Towards an improved understanding of molecular evolution: the relative roles of selection, drift, and everything in between

Fanny Pouyet 1,2,3,4 and Kimberbly J. Gilbert 5,1,6,4

Volume 1 (2021), article e27

https://doi.org/10.24072/pcjournal.16

Abstract

A major goal of molecular evolutionary biology is to identify loci or regions of the genome under selection versus those evolving in a neutral manner. Correct identification allows accurate inference of the evolutionary process and thus comprehension of historical and contemporary processes driving phenotypic change and adaptation. A fundamental difficulty lies in distinguishing sites targeted by selection from both sites linked to these targets and sites fully independent of selection. These three categories of sites necessitate attention in light of the debate over the relative importance of selection versus neutrality and the neutral theory. Modern genomic insights have proved that complex processes such as linkage, demography, and biased gene conversion complicate our understanding of the role of neutral versus selective processes in evolution. In this perspective, we first highlight the importance of the genomic and (a)biotic context of new mutations to identify the targets of natural selection. We then present mechanisms that may constrain the evolution of genomes and bias the inference of selection. We discuss these mechanisms within the two critical levels that they occur: the population level and the molecular level. We highlight that they should be taken into account to correctly distinguish sites across the genome subject to selective or non-selective forces and stress that a major current field-wide goal is to quantify the absolute importance of these mechanisms.

1Institute of Ecology and Evolution, University of Bern, Switzerland, 2Interdisciplinary laboratory of Digital Sciences, Universite Paris-Saclay, France, 3Sorbonne Universite, CNRS, IBPS, Laboratory of Computational and Quantitative Biology, France, 4Swiss Institute of Bioinformatics, Lausanne, Switzerland, 5Department of Computational Biology, University of Lausanne (UNIL), Switzerland, 6Institute of Plant Sciences, University of Bern, Altenbergrain 21, 3013 Bern, Switzerland
1. Introduction

Understanding the relative importance of evolutionary forces in driving adaptive change has been a longstanding goal of evolutionary biology. In today’s genomic era, accurately and precisely addressing this question has become more feasible than ever before. Genomic data has allowed, for example, quantification of introgression rates between populations or species (e.g. Ellstrand et al., 2013) and accurate estimation of mutation rates within species or across the genome (Besenbacher et al., 2019; Ellegren et al., 2003; Hodgkinson and Eyre-Walker, 2011; Zhu et al., 2014). Yet the interplay between neutral evolution and selective forces has remained a difficult problem to address. Since the advent of population genetics as a field, debate over the relative importance of these processes has arisen, been resolved, and re-arisen (e.g. Gillespie, 1995; Kimura et al., 1968; Kreitman, 1996; Ohta and Kimura, 1971). Most recently, 50 years since the advent of the neutral theory, this debate has been rekindled in light of emerging genomic data (Jensen et al., 2018; Kern and Matthew W Hahn, 2018). In an era of limited genetic tools and data, the neutral theory aimed to explain the greater than expected genetic diversity observed based on the actions of natural selection alone. Kern and Hahn (2018) have most recently argued that modern genomic data allows us to reject the applicability of neutral theory for understanding molecular evolution, while Jensen et al. (2018) have replied that this is not the case. A major dividing view on this point is whether a large proportion of the genome is affected by adaptive natural selection (directly or indirectly), and there is ample space for additional data across a wider range of species to contribute towards these investigations and our understanding of molecular evolution.

Natural selection functions in a diversity of modes. Negative selection – also termed purifying selection – acts to reduce the frequency of deleterious mutations (i.e. mutations that reduce an individual’s fitness, with selection coefficient $s < 0$) while positive selection favors the fixation of beneficial mutations ($s > 0$). Both these modes of selection reduce diversity by favoring or disfavoring specific alleles, but selection can also maintain genetic diversity when there is a selective advantage to being in the heterozygous state, e.g. balancing selection. There is also the case of sexual selection that we do not address here as its many and complex cases merit a review of their own.

Genetic drift is the neutral corollary to natural selection, where allele frequencies change due to random chance and sampling effects. Sites that are truly neutral are defined as those with $s = 0$. If we define the effective size of a population as the size of an ideal population that
experiences the same amount of genetic drift as the observed population (Wright, 1931), then the appropriate measure of the strength of selection is \((Ns)\). This is key because genetic drift can act across a range of weakly selected sites when \(N\) is sufficiently large, while at a smaller \(N\), these sites may behave in a neutral manner.

Inference methods making use of empirical genomic data often require data solely originating from neutral processes, e.g. to infer demography with SFS-based methods such as FastSimCoal (Excoffier et al., 2013), or to infer the distribution of fitness effects (DFE) which requires contrasting SFS from neutral versus selected sites (Keightley and Eyre-Walker, 2007; Tataru and Bataillon, 2019). A priori misidentification of selected versus neutral sites may strongly bias resultant inferences, having a major impact on downstream interpretations. Approaches that search for signatures of selection and identify causal variants for adaptation or other phenotypic change mainly rely on identifying outlier regions of genomic differentiation (e.g. GWAS, \(F_{ST}\) outlier tests, or environment-allele correlations; Beaumont and Nichols, 1996; Foll et al., 2008; Luu et al., 2017; Whitlock and Lotterhos, 2015), sometimes incorporating the signature across a stretch of the genome (e.g. Schrider and Kern, 2016).

Yet, a departure from genetic drift alone is not sufficient to merit a conclusion of selection. Even supervised machine learning methods that use summary statistics to infer a history of selective sweeps, such as S/HIC (Schrider and Kern, 2016), are sensitive to confounding factors such as complex demographic history to accurately identify the variants under selection. In this review we highlight both the importance of considering the genomic, biotic, and abiotic context in which new mutations occur and the major evolutionary processes that can change allele frequencies, creating a major confounding factor for evolutionary inference of natural selection.

### 2. Genomic and environmental context

The context in which mutations occur plays an important role in the actions of selection versus drift. This context encapsulates both the genomic environment as well as the biotic and abiotic environment of an organism containing those genes, therefore becoming relevant at both the molecular and population levels of interaction. When interactions between loci occurring together in the genome create non-additive phenotypic changes (Cordell, 2002; Fisher, 1918; G Martin et al., 2007) this can greatly complicate the inference of selection. In the presence of epistatic interactions, an allele at a given site may only be beneficial to an organism when its genomic environment contains another mutant allele at a different site in the genome, so that when together these alleles generate a phenotypic change. The concept of epistasis is tightly linked to the relative fitness effects of alleles, where a specific allele at one locus might change the sign of the selection coefficient at another locus. This phenomenon may create strong patterns of linkage disequilibrium and interfere with detection of selection since the selective effect is dependent on the combinations of alleles across loci. Such an example of epistasis is provided in the Segregation distortion paragraph.

In some cases, interference between selected sites of differential fitness effects can alter the strength of selection on a genomic region. More positively (or negatively) selected sites with physical linkage between each other can behave as a larger multi-site locus with an amplified selection coefficient representative of all the selected alleles in the region. In other cases, the strength of selection may be reduced when sites have competing impacts on fitness, termed Hill-Robertson interference (Felsenstein, 1974). Amplification of the strength of selection by tightly-linked, jointly selected sites may simplify detection of selection but complicate the identification of precise sites under selection. Conversely, Hill-Robertson interference may complicate identifying both the presence of selection and the sites it targets. All of these effects depend on many parameters, for instance, with a single population undergoing partial self-fertilization, selective interference on deleterious alleles tends to reduce mean fitness and increase inbreeding depression. This effect is stronger when deleterious alleles are more recessive but only weakly dependent on the strength of selection against deleterious alleles and the recombination rate (Roze, 2015). Selective interference thus affects the relative impact of adaptive and non-adaptive processes in the genome.
Equally complicating is the individual-level scenario where fitness is dependent on the local community of organisms and whatever traits they exhibit (e.g. frequency-dependent selection) or on the abiotic local conditions exerted by variable environments on the phenotype (spatially or temporally varying selection). For example, in scale-eating cichlids, frequency-dependent selection can drive the handedness of individual phenotypes, where it is advantageous to be the rarer morph (Hori, 1993). Interestingly, competition between alleles at the genomic level can also lead to frequency-dependent selection, with the most famous example being the amino acid polymorphism at position 6 of $\beta$-globin in Africa, which is associated with resistance to malaria (see Taylor et al., 2012, for review and meta-analyses). Spatially varying selection results from environmental variation across geographic space (e.g. Gagnaire et al., 2012). When populations exist across variable environments, distinct combination of alleles may arise within subpopulations locally adapted to their environments. Thus, the strength of selection on non-neutral sites can vary over time and space as an organism’s environment changes. For the purposes of evolutionary inference and understanding the action of selection, experimental design is key so that empirical analyses can be conducted where the time point is equivalent and the environment is equivalent (or otherwise controlled for) to as much of an extent as possible (e.g. Gorter et al., 2018). Spatially and temporally varying selection may be best accounted for by this direct approach or by studying natural clines of allele frequencies (Endler, 1977; Machado et al., 2016). However, for sites in the genome that are not targets of selection, nor purely under genetic drift, remaining processes can change allele frequencies in ways indicative of selection and are thus essential to bear in mind for evolutionary inference.

In the remainder of this perspective, we consider evolutionary processes at both the population and molecular levels that have the potential to bias the inference of selection. Without consideration of such processes, inaccurate conclusions may be drawn about the role of selection in the evolutionary process. An extreme example of such mis-inference is Evans et al. (2005) and Mekel-Bobrov et al. (2005), where the result that selection drove brain size in humans was shown to be equivalently explained by neutral demographic processes (Currat et al., 2006). Scenarios involved at the molecular level can also interfere with correct inference of selection, including transmission bias (e.g. meiotic driver genes; Bravo Núñez et al., 2018, for a review), biased gene conversion (e.g. human accelerated genomic regions, HAR; N. Galtier and Duret, 2007; Pollard et al., 2006), or inference of the DFE (e.g. in flycatchers; Bolivar et al., 2018).

Though these are few examples, we argue that these population and molecular processes are prevalent enough to act as major determinants of genomic diversity and are particularly easy to confound with selection. We also argue that even if some cases of demographic processes may be versions of genetic drift, in that they are due to sampling process, the signatures that they leave in populations differs from that of single, ideal populations undergoing drift, emphasizing their importance in distinguishing selection from neutrality. We define these as non-adaptive processes since they are neither the direct action of selection nor are they purely subject to drift (Table 1). This non-adaptive category should help to improve studies of both neutral or selective processes impacting the genome by considering the relevant sites impacted by either process.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>yes</td>
<td>yes</td>
<td>Adaptive processes increase fitness</td>
</tr>
<tr>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balancing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic drift</td>
<td>yes</td>
<td>yes</td>
<td>Neutral and non-adaptive as independent of fitness; depending on N, may impact site(s) with s ≠ 0</td>
</tr>
<tr>
<td>Population-level processes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bottleneck, demographic expansion</td>
<td>yes</td>
<td>yes</td>
<td>Lead to rapid genetic drift, can mimic positive or negative selection</td>
</tr>
<tr>
<td>Population structure</td>
<td>yes</td>
<td>yes</td>
<td>Can mimic local adaptation, responsible for many GWAS false-positives</td>
</tr>
<tr>
<td>Spatial expansion (gene surfing)</td>
<td>yes</td>
<td>yes</td>
<td>Strong genetic drift over space, can mimic selection</td>
</tr>
<tr>
<td>Molecular-level processes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linked selection</td>
<td>yes</td>
<td>yes</td>
<td>Impact neutral diversity, can change fitness, and act on selected sites depending on the relative selection coefficients of nearby sites</td>
</tr>
<tr>
<td>Selective sweeps</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Associative overdominance</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>CpG hyper-mutability</td>
<td>yes</td>
<td>yes</td>
<td>Methylation Cytosines deaminate to Thymines, increasing the frequency of T within the genome</td>
</tr>
<tr>
<td>Segregation distortion</td>
<td>yes</td>
<td>yes</td>
<td>Meiotic gene drivers manipulate their transmission even though they can be detrimental</td>
</tr>
<tr>
<td>Gene conversion</td>
<td>yes</td>
<td>yes</td>
<td>Increases the frequency of certain variants deterministically; regardless of fitness</td>
</tr>
</tbody>
</table>

**Table 1** - Evolutionary processes discussed in this paper that impact the fate of genetic variants. We focus on SNPs in this table and omit a category on the mobility of genetic elements, but transposable elements are also impacted by molecular-level genomic linkage processes.
3. Population-level processes


A major difficulty in distinguishing neutral sites from those impacted by selection is demographic history. The term demography incorporates several factors, but at its core is defined as change in population size, $N$. Population size combined with the selection coefficient determines the effective strength of selection on existing genomic diversity and therefore has major effects on the evolutionary process. Changes in population size result during population bottlenecks, population expansions (within one locale), spatial population expansions over geographic space (e.g. range expansion), or migration among populations (a more complicated case where a larger gene pool becomes relevant).

While selection or drift may act on specific variants or regions of the genome, demographic change affects the whole genome equally. Population bottlenecks have long been known to impact genetic diversity and change the efficiency of selection acting on alleles with $s \neq 0$ within the population. In such a case, stronger drift impacts sites of both $s = 0$ and $s$ close to zero (nearly neutral mutations), with an increasing range of $s$ values as population size decreases and selection becomes less efficient. These alleles are driven to more rapid fixation or more rapid loss, a pattern of allele frequency change which can mimic that of positive selection and selective sweeps (see the Linked selection paragraph).

Inferring past population size bottlenecks is a rich field with many methods to do so from genomic data (Heled and Drummond, 2008; Li and Durbin, 2011; Liu and Fu, 2015; Terhorst et al., 2017). Importantly, these methods rely on the use of neutral variants to obtain a proper inference (Gattepaille et al., 2013), and are perhaps particularly important in conservation genetics to identify species at risk due to a recent bottleneck (rather than, e.g. an incorrect inference suggesting low diversity is due to a selective sweep; Peery et al., 2012). The bias that results from a demographic history of range expansion after a bottleneck has been particularly notable for humans having expanded out of Africa and more troublesome for distinguishing demography from selection. Studies attempting to find signatures of selection in humans may suffer from biased inferences due to these neutral historic processes (Amos and Bryant, 2011; Heller et al., 2013; AR Martin et al., 2017). Such a demographic history is particularly intriguing as it combines not only impactful changes in population size, but movement over geographic space which includes complications of population structure and spatially varying selection.

3.2. Population structure.

Spatial population genetic structure and migration among subpopulations also plays an important role in the inference of selection. Population substructure can mimic a signal of local adaptation, where some populations which happen to exist in different environments possess different genetic signatures, leading methods to identify these differentiated loci as targets of selection. For instance, several recent studies have encountered this difficulty where signals previously thought to be selective were instead due to the lack of accounting for genetic structure among populations (Berg et al., 2019; Tian et al., 2008). Additionally, the process of migration into populations or admixture among species can create an influx of novel genetic material. Even if fully neutral, the presence of such heterozygosity in the population leaves a signal indicative of either adaptive processes (e.g. balancing selection) or non-adaptive processes (e.g. secondary contact or gene flow among structured populations) (Matthew William Hahn, 2018).

It is again of vital importance for studies inferring selection that population structure be identified and accounted for. Many such approaches exist and vary depending on the form in which structure presents itself: isolation by distance versus more distinct populations over space, with varying degrees of migration occurring across the landscape. For example, isolation-with-migration models aim to infer the amount of migration between isolated populations leading to the level of polymorphism observed (Hey, 2010). Isolation by distance can also be difficult in the face of inferring selection due to the correlation of allele frequency changes over space...
with environmental changes. Fortunately, much work has been done to correct for this population structure when inferring selection over landscapes (Caye et al., 2019; de Villemereuil and Gaggiotti, 2015; Gautier, 2015; Günther and Coop, 2013).

3.3. Spatial and temporal variation.

Spread or growth of populations across geographic space also introduces the complexity of changing environmental conditions (or analogously temporally changing environments may have similar impacts). Many populations and species are known to have undergone or are expected to undergo this demographic change: from post-glacial recolonizations, to species invasions, to shifting species ranges in response to climate change (Davis and Shaw, 2001; Thomas, 2010). During spatial expansions, not only does population size change with repeated bottlenecks of founder individuals, but these populations colonize new geographic space, resulting in a process termed gene surfing. Gene surfing is a unique genetic process that can leave genomic signatures similar to those of selection, yet are due entirely to demographic processes. Sequential founder events reduce the effective population size in colonizing populations and thus the efficiency of selection, thereby allowing alleles that might otherwise be subject to strong selection to surf to high frequency at the expanding wave front of a population (Edmonds et al., 2004; Klopfstein et al., 2005). Because surfing can lead to the increase or even fixation of a given allele (be it neutral or not), it is easily mistaken for the product of selective forces. Yet unlike selection, surfing can also cause deleterious variants to increase and result in severe fitness loss at expanding fronts, termed expansion load (Peischl et al., 2013). This demographic process alters the actions of natural selection and genetic drift within the genome and has potentially large effects on population fitness, emphasizing its importance as a non-adaptive force in evolution.

4. Molecular-level processes

4.1. Linked selection.

The fact that recombination breaks apart combinations of alleles at an increasing probability with greater distance along the genome results in many sites being physically linked and evolving in a non-independent manner. The background where a new mutation occurs therefore influences that variant’s probability of fixation, as any more strongly-selected target sites nearby will influence that linked site, as first pointed out by Fisher (1930) and Muller (1932) (Figure 1). A extended review on this topic was written by Gordo and Charlesworth (2001).

Neutral sites in a background with one or more sites under negative selection will have a lower probability of fixation than unlinked neutral sites. This is due to background selection (BGS), where negative selection against a variant reduces the frequency of nearby neutral variants (B. Charlesworth et al., 1993). Sites subject to BGS fall in the category of non-adaptive evolution because these variants are not directly selected against nor are they evolving neutrally since selection indirectly impacts them. These linked sites evolve in a non-neutral fashion, so even if phenotypically and adaptively they confer no change in phenotype, they must not be considered neutral for inferential purposes, a point which is widely recognized in the field.

Similarly, the occurrence of a mutation conferring a fitness benefit can also result in the reduction of genetic diversity through a selective sweep (Figure 1). When selection increases the frequency of a beneficial allele in the population, nearby neutral variants likewise increase in frequency, hitchhiking along to fixation with the beneficial variant. Whether selection acts on a single novel variant (hard sweeps; Messer and Neher, 2012) or on standing genetic variation (soft sweeps; Hermisson and Pleuni S. Pennings, 2005; Pleuni S Pennings and Hermisson, 2006a,b) can influence the extent of the impact on allele frequency change for linked neutral sites. Several population level processes may even be contributors to instances where standing genetic variation results in a soft sweep, for example if existing diversity shifts to become beneficial, perhaps due to environmental change.

Finally, genetic linkage can also lead to an increase in genetic diversity when neutral sites fall near a partially deleterious recessive allele or near an allele under balancing selection. In the presence of partially recessive deleterious alleles, this increase in diversity is termed associative...
overdominance (AOD; Becher et al., 2020; Gilbert et al., 2020; Ohta and Kimura, 1970; Zhao and Brian Charlesworth, 2016), and is limited to regions of low recombination (Figure 1). In contrast, sites linked to those under classical balancing selection should increase in diversity across regions spanning the range of possible recombination rates.

Most methods to detect selection on linked polymorphisms are based on the site frequency spectrum (SFS), such as Tajima’s D which depends on the pairwise nucleotide diversity and the number of segregating sites (Tajima, 1989) and other methods which are based on the haplotype structure (Hudson and Kaplan, 1995; Sabeti, Reich, et al., 2002; Sabeti, Varilly, et al., 2007; Voight et al., 2006). Methods based on haplotype structure are most effective to detect recent episodes of hitchhiking (Garud et al., 2015).

![Figure 1](image-url)

**Figure 1** - Genetic linkage can drastically change the frequency of neutral alleles in a population, falling into three categories depending on the manner of selection towards a focal site. For each form of selection shown, the top row shows the haplotypes of a non-recombining region in the initial population. The bottom row shows the resultant haplotypes after an episode of selection. For illustrative purpose, each haplotype contains 3 derived alleles (circles): neutral ones are in blue, beneficial in green, and deleterious alleles are in red or orange to indicate dominant or recessive, respectively. For background selection (left), neutral diversity is reduced due to negative selection on nearby linked deleterious alleles and homozygosity increases at the population level. Associative overdominance (center) prevents combinations of homozygous neutral alleles from accumulating in the population in regions of low recombination. At the beginning, each haplotype contains one deleterious recessive allele (orange circle) and loci carrying such alleles are genetically linked as the region is non-recombining. Selection favors combinations of heterozygous deleterious alleles as they are recessive. Neutral heterozygosity is favored at the population level and diversity increases. Selective sweeps (right) reduce diversity and increase allele frequency through the hitchhiking of neutral variants that are linked to beneficial mutations under positive selection. Figure inspired from Alves et al. (2012).

### 4.2. Hypermutability of CpG sites and mutation rate variation.

CpG sites in which a cytosine and a guanine appear consecutively, can experience high levels of mutational pressure. Cytosines at CpG sites are one of the preferential targets of methylation in vertebrates and some other species. Methylated cytosines spontaneously deaminate to thymines, leading to an increase in the frequency of TpG sites within the genome and a relative deficit of CpG (reviewed in Hodgkinson and Eyre-Walker, 2011), potentially leaving a signature indicative of selection.

A recent paper by Laurin-Lemay et al. (2018) found that a large proportion of mammalian codon usage, such as the preferential usage of the GCC Alanine codon compared to its synonymous GCG in humans, can be explained by the hypermutability of CpG sites, even though this is often unaccounted for in codon substitution models. The authors advocate for evaluating the impact of such model violations on statistical tests in phylogenetic analyses. Interestingly, CpG hypermutability is also an underappreciated process in the field of population genomics where...
the favored strategy has been to filter out hypermutable sites before performing evolutionary inferences (Pouyet, Aeschbacher, et al., 2018). The hypermutability of these sites and the subsequent bias to certain alleles has been shown to shift site frequency spectra in ways that might interfere with population genetic inferences that are based on the SFS (Harpak et al., 2016).

4.3. Segregation distortion.

Meiotic gene drivers manipulate the transmission process during meiosis to their own advantage, leading to their over-representation in gametes despite the lack of advantage to the carrier. As segregation distortion encompasses a wide variety of mechanisms, we provide three such examples. The first one is the segregation distoriter gene complex (Figure 2; Larracuente and Presgraves, 2012), present at low frequency in all natural populations of *Drosophila melanogaster*. This complex involves the \( Sd \) locus and its target responder locus \( (Rsp) \). There is variable transmission advantage between the different \( Sd \) alleles, and one of these alleles recently swept to fixation in Africa causing strong linkage disequilibrium and loss of genetic diversity (Presgraves et al., 2009). A second example is *Wolbachia*, where a maternally inherited bacteria of arthropods manipulates host reproduction. *Wolbachia* is known to be a selfish element favoring its own propagation through, for instance, the inability of infected males to successfully reproduce with uninfected females (Turelli and Hoffmann, 1991). At the level of the host, the presence of *Wolbachia* is also associated to segregation distortion: the maternal inheritance induces genetic linkage on host mitochondria and mitochondria in infected females are over-represented in the next generation. This effect reduces the effective population size and the efficacy of selection in mitochondria and could drive fixation of mitochondrial haplotypes (Cariou et al., 2017; GD Hurst and Jiggins, 2005). A third example comes from centromere evolution, where asymmetry at female meiosis causes only one of the four products of meiosis to become the oocyte nucleus and can lead to a kind of segregation distortion (Henikoff et al., 2001). Centromeres have a central role in preventing aneuploidy by facilitating the assembly of several components required for chromosome separation (see Tanaka et al., 2013, for a review). This centromere drive model includes proteins such as Cid and highly repetitive satellite sequences that bind to microtubules during meiosis I (Henikoff et al., 2001). Centromeres which preferentially transmit to the oocyte nucleus can rapidly drive to fixation even with a slight advantage at each meiosis (Henikoff et al., 2001).

Meiotic drivers are predicted to be evolutionarily labile by favoring their fixation in the population even though they are detrimental for their carriers (Lindholm et al., 2016). Identifying loci under selection in these cases is anything but straightforward, as the signal of meiotic drive might be easily confounded with selective sweeps and positive selection (Presgraves et al., 2009).

4.4. Mobility of genetic elements.

Detecting the signature of selection is often restricted to SNPs even though transposable elements (TE) and other structural genomic changes such as inversions are likely also prevalent biases for adaptive evolution. Until further studies improve our understanding of mobile genetic elements and our ability to identify them, it is certainly possible that these regions of the genome may play a role in changing diversity and biasing or interrupting our ability to infer selection or neutral demographic parameters, e.g. in inverted regions that can no longer recombine, deleterious variation can become masked or beneficial variation may be maintained in tight linkage.

TEs are widespread across the tree of life and in some species can represent a major fraction of the genome (de Koning et al., 2011; Wicker et al., 2007). TEs are associated with the creation of new mutations and changes in recombination patterns (e.g. Bartolome et al., 2002). McClintock (1950) first discovered that mobile elements were associated with phenotypic changes in maize. However, these sorts of structural genomic changes are often disregarded in favor of SNPs when inferring sources of adaptive evolution mainly because of methodological limitations (Villanueva-Canas et al., 2017). First, TE families are difficult to identify as they are repetitive elements spread throughout the genome with limited descriptive features such as target site duplications or terminal repeats and transposases (Xiong et al., 2014). They are also commonly associated with genetic load as they can lead to diseases if inserted into genes (Chen et al.,...
An example of segregation distortion of meiotic gene drivers with the “killer-target” strategy is shown for the SD gene complex in Drosophila melanogaster. This complex involves the Sd and Rsp loci. The “target” locus Rsp harbours two alleles: the Rsp\(^s\) and Rsp\(^i\) that are respectively sensitive and insensitive to the meiotic driver. Sd produces a “killer” element (red dots, a protein in this example) which interferes with Rsp\(^s\) and kills the meiotic products that inherit Rsp\(^s\).
Figure 3 – Figure adapted from Glémin et al. (2014). GC-biased gene conversion (gBGC) occurs after the formation of hetero-duplexes during a meiotic recombination event. Heteroduplexes can result from crossing-over (shown here) as well as from non-crossover events. In the presence of gBGC, mismatches are repaired more than half of the time in favor of guanine and cytosine rather than adenine and thymine. This bias towards enrichment of guanine and cytosine is of few percent in yeast or humans, leading to the accumulation of GC content within genomes over time.

Lachance and Tishkoff, 2014; Pouyet, Aeschbacher, et al., 2018). One way to consider the consequences of gBGC is to contrast genetic diversity of a region between weak-to-strong and strong-to-weak mutations. Additionally, gBGC should not leave the same signature as linked selection because gBGC acts solely on its targets sites and does not affect the surrounding diversity. There is no inherent selective bias driving the genomic changes resulting from gene conversion, nor can a biased process by definition be considered neutral. Instead, this process is best considered as non-adaptive evolution, where fitness is not impacted, but there is clearly a non-neutral change in allele frequencies over time. As illustrated by Hurst (2019), segregation distortion and BGC are rarely considered simultaneously even though they act similarly to increase their transmission to the detriment of other alleles. Both BGC and segregation distortion along with structural genomic changes are common in nature, and including these processes in studies of adaptive evolution will allow us to properly identify targets of positive selection (Villanueva-Canas et al., 2017).

5. Conclusion

In the modern genomic era, emerging data will hopefully allow us to build a complete picture of the relevant genomic, demographic, and environmental scenarios where different evolutionary processes are expected to dominate changes in molecular diversity over time. Identifying a variant as subject to natural selection is difficult since the selective environment of an allele is a combination of its own innate properties impacting the genome, along with epistatic effects due to its genomic environment, as well as its demographic situation (how efficient selection is in the population where this individual exists), and lastly its (a)biotic environment (e.g. stressful environments for the organism harboring this variant, or the frequency of conspecific phenotypes). Understanding the processes that may bias our inferences of sites under selection is paramount to better understanding the evolutionary forces leading to genomic change. To a large extent these biases result from the difficulty or inability to distinguish non-adaptive sites from sites under direct selection. As discussed, this is largely due to the demographic processes in a population’s past that make otherwise neutral sites mimic selection and fall into the category of non-adaptive, as well as due to the molecular processes that change neutral allele frequencies in biased ways and can even counteract the effect of selection on selected sites (gBGC, for instance). There is an implied distinction worth explicitly stating: even if the majority of sites within the genome may be neutral in terms of their selection coefficient, it is very likely the case that the majority of the genome evolves due to the impact of selective forces, even if that targets few specific sites, due to the degree of linkage within the genome. This still does not, however, discount the fact
that non-adaptive evolutionary processes have an important impact on genomic change. Incorporating all of this information in future studies is a tall task, particularly since empirical study of biology is further complicated by changing environments and demographics that are not always apparent to observers, nor always sufficiently sampled. We hope that this perspective has highlighted the importance of recognizing and distinguishing the complex interactions of selective, non-adaptive, and neutral processes acting within and among genomes and serves to move the field of evolutionary genomics forward in understanding the drivers of molecular diversity.

6. Conflict of interest disclosure

The authors of this article declare that they have no financial conflict of interest with the content of this article.

7. Acknowledgements

We would like to thank Benoit Nabholz, an anonymous reviewer, and Guillaume Achaz for critical feedback, as well as Tyler Kent, Stephan Peischl, and Marie Cariou for their input on earlier versions, all of which greatly improved this manuscript. A previous version of this article has been peer-reviewed and recommended by Peer Community In Evol Biol (https://doi.org/10.24072/pci.evolbiol.100103).

References


Kimura M et al. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626. [https://doi.org/10.1038/217624a0](https://doi.org/10.1038/217624a0).


