sortField=&sortOrder=&agrovocString=&advQuery=¢erString=&enableField=
63. Mohan J, Ritossa FM. Regulation of ribosomal RNA synthesis and its bearing on the bobbed phenotype in *Drosophila melanogaster*. *Dev Biol*. 1970;22: 495–512. doi:10.1016/0012-1606(70)90165-X
64. James SA, O’Kelly MJT, Carter DM, Davey RP, Oudenaarden A van, Roberts IN. Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing. *Genome Res*. 2009;19: 626–635. doi:10.1101/gr.084517.108
65. Kwan EX, Wang XS, Amemiya HM, Brewer BJ, Raghuraman MK. rDNA Copy Number Variants Are Frequent Passenger Mutations in *Saccharomyces cerevisiae* Deletion Collections and de Novo Transformants. *G3*. 2016;6: 2829–2838. doi:10.1534/g3.116.030296
66. Preuss S, Pikaard CS. rRNA gene silencing and nucleolar dominance: Insights into a chromosome-scale epigenetic on/off switch. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*. 2007;1769: 383–392. doi:10.1016/j.bbaexp.2007.02.005
67. The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences*. 1999;24: 437–440. doi:10.1016/S0968-0004(99)01460-7
68. Sanchez JC, Kwan EX, Pohl TJ, Amemiya HM, Raghuraman MK, Brewer BJ. Defective replication initiation results in locus specific chromosome breakage and a ribosomal RNA deficiency in yeast. *PLOS Genetics*. 2017;13: e1007041. doi:10.1371/journal.pgen.1007041
69. Kobayashi T, Heck DJ, Nomura M, Horiuchi T. 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*. 1998;12: 3821–3830.
70. Lu KL, Nelson JO, Watase GJ, Warsinger-Pepe N, Yamashita YM. Transgenerational dynamics of rDNA copy number in *Drosophila* male germline stem cells. In: *eLife* [Internet]. 13 Feb 2018 [cited 19 Aug 2018]. doi:10.7554/eLife.32421
71. Kwan EX, Foss EJ, Tsuchiyama S, Alvino GM, Kruglyak L, Kaerberlein M, et al. A Natural Polymorphism in rDNA Replication Origins Links Origin Activation with Calorie Restriction and Lifespan. *PLoS Genet*. 2013;9. doi:10.1371/journal.pgen.1003329
72. Takeuchi Y, Horiuchi T, Kobayashi T. Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes Dev*. 2003;17: 1497–1506. doi:10.1101/gad.1085403

- 1 73. Sinclair DA, Guarente L. Extrachromosomal rDNA circles--a cause of aging in yeast. *Cell*.
2 1997;91: 1033–1042.
- 3 74. Stults DM, Killen MW, Pierce HH, Pierce AJ. Genomic architecture and inheritance of human
4 ribosomal RNA gene clusters. *Genome Res*. 2008;18: 13–18. doi:10.1101/gr.6858507
- 5 75. Ritossa FM, Atwood KC, Spiegelman S. A molecular explanation of the bobbed mutants of
6 *Drosophila* as partial deficiencies of “ribosomal” DNA. *Genetics*. 1966;54: 819–834.
- 7 76. French SL, Osheim YN, Cioci F, Nomura M, Beyer AL. In exponentially growing
8 *Saccharomyces cerevisiae* cells, rRNA synthesis is determined by the summed RNA
9 polymerase I loading rate rather than by the number of active genes. *Mol Cell Biol*. 2003;23:
10 1558–1568.
- 11 77. Li B, Kremling K, Wu P, Bukowski R, Romay M, Xie E, et al. Co-regulation of ribosomal RNA
12 with hundreds of genes contributes to phenotypic variations. *Genome Res*. 2018;
13 gr.229716.117. doi:10.1101/gr.229716.117
- 14 78. Ganley ARD, Ide S, Saka K, Kobayashi T. The effect of replication initiation on gene
15 amplification in the rDNA and its relationship to aging. *Mol Cell*. 2009;35: 683–693.
16 doi:10.1016/j.molcel.2009.07.012
- 17 79. Saka K, Ide S, Ganley ARD, Kobayashi T. Cellular senescence in yeast is regulated by rDNA
18 noncoding transcription. *Curr Biol*. 2013;23: 1794–1798. doi:10.1016/j.cub.2013.07.048
- 19 80. Wang M, Lemos B. Ribosomal DNA copy number amplification and loss in human cancers is
20 linked to tumor genetic context, nucleolus activity, and proliferation. *PLoS Genet*. 2017;13:
21 e1006994. doi:10.1371/journal.pgen.1006994
- 22 81. Xu B, Li H, Perry JM, Singh VP, Unruh J, Yu Z, et al. Ribosomal DNA copy number loss and
23 sequence variation in cancer. *PLoS Genet*. 2017;13: e1006771.
24 doi:10.1371/journal.pgen.1006771
- 25 82. Udugama M, Sanij E, Voon HPJ, Son J, Hii L, Henson JD, et al. Ribosomal DNA copy loss and
26 repeat instability in ATRX-mutated cancers. *PNAS*. 2018;115: 4737–4742.
27 doi:10.1073/pnas.1720391115
- 28 83. Kobayashi T. A new role of the rDNA and nucleolus in the nucleus--rDNA instability
29 maintains genome integrity. *Bioessays*. 2008;30: 267–272. doi:10.1002/bies.20723
- 30 84. Paredes S, Branco AT, Hartl DL, Maggert KA, Lemos B. Ribosomal DNA deletions modulate
31 genome-wide gene expression: “rDNA-sensitive” genes and natural variation. *PLoS Genet*.
32 2011;7: e1001376. doi:10.1371/journal.pgen.1001376
- 33 85. Waddington CH. CANALIZATION OF DEVELOPMENT AND THE INHERITANCE OF ACQUIRED
34 CHARACTERS : Abstract : *Nature*. *Nature*. 1942;150: 563–565.

- 1 86. Paredes S, Maggert KA. Ribosomal DNA contributes to global chromatin regulation. *Proc*
2 *Natl Acad Sci USA*. 2009;106: 17829–17834. doi:10.1073/pnas.0906811106
- 3 87. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil:
4 Quantification of genome dynamics in prokaryote supergenomes. *BMC biology*. 2014;12:
5 66. doi:10.1186/s12915-014-0066-4
- 6 88. Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia*
7 *coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci USA*.
8 2000;97: 10567–10572. doi:10.1073/pnas.180094797
- 9 89. Zhang Y, Lin K. A phylogenomic analysis of *Escherichia coli* / *Shigella* group: implications of
10 genomic features associated with pathogenicity and ecological adaptation. *BMC*
11 *Evolutionary Biology*. 2012;12: 174. doi:10.1186/1471-2148-12-174
- 12 90. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised Genome
13 Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths. *PLOS*
14 *Genetics*. 2009;5: e1000344. doi:10.1371/journal.pgen.1000344
- 15 91. Press MO, Queitsch C, Borenstein E. Evolutionary assembly patterns of prokaryotic
16 genomes. *Genome Research*. 2016; gr.200097.115. doi:10.1101/gr.200097.115
- 17 92. Johnson CM, Grossman AD. Integrative and Conjugative Elements (ICEs): What They Do and
18 How They Work. *Annu Rev Genet*. 2015;49: 577–601. doi:10.1146/annurev-genet-112414-
19 055018
- 20 93. Haskett TL, Terpolilli JJ, Bekuma A, O’Hara GW, Sullivan JT, Wang P, et al. Assembly and
21 transfer of tripartite integrative and conjugative genetic elements. *PNAS*. 2016;113: 12268–
22 12273. doi:10.1073/pnas.1613358113
- 23 94. Haskett TL, Terpolilli JJ, Ramachandran VK, Verdonk CJ, Poole PS, O’Hara GW, et al.
24 Sequential induction of three recombination directionality factors directs assembly of
25 tripartite integrative and conjugative elements. *PLOS Genetics*. 2018;14: e1007292.
26 doi:10.1371/journal.pgen.1007292
- 27 95. Haskett TL, Ramsay JP, Bekuma AA, Sullivan JT, O’Hara GW, Terpolilli JJ. Evolutionary
28 persistence of tripartite integrative and conjugative elements. *Plasmid*. 2017;92: 30–36.
29 doi:10.1016/j.plasmid.2017.06.001
- 30 96. MAYNARD SMITH J. Natural Selection and the Concept of a Protein Space. *Nature*.
31 1970;225: 563–564. doi:10.1038/225563a0
- 32 97. Felsenstein J. Cases in which Parsimony or Compatibility Methods Will be Positively
33 Misleading. *Systematic Zoology*. 1978;27: 401–410. doi:10.2307/2412923

98. Yang Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*. 1996;11: 367–372. doi:10.1016/0169-5347(96)10041-0
99. Fitch WM, Markowitz E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet*. 1970;4: 579–593. doi:10.1007/BF00486096

BOX 1. Mutational modes and molecular evolution.

John Maynard Smith [96] 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 allowing many more kinds of transitions. 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; 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 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 [97]. Firstly, while 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 equal probability [98]; or even that transition probabilities are constant along the path [99]. 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.

GLOSSARY.

- **Transposable element (TE):** DNA elements which 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”.
- **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 or centromeres tend to consist of satellite DNA. Short tandem repeats, also called microsatellites, consist of very short repeat units (<10 nt).
- **Parallel mutation:** A mutation which 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.
- **Substitution:** A mutation which replaces a nucleotide at a single position (A, C, G, T) with one of the other three nucleotides.
- **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.

HIGHLIGHTS.

- Single nucleotide variants or mutations (*e.g.* point mutations) are less common than other variations and mutations, and cannot generate observed genomic diversity. Focus on single nucleotide variations is an ascertainment bias of genomic technology.
- 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. There is strong evidence that these mutations have functional consequences.

- 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.

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 inclusive of 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?