Genome Evolution in Outcrossing vs. Selfing vs. Asexual Species

Sylvain Glémin, Clémentine M. François, Nicolas Galtier

Chapter 11

Genome Evolution in Outcrossing vs. Selfing vs. Asexual Species

Sylvain Glémin, Clémentine M. François, and Nicolas Galtier

Abstract

A major current molecular evolution challenge is to link comparative genomic patterns to species' biology and ecology. Breeding systems are pivotal because they affect many population genetic processes and thus genome evolution. We review theoretical predictions and empirical evidence about molecular evolutionary processes under three distinct breeding systems—outcrossing, selfing, and asexuality. Breeding systems may have a profound impact on genome evolution, including molecular evolutionary rates, base composition, genomic conflict, and possibly genome size. We present and discuss the similarities and differences between the effects of selfing and clonality. In reverse, comparative and population genomic data and approaches help revisiting old questions on the long-term evolution of breeding systems.

Key words Breeding systems, GC-biased gene conversion, Genome evolution, Genomic conflicts, Selection, Transposable elements

1 Introduction

In-depth investigations on genome organization and evolution are increasing and have revealed marked contrasts between species, e.g., evolutionary rates, nucleotide composition, and gene repertoires. However, little is still known on how to link this "genomic diversity" to the diversity of life history traits or ecological forms. Synthesizing previous works in a provocative and exciting book, M. Lynch asserts that variations in fundamental population genetic processes are essential for explaining the diversity of genome architectures while emphasizing the role of the effective population size (N_e) and nonadaptive processes [1]. Life history and ecological traits may influence population genetic parameters, including N_e, making it possible to link species' biology and their genomic organization and evolution (e.g., [2–7])

Among life history traits affecting population genetic processes, breeding systems are pivotal as they determine the way genes are transmitted to the next generation (Fig. 1). Outcrossing,

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_11,

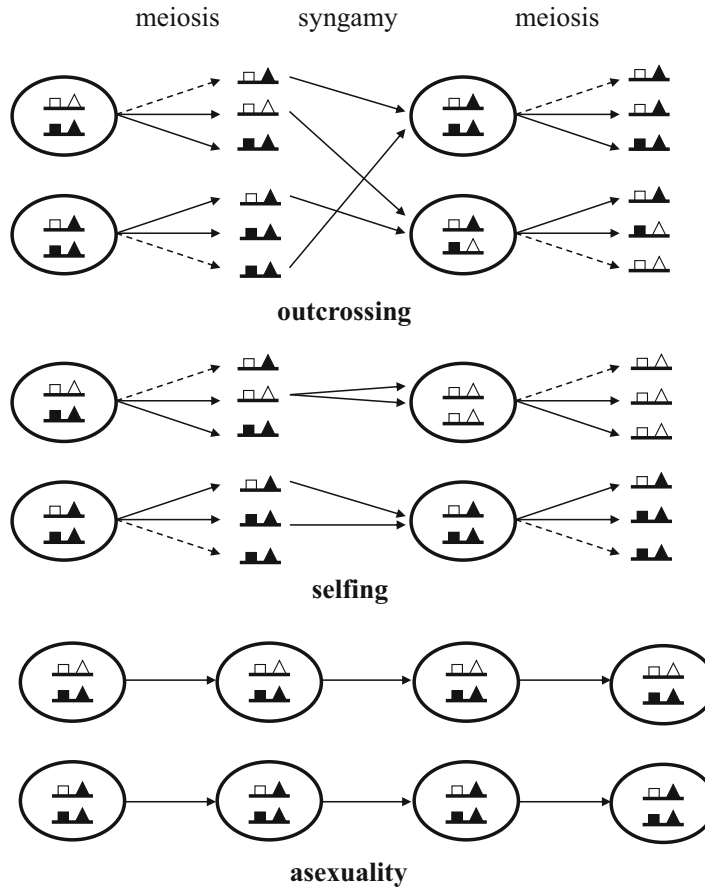

Fig. 1 Reproduction and genotype transmission in outcrossing, selfing, and asexual species. In outcrossers, parental and recombinant (dotted lines) gametes from distinct zygotes are shuffled at generation $n + 1$. In selfers, only gametes produced by a given zygote can mate, which quickly increases homozygosity and reduces the recombination efficacy. Asexuals do not undergo meiosis or syngamy. They reproduce clonally

sexual species (outcrossers) reproduce through the alternation of syngamy (from haploid to diploid) and meiosis (from diploid to haploid), with random mating of gametes from distinct individuals at each generation. Outcrossing is a common breeding system that is predominant in vertebrates, arthropods, and many plants, especially perennials, etc. [8, 9]. Selfing species (selfers) also undergo meiosis, but fertilization only occurs between gametes produced by the same hermaphrodite individual. Consequently, diploid individuals from selfing species are highly homozygous (FIS ~ 1; *see*, for instance, ref. 10)—heterozygosity is divided by two at each generation, and the two gene copies carried by an individual have a high probability of being identical by descent. Selfing is common in various plant families (e.g., *Arabidopsis thaliana*), mollusks,

nematodes (e.g., *Caenorhabditis elegans*), and platyhelminthes, among others [8, 9]. Note that many sexual species have intermediate systems in which inbreeding and outbreeding coexist. In organisms with a prolonged haploid phase (such as mosses, ferns, or many algae and fungi), a more extreme form of selfing can occur by taking place during the haploid phase (haploid selfing or intragametophytic selfing), leading instantaneously to genome-wide homozygosity [11]. Clonal asexual species, finally, only reproduce via mitosis, so that daughters are genetically identical to mothers unless a mutation occurs. In diploid asexuals, homologous chromosomes associated in a given zygote do not segregate in distinct gametes—they are co-transmitted to the next generation in the absence of any haploid phase. In contrast to selfing species, individuals from asexual diploid species tend to be highly heterozygous (FIS ~ −1, [12]), since any new mutation will remain at the heterozygote stage forever, unless the same mutation occurs in the homologous chromosome. Clonality is documented in insects (e.g., aphids), crustaceans (e.g., daphnia), mollusks, vertebrates, and angiosperms, among others [13–16]. As for selfing, clonality can also be partial, with sexual reproduction occurring in addition or in alternation with asexual reproduction. In addition to this common form of asexuality, other forms such as automixis imply a modified meiosis in females where unfertilized diploid eggs produce offspring potentially diverse and distinct from their mother, leading to different levels of heterozygosity [13]. This diversity of reproductive systems should be kept in mind, but for clarity we will mainly compare outcrossing, diploid selfing, and clonality.

Through the occurrence, or not, of syngamy, recombination, and segregation, breeding systems affect population genetic parameters (effective population size, recombination rate, efficacy of natural selection; Fig. 2) and thus, potentially, genomic patterns. A large corpus of population genetic theory has been developed to study the causes and consequences of the evolution of breeding systems (Table 1). Thanks to the exponentially growing amount of genomic data, and especially data from closely related species with contrasted breeding systems, it is now possible to test these theoretical predictions. Conversely, genomic data may help in understanding the evolution of breeding systems. Genomes should record the footprints of transitions in breeding systems and help in testing the theory of breeding system evolution in the long run, e.g., the "dead-end hypothesis," which posits that selfers and asexuals are doomed to extinction because of their inefficient selection and low adaptive potential [17, 18]. Since the first edition of this book, several theoretical developments have clarified the population genetics consequences of the different breeding systems, and empirical evidences have been accumulating, partly changing our view of breeding system evolution and consequences, especially for asexual organisms. We first review and update the consequences of

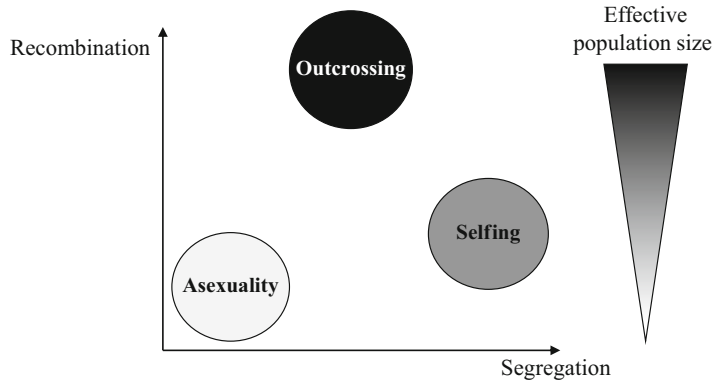

Fig. 2 A schematic representation of the effect of breeding systems on population genetic parameters

Table 1
Summary of the major theoretical predictions regarding breeding systems and evolutionary genomic variables, with outcrossing being taken as reference

	F_{IS}	πS	dN/dS	Codon usage	TE	LD	GC-content
Outcrossing	~0	+	+	+	+	+	+
Selfing	~1	−	++	−	Unclear	++	−
Asexuality	~−1	−	+++	−	Unclear	+++	−

TE transposable element abundance, *LD* linkage disequilibrium

breeding systems on genome evolution and then discuss and re-evaluate how evolutionary genomics shed new light on the old question of breeding system evolution.

2 Contrasted Genomic Consequences of Breeding Systems

2.1 Consequences of Breeding Systems on Population Genetics Parameters

Sex involves an alternation of syngamy and meiosis. In outcrossing sexual species, random mating allows alleles to spread across populations, while segregation and recombination (here in the sense of crossing-over) associated with meiosis generate new genotypic and haplotypic combinations. This strongly contrasts with the case of selfing and asexual species. In such species, alleles cannot spread beyond the lineage they originated from because mating occurs within the same lineage (selfers) or because syngamy is suppressed (asexuals). Recombination, secondly, is not effective in non-outcrossers. In selfers, while physical recombination does occur (r_0), effective recombination (r_e) is reduced because it mainly occurs between homozygous sites, and it completely vanishes under complete selfing: for tight linkage, $r_e = r_0(1 - F_{IS})$, where F_{IS} is the Wright's fixation index [19], whereas for looser linkage,

effective recombination is more reduced than predicted by this simple expression [20–22]. In asexuals, physical recombination is suppressed ($r_0 = r_e = 0$). High levels of linkage disequilibrium (nonrandom association of alleles between loci) could therefore be expected in selfers and asexuals. The observed data are mainly consistent with these predictions. In the selfing model species *Arabidopsis thaliana*, LD extends over a few hundreds of kb, while in maize, an outcrosser, LD quickly vanishes beyond a few kb [23]. In a meta-analysis, Glémin et al. [24] also found higher LD levels in selfers than in outcrossers. Beyond pairwise LD, selfing also generates higher-order associations, such as identity disequilibria (the excess probability of being homozygote at several loci, [25]) that alter population genetics functioning compared to outcrossing populations (e.g., [26]).

Theory also predicts that the effective population size, N_e, depends on the breeding system (Fig. 2). First, compared to outcrossers, selfing is expected to directly lower N_e by a factor $1 + F_{IS}$ by reducing the number of independent gametes sampled for reproduction [27]. From a coalescent point of view, selfing reduces coalescent time (again by the same factor $1 + F_{IS}$). Under outcrossing, two gene copies gathered in a same individual either directly coalesce or move apart at the preceding generation. Selfing prolongs the time spent within an individual, hence the probability of coalescing [19, 28]. In diploid asexuals, the picture is less obvious. Since genotypes, not alleles, are sampled, Balloux et al. [12] distinguished between the genotypic and allelic effective size. The genotypic effective size equals N, not $2N$, i.e., the actual population size, similarly to the expectation under complete selfing. On the contrary, the allelic effective size tends toward infinity under complete clonality because genetic diversity within individuals cannot be lost [12]. This corresponds to preventing coalescence as long as gene copies are transmitted clonally [29, 30]. However, very low level of sex (higher than $1/2N$) is sufficient to retrieve standard outcrossing coalescent behavior [29, 30], and as far as natural selection is concerned (see below), the genotypic effective size is what matters [31]. The ecology of selfers and asexuals may also contribute to decreasing N_e as they supposedly experience more severe bottlenecks than outcrossers [32, 33]. On the contrary, higher population subdivision in selfers could contribute to increasing N_e at the species scale. However, Ingvarsson [34] showed that, under most conditions, the extinction/recolonization dynamics is predicted to decrease N_e in selfers, at both the local and metapopulation scale. Finally, because of low or null effective recombination, hitchhiking effects—the indirect effects of selection at a locus on other linked loci—reduce N_e further [35]. Under complete selfing or clonality, because of full genetic linkage, selection at a given locus affects the whole genome. Most forms of selection, and especially directional selection, reduce the number of gene copies

contributing to the next generation by removing deleterious alleles to the benefit of advantageous ones. Because of linkage, such a reduction spreads over the rest of the genome, globally reducing the effective population size (sensu lato) in non-outcrossing species. Background selection, the reduction in N_e due to the removal of deleterious mutations at linked loci, can be particularly severe in highly selfing and clonal population, potentially reducing N_e by one order of magnitude or more [22, 36]. And this effect is expected to be stronger in asexuals than in selfers [36]. In the predominantly selfing nematode *C. elegans*, nucleotide diversity has been shown to be reduced genome wide by both background selection [37] and selective sweeps [38], and in a comparative analysis, the effect of linked selection has shown to be more pronounced in selfing than in outcrossing species [39].

As genetic diversity scales positively with $N_e\mu$, where μ is the mutation rate, selfers are expected to be less polymorphic than outcrossers. Asexuals should also exhibit lower genotypic diversity, but the prediction is not clear for allelic diversity (see above). However, because of the lack of recombination, haplotype diversity should be lower for both breeding systems. The effect of selfing on the polymorphism level is well documented, and empirical data mainly agree with the theoretical predictions. Selfing species tend to be more structured, less diverse, and straightforwardly more homozygotes than outcrossers [6, 24, 40, 41]. Much fewer data exist regarding diversity levels in asexuals, but the available datasets confirm that genotypic diversity, at least, is usually low in such species (*see* discussion in ref. 12). At the population level, a recent comparative analysis of sexual and asexual *Aptinothrips rufus* grass thrips confirmed the expected lower nuclear genetic diversity of asexual populations while also evidencing that some asexuals with extensive migration can feature very high mitochondrial genetic diversity [42].

These predictions concerning polymorphism patterns implicitly assumed that mutation rates are the same among species with contrasted breeding systems. However, modifications in breeding systems can also affect various aspects of the species life cycle potentially related to the mutation rate. In asexuals, for instance, loss of spermatogenesis can reduce mutation rates, while loss of the dormant sexual phase can increase them (reviewed in [43]). Mutation rates can also be decreased in non-outcrossers due to the loss of recombination, which can be mutagenic [44, 45]. In selfers, meiosis and physical recombination do occur. However, the specific mutagenic process during meiosis depends on the level of heterozygosity, such as indel-associated mutations (IDAM): heterozygote indels could increase the point mutation rate at nearby nucleotides because of errors during meiosis [46, 47]. Consistent with this prediction, the IDAM process more strongly affects the outcrossing wild rice, *Oryza rufipogon*, than the very recent selfer and weakly

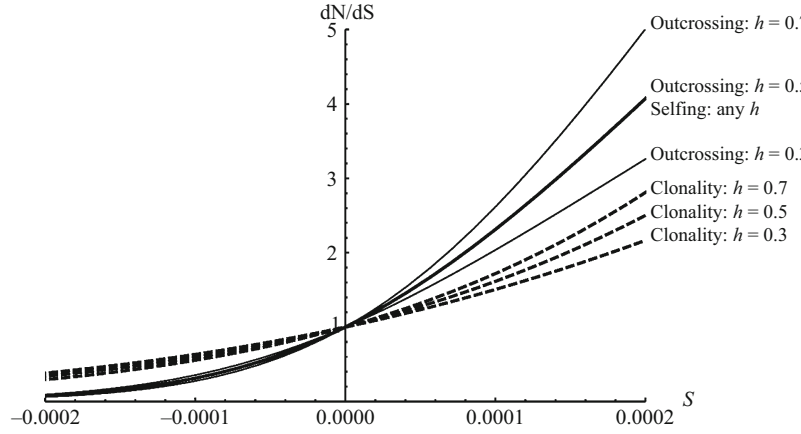

Fig. 3 Substitution rates relative to the neutral case (dN/dS) in outcrossers (thin lines), selfers (bold line), and asexuals (dotted lines) for different mutation dominance levels. The fitness of the resident, heterozygote, and homozygote mutant genotypes are 1, $1 - hs$, and $1 - s$, respectively. For asexuals, it is necessary to consider two substitution rates corresponding to the initial fixation of heterozygotes and the ultimate fixation of complete homozygote mutants from an initially heterozygote population [31]. Population size: $N = 10{,}000$. To highlight the difference between selfers and asexuals due to segregation, demographic and hitchhiking effects reducing N_e in asexuals and selfers are not taken into account

heterozygous domesticated rice, *O. sativa*. *A. thaliana*, a more ancient and mostly homozygous selfer, is very weakly affected by IDAM [48]. Overall, these processes should globally contribute to lowering mutation rates, and thus polymorphism, in selfing and asexual species.

2.2 Breeding Systems and Selection Efficacy

2.2.1 Drift and Recombination: Parallel Reduction in Selection Efficacy in Selfers and Asexuals?

The effective population size strongly affects the outcome of natural selection. The probability of fixation of a new mutation is a function of the $N_e s$ product, where s is the selection coefficient ([49] and *see* Fig. 3). As N_e is reduced, a higher proportion of mutations behave almost neutrally. Weakly deleterious alleles can thus be fixed, while weakly advantageous ones can be lost. Genetic associations among loci generated by selfing and clonality also induce selective interferences [26, 50]. Because of their reduced effective population size and recombination rate, selection is thus expected to be globally less effective in selfers and asexuals than in outcrossers, which should result in various footprints at the molecular level (Table 1). Assuming that most mutations are deleterious (with possible back compensatory mutations), both the ratio of non-synonymous to synonymous polymorphism, $\pi N/\pi S$, and the ratio of non-synonymous to synonymous substitutions, dN/dS, are predicted to be higher in selfers and asexuals than in outcrossers. Codon usage should also be less optimized in selfers and asexuals than in outcrossers.

Contrary to polymorphism surveys, few studies have tested these predictions empirically (Table 2). In the few available

Table 2
Summary of studies comparing patterns of molecular evolution between different breeding systems

	Taxonomic group	Groups compared	Dataset	dN/dS	πN/πS	Positive selection	Codon usage	GC-content	Reference
Outcrossing/selfing	Angiosperms	29 selfers/42 outcrossers	Meta-analysis (polymorphism)			+/−		+/−	Glémin et al. [24]
	Angiosperms	10 selfers/18 outcrossers				+			Chen et al. [6]
	Angiosperms (17 phylogenetic datasets)	4–61 selfers or SC/5–125 outcrossers or SI	Chloroplastic genes (matK and rbcL)	+/−[a]					Glémin and Muyle [64]
	Arabidopsis	1 selfer/1 outcrosser	23 nuclear genes + 1 chloroplastic gene	−	−				Wright et al. [61]
	Arabidopsis	1 selfer/1 outcrosser	675/62 nuclear genes			+/−			Foxe et al. [218]
	Arabidopsis/Brassica	1 selfer/2 outcrossers	185 nuclear genes		−				Wright et al. [118]
	Arabidopsis/Capsella	1 selfer/1 outcrosser	257 nuclear genes			+		+	Slotte et al. [83]
	Arabidopsis/Capsella	2 selfers/2 outcrossers	780, 89/120, 257 nuclear genes				+		Qiu et al. [223]
	Capsella	1 selfer/1 outcrosser	Complete genome/transcriptomes			+			Slotte et al. [224], [225]
	Collinsia	1 selfer/1 outcrosser	17 nuclear genes + transcriptomes		−	+			Hazzouri et al. [123]
	Eichornia	3 selfers/1 outcrosser	~8000 nuclear genes		+	+			Ness et al. [221]
	Eichornia paniculata (several populations)	10 selfers/10 outcrossers	Transcriptomes (>16,000 contigs)			+			
	Triticeae	2 selfers/2 outcrossers	52 nuclear genes + 1 chloroplastic gene	−				+/−	Haudry et al. [84]
	Triticeae	9 selfers/10 outcrossers	27 nuclear genes	−					Escobar et al. [63]
	Neurospora	32 homothalics/17 heterothallics	7 nuclear genes	+[b]					Nygren et al. [222]
	Neurospora	1 homothalic/1 heterothallic	>2000 nuclear genes		−				Whittle et al. [226]

	Neurospora	4 homothalic/31 heterothallic	>2700 nuclear genes	+	+/−				Gioti et al. [219]
	Caenorhabditis	2 selfers/4 outcrossers	>1000 nuclear genes	−	+/−				Cutter et al. [62]
	Galba/Physa	2 selfers/2 outcrossers	Transcriptomes (>1000 contigs)	+		+	+	+	Burgarella et al. [66]
Sexuals/ asexuals	*Boechera*	8 asexuals/8 sexuals	Complete genome	+					Lovell et al. [55]
	Oenothera	16 asexuals/16 sexuals	1 nuclear gene (*chiB*)					+	Hersch-Green et al. [85]
	Oenothera	16 asexuals/13 sexuals	Transcriptomes (>4000 contigs)	+		+			Hollister et al. [58]
	Ranunculus	2 sexuals/3 asexuals	Transcriptomes	−					Pellino et al. [57]
	Aphids	4 sexuals/4 asexuals	255 nuclear genes + 10 mitochondrial genes	−					Ollivier et al. [56]
	Campeloma	6 asexuals/12 sexuals	1 mitochondrial gene (*Cytb*)	+					Johnson et al. [53]
	Daphnia	14 asexuals/14 sexuals	Complete mitochondrial genome	+					Paland and Lynch [51]
	Daphnia	11 asexuals/11 sexuals	Complete mitochondrial genome	−					Tucker et al. [60]
	Lineus	2 asexuals/7 sexuals	Transcriptomes (>2800 contigs)			+/−[c]			Ament-Velasquez et al. [59]
	Potamopyrgus	14 asexuals/14 sexuals	Complete mitochondrial genome	+					Neiman et al. [54]
	Rotifers	3 asexuals/2 sexuals	1 nuclear gene (*Hsp 82*)	−					Mark-Welch and Meselson [220]
	Rotifers	3 asexuals/4 sexuals	1 mitochondrial gene (*Cox I*)	+/−		+			Barraclough et al. [217]
	Timema	6 asexuals/7 sexuals	2 nuclear genes + 1 mitochondrial gene	+					Henry et al. [52]

+ and − indicate if theoretical predictions are confirmed or not. Empty cells correspond to nonavailable data

[a]No more significant after controlling for terminal vs. internal branches

[b]Terminal vs. internal branches not controlled

[c]Possible confounding effect of hybrid origin of one asexual lineage

comparative studies, contrasted patterns were observed between selfers and asexuals. Compared to sexual ancestors, recent asexual lineages show a marked increase in the dN/dS ratio in *Daphnia* ([51] but see below), *Timema* stick insects [52], gastropods *Campeloma* [53] and *Potamopyrgus* [54], and the plant *Boechera* [55], in agreement with theoretical predictions (Table 2). However, no significant effect of asexuality on dN/dS was found in four aphid species [56] and in the plant *Ranunculus auricomus* [57]. Bdelloid rotifers, long considered as ancient asexuals (see below), exhibit a higher $\pi N/\pi S$ ratio but not a higher dN/dS ratio than comparable sexual groups, suggesting that mildly deleterious mutations can segregate at a higher frequency in asexuals but are eventually removed. A higher $\pi N/\pi S$ ratio in asexual lineages than in sexual relatives was reported from transcriptome data in *Oenothera* primroses [58] and *Lineus* nemerteans [59]. Note however that in the latter case, the increased $\pi N/\pi S$ is primarily explained by the hybrid nature of the asexual *Lineus pseudolacteus* (Table 2). The recent origin of asexuality through introgression also challenges the interpretation of elevated dN/dS ratio in the mitochondrial genome of asexual lineages of *Daphnia pulex* [51], as less than 1% of mutations on the branches leading to asexual lineages would have arisen after the transition to asexuality [60]. Here, rather than being the direct cause of genomic degradation, asexuality may have evolved in already-degraded lineages.

All predictions are not equally supported by data in selfers. Polymorphism-based measures mostly support reduction in selection efficiency in selfers in various plant species, and this was recently confirmed by a meta-analysis of genome-wide polymorphism data ([6] and *see* Table 2). On the contrary, as far as dN/dS or base composition are compared, most studies, in plants, fungi, and animals, did not find evidence of relaxed selection in selfers (Table 2). A recent origin of selfing is often invoked to explain that effect of selfing is rarely observed in species divergence (e.g., [61, 62–64]), whereas a recent transition to selfing can leave a clear signature of relaxed selection at the polymorphism level [65]. In contrast, in the freshwater snail *Galba truncatula* where selfing is supposed to be old and ancestral to a clade of several species, relaxed selection in the selfing lineage was also observed at the divergence level [66]. The same rationale should apply to asexual species. However, in *Campeloma*, *Potamopyrgus*, *Timema*, and *Boechera*, clonality is also recent, yet the expected patterns are observed at the divergence level. The reduction in N_e could simply be less severe in selfers than in asexuals as predicted by background selection models [36]. Furthermore, complete selfing is hardly ever noted in natural populations; residual outcrossing typically occurs. Among hitchhiking effects, some are very sensitive to the recombination level, such as Muller's ratchet [67], weak Hill-Robertson interferences [50], or hitchhiking of deleterious mutations during

selective sweeps [68, 69]. If such mechanisms are the main cause of reduction of N_e in selfers, then even a low recombination rate could be enough to maintain the selection efficacy. This is suggested by genomic patterns across recombination gradients in outcrossing species. In primates, no effect of recombination on the selection efficacy has been detected [70]. In *Drosophila*, Haddrill et al. [71] found little evidence of reduced selection in low recombining regions, except when recombination was fully suppressed, as in Y chromosomes. Differences between selfers and asexuals could thus simply result from different degrees of residual outcrossing. However, as stated above, selfers and asexuals also fundamentally differ as far as segregation is concerned, as we now discuss in more detail.

2.2.2 Segregation: Dealing with Heterozygotes

Selfing affects the selection efficacy by increasing homozygosity and thus exposing recessive alleles to selection. This effect can counteract the effect of reducing N_e. Considering the sole reduction in N_e due to non-independent gamete sampling, selection is less efficient under partial selfing for dominant mutations but more efficient for recessive ones (Fig. 3, and *see* ref. 72). More precisely, Glémin [73] determined the additional reduction in N_e (due to hitchhiking and demographic effects) necessary to overcome the increased selection efficacy due to homozygosity. This additional reduction can be high for recessive mutations. On the contrary, the lack of segregation in asexuals reduces selection efficacy and increases the drift load, as heterozygotes can fix [31]. The effects of selfing and clonality on the fixation probability of codominant, recessive, or dominant mutations are summarized in Fig. 3. Note that segregation may also have indirect effects. When recombination is suppressed, Muller's ratchet is supposed to reduce N_e and contribute to the fixation of weakly deleterious alleles [74]. In selfers, the purging of partially recessive deleterious alleles slows down the ratchet [67], which suggests that the fixation of deleterious alleles at linked loci would be lower in selfers than in asexuals. The same mechanism also contributes to weaker background selection in selfers than in asexuals (see above, [36]). In the extreme case of intra-gametophytic selfing, purging could be even more efficient at removing deleterious alleles [11], as it has been suggested for moss species [75]. Segregation at meiosis could thus partly explain the differences between selfers and asexuals, but more data are clearly needed to confirm this hypothesis.

The two opposite effects of drift and segregation in selfers should also affect adaptive evolution. In outcrossers, new beneficial mutations are more likely to be rapidly lost if recessive, as they are initially present in heterozygotes and masked to selection—a process known as Haldane's sieve [76]. By unmasking these mutations in homozygotes, selfing could help adaptive evolution from recessive mutations [72, 73]. However, this advantage of selfing disappears when adaptation proceeds from pre-existing variation because

homozygotes can also be present in outcrossers [77]. Selective interference in selfers also reduces their advantage of not experiencing Haldane's sieve, especially for weakly beneficial mutations [21], and the effect of background should globally reduce the rate of adaptation [73, 77, 78]. Conversely, the lack of segregation in asexuals delays the complete fixation of an advantageous mutation. Once a new advantageous mutation gets fixed in the heterozygotic state, additional lag time until occurrence and fixation of a second mutation is necessary to ensure fixation [79]. Little is known about the dominance levels of new adaptive mutations, but a survey of QTL fixed during the domestication process in several plant species confirmed the absence of Haldane's sieve in selfers compared to outcrossers [80]. This mostly corresponds to strong selection on new mutations or mutations in low initial frequencies in the wild populations. More generally, the effect of selfing on adaptive evolution will depend on the distribution of dominance and selective effects of mutations and the magnitude of genetic drift and linkage.

Few studies have tested for difference in positive selection between selfers and outcrossers. In their survey of sequence polymorphism data in flowering plants, Glémin et al. [24] found, on average, more genes with a signature of positive selection in outcrossers than in selfers assessed by the McDonald-Kreitman test [81]. An extension of this method—where non-synonymous vs. synonymous polymorphism data are used to calibrate the distribution of the deleterious effects of mutations and then attribute the excess non-synonymous divergence observed to positive selection [82]—was applied to one plant [83] and one freshwater snail dataset. In both studies, a large fraction of non-synonymous substitutions was estimated to be adaptive in the outcrossing species (~40% in the plant *Capsella grandiflora* and ~55% in the snail *Physa acuta*), whereas this proportion was not significantly different from zero in the selfer (*Arabidopsis thaliana* and *Galba truncatula*, respectively). Based on methods where the dN/dS ratio is allowed to vary both among branches and sites, a comparative analysis of two outcrossing and two selfing Triticeae species [84] suggested that adaptive substitutions may have specifically occurred in the outcrossing lineages. This would contribute to explaining why selfing lineages did not show a higher dN/dS ratio than outcrossing ones (see above and Table 2). So the data available so far support an increased rate of adaptation in outcrossing species, suggesting that the effects of drift and linkage overwhelm the advantage of avoiding Haldane's sieve. A similar approach was used in *Oenothera* species suggesting also reduced adaptive evolution in clonal compared to sexual lineages [85].

Finally, the classical assumption of a lack of segregation in asexuals must be modulated. First, in some form of asexuality,

such as automixis, female meiosis is retained, and diploidy restoration occurs by fusion or duplication of female gametes. Depending on how meiosis is altered, automixis generates a mix of highly heterozygous and highly homozygous regions along chromosomes. The genomes of such species could thus exhibit a gradient of signatures of selfing and diploid clonal evolution [86]. Secondly, mitotic recombination and gene conversion in the germline of asexual lineages can also reduce heterozygosity at a local genomic scale. Mitotic recombination has been well documented in yeast (*see* review in ref. 87) and also occurs in the asexual trypanosome *T. b. gambiense* [88] and in asexual *Daphnia* lineages [60, 89, 90]. If its frequency is of the order or higher than mutation rates, as reported in yeast and *Daphnia*, asexuals would not suffer much from the lack of segregation at meiosis. Especially, during adaptation, the lag time between the appearance of a first beneficial mutation and the final fixation of a mutant homozygote could be strongly reduced [87]. However, such mechanisms of loss of heterozygosity also rapidly expose recessive deleterious alleles in heterozygotes and generate inbreeding-depression-like effects [60].

2.2.3 Selection on Genetic Systems

So far, we have only considered the immediate, mechanistic effects of breeding systems on population genetic parameters. Breeding systems, however, can also affect the evolution of genetic systems themselves, which modulates previous predictions. Theoretical arguments suggested that selfing, even at small rates, greatly increases the parameter range under which recombination is selected for [91–93]. These predictions have been confirmed in a meta-analysis in angiosperms in which outcrossers exhibited lower chiasmata counts per bivalent than species with mixed or selfing mating systems [94]. Higher levels of physical recombination (r_0) could thus help break down LD and reduce hitchhiking effects. This could contribute to explaining why little evidence of long-term genomic degradation has been observed in selfers, compared to asexuals.

Breeding systems may also affect selection on mutation rates. Since the vast majority of mutations are deleterious, mutation rates should tend toward zero, up to physiological costs of further reducing mutation rates being too high (e.g., [95, 96]). Under complete linkage, a modifier remains associated with its "own" mutated genome. Selection should thus favor lower mutation rates in asexuals and selfers (e.g., [95, 96]). However, Lynch recently challenged this view and suggested a lower limit to DNA repair may be set by random drift, not physiological cost [97]. Such a limit should thus be higher in asexuals and selfers. Asexuality is often associated with very efficient DNA repair systems (reviewed in [43]), supporting the view that selection for efficient repair may overwhelm drift in asexual lineages. Alternatively, only groups

having high-fidelity repair mechanisms could maintain asexuality in the long run. More formal tests of mutation rate differences between breeding systems are still scarce. The phylogenetic approach revealed no difference in dS, as a proxy of the neutral mutation rate, between *A. thaliana* and *A. lyrata* [61], nor did a mutation accumulation experiment that compared the deleterious genomic mutation rate between *Amsinckia* species with contrasted mating systems [98]. A similar experiment in *Caenorhabditis* showed that the rate of mutational decay was, on average, fourfold greater in gonochoristic outcrossing taxa than in the selfer *C. elegans* [99]. Recent mutation accumulation experiments on *Daphnia pulex* suggested a slightly lower mutation rate in obligate than in facultative asexual genotypes, except for one mutator phenotype which evolved in an asexual subline [90]. Overall, these results do not support Lynch's hypothesis of mutation rates being limited by drift in asexual and selfing species. However, such experiments are still too scarce, and quantifying how mutation rates vary or not with breeding systems is a challenging issue that requires more genomic data.

2.3 Breeding Systems and Genomic Conflicts

Outcrossing species undergo various sorts of genetic conflict. Sexual reproduction directly leads to conflicts within (e.g., for access to mating) and between sexes (e.g., for resource allocations between male and female functions or between offspring). In selfers and asexuals, such conflicts occur because mates are akin or because mating is absent [100, 101]. Outcrossers are also sensitive to epidemic selfish element proliferation and to meiotic drive, because alleles can easily spread over the population through random mating. In contrast, selfers and asexuals should be immune to such genomic conflicts because selection only occurs between selfing or asexual lineages so that selfish elements should be either lost or evolve into commensalists or mutualists [102].

2.3.1 Relaxation of Sexual Conflicts in Selfers and Asexuals

Some genes involved in sexual reproduction are known to evolve rapidly because of recurrent positive selection [103]. Arm races for mating or for resource allocation to offspring are the most likely causes of this accelerated evolution. In selfers and asexuals, selection should be specifically relaxed on these genes, not only because of low recombination and effective size but mainly because the selection pressure per se should be suppressed. According to this prediction, in the outcrosser *C. grandiflora*, 6 out of the 20 genes that show the strongest departure from neutrality are reproductive genes and under positive selection. This contrasts with the selfer *A. thaliana*, for which no reproductive genes are under positive selection [83].

More specifically, two detailed analyses provided direct evidence of relaxed selection associated with sexual conflict reduction. In the predominantly selfer *C. elegans*, some males deposit a copulatory plug that prevents multiple matings. However, other males do not deposit this plug. A single gene (plg-1), which encodes a major structural component of this plug, is responsible for this dimorphic reproductive trait [104]. Loss of the copulatory plug is caused by the insertion of a retrotransposon into an exon of plg-1. This same allele is present in many populations worldwide, suggesting a single origin. The strong reduction in male-male competition following hermaphroditism and selfing evolution explains that no selective force opposes the spread of this loss-of-function allele [104, 105]. In *A. thaliana*, similar relaxed selection has been documented in the MEDEA gene, an imprinted gene directly involved in the male vs. female conflict. MEDEA is expressed before fertilization in the embryo sac and after fertilization in the embryo and the endosperm, a tissue involved in nutrient transfer to the embryo. In *A. lyrata*, an outcrossing relative to *A. thaliana*, MEDEA could be under positive [106] or balancing selection [107], in agreement with permanent conflicting pressures for resource acquisition into embryos between males and females. Conversely, this gene evolved under purifying selection in *A. thaliana*, where the level of conflict is reduced.

Male vs. female diverging interests are also reflected by cyto-nuclear conflicts. When cytoplasmic inheritance is uniparental, as in most species, cytoplasmic male sterility (CMS) alleles favoring transmission via females at the expense of males can spread in hermaphroditic outbreeding species, leaving room for coevolution with nuclear restorers. Maintenance of CMS/non-CMS polymorphism leads to stable gynodioecy [108]. In selfers, CMS mutants also reduce female fitness—because ovules cannot be fertilized—and are thus selected against. In the genus *Silene*, the mitochondrial genome of gynodioecious species exhibits molecular signatures of adaptive and/or balancing selection. This is likely due to cyto-nuclear conflicts as this is not, or is less, observed in hermaphrodites and dioecious [109–111]. Although less studied, cyto-nuclear conflicts are also expected in purely hermaphroditic species. In a recent study in *A. lyrata*, Foxe and Wright [112] found evidence of diversifying selection on members of a nuclear gene family encoding transcriptional regulators of cytoplasmic genes. Some of them show sequence similarity with CMS restorers in rice. Given the putative function of these genes, such selection could be due to ongoing cyto-nuclear coevolution. Interestingly, in *A. thaliana*, these genes do not seem to evolve under similar diversifying selection, as expected in a selfing species where conflicts are reduced.

2.3.2 Biased Gene Conversion as a Meiotic Drive Process: Consequences for Nucleotide Landscape and Protein Evolution

GC-biased gene conversion (gBGC) is a kind of meiotic drive at the base pair scale that can also be strongly influenced by breeding systems. In many species, gene conversion occurring during double-strand break recombination repair is biased toward G and C alleles (reviewed in [113]). This process mimics selection and can rapidly increase the GC content, especially around recombination hotspots [114, 115], and, more broadly, can affect genome-wide nucleotide landscapes. For instance, it is thought to be the main force that shaped the isochore structure of mammals and birds [116]. gBGC has been mostly studied by comparing genomic regions with different rates of (crossing-over) recombination (reviewed in [116]). However, comparing species with contrasted breeding systems offers a broader and unique opportunity to study gBGC. gBGC cannot occur in asexuals because recombination is lacking. Selfing is also expected to reduce the gBGC efficacy because meiotic drive does not occur in homozygotes [117]. To our knowledge, GC content has never been compared between sexual and asexual taxa, but there have been comparisons between outcrossers and selfers.

As expected, no relationship was found between local recombination rates and GC-content in the highly selfing *Arabidopsis thaliana* [117], and Wright et al. [118] suggested that the (weak) differences observed with the outcrossing *A. lyrata* and *Brassica oleracea* could be due to gBGC. Much stronger evidence has been obtained in grasses. Grasses are known to exhibit unusual genomic base composition compared to other plants, being richer and more heterogeneous in GC-content [119], and direct and indirect evidences of gBGC have been accumulating [119, 120–122]. Accordingly, GC-content or equilibrium GC values were found to be higher in outcrossing than in selfing species [24, 84, 120]. Difference in gBGC between outcrossing and selfing lineages has also been found in the plant genus *Collinsia* [123] and in freshwater snails [66], although difference in selection on codon usage cannot be completely ruled out.

gBGC can also affect functional sequence evolution, leaving a spurious signature of positive selection and increasing the mutation load through the fixation of weakly deleterious AT→GC mutations: gBGC would represent a genomic Achilles' heel [124]. Once again, comparing outcrossing and selfing species is useful for detecting interference between gBGC and selection. gBGC is expected to counteract selection in outcrossing species only. The Achilles' heel hypothesis could explain why relaxed selection was not detected in four grass species belonging to the Triticeae tribe [84]. In outcrossing species, but not in selfing ones, dN/dS was found to be significantly higher for genes exhibiting high than low equilibrium GC-content, suggesting that selection efficacy could be reduced because of high substitution rates in favor of GC alleles in these outcrossing grasses. In outcrossing species,

gBGC can maintain recessive deleterious mutations for a long time at intermediate frequency, in a similar way to overdominance [125]. This could generate high inbreeding depression in outcrossing species, preventing the transition to selfing. In reverse, recurrent selfing would reduce the load through both purging and the avoidance of gBGC, thus reducing the deleterious effects of inbreeding. Under this scenario, gBGC would reinforce disruptive selection on mating systems. In the long term, gBGC could be a new cost of outcrossing: because of gBGC, not drift, outcrossing species could also accumulate weakly deleterious mutations, to an extent which could be substantial given current estimates of gBGC and deleterious mutation parameters [125]. Whether this gBGC-induced load could be higher than the drift load experienced by selfing species remains highly speculative. Both theoretical works, to refine predictions, and empirical data, to quantify the strength of gBGC and its impact on functional genomic regions, are needed in the future. Grasses are clearly an ideal model for investigating these issues, but comparisons with groups having lower levels of gBGC would also be helpful.

2.3.3 Transposable Elements in Selfers and Asexuals: Purging or Accumulation?

Considering the role of sex in the spread of selfish elements, TEs should be less frequent in selfers and asexuals than in outcrossers because they cannot spread from one genomic background to another through syngamy. However, highly selfing and asexual species derive from sexual outcrossing ancestors, from which they inherit their load of TEs. TE distribution eventually depends on the balance between additional transposition within selfing/clonal lineages on one hand and selection or excision on the other. Following the abandonment of sex, large asexual populations are expected to purge their load of TEs, provided excision occurs, even at very low rates. However, purging can take a very long time, and, without excision, TEs should slowly accumulate, not decline [126]. In small populations, even with excision, a Muller's ratchet-like process drives TE accumulation throughout the genome [126]. Transition from outcrossing to selfing should also rapidly purge TEs, but as for asexuals, in small fully selfing populations, TEs can be retained [127]. Using yeast populations, it was experimentally confirmed that sex increases the spread of TEs [128, 129]. TE numbers were also found to be higher in cyclically sexual than in fully asexual populations of *Daphnia pulex* [130–132] (Table 3), contrary to what was described in the parasitoid wasp *Leptopilina clavipes* and in root knot nematode species (Table 3). It should be noted that several comparative studies on asexual arthropods, nematodes, primroses, and green algae did not evidence any significant effect of breeding system on TE content or evolution (Table 3). At larger evolutionary scales, the putatively ancient asexual bdelloid rotifers strikingly exemplify the fact that

Table 3
Summary of studies comparing transposable element distribution and dynamics between different breeding systems

	Taxonomic group	Groups compared	Age of selfing/ asexuality	TE types	Effect of breeding system	References
Outcrossing/ selfing	*Arabidopsis*	*A. thaliana* (selfer)/ *A. lyrata* (outcrosser)	Recent (0.5–1 M years)	Ac-like DNA TE		
	Arabidopsis	*A. thaliana* (selfer)/ *A. lyrata* (outcrosser)	Recent (0.5–1 M years)	DNA TE, LTR and non-LTR RNA TE (no full genome reference)	No difference in insertion number, purifying selection lower in the selfing species	Lockton and Gaut [140]
	Arabidopsis	*A. thaliana* (selfer)/ *A. lyrata* (outcrosser)	Recent (0.5–1 M years)	DNA TE, LTR and non-LTR RNA TE (full genome reference)	Three times more copies and ten times more specific families in the outcrosser. Recent decrease in TE in number in the selfer	de la Chaux et al. [141]
	Capsella	*C. orientalis* and *C. rubella* (selfers)/ *C. grandiflora* (outcrosser)	Recent (~<1 M years) for *C. orientalis*, very recent (~20,000 years) for *C. rubella*	DNA TE, LTR and non-LTR RNA TE	Slight insertion increase in the recent selfer, strong decrease in the older selfer	Agren et al. [142]
	Solanum	SI and SC species of the *S. lycopersicum* section	Recent (~<1 M years)	*copia*-type RNA TE (ToRTL1, T135, Tnt1)	No effect of MS on TE insertions. Reduced TE sequence diversity in SC lineages. Compatible with a nearly neutral model	Tam et al. [138]
	Caenorhabditis	*C. elegans* (selfer)/ *C. remanei* (outcrosser)	Rather recent (<~4 M years)	Tc1 DNA TE	Purifying selection against TEs less efficient in the selfing species	Dolgin et al. [229]

Sexuals/asexuals	Four asexual angiosperm species	Comparison with sexual plants	Uncertain, maybe between 1 and 10 M years	Ty1/*copia*, Ty3/*gypsy*, and LINE-like RNA TE	Presence of conserved TE in asexuals	Docking et al. [228]
	Oenothera	17 asexual/13 sexual lineages	Unknown	DNA TE, LTR and non-LTR RNA TE	No significant effect	Agren et al. [160]
	Chlamydomonas reinhardtii	Asexual experimental lines	800 asexual generations/100 asexual generations vs. 11 sex events	Two DNA TE (TOC1, *Gulliver*)	No significant effect	Zeyl et al. [235]
	Saccharomyces cerevisiae	Sexual and asexual experimental lines with TE at initial frequency 1%	200–300 asexual generations/8 sex events	Ty3 RNA TE	Higher increase in TE frequency in sexual lines	Zeyl et al. [128]
	Candida albicans	Asexual species, compared with *S. cerevisiae*	Unknown. Rare sex events	LTR RNA TE	More TE families but most of them inactive and lower copy number than in *S. cerevisiae*	Goodwin and Poulter [230]
	Arthropods	Five pairs of asexual/sexual lineages	From very recent (~22 yrs., 10,000–40,000 generations) to old (~10 Myrs)	DNA TE, LTR and non-LTR RNA TE	No difference in any of the five pairs	Bast et al. [226]
	Bdelloid rotifers	Comparison with many other sexual metazoan	Old	LINE-like and *gypsy*-like RNA TE, *Mariner*/TC1-like DNA TE	Absence of RNA TE in asexuals	Arkhipova and Meselson [133]
	Daphnia pulex	Different isolates of the same species	Recent (<200,000 years)	One DNA TE (*Pokey*)	Lower TE insertion in asexuals	Sullender and Crease [130], Valizadeh and Crease [131]
	Daphnia pulex	20 asexuals/20 sexuals isolates	Recent (<200,000 years)	DNA TE, LTR and non-LTR RNA TE	Lower TE insertion but more fixed ones in asexuals. Substantial fraction of TE in asexuals inherited directly from sexuals	Jiang et al. [231]

(continued)

Table 3
(continued)

Taxonomic group	Groups compared	Age of selfing/ asexuality	TE types	Effect of breeding system	References
Daphnia pulex	Asexual/sexual mutation-accumulation experimental lines	40 asexual generations/at least one sex event	6 DNA TE families	Higher rate of DNA TE loss in cyclical than in obligate parthenogenous lineages	Schaack et al. [233]
Root-knot nematodes	3 obligate asexuals/1 facultative asexual	Uncertain, maybe between 17 and 40 Myrs	DNA TE, LTR and non-LTR RNA TE	Higher TE content in asexuals	Blanc-Mathieu et al. [157]
Nematodes	42 species (dioecy, androdioecy, facultative parthenogenesis, strict apomixis)	Uncertain, maybe between 17 and 40 M years	DNA TE, LTR and non-LTR RNA TE	No significant effect of breeding system	Szitenberg et al. [234]
Leptopilina clavipes	1 sexual/1 asexual (*Wolbachia*-induced) lineages	Recent (<12,000–43,000 generations)	DNA TE, LTR and non-LTR RNA TE	Proliferation of DNA TE and *gypsy*-like RNA TE in asexual lineages	Kraaijeveld et al. [232]

asexuals can purge their load of TEs. Unlike all sexual eukaryotes, they appear to be free of vertically transmitted retrotransposon, while their genome contains DNA transposons, probably acquired via horizontal transfers [133, 134]. Examples of TE accumulation in asexuals are less common, maybe because species are doomed to extinction under this evolutionary scenario [135]. However, the increase in genome size in some apomictic lineages of *Hypericum* species may result from this process [136].

In selfers, the distribution of TEs depends not only on the population size but also on the mode of selection against TEs [127, 137]. Under the "deleterious" model, TE insertions are selected against because they disrupt gene functions. According to the "ectopic exchange" model, TEs are selected against because they generate chromosomal rearrangements through unequal crossing-over between TE at nonhomologous insertion sites. Under the first of these two models, homozygosity resulting from selfing increases the selection efficacy against TEs, while under the second one, under-dominant chromosomal rearrangements are less selected against in selfing than in outcrossing populations [127, 137]. A survey of Ty1-copia-like elements in plants suggests that they are less abundant in self-fertilizing than in outcrossing plants, thus supporting the "deleterious" rather than the "ectopic" exchange model [127]. The distribution of retrotransposons in self-incompatible and self-compatible *Solanum* species also supports the "deleterious" model, even though most insertions are probably neutral [138] (Table 3). In the selfer *Arabidopsis thaliana*, selection efficacy against TEs seems to be reduced compared to its outcrossing sister species *A. lyrata* [139, 140], but comparison of the two complete genomes revealed a higher load of TE in *A. lyrata* and a recent decrease in TE in number in *A. thaliana*, in agreement with the date of transition to selfing [141]. In the *Capsella* genus, while the very recent selfer *C. rubella* possesses a slightly higher number of TEs than the outcrossing *C. grandiflora*, the oldest selfer *C. orientalis* exhibits a significantly reduced load of TE [142] (Table 3). Other selfish elements, such as B chromosomes, are also less frequent in selfers, in support of the view that inbreeding generally prevents selfish element transmission [102].

2.4 Breeding Systems, Ploidy, and Hybridization

Atypical breeding systems are often associated with polyploidy [143], and the reasons for this association are not entirely clear. Polyploid mutants might be more likely to establish as new lineages in selfers and asexuals than in obligate outcrossers if crosses between polyploids and diploids are unfertile or counterselected. This is because at low population frequency a polyploid mutant will experience the disadvantage of mostly mating with diploids—the minority cytotype exclusion principle [144, 145]—unless it reproduces asexually or via selfing. In addition, by doubling gene copy number, polyploidy might alleviate the fitness cost of recessive

deleterious mutations being exposed at homozygous state in selfers [146]. Kreiner et al. [147] reported that in Brassicaceae the rate of production of unreduced gametes is higher in asexuals than in outcrossers, suggesting that mating systems can influence not only the establishment rate but also the mutation rate to polyploidy.

Recent genome-wide data analyses have revealed that a number of polyploid selfers or asexuals actually correspond to allopolyploids (e.g., [59, 148–151]), highlighting the possibility that hybridization plays a role in breeding system and ploidy evolution. Hybridization between facultative asexuals might cause immediate transition to obligate asexuality if the two progenitor genomes are so divergent that meiosis is impaired—e.g., due to chromosomal rearrangements, or in case of genetic incompatibilities affecting genes involved in sexual reproduction [16]. Numerous selfing or asexual lineages, either diploid or polyploid, are known to be of hybrid origin (e.g., [13, 152–157]). Hybridization would therefore appear as a potential cause, and polyploidy a potential consequence, of atypical breeding systems [16], but more genome-wide data are obviously needed to draw firm conclusions on these complex relationships.

2.5 Breeding Systems and Genome Size Evolution

As argued above, breeding systems can affect many aspects of genome content and organization. They should also affect the whole genome size. Following Lynch's theory [1], genome size should be higher in selfers and asexuals because of their reduced effective population size, hence reduced ability to get rid of useless, slightly costly sequences. However, the picture is probably more complex. First, because of the recent origin of many selfing and (at least some) asexual lineages, relaxed selection may not have operated longly enough to impact genome size. Second, because of their immunity to selfish element transmission, selfers and asexuals should exhibit lower genome size, especially in groups where TEs are major determinants of genome size. Hence, it is not clear whether genetic drift or resistance to selfish elements (or other processes) is the most important in governing genome size evolution in various breeding systems.

Meta-analyses performed in plants provided equivocal answers. Analysis of the distribution of B chromosomes showed a strong and significant positive association between outcrossing, the occurrence of B chromosomes, and genome size [102, 158]. However, after phylogenetic control, only the association between breeding systems and B chromosomes remains. Whitney et al. [159] simultaneously tested the effect of breeding systems (using outcrossing rate estimates) and genetic drift (using polymorphism data) on genome size in seed plants. Raw data showed a significant effect of both breeding systems and genetic drift, according to theoretical predictions. However, no effect was observed after phylogenetic control, leading the authors to reconsider the hypothesis of a role

of nonadaptive processes in genome size evolution. Similarly, phylogenetic comparative analysis of 30 primrose species (*Oenothera*) covering several transitions to asexuality showed no significant relationship between reproductive mode and genome size [160].

Because breeding systems can evolve quickly, more detailed analyses at a short phylogenetic scale are needed to get a clearer picture of their effects on genome size evolution. Moreover, breeding systems are often correlated with other life history traits, such as lifespan, which can make it hard to clarify the causes and consequences of the observed correlations. A detailed analysis of genome size in the *Veronica* genus suggests that selfing, not annuality, is associated with genome size reduction [161]. A comparison of 14 pairs of plant congeneric species with contrasted mating systems also suggested a genome size reduction in selfers [162]. However, this could partly have been due to the four polyploid selfing species of the dataset—polyploidy can lead to haploid genome size reduction because of the loss of redundant DNA following polyploidization. A better understanding can be gained from the comparative analysis of genome composition and organization, not only genome size. In *Caenorhabditis* nematodes, the observed reduction in genome size is not driven by reduction in TEs but by a global loss of all genomic compartments [163]. This pattern contradicts the hypothesis of relaxed selection in selfers against the accumulation of deleterious genomic elements. Alternatively, it could be explained by deletion bias and high genetic drift in selfers. However, in mutation accumulation lines, insertions predominate over deletion in the selfing *C. elegans*, and deletions occurred at the whole gene level instead of being at random among genomic compartments, as predicted under a general deletion bias (*see* discussion in ref. 163). In this genus, Lynch's hypothesis that evolution of genome size should be driven by changes in N_e does not apply. Alternatively, the authors suggested that it is a more direct consequence or even an adaptation to the selfing lifestyle, although the underlying mechanisms still remain unclear.

3 A Genomic View of Breeding System Evolution

Because breeding systems can strongly affect genome structure and evolution, conversely, genomic approaches offer new powerful tools to reconstruct breeding system evolution and to test evolutionary hypotheses, especially concerning long-term evolution.

3.1 Genomic Approaches to Infer Breeding System Evolution

3.1.1 Genomic Characterization of Breeding Systems

Genetic markers have long been used to determine breeding systems and quantify selfing rates or degrees of asexuality. For instance, current selfing rates can be inferred using molecular markers through F_{IS} estimates or preferably—although more time consuming—through progeny analyses [164–166]. Multilocus-based estimates that take identity disequilibrium into account greatly improve the simple F_{IS}-based method that is sensitive to several artifacts such as null alleles ([167], *see* also refs. 168, 169). This method, implemented in the RMES software [167], has proven to give results very similar to progeny-based methods [170]. To take advantage of the information potentially available in sequence data, coalescence-based estimators have also been proposed to infer long-term selfing rates, and they have been implemented more recently in a Bayesian clustering approach in the INSTRUCT software package [171]. However, this approach mostly captures information from recent coalescence events so that such approaches still estimate recent selfing rates [28]. Much more information about long-term selfing rates can be derived from LD patterns [19], but this has not been fully exploited for selfing rate estimators (for instance, LD is not taken into account in INSTRUCT). Similarly, recombination can be inferred using genetic markers or sequence data, and more generally, various methods have been proposed to characterize the degree of clonality in natural populations (for review *see* ref. 172) and recently implemented in the R package RClone [173].

Initially, such methods were applied with few markers, from which only global descriptions of breeding systems were deducible. Thanks to the considerable increase in sequencing facilities, it has become possible to finely characterize temporal and spatial variations in breeding systems. In *A. thaliana*, an analysis of more than 1000 individuals in 77 local stands using more than 400 SNP markers revealed spatial heterogeneity in outcrossing rates. Local "hotspots" of recent outcrossing (up to 15%) were identified, while other stands exhibited complete homozygosity with no detectable outcrossing [174]. Interestingly, at this local scale (from 30 m to 40 km), outcrossing rates have been found to be twofold higher on average in rural than in urban stands; hence, selfing could be associated with higher disturbance in urban stands.

Genomic data may also help characterize breeding systems in species with unknown or ill-characterized life cycles. In yeasts *Saccharomyces cerevisiae* and *S. paradoxus*, the analyses of linkage disequilibrium patterns allowed to quantify the frequency of (rare) sexual reproduction events and the proportion of inbreeding and outcrossing during these events [175, 176]. For instance, in the pico-algae *Ostreococcus*, no sexual form or process has been detected in the lab. However, the occurrence of infrequent recombination (about 1 meiosis for 10 mitoses) inferred from a population genomics approach and the presence of meiosis genes in the genome

support the existence of a sexual life cycle [177]. Moreover, a strong negative correlation between chromosome size and GC-content has been observed [178]. In mammals and birds (among others), such a pattern has been interpreted as a long-term effect of gBGC acting on chromosomes with different average recombination rates [116]—small chromosomes having higher recombination rates because of the constraint of at least one chiasmata per chromosome arm. A similar interpretation for *Ostreococcus* is thus appealing. Genomic data also allow to test whether the theoretical signatures of long-term asexuality are observed in putative asexuals. As an example, whole-genome analyses of the trypanosome *T. b. gambiense* demonstrated an independent evolution and divergence of alleles on each homologous chromosome (the "Meselson effect" [179, 180]), which is indicative of strict asexual evolution [88]. In contrast, genomic studies of the putatively ancient asexual bdelloids recently uncovered the occurrence of inter-individual genetic exchanges ([181, 182] *see* below Subheading 3.2.2).

3.1.2 Inferring and Dating Breeding System Transitions

Genomic approaches are also useful for analyzing the dynamics of breeding system evolution. A simple way is to map breeding system evolution on phylogenies, which could provide a raw picture of the frequency and relative timing of breeding system transitions (e.g., [183]). However, these approaches, based on ancestral character reconstruction, are hampered by numerous uncertainties. For instance, in the case of two sister species with contrasting breeding systems, such as *A. thaliana* and *A. lyrata*, it is impossible to know whether *A. thaliana* evolved toward selfing just after divergence (about five million years ago) or only very recently. At a larger phylogenetic scale, inferring rates of transition between characters and ancestral states can be biased if diversification rates differ between characters—this is typically expected with breeding systems for which asexuals and selfers should exhibit higher extinction rates than outcrossers [184].

Thanks to the genomic signatures left by contrasted breeding systems, it is possible to trace back transitions in the past and to date them more precisely. In diploid asexual species, because of the arrest of recombination, the two copies of each gene have diverged independently since the origin of asexuality. After having calibrated the molecular clock, it is thus possible to date this origin from the level of sequence divergence between the two copies. This so-called Meselson effect was observed and quantified in the trypanosome *T. b. gambiense*, suggesting that this species evolved asexually about 10,000 years ago [88]. However, no Meselson effect has been observed in other presumably ancient asexual species such as oribatid mites [185] or darwinulid ostracods [186], while data refute the possibility of cryptic sex. In such cases, it is thus not possible to infer when recombination actually stopped, presumably because of

homogenizing processes such as very efficient DNA repair or automixis. Mitotic recombination could also obscure the pattern predicted under this Meselson effect. Of note, when asexuality originates by hybridization (*see* above Subheading 2.4), the last common ancestor of the two copies of a gene dates back to the ancestor of the two parental lineages, which can be much older than the hybridization date, faulting the above-described rationale.

Past transitions from outcrossing to selfing have also been investigated, through either population genomics approaches or the evolutionary analysis of self-incompatibility (SI) genes, which are directly involved in the transition to selfing. Since the evolution of selfing requires the breakdown of SI systems, initially constrained S-locus genes are expected to evolve neutrally after a shift to selfing. In *A. thaliana*, Bechsgaard et al. [187] reasoned that the dN/dS ratio in the selfing lineage should be the average of the neutral dN/dS (i.e., 1) and the outcrossing dN/dS—inferred from sister lineages—weighted by the time spent in the selfing vs. the outcrossing state. They deduced that SRK, one of the major SI genes, became a pseudogene less than 400,000 years ago. SRK, however, is not the only gene involved in SI. Mutations in other genes may have previously disrupted the SI system, thus confusing SRK-based dating. Indeed, coalescence simulations showed that the observed genome-wide pattern of linkage disequilibrium is compatible with the transition to selfing one million years ago [188], suggesting a possible but debated two-step scenario in the evolution of selfing [189, 190]. The persistence of three distinct divergent SRK haplotypes among extant *A. thaliana* individuals also suggests multiple loss of SI [191], but the recent discovery of the co-occurrence of the three haplotypes in Moroccan populations makes possible the evolution of selfing in a single geographic region [192]. In another Brassicaceae, i.e., *Capsella rubella*, analyses of both S-locus and genome-wide genes coupled with coalescence simulations suggested that selfing evolved very recently from the outcrosser *C. grandiflora*, around 50,000 years ago [193, 194] from a potentially large number of founding individuals followed by a strong reduction in N_e [195]. In the tetraploid selfer *Arabidopsis suecica*, which originated as a hybrid between *A. thaliana* and the outcrossing *A. arenosa*, the genomic analysis of the S-locus also revealed the origin of selfing, suggesting an instantaneous loss of SI due to the fixation of nonfunctional alleles from both parents around 16,000 years ago [150].

3.2 Matching Breeding System Evolution Theories with Genomic Data

3.2.1 Testing the Dead-End Hypothesis: Comparison Between Selfing and Asexuality

The expected reduction in N_e in selfers and asexuals may increase the drift load (accumulation of slightly deleterious mutations) and preclude adaptation. Selfing and clonality are thus supposed to be evolutionary dead ends [17, 18]. The twiggy phylogenetic distributions of asexuals [196] and selfers [183] or self-compatible species [197] suggest they are mostly derived recently from outcrossing ancestors (but *see* ref. 198). However, this observation may not be sufficient to support the dead-end hypothesis, and neutral models can also explain this pattern [199–201]. In a comprehensive and epochal phylogenetic study of several Solanaceae genera, Goldberg et al. [202] went further by testing the irreversibility of transitions. Using a phylogenetic method developed for estimating the character effect on speciation and extinction [203, 204], they showed that self-compatible species have both higher speciation and extinction rates—with the resulting net diversification rates being lower—than self-incompatible species. This was the first direct demonstration of the dead-end hypothesis, and additional results have been obtained in *Primula* species [205]. On the contrary, in the *Oenothera* genus, asexuality has been found associated with increased diversification but frequent reversion toward the sexual system, suggesting that the form of asexuality in this group is not an evolutionary dead end [206].

Genomic data also provide an opportunity to investigate the genetic causes of such long-term evolutionary failures. The increased dN/dS ratios reported in asexuals (see above) suggest that deleterious point mutations contribute to the load. However, in *Daphnia* rapid exposure of recessive deleterious alleles through mitotic recombination or gene conversion likely has a much stronger effect on clone persistence than their long-term accumulation under Muller's ratchet [60]. TE could also contribute to the load and to the extinction of asexuals [135], though more data are still needed to unambiguously support this hypothesis (but *see* ref. 136). The pattern in selfers is less clear. While theory globally predicts a reduction in selection efficacy in selfers, models also highlight conditions under which selection can be little affected or even enhanced in selfers [72, 73, 207], especially regarding TE accumulation [127, 137]. Empirical data on both protein and TE evolution have not revealed any strong evidence of long-term accumulation of deleterious mutation in selfers, as compared to outcrossers, whereas polymorphism data mainly support relaxation of selection in selfers (Table 2). This is in agreement with the recent origin of selfing but makes difficult further inference of the underlying causes of higher extinction in selfers as trait-dependent diversification processes alter the relationship between life history traits and rate of molecular evolution [208]. A reduced ability to respond to environmental changes through adaptive evolution could also contribute to long-term extinction in asexuals (but *see* ref. 209) and selfers, especially if standing variation is needed to rescue

populations experiencing environmental challenges [77, 210]. Few studies, however, have compared the rate of adaptation in selfers and outcrossers (*see* Table 2). Theoretical predictions regarding this effect, moreover, critically depend on the dominance level of new favorable mutations [72, 73, 77, 210], which are poorly known (but *see* ref. 80).

While several issues remain open, current knowledge suggests that selfers are less prone to extinction than asexuals. The wider distribution of selfing than clonality in plants supports this view [211, 212]. Selfers could go toward extinction more slowly than asexuals, and the causes of their extinction could differ. Since deleterious mutations should accumulate at a slower rate in selfers than in asexuals, as suggested by theory and current data, this process would likely not be sufficient to drive them to extinction. The reduced adaptive potential could be the very cause of their ultimate extinction as initially proposed by Stebbins [18], which could generally occur before sufficient deleterious mutations have accumulated to be detected via molecular measures of divergence. On the contrary, in asexuals, the accumulation of deleterious mutations could be fast enough to leave a molecular signature and contribute to extinction. Alternatively, demographic characteristics associated with uniparental reproduction, such as recurrent bottlenecks, fragmented populations, and extinction/recolonization dynamics, could be sufficient to drive population extension simply because of higher sensitivity to demographic stochasticity (*see* also ref. 213). Genomic degradation would only be the witness of the evolution toward selfing and clonality without being the ultimate cause of their extinctions. These hypotheses need to be further investigated by building more realistic demo-genetic model and by better integrating genomic and ecological approaches.

The literature reviewed above focuses on intrinsic factors that may affect the extinction rate of selfing and asexual species, taken as established lineages, compared to their sexual relatives. Alternatively, Janko et al. [199] suggested that if asexual mutants are produced at a relatively high rate and compete with each other, this would imply a rapid turnover between clonal lineages and a young expected age for extant asexuals, without the need to invoke any fitness effect (*see* also refs. 200, 201). Of note, this model invokes competitive exclusion among clonal lineages, but not between clonal and sexual ones—the ancestral sexual gene pool is assumed to be immune from extinction.

3.2.2 Evading the "Dead End"

The few putatively ancient asexuals known so far seem to escape the mutational load predicted by the dead-end hypothesis and avoid extinction over long evolutionary time scales. For example, fossil evidence and decades of microscopic observations indicate that bdelloid rotifers have apparently persisted for over 40 million years without meiosis, males, or conventional sexual reproduction

[15, 214]. As a matter of fact, the first genome assembly published for these organisms confirmed that their genome structure is incompatible with conventional meiosis [215]. However, two independent studies recently demonstrated that bdelloids could experience genetic exchanges between individuals.

A first article by Debortoli et al. [182] evidenced frequent horizontal exchanges of genetic fragments between individuals of the species *Adineta vaga* (Adinetidae). Such horizontal transfers could be promoted by the peculiar ecology of these rotifers, which experience frequent desiccations damaging their cell and nucleus membranes and thus allowing for the entry of foreign DNA in the cells. In addition, desiccation induces multiple DNA double-strand breaks, facilitating the integration of foreign DNA during repair processes.

Another study by Signorovitch et al. [181] identified a pattern of allele sharing between individuals of the species *Macrotrachela quadricornifera* (Philodinidae) that was incompatible with strict asexual evolution. The authors suggested that bdelloids had evolved an atypical meiotic mechanism similar to what has been described in some species of primroses (*Oenothera*), in which chromosomes organize into a ring during meiosis without requiring homologous chromosome pairing [216]. They advocated that even rare events of such unconventional sex could be enough to generate the observed pattern of allele sharing.

In the absence of conventional meiosis and syngamy, bdelloid rotifers might thus have escaped extinction by maintaining some level of genetic exchanges between individuals, either through horizontal gene transfers or unconventional *Oenothera*-like meiosis. Regardless of the underlying molecular mechanisms, bdelloids should not be considered as "ancient asexual scandals" anymore. These recent results call for a reassessment of the reproductive mode of all supposedly ancient asexuals (*see* Subheading 3.1.1 above). The rise of genomic studies in recent years will greatly contribute to decipher whether putative asexuals evolve as strict asexuals or have developed new alternatives to sex.

4 Conclusion and Prospects

There is a large body of theory on the effects of breeding systems on molecular evolution. However, some of them have not been clearly verified by empirical data, and numerous questions remain. Genomic data have also partly unveiled the complexity of breeding systems, especially in asexual or presumably asexual species. Promising prospects include (1) analysis of the rate and pattern of transition to selfing/asexuality using densely sampled phylogenies with appropriate breeding system distributions combined with

genome-wide molecular data, (2) distinguishing between the different forms of selection with a better characterization of the fitness effect of mutations, (3) explicitly accounting for the possible association between breeding system shifts and non-equilibrium demographic dynamics (e.g., bottlenecks in selfers, clone turnover in asexuals). A large theoretical corpus has already been developed, and thanks to the increasing availability of genomic data, qualitative patterns are now rather well described and partly understood. Another challenge in the future is also to make our predictions and tests more quantitative.

5 Questions

1. What population genetic parameters are affected, and how, by selfing and asexuality?
2. What are the potential problems when comparing the dN/dS ratio between selfers and outcrossers or sexuals and asexuals?
3. What is the evolutionary "dead-end hypothesis," and how can we test it using phylogenetic and evolutionary genomic tools?

Acknowledgments

This work was supported by ARCAD, a flagship project of Agropolis Fondation, by an ERC grant (PopPhyl) to N.G. and by the CoGeBi program (grant number ANR-08-GENM-036-01) and a Swiss National Research Found SINERGIA grant.

References

1. Lynch M (2007) The origin of genome architecture, 1st edn. Sinauer, Sunderland, MA
2. Smith SA, Donoghue MJ (2008) Rates of molecular evolution are linked to life history in flowering plants. Science 322(5898):86–89
3. Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, Loire E, Lourenco JM, Nabholz B, Roux C, Tsagkogeorga G, Weber AA, Weinert LA, Belkhir K, Bierne N, Glemin S, Galtier N (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515(7526):261–263
4. Bromham L, Hua X, Lanfear R, Cowman PF (2015) Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. Am Nat 185(4):507–524
5. Figuet E, Nabholz B, Bonneau M, Mas Carrio E, Nadachowska-Brzyska K, Ellegren H, Galtier N (2016) Life history traits, protein evolution, and the nearly neutral theory in amniotes. Mol Biol Evol 33 (6):1517–1527
6. Chen J, Glemin S, Lascoux M (2017) Genetic diversity and the efficacy of purifying selection across plant and animal species. Mol Biol Evol 34(6):1417–1428
7. Lefebure T, Morvan C, Malard F, Francois C, Konecny-Dupre L, Gueguen L, Weiss-Gayet-M, Seguin-Orlando A, Ermini L, Sarkissian C, Charrier NP, Eme D, Mermillod-Blondin F, Duret L, Vieira C, Orlando L, Douady CJ (2017) Less effective selection leads to larger genomes. Genome Res 27(6):1016–1028
8. Jarne P, Auld JR (2006) Animals mix it up too: the distribution of self-fertilization

among hermaphroditic animals. Evolution 60 (9):1816–1824
9. Vogler DW, Kaliz S (2001) Sex among the flowers: the distribution of plant mating systems. Evolution 55(1):202–204
10. Haldane JBS (1932) The causes of evolution, vol 1, 1st edn. Princeton University Press, Princeton
11. Hedrick PW (1987) Population genetics of intragametophytic selfing. Evolution 41 (1):137–144
12. Balloux F, Lehmann L, de Meeus T (2003) The population genetics of clonal and partially clonal diploids. Genetics 164(4):1635–1644
13. Simon JC, Delmotte F, Rispe C, Crease TJ (2003) Phylogenetic relationships between parthenogens and their sexual relatives: the possible routes to parthenogenesis in animals. Biol J Lin Soc 79:151–163
14. Whitton J, Sears CJ, Baack EJ, Otto SP (2008) The dynamic nature of apomixis in the angiosperms. Int J Plant Sci 169 (1):169–182
15. Schurko AM, Neiman M, Logsdon JM (2009) Signs of sex: what we know and how we know it. Trends Ecol Evol 24(4):208–217
16. Neiman M, Sharbel TF, Schwander T (2014) Genetic causes of transitions from sexual reproduction to asexuality in plants and animals. J Evol Biol 27(7):1346–1359
17. Maynard-Smith J (1978) The evolution of sex. Cambridge University Press, Cambridge
18. Stebbins GL (1957) Self fertilization and population variability in higher plants. Am Nat 91:337–354
19. Nordborg M (2000) Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics 154(2):923–929
20. Padhukasahasram B, Marjoram P, Wall JD, Bustamante CD, Nordborg M (2008) Exploring population genetic models with recombination using efficient forward-time simulations. Genetics 178(4):2417–2427
21. Hartfield M, Glemin S (2016) Limits to adaptation in partially selfing species. Genetics 203 (2):959–974
22. Roze D (2016) Background selection in partially selfing populations. Genetics 203 (2):937–957
23. Flint-Garcia SA, Thornsberry JM, Buckler ES IV (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374
24. Glémin S, Bazin E, Charlesworth D (2006) Impact of mating systems on patterns of sequence polymorphism in flowering plants. Proc R Soc Lond B Biol Sci 273 (1604):3011–3019
25. Golding GB, Strobeck C (1980) Linkage disequilibrium in a finite population that is partially selfing. Genetics 94(3):777–789
26. Roze D (2015) Effects of interference between selected loci on the mutation load, inbreeding depression, and heterosis. Genetics 201(2):745–757
27. Pollak E (1987) On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics 117(2):353–360
28. Nordborg M, Donnelly P (1997) The coalescent process with selfing. Genetics 146 (3):1185–1195
29. Ceplitis A (2003) Coalescence times and the Meselson effect in asexual eukaryotes. Genet Res 82(3):183–190
30. Hartfield M, Wright SI, Agrawal AF (2016) Coalescent times and patterns of genetic diversity in species with facultative sex: effects of gene conversion, population structure, and heterogeneity. Genetics 202(1):297–312
31. Haag CR, Roze D (2007) Genetic load in sexual and asexual diploids: segregation, dominance and genetic drift. Genetics 176 (3):1663–1678
32. Schoen DJ, Brown AHD (1991) Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. Proc Natl Acad Sci U S A 88:4494–4497
33. Haag CR, Ebert D (2004) A new hypothesis to explain geographic parthenogenesis. Ann Zool Fennici 41:539–544
34. Ingvarsson PK (2002) A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. Evolution 56(12):2368–2373
35. Gordo I, Charlesworth B (2001) Genetic linkage and molecular evolution. Curr Biol 11(17):R684–R686
36. Agrawal AF, Hartfield M (2016) Coalescence with background and balancing selection in systems with bi- and uniparental reproduction: contrasting partial asexuality and selfing. Genetics 202(1):313–326
37. Thomas CG, Wang W, Jovelin R, Ghosh R, Lomasko T, Trinh Q, Kruglyak L, Stein LD, Cutter AD (2015) Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis. Genome Res 25(5):667–678
38. Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, Felix MA, Kruglyak L (2012) Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. Nat Genet 44(3):285–290

39. Coop G (2016) Does linked selection explain the narrow range of genetic diversity across species? bioArxiv. https://doi.org/10.1101/042598
40. Hamrick JL, Godt MJW (1996) Effects of life history traits on genetic diversity in plant species. Philos Trans R Soc Lond B 351 (1345):1291–1298
41. Nybom H (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Mol Ecol 13 (5):1143–1155
42. Fontcuberta Garcia-Cuenca A, Dumas Z, Schwander T (2016) Extreme genetic diversity in asexual grass thrips populations. J Evol Biol 29(5):887–899
43. Normark BB, Judson OP, Moran NA (2003) Genomic signatures of ancient asexual lineages. Biol J Lin Soc 79:69–84
44. Lercher MJ, Hurst LD (2002) Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet 18(7):337–340
45. Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet 72 (6):1527–1535
46. Longman-Jacobsen N, Williamson JF, Dawkins RL, Gaudieri S (2003) In polymorphic genomic regions indels cluster with nucleotide polymorphism: quantum genomics. Gene 312:257–261
47. Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen JQ (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455(7209):105–108
48. Hollister JD, Ross-Ibarra J, Gaut BS (2010) Indel-associated mutation rate varies with mating system in flowering plants. Mol Biol Evol 27(2):409–416
49. Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47:713–719
50. Hill WG, Robertson AW (1966) The effect of genetic linkage on the limits to artificial selection. Genet Res 8:269–294
51. Paland S, Lynch M (2006) Transitions to asexuality result in excess amino acid substitutions. Science 311(5763):990–992
52. Henry L, Schwander T, Crespi BJ (2012) Deleterious mutation accumulation in asexual Timema stick insects. Mol Biol Evol 29 (1):401–408
53. Johnson SG, Howard RS (2007) Contrasting patterns of synonymous and nonsynonymous sequence evolution in asexual and sexual freshwater snail lineages. Evolution 61 (11):2728–2735
54. Neiman M, Hehman G, Miller JT, Logsdon JM Jr, Taylor DR (2010) Accelerated mutation accumulation in asexual lineages of a freshwater snail. Mol Biol Evol 27 (4):954–963
55. Lovell JT, Williamson RJ, Wright SI, McKay JK, Sharbel TF (2017) Mutation accumulation in an asexual relative of Arabidopsis. PLoS Genet 13(1):e1006550
56. Ollivier M, Gabaldon T, Poulain J, Gavory F, Leterme N, Gauthier JP, Legeai F, Tagu D, Simon JC, Rispe C (2012) Comparison of gene repertoires and patterns of evolutionary rates in eight aphid species that differ by reproductive mode. Genome Biol Evol 4 (2):155–167
57. Pellino M, Hojsgaard D, Schmutzer T, Scholz U, Horandl E, Vogel H, Sharbel TF (2013) Asexual genome evolution in the apomictic Ranunculus auricomus complex: examining the effects of hybridization and mutation accumulation. Mol Ecol 22 (23):5908–5921
58. Hollister JD, Greiner S, Wang W, Wang J, Zhang Y, Wong GK, Wright SI, Johnson MT (2015) Recurrent loss of sex is associated with accumulation of deleterious mutations in Oenothera. Mol Biol Evol 32(4):896–905
59. Ament-Velasquez SL, Figuet E, Ballenghien M, Zattara EE, Norenburg JL, Fernandez-Alvarez FA, Bierne J, Bierne N, Galtier N (2016) Population genomics of sexual and asexual lineages in fissiparous ribbon worms (Lineus, Nemertea): hybridization, polyploidy and the Meselson effect. Mol Ecol 25(14):3356–3369
60. Tucker AE, Ackerman MS, Eads BD, Xu S, Lynch M (2013) Population-genomic insights into the evolutionary origin and fate of obligately asexual Daphnia pulex. Proc Natl Acad Sci U S A 110(39):15740–15745
61. Wright SI, Lauga B, Charlesworth D (2002) Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol Biol Evol 19(9):1407–1420
62. Cutter AD, Wasmuth JD, Washington NL (2008) Patterns of molecular evolution in *Caenorhabditis* preclude ancient origins of selfing. Genetics 178(4):2093–2104
63. Escobar JS, Cenci A, Bolognini J, Haudry A, Laurent S, David J, Glémin S (2010) An integrative test of the dead-end hypothesis of

selfing evolution in Triticeae (poaceae). Evolution 64(10):2855–2872

64. Glémin S, Muyle A (2014) Mating systems and selection efficacy: a test using chloroplastic sequence data in Angiosperms. J Evol Biol 27(7):1386–1399
65. Arunkumar R, Ness RW, Wright SI, Barrett SC (2015) The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. Genetics 199(3):817–829
66. Burgarella C, Gayral P, Ballenghien M, Bernard A, David P, Jarne P, Correa A, Hurtrez-Bousses S, Escobar J, Galtier N, Glémin S (2015) Molecular evolution of freshwater snails with contrasting mating systems. Mol Biol Evol 32(9):2403–2416
67. Charlesworth D, Morgan MT, Charlesworth B (1993) Mutation accumulation in finite outbreeding and inbreeding populations. Genet Res 61:39–56
68. Hartfield M, Glemin S (2014) Hitchhiking of deleterious alleles and the cost of adaptation in partially selfing species. Genetics 196 (1):281–293
69. Hartfield M, Otto SP (2011) Recombination and hitchhiking of deleterious alleles. Evolution 65(9):2421–2434
70. Bullaughey K, Przeworski M, Coop G (2008) No effect of recombination on the efficacy of natural selection in primates. Genome Res 18 (4):544–554
71. Haddrill PR, Halligan DL, Tomaras D, Charlesworth B (2007) Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. Genome Biol 8(2):R18
72. Charlesworth B (1992) Evolutionary rates in partially self-fertilizing species. Am Nat 140 (1):126–148
73. Glémin S (2007) Mating systems and the efficacy of selection at the molecular level. Genetics 177(2):905–916
74. Charlesworth B, Charlesworth D (1997) Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genet Res 70 (1):63–73
75. Szovenyi P, Devos N, Weston DJ, Yang X, Hock Z, Shaw JA, Shimizu KK, McDaniel SF, Wagner A (2014) Efficient purging of deleterious mutations in plants with haploid selfing. Genome Biol Evol 6(5):1238–1252
76. Haldane JBS (1937) The effect of variation on fitness. Am Nat 71:337–349
77. Glémin S, Ronfort J (2013) Adaptation and maladaptation in selfing and outcrossing species: new mutations versus standing variation. Evolution 67(1):225–240
78. Kamran-Disfani A, Agrawal AF (2014) Selfing, adaptation and background selection in finite populations. J Evol Biol 27 (7):1360–1371
79. Kirkpatrick M, Jenkins CD (1989) Genetic segregation and the maintenance of sexual reproduction. Nature 339(6222):300–301
80. Ronfort J, Glémin S (2013) Mating system, Haldane's sieve, and the domestication process. Evolution 67(5):1518–1526
81. McDonald JH, Kreitman M (1991) Adaptative protein evolution at the ADH locus in Drosophila. Nature 351:652–654
82. Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26 (9):2097–2108
83. Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010) Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol Biol Evol 27 (8):1813–1821
84. Haudry A, Cenci A, Guilhaumon C, Paux E, Poirier S, Santoni S, David J, Glémin S (2008) Mating system and recombination affect molecular evolution in four Triticeae species. Genet Res 90(1):97–109
85. Hersch-Green EI, Myburg H, Johnson MT (2012) Adaptive molecular evolution of a defence gene in sexual but not functionally asexual evening primroses. J Evol Biol 25 (8):1576–1586
86. Engelstadter J (2017) Asexual but not clonal: evolutionary processes in automictic populations. Genetics 206(2):993–1009
87. Mandegar MA, Otto SP (2007) Mitotic recombination counteracts the benefits of genetic segregation. Proc Biol Sci 274 (1615):1301–1307
88. Weir W, Capewell P, Foth B, Clucas C, Pountain A, Steketee P, Veitch N, Koffi M, De Meeus T, Kabore J, Camara M, Cooper A, Tait A, Jamonneau V, Bucheton B, Berriman M, MacLeod A (2016) Population genomics reveals the origin and asexual evolution of human infective trypanosomes. Elife 5:e11473
89. Omilian AR, Cristescu ME, Dudycha JL, Lynch M (2006) Ameiotic recombination in asexual lineages of Daphnia. Proc Natl Acad Sci U S A 103(49):18638–18643
90. Keith N, Tucker AE, Jackson CE, Sung W, Lledo JIL, Schrider DR, Schaack S, Dudycha JL, Ackerman M, Younge AJ, Shaw JR, Lynch M (2016) High mutational rates of large-scale

duplication and deletion in Daphnia pulex. Genome Res 26(1):60–69

91. Charlesworth D, Charlesworth B, Strobeck C (1979) Selection for recombination in self-fertilizing species. Genetics 93:237–244
92. Charlesworth D, Charlesworth B, Strobeck C (1977) Effects of selfing on selection for recombination. Genetics 68:213–226
93. Roze D, Lenormand T (2005) Self-fertilization and the evolution of recombination. Genetics 170:841–857
94. Ross-Ibarra J (2007) Genome size and recombination in angiosperms: a second look. J Evol Biol 20(2):800–806
95. Dawson KJ (1998) Evolutionarily stable mutation rates. J Theor Biol 194(1):143–157
96. Kondrashov AS (1995) Modifiers of mutation-selection balance - general-approach and the evolution of mutation-rates. Genet Res 66(1):53–69
97. Lynch M (2010) Evolution of the mutation rate. Trends Genet 26(8):345–352
98. Schoen DJ (2005) Deleterious mutation in related species of the plant genus *Amsinckia* with contrasting mating systems. Evolution 59(11):2370–2377
99. Baer CF, Joyner-Matos J, Ostrow D, Grigaltchik V, Salomon MP, Upadhyay A (2010) Rapid decline in fitness of mutation accumulation lines of gonochoristic (outcrossing) Caenorhabditis nematodes. Evolution 64(11):3242–3253
100. Brandvain Y, Haig D (2005) Divergent mating systems and parental conflict as a barrier to hybridization in flowering plants. Am Nat 166(3):330–338
101. Tazzyman SJ, Abbott JK (2015) Self-fertilization and inbreeding limit the scope for sexually antagonistic polymorphism. J Evol Biol 28(3):723–729
102. Burt A, Trivers R (1998) Selfish DNA and breeding systems in plants. Proc R Soc Lond B 265:141–146
103. Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. Nat Rev Genet 3(2):137–144
104. Palopoli MF, Rockman MV, TinMaung A, Ramsay C, Curwen S, Aduna A, Laurita J, Kruglyak L (2008) Molecular basis of the copulatory plug polymorphism in Caenorhabditis elegans. Nature 454 (7207):1019–1022
105. Cutter AD (2008) Reproductive evolution: symptom of a selfing syndrome. Curr Biol 18(22):R1056–R1058
106. Spillane C, Schmid KJ, Laoueille-Duprat S, Pien S, Escobar-Restrepo JM, Baroux C, Gagliardini V, Page DR, Wolfe KH, Grossniklaus U (2007) Positive darwinian selection at the imprinted MEDEA locus in plants. Nature 448(7151):349–352
107. Kawabe A, Fujimoto R, Charlesworth D (2007) High diversity due to balancing selection in the promoter region of the Medea gene in *Arabidopsis lyrata*. Curr Biol 17 (21):1885–1889
108. Budar F, Touzet P, De Paepe R (2003) The nucleo-mitochondrial conflict in cytoplasmic male sterilities revisited. Genetica 117 (1):3–16
109. Houliston GJ, Olson MS (2006) Nonneutral evolution of organelle genes in Silene vulgaris. Genetics 174(4):1983–1994
110. Ingvarsson PK, Taylor DR (2002) Genealogical evidence for epidemics of selfish genes. Proc Natl Acad Sci U S A 99 (17):11265–11269
111. Touzet P, Delph LF (2009) The effect of breeding system on polymorphism in mitochondrial genes of Silene. Genetics 181 (2):631–644
112. Foxe JP, Wright SI (2009) Signature of diversifying selection on members of the pentatricopeptide repeat protein family in Arabidopsis lyrata. Genetics 183(2):663–672, 661SI–668SI
113. Marais G (2003) Biased gene conversion: implications for genome and sex evolution. Trends Genet 19(6):330–338
114. Clement Y, Arndt PF (2013) Meiotic recombination strongly influences GC-content evolution in short regions in the mouse genome. Mol Biol Evol 30(12):2612–2618
115. Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L (2015) Quantification of GC-biased gene conversion in the human genome. Genome Res 25(8):1215–1228
116. Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10:285–311
117. Marais G, Charlesworth B, Wright SI (2004) Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol 5(7):R45
118. Wright SI, Iorgovan G, Misra S, Mokhtari M (2007) Neutral evolution of synonymous base composition in the Brassicaceae. J Mol Evol 64(1):136–141
119. Serres-Giardi L, Belkhir K, David J, Glémin S (2012) Patterns and evolution of nucleotide

landscapes in seed plants. Plant Cell 24 (4):1379–1397

120. Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glémin S (2011) GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). Mol Biol Evol 28 (9):2695–2706
121. Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES (2016) Open chromatin reveals the functional maize genome. Proc Natl Acad Sci U S A 113(22):E3177–E3184
122. Li X, Li L, Yan J (2015) Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. Nat Commun 6:6648
123. Hazzouri KM, Escobar JS, Ness RW, Killian Newman L, Randle AM, Kalisz S, Wright SI (2013) Comparative population genomics in *Collinsia* sister species reveals evidence for reduced effective population size, relaxed selection, and evolution of biased gene conversion with an ongoing mating system shift. Evolution 67(5):1263–1278
124. Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 23(6):273–277
125. Glémin S (2010) Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. Genetics 185(3):939–959
126. Dolgin ES, Charlesworth B (2006) The fate of transposable elements in asexual populations. Genetics 174(2):817–827
127. Morgan MT (2001) Transposable element number in mixed mating populations. Genet Res 77(3):261–275
128. Zeyl C, Bell G, Green DM (1996) Sex and the spread of retrotransposon Ty3 in experimental populations of *Saccharomyces cerevisiae*. Genetics 143(4):1567–1577
129. Goddard MR, Greig D, Burt A (2001) Outcrossed sex allows a selfish gene to invade yeast populations. Proc Biol Sci 268 (1485):2537–2542
130. Sullender BW, Crease TJ (2001) The behavior of a *Daphnia pulex* transposable element in cyclically and obligately parthenogenetic populations. J Mol Evol 53(1):63–69
131. Valizadeh P, Crease TJ (2008) The association between breeding system and transposable element dynamics in *Daphnia pulex*. J Mol Evol 66(6):643–654
132. Schaack S, Pritham EJ, Wolf A, Lynch M (2010) DNA transposon dynamics in populations of Daphnia pulex with and without sex. Proc R Soc B Biol Sci 277(1692):2381–2387
133. Arkhipova I, Meselson M (2000) Transposable elements in sexual and ancient asexual taxa. Proc Natl Acad Sci U S A 97 (26):14473–14477
134. Arkhipova IR, Meselson M (2005) Diverse DNA transposons in rotifers of the class Bdelloidea. Proc Natl Acad Sci U S A 102 (33):11781–11786
135. Arkhipova I, Meselson M (2005) Deleterious transposable elements and the extinction of asexuals. BioEssays 27(1):76–85
136. Matzk F, Hammer K, Schubert I (2003) Coevolution of apomixis and genome size within the genus *Hypericum*. Sex Plant Reprod 16:51–58
137. Wright SI, Schoen DJ (1999) Transposon dynamics and the breeding system. Genetica 107(1–3):139–148
138. Tam SM, Causse M, Garchery C, Burck H, Mhiri C, Grandbastien MA (2007) The distribution of copia-type retrotransposons and the evolutionary history of tomato and related wild species. J Evol Biol 20 (3):1056–1072
139. Wright SI, Le QH, Schoen DJ, Bureau TE (2001) Population dynamics of an Ac-like transposable element in self- and cross-pollinating Arabidopsis. Genetics 158 (3):1279–1288
140. Lockton S, Gaut BS (2010) The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. BMC Evol Biol 10:10
141. de la Chaux N, Tsuchimatsu T, Shimizu KK, Wagner A (2012) The predominantly selfing plant Arabidopsis thaliana experienced a recent reduction in transposable element abundance compared to its outcrossing relative Arabidopsis lyrata. Mob DNA 3(1):2
142. Agren JA, Wang W, Koenig D, Neuffer B, Weigel D, Wright SI (2014) Mating system shifts and transposable element evolution in the plant genus Capsella. BMC Genomics 15:602
143. Otto SP, Whitton J (2000) Polyploid incidence and evolution. Annu Rev Genet 34:401–437
144. Fowler NL, Levin DA (1984) Ecological constraints on the establishment of a novel polyploid in competition with its diploid progenitor. Am Nat 124:703–711
145. Husband BC (2000) Constraints on polyploid evolution: a test of the minority cytotype exclusion principle. Proc Biol Sci 267 (1440):217–223

146. Husband BC (2016) Effect of inbreeding on pollen tube growth in diploid and tetraploid Chamerion angustifolium: do polyploids mask mutational load in pollen? Am J Bot 103(3):532–540
147. Kreiner JM, Kron P, Husband BC (2017) Frequency and maintenance of unreduced gametes in natural plant populations: associations with reproductive mode, life history and genome size. New Phytol 214(2):879–889
148. Xu S, Innes DJ, Lynch M, Cristescu ME (2013) The role of hybridization in the origin and spread of asexuality in Daphnia. Mol Ecol 22(17):4549–4561
149. Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, Arunkumar R, Agren JA, Hazzouri KM, Wang W, Platts AE, Williamson RJ, Neuffer B, Lascoux M, Slotte T, Wright SI (2015) Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. Proc Natl Acad Sci U S A 112 (9):2806–2811
150. Novikova PY, Tsuchimatsu T, Simon S, Nizhynska V, Voronin V, Burns R, Fedorenko OM, Holm S, Sall T, Prat E, Marande W, Castric V, Nordborg M (2017) Genome sequencing reveals the origin of the allotetraploid Arabidopsis suecica. Mol Biol Evol 34 (4):957–968
151. Szitenberg A, Salazar-Jaramillo L, Blok VC, Laetsch DR, Joseph S, Williamson VM, Blaxter ML, Lunt DH (2017) Comparative genomics of apomictic root-knot nematodes: hybridization, ploidy, and dynamic genome change. Genome Biol Evol 9(10):2844–2861
152. Svensson O, Smith A, Garcia-Alonso J, van Oosterhout C (2016) Hybridization generates a hopeful monster: a hermaphroditic selfing cichlid. R Soc Open Sci 3(3):150684
153. Beck JB, Alexander PJ, Allphin L, Al-Shehbaz IA, Rushworth C, Bailey CD, Windham MD (2012) Does hybridization drive the transition to asexuality in diploid Boechera? Evolution 66(4):985–995
154. Janko K, Kotusz J, De Gelas K, Slechtova V, Opoldusova Z, Drozd P, Choleva L, Popiolek M, Balaz M (2012) Dynamic formation of asexual diploid and polyploid lineages: multilocus analysis of Cobitis reveals the mechanisms maintaining the diversity of clones. PLoS One 7(9):e45384
155. Lampert KP, Schartl M (2008) The origin and evolution of a unisexual hybrid: Poecilia formosa. Philos Trans R Soc Lond B Biol Sci 363(1505):2901–2909
156. Moon CD, Craven KD, Leuchtmann A, Clement SL, Schardl CL (2004) Prevalence of interspecific hybrids amongst asexual fungal endophytes of grasses. Mol Ecol 13 (6):1455–1467
157. Blanc-Mathieu R, Perfus-Barbeoch L, Aury J-M, Da Rocha M, Gouzy J, Sallet E, Martin-Jimenez C, Bailly-Bechet M, Castagnone-Sereno P, Flot J-F, Kozlowski DK, Cazareth J, Couloux A, Da Silva C, Guy J, Kim-Jo Y-J, Rancurel C, Schiex T, Abad P, Wincker P, Danchin EGJ (2017) Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. PLoS Genet 13 (6):e1006777
158. Trivers R, Burt A, Palestis BG (2004) B chromosomes and genome size in flowering plants. Genome 47(1):1–8
159. Whitney KD, Baack EJ, Hamrick JL, Godt MJ, Barringer BC, Bennett MD, Eckert CG, Goodwillie C, Kalisz S, Leitch IJ, Ross-Ibarra J (2010) A role for nonadaptive processes in plant genome size evolution? Evolution 64 (7):2097–2109
160. Agren JA, Greiner S, Johnson MT, Wright SI (2015) No evidence that sex and transposable elements drive genome size variation in evening primroses. Evolution 69(4):1053–1062
161. Albach DC, Greilhuber J (2004) Genome size variation and evolution in *Veronica*. Ann Bot 94(6):897–911
162. Wright S, Ness RW, Foxe JP, Barrett SC (2008) Genomic consequences of outcrossing and selfing in plants. Int J Plant Sci 169 (1):105–118
163. Fierst JL, Willis JH, Thomas CG, Wang W, Reynolds RM, Ahearne TE, Cutter AD, Phillips PC (2015) Reproductive mode and the evolution of genome size and structure in Caenorhabditis nematodes. PLoS Genet 11 (6):e1005323
164. Ritland K (2002) Extensions of models for the estimation of mating systems using n independent loci. Heredity 88(4):221–228
165. Ritland K, Jain S (1981) A model for the estimation of outcrossing rate and gene-frequencies using N independent loci. Heredity 47(1):35–52
166. Koelling VA, Monnahan PJ, Kelly JK (2012) A Bayesian method for the joint estimation of outcrossing rate and inbreeding depression. Heredity (Edinb) 109(6):393–400
167. David P, Pujol B, Viard F, Castella V, Goudet J (2007) Reliable selfing rate estimates from imperfect population genetic data. Mol Ecol 16(12):2474–2487

168. Redelings BD, Kumagai S, Tatarenkov A, Wang L, Sakai AK, Weller SG, Culley TM, Avise JC, Uyenoyama MK (2015) A Bayesian approach to inferring rates of selfing and locus-specific mutation. Genetics 201 (3):1171–1188
169. McClure NS, Whitlock MC (2012) Multilocus estimation of selfing and its heritability. Heredity (Edinb) 109(3):173–179
170. Burkli A, Sieber N, Seppala K, Jokela J (2017) Comparing direct and indirect selfing rate estimates: when are population-structure estimates reliable? Heredity (Edinb) 118 (6):525–533
171. Gao H, Williamson S, Bustamante CD (2007) A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176 (3):1635–1651
172. Halkett F, Simon JC, Balloux F (2005) Tackling the population genetics of clonal and partially clonal organisms. Trends Ecol Evol 20(4):194–201
173. Bailleul D, Stoeckel S, Arnaud-Haond S (2016) RClone: a package to identify MultiLocus Clonal Lineages and handle clonal data sets in R. Methods Ecol Evol 7:966–970
174. Bomblies K, Yant L, Laitinen RA, Kim ST, Hollister JD, Warthmann N, Fitz J, Weigel D (2010) Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. PLoS Genet 6(3):e1000890
175. Tsai IJ, Bensasson D, Burt A, Koufopanou V (2008) Population genomics of the wild yeast Saccharomyces paradoxus: quantifying the life cycle. Proc Natl Acad Sci U S A 105 (12):4957–4962
176. Ruderfer DM, Pratt SC, Seidel HS, Kruglyak L (2006) Population genomic analysis of outcrossing and recombination in yeast. Nat Genet 38(9):1077–1081
177. Grimsley N, Pequin B, Bachy C, Moreau H, Piganeau G (2010) Cryptic sex in the smallest eukaryotic marine green alga. Mol Biol Evol 27(1):47–54
178. Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, Saeys Y, Wuyts J, Jabbari K, Bowler C, Panaud O, Piegu B, Ball SG, Ral JP, Bouget FY, Piganeau G, De Baets B, Picard A, Delseny M, Demaille J, Van de Peer Y, Moreau H (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci U S A 103(31):11647–11652
179. Birky CW (1996) Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. Genetics 144(1):427–437
180. Welch DM, Meselson M (2000) Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. Science 288(5469):1211–1215
181. Signorovitch A, Hur J, Gladyshev E, Meselson M (2015) Allele sharing and evidence for sexuality in a mitochondrial clade of bdelloid rotifers. Genetics 200(2):581–590
182. Debortoli N, Li X, Eyres I, Fontaneto D, Hespeels B, Tang CQ, Flot JF, Van Doninck K (2016) Genetic exchange among bdelloid rotifers is more likely due to horizontal gene transfer than to meiotic sex. Curr Biol 26 (6):723–732
183. Takebayashi N, Morrell PL (2001) Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. Am J Bot 88(7):1143–1150
184. Goldberg EE, Igic B (2008) On phylogenetic tests of irreversible evolution. Evolution 62 (11):2727–2741
185. Schaefer I, Domes K, Heethoff M, Schneider K, Schon I, Norton RA, Scheu S, Maraun M (2006) No evidence for the 'Meselson effect' in parthenogenetic oribatid mites (Oribatida, Acari). J Evol Biol 19 (1):184–193
186. Schon I, Martens K (2003) No slave to sex. Proc R Soc Lond B 270(1517):827–833
187. Bechsgaard JS, Castric V, Charlesworth D, Vekemans X, Schierup MH (2006) The transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 Myr. Mol Biol Evol 23 (9):1741–1750
188. Tang C, Toomajian C, Sherman-Broyles S, Plagnol V, Guo YL, Hu TT, Clark RM, Nasrallah JB, Weigel D, Nordborg M (2007) The evolution of selfing in *Arabidopsis thaliana*. Science 317(5841):1070–1072
189. Shimizu KK, Tsuchimatsu T (2015) Evolution of selfing: recurrent patterns in molecular adaptation. Annu Rev Ecol Syst 46:593–622
190. Castric V, Billiard S, Vekemans X (2014) Trait transitions in explicit ecological and genomic contexts: plant mating systems as case studies. Adv Exp Med Biol 781:7–36
191. Tsuchimatsu T, Goubet PM, Gallina S, Holl AC, Fobis-Loisy I, Berges H, Marande W, Prat E, Meng D, Long Q, Platzer A, Nordborg M, Vekemans X, Castric V (2017) Patterns of polymorphism at the self-incompatibility locus in 1,083 Arabidopsis

thaliana genomes. Mol Biol Evol 34 (8):1878–1889

192. Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Tsuchimatsu T, Burbano HA, Pico FX, Alonso-Blanco C, Hancock AM (2017) African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. Proc Natl Acad Sci U S A 114 (20):5213–5218
193. Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI (2009) Recent speciation associated with the evolution of selfing in *Capsella*. Proc Natl Acad Sci U S A 106 (13):5241–5245
194. Guo YL, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D, Schierup MH (2009) Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proc Natl Acad Sci U S A 106 (13):5246–5251
195. Brandvain Y, Slotte T, Hazzouri KM, Wright SI, Coop G (2013) Genomic identification of founding haplotypes reveal the history of the selfing species *Capsella rubella*. PLoS Genet 9 (9):e1003754
196. Judson OP, Normark BB (1996) Ancient asexual scandals. Trends Ecol Evol 11 (2):41–46
197. Igic B, Bohs L, Kohn JR (2006) Ancient polymorphism reveals unidirectional breeding system shifts. Proc Natl Acad Sci U S A 103 (5):1359–1363
198. Ferrer MM, Good-Avila SV (2007) Macrophylogenetic analyses of the gain and loss of self-incompatibility in the Asteraceae. New Phytol 173(2):401–414
199. Janko K, Drozd P, Flegr J, Pannell JR (2008) Clonal turnover versus clonal decay: a null model for observed patterns of asexual longevity, diversity and distribution. Evolution 62(5):1264–1270
200. Janko K (2014) Let us not be unfair to asexuals: their ephemerality may be explained by neutral models without invoking any evolutionary constraints of asexuality. Evolution 68 (2):569–576
201. Schwander T, Crespi BJ (2009) Twigs on the tree of life? Neutral and selective models for integrating macroevolutionary patterns with microevolutionary processes in the analysis of asexuality. Mol Ecol 18(1):28–42
202. Goldberg EE, Kohn JR, Lande R, Robertson KA, Smith SA, Igic B (2010) Species selection maintains self-incompatibility. Science 330 (6003):493–495
203. Fitzjohn RG, Maddison WP, Otto SP (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst Biol 58(6):595–611
204. Maddison WP, Midford PE, Otto SP (2007) Estimating a binary character's effect on speciation and extinction. Syst Biol 56 (5):701–710
205. de Vos JM, Hughes CE, Schneeweiss GM, Moore BR, Conti E (2014) Heterostyly accelerates diversification via reduced extinction in primroses. Proc Biol Sci 281 (1784):20140075
206. Johnson MT, Fitzjohn RG, Smith SD, Rausher MD, Otto SP (2011) Loss of sexual recombination and segregation is associated with increased diversification in evening primroses. Evolution 65(11):3230–3240
207. Glémin S (2003) How are deleterious mutations purged? Drift versus nonrandom mating. Evolution 57(12):2678–2687
208. Tahir D, Glémin S, Lascoux M, Kaj I (2019) Modeling a trait-dependent diversification process coupled with molecular evolution on a random species tree. J Theor Biol 461:189–203
209. Dalrymple RL, Buswell JM, Moles AT (2015) Asexual plants change just as often and just as fast as do sexual plants when introduced to a new range. Oikos 124(2):196–205
210. Uecker H (2017) Evolutionary rescue in randomly mating, selfing, and clonal populations. Evolution 71(4):845–858
211. Richards AJ (1997) Plant breeding systems, 2nd edn. Chapman & Hall Ltd, London
212. Igic B, Kohn JR (2006) The distribution of plant mating systems: study bias against obligately outcrossing species. Evolution 60 (5):1098–1103
213. Wright SI, Kalisz S, Slotte T (2013) Evolutionary consequences of self-fertilization in plants. Proc Biol Sci 280(1760):20130133
214. Fontaneto D, Barraclough TG (2015) Do species exist in asexuals? Theory and evidence from bdelloid rotifers. Integr Comp Biol 55 (2):253–263
215. Flot JF, Hespeels B, Li X, Noel B, Arkhipova I, Danchin EGJ, Hejnol A, Henrissat B, Koszul R, Aury JM, Barbe V, Barthelemy RM, Bast J, Bazykin GA, Chabrol O, Couloux A, Da Rocha M, Da Silva C, Gladyshev E, Gouret P, Hallatschek O, Hecox-Lea B, Labadie K, Lejeune B, Piskurek O, Poulain J, Rodriguez F, Ryan JF, Vakhrusheva OA, Wajnberg E, Wirth B, Yushenova I, Kellis M, Kondrashov AS, Welch DBM, Pontarotti P,

Weissenbach J, Wincker P, Jaillon O, Van Doninck K (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. Nature 500(7463):453–457

216. Golczyk H, Massouh A, Greiner S (2014) Translocations of chromosome end-segments and facultative heterochromatin promote meiotic ring formation in evening primroses. Plant Cell 26(3):1280–1293
217. Barraclough TG, Fontaneto D, Ricci C, Herniou EA (2007) Evidence for inefficient selection against deleterious mutations in cytochrome oxidase I of asexual bdelloid rotifers. Mol Biol Evol 24:1952–1962
218. Foxe JP, Dar VU, Zheng H, Nordborg M, Gaut BS et al (2008) Selection on amino acid substitutions in Arabidopsis. MBE 25:1375–1383
219. Gioti A, Stajich J, Johannesson H (2013) Neurospora and the dead-end hypothesis: genomic consequences of selfing in the model genus. Evolution 67(12):3600–3616
220. Mark Welch DB, Meselson MS (2001) Rates of nucleotide substitution in sexual and anciently asexual rotifers. Proc Natl Acad Sci USA 98:6720–6724
221. Ness RW, Siol M, Barrett SC (2012) Genomic consequences of transitions from cross- to self-fertilization on the efficacy of selection in three independently derived selfing plants. BMC Genomics 13:611
222. Nygren K, Strandberg R, Wallberg A, Nabholz B, Gustafsson T et al (2011) A comprehensive phylogeny of Neurospora reveals a link between reproductive mode and molecular evolution in fungi. Mol Phylogenet Evol 59:649–663
223. Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D (2011) Reduced efficacy of natural selection on codon usage bias in selfing Arabidopsis and Capsella species. Genome Biol Evol 3:868–880
224. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F et al (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet 45:831–835
225. Brandvain Y, Slotte T, Hazzouri KM, Wright SI, Coop G (2013) Genomic Identification of Founding Haplotypes Reveals the History of the Selfing Species Capsella rubella. PLOS Genetics 9(9): e1003754
226. Whittle CA, Sun Y, Johannesson H (2011) Evolution of synonymous codon usage in Neurospora tetrasperma and Neurospora discreta. Genome Biol Evol 3:332–343
227. Bast J, Schaefer I, Schwander T, Maraun M, Scheu S, Kraaijeveld K (2016) No accumulation of transposable elements in asexual arthropods. Mol Biol Evol 33:697–706
228. Docking TR, Saade FE, Elliott MC, Schoen DJ (2006) Retrotransposon sequence variation in four asexual plant species. J Mol Evol 62:375–387
229. Dolgin ES, Charlesworth B, Cutter AD (2008) Population frequencies of transposable elements in selfing and outcrossing Caenorhabditis nematodes. Genet Res 90:317–329
230. Goodwin TJ, Poulter RT (2008) Multiple LTR-Retrotransposon families in the asexual yeast Candida albicans. Genome Res 10:174–191
231. Jiang X, Tang H, Ye Z, Lynch M (2017) Insertion polymorphisms of mobile genetic elements in sexual and asexual populations of Daphnia pulex. Genome Biol Evol 9:362–374
232. Kraaijeveld K, Zwanenburg B, Hubert B, Vieira C, De Pater S, Alphen V et al (2012) Transposon proliferation in an asexual parasitoid. Mol Ecol 21:3898–3906
233. Schaack S, Choi E, Lynch M, Pritham EJ (2010) DNA transposons and the role of recombination in mutation accumulation in Daphnia pulex. Genome Biol 11:R46
234. Szitenberg A, Cha S, Opperman CH, Bird DM, Blaxter ML, Lunt DH (2016) Genetic drift, not life history or RNAi, determine long-term evolution of transposable elements. Genome Biol Evol 8:2964–2978
235. Zeyl C, Bell G, Da Silva J (1994) Transposon abundance in sexual and asexual populations of Chlamydomonas reinhardtii. Evolution 48:1406–1409

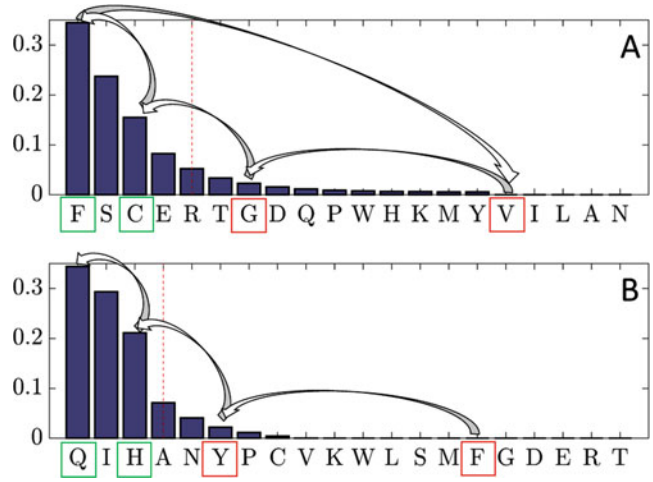

Fig. 1 It can be useful to think of the substitution process at a site as movement on a site-specific fitness landscape. The horizontal axis in each figure shows the amino acids at a hypothetical site in order of their stationary frequencies indicated by the height of the bars. Frequency is a function of mutation and selection, but can be construed as a proxy for fitness. The site-specific *dN*/*dS* ratio [25] is a function of the amino acid that occupies the site, and can be <1 (left of the red dashed line) or >1 (right of the dashed red line). (**a**) Suppose phenylalanine (F, TTT) is the fittest amino acid. The site-specific *dN*/*dS* ratio is much less than one when occupied by F because any nonsynonymous mutation will always be to an amino acid that is less fit. Nevertheless, it is possible for an amino acid such as valine (V, GTT) to be fixed on occasion, provided that selection is not too stringent. When this happens, *dN*/*dS* at the site is temporarily elevated to a value greater than one as positive selection moves the site back to F by a series of replacement substitutions, e.g., V (GTT) $\rightarrow$ G (GGT) $\rightarrow$ C (TGT) $\rightarrow$ F (TTT). We call the episodic recurrence of this process **shifting balance** on a static fitness landscape. Shifting balance on a landscape for which all frequencies are approximately equal corresponds to **nearly neutral** evolution (not depicted), when *dN*/*dS* is always ≈ 1. (**b**) Now, consider what happens following a change in one or more external factors that impact the functional significance of the site. The relative fitnesses of the amino acids might change from that depicted in **a** to that in **b** for instance, where glutamine (Q) is fittest. If at the time of the change the site is occupied by F (as is most likely), then *dN*/*dS* would be temporarily elevated as positive selection moves the site toward its new peak at Q, e.g., F (TTT) $\rightarrow$ Y (TAT) $\rightarrow$ H (CAT) $\rightarrow$ Q (CAA). This process of **adaptive evolution** is followed by a return to shifting balance once the site is occupied by Q

for 61^N categories). We refer to $\boldsymbol{P} = P_{\mathrm{M}}(\theta_{\mathrm{M}})$ as the site-pattern distribution for that model. Figure 2 depicts the space of all possible site-pattern distributions for an N-taxon alignment. Each ellipse represents the family of distributions $\{P_{\mathrm{M}}(\theta_{\mathrm{M}})|\theta_{\mathrm{M}} \in \Omega_{\mathrm{M}}\}$, where Ω_{M} is the vector space of all possible values of θ_{M}. For example, $\{P_{\mathrm{M0}}(\theta_{\mathrm{M0}})|\theta_{\mathrm{M0}} \in \Omega_{\mathrm{M0}}\}$ is the family of distributions that can be

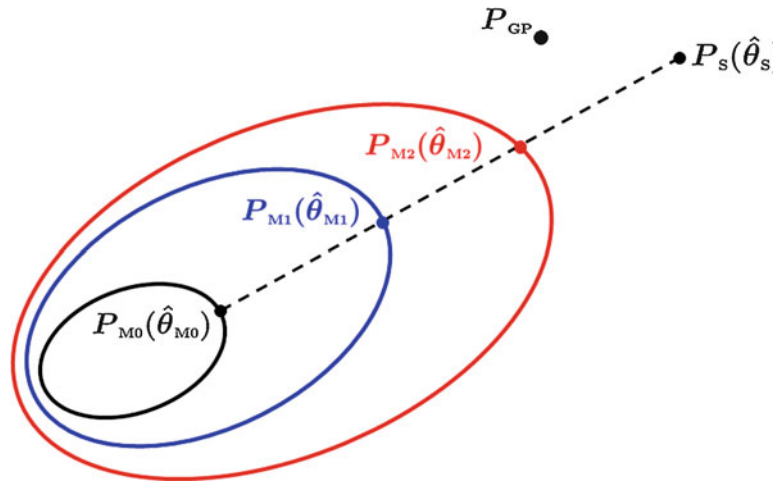

Fig. 2 The $(61^N - 1)$-dimensional simplex containing all possible site-pattern distributions for an *N*-taxon alignment is depicted. The innermost ellipse represents the subspace $\{P_{M0}(\theta_{M0})|\theta_{M0} \in \Omega_{M0}\}$ that is the family of distributions that can be specified using M0, the simplest of CSMs. This is nested in the family of distributions that can be specified using M1 (blue ellipse), a hypothetical model that has the same parameters as M0 plus some extra parameters. Similarly, M1 is nested in M2 (red ellipse). Whereas models are represented by subspaces of distributions, the true generating process is represented by a single point P_{GP}, the location of which is unknown. The empirical site-pattern distribution $P_S(\hat{\theta}_S)$ corresponds to the saturated model fitted to the alignment; with large samples, $P_S(\hat{\theta}_S) \approx P_{GP}$. For any other model M, the member $P_M(\hat{\theta}_M) \in \{P_M(\theta_M) \mid \theta_M \in \Omega_M\}$ most consistent with *X* is the one that minimizes deviance, which is twice the difference between the maximum log-likelihood of the data under the saturated model and the maximum log-likelihood of the data under M

specified using M0, the simplest CSM that assumes a common substitution rate matrix Q for all sites and branches. This is nested inside $\{P_{M1}(\theta_{M1})|\theta_{M1} \in \Omega_{M1}\}$, where M1 is a hypothetical model that is the same as M0 but for a few extra parameters. Likewise, M1 is nested in M2. The location of the site-pattern distribution for the true generating process is represented by P_{PG}. Its location is fixed but unknown. It is therefore not possible to assess the distance between it and any other distribution. Instead, comparisons are made using the site-pattern distribution inferred under the saturated model.

Whereas a CSM $\{P_M(\theta_M)|\theta_M \in \Omega_M\}$ can be thought of as a family of multinomial distributions for the 61^N possible site patterns, the fitted saturated model $P_S(\hat{\theta}_S)$ is the unique distribution defined by the MLE $\hat{\theta}_S = (y_1/n, \ldots, y_m/n)^T$, where $y_i > 0$ is the observed frequency of the ith site pattern, m is the number of unique site patterns, and n is the number of codon sites. In other

words, the fitted saturated model is the empirical site-pattern distribution for a given alignment. Because it takes none of the mechanisms of mutation or selection into account, ignores the phylogenetic relationships between sequences, and excludes the possibility of site patterns that were not actually observed (i.e., $y_i/n = 0$ for site patterns i not observed in X), $P_S(\hat{\theta}_S)$ can be construed as the maximally phenomenological explanation of the observed alignment. An alignment is always more likely under the saturated model than it is under any other CSM. $P_S(\hat{\theta}_S)$ therefore provides a natural benchmark for model improvement.

For any alignment, the MLE over the family of distributions $\{P_M(\theta_M)|\theta_M \in \Omega_M\}$ is represented by a fixed point $P_M(\hat{\theta}_M)$ in Fig. 2. $P_M(\hat{\theta}_M)$ is the distribution that minimizes the statistical deviance between $P_M(\theta_M)$ and $P_S(\hat{\theta}_S)$. Deviance is defined as twice the difference between the maximum log-likelihood (LL) of the data under the saturated model and the maximum log-likelihood of the data under M:

$$D(\hat{\theta}_M,\hat{\theta}_S) = 2\{\ell(\hat{\theta}_S \mid X) - \ell(\hat{\theta}_M \mid X)\} \tag{2}$$

A key feature of deviance is that it always decreases as more parameters are added to the model, corresponding to an increase in the probability of the data under that model. For example, suppose $\{P_{M2}(\theta_{M2})|\theta_{M2} \in \Omega_{M2}\}$ is the same family of distributions as $\{P_{M1}(\theta_{M1})|\theta_{M1} \in \Omega_{M1}\}$ but for the inclusion of one additional parameter ψ, so that $\theta_{M2} = (\theta_{M1}, \psi)$. The improvement in the probability of the data under $P_{M2}(\hat{\theta}_{M2})$ over its probability under $P_{M1}(\hat{\theta}_{M1})$ is assessed by the size of the reduction in deviance induced by ψ:

$$\begin{aligned}\Delta D(\hat{\theta}_{M1},\hat{\theta}_{M2}) &= D(\hat{\theta}_{M1},\hat{\theta}_S) - D(\hat{\theta}_{M2},\hat{\theta}_S) \\ &= 2\{\ell(\hat{\theta}_{M2} \mid X) - \ell(\hat{\theta}_{M1} \mid X)\}\end{aligned} \tag{3}$$

Equation 3 is just the familiar log-likelihood ratio (LLR) used to compare nested models under the maximum likelihood framework.

Given this measure of model improvement, the de facto objective of model building is not to provide a mechanistic explanation of the data that more accurately represents the true generating process, but only to move closer to the site-pattern distribution of the fitted saturated model. Real alignments are limited in size, so there will always be some distance between $P_S(\hat{\theta}_S)$ and P_{GP} due to sampling error (as represented in Fig. 2). But even with an infinite number of codon sites, when $P_S(\hat{\theta}_S)$ converges to P_{GP}, the criterion of minimizing deviance does not inevitably lead to a better explanation of the data because of the possibility of confounding. Two processes are said to be confounded if they can produce similar patterns in the data. Hence, if ψ represents a process E that did not actually occur when the data was generated, and if E is confounded

with another process that did occur, the LLR in Eq. 3 can still be significant. Under this scenario, the addition of ψ to M1 would engender movement toward $P_S(\hat{\theta}_S)$ and P_{GP}, but the new model M2 would also provide a worse mechanistic explanation of the data because it would falsely indicate that E occurred. The possibility of confounding and its impact on inference is demonstrated in Case Study D.

3 Phase I: Pioneering CSMs

The first effort to detect positive selection at the molecular level [24] relied on heuristic counting methods [43]. Phase I of CSM development followed with the introduction of formal statistical approaches based on ML [16, 42]. The first CSMs were used to infer whether the estimate $\hat{\omega}$ of a single nonsynonymous to synonymous substitution rate ratio averaged over all sites and branches was significantly greater than one. Such CSMs were found to have low power due to the pervasiveness of synonymous substitutions at most sites within a typical gene [76]. An early attempt to increase the statistical power to infer positive selection was the CSM designed to detect $\hat{\omega} > 1$ on specific branches [78]. Models accounting for variations in ω across sites were subsequently developed, the most prominent of which are the M-series models [78, 81]. These were accompanied by methods to identify individual sites under positive selection. The quest for power culminated in the development of models that account for variations in the rate ratio across both sites and branches. The appearance of various branch-site models (e.g., [4, 10, 79, 86]) marks the end of Phase I of CSM development.

Two case studies are employed in this section to illustrate some of the inferential challenges associated with Phase I models. We use Case Study A to examine the impact of low information content on the inference of positive selection at individual codon sites. The subject of this study is the M1a vs M2a model contrast applied to the *tax* gene of the human T-cell lymphotropic virus type I (HTLV-I; [63, 82]). We use Case Study B to illustrate how model misspecification (i.e., differences between the fitted model and the generating process) can lead to false inferences. The subject of this study is the Yang–Nielsen branch-site model (YN-BSM; [79]) applied to simulated data.

3.1 Case Study A: Low Information Content

To study the impact of low information content on inference, we use a pair of nested M-series models known as M1a and M2a [70, 82]. Under M1a, sites are partitioned into two rate-ratio categories, $0 < \omega_0 < 1$ and $\omega_1 = 1$ in proportions p_0 and $p_1 = 1 - p_0$. M2a includes an additional category for the proportion of sites $p_2 = 1 - p_0 - p_2$ that evolved under positive selection with

$\omega_2 > 1$. The use of multiple categories permits two levels of inference. The first is an omnibus likelihood ratio test (LRT) for evidence of positive selection somewhere in the gene, which is conducted by contrasting a pair of nested models. For example, the contrast of M1a vs M2a is made by computing the distance $\mathrm{LLR} = \Delta D(\hat{\theta}_{\mathrm{M1a}}, \hat{\theta}_{\mathrm{M2a}})$ between the two models and comparing the result to the limiting distribution of the LLR under the null model. In this case, the limiting distribution of LLR is often taken to be χ^2_2 [75], which would be correct under regular likelihood theory because the models differ by two parameters. The second level of inference is used to identify individual sites that underwent positive selection. This is conducted only if positive selection is inferred by the omnibus test (e.g., if LLR > 5.99 for the M1a vs M2a contrast at the 5% level of significance). Let c_0, c_1, and c_2 represent the event that a given site pattern x falls into the stringent ($0 < \hat{\omega}_0 < 1$), neutral ($\hat{\omega}_1 = 1$), or positive ($\hat{\omega}_2 > 1$) selection category, respectively. Applying Bayes' rule:

$$\Pr(c_2 \mid x, \hat{\theta}_{\mathrm{M2a}}) = \frac{\Pr(x \mid c_2, \hat{\theta}_{\mathrm{M2a}})\hat{p}_2}{\sum_{k=0}^{2} \Pr(x \mid c_k, \hat{\theta}_{\mathrm{M2a}})\hat{p}_k} \tag{4}$$

Sites with a sufficiently high posterior probability (e.g., $\Pr(c_2 \mid x, \hat{\theta}_{\mathrm{M2a}}) > 0.95$) are inferred to have undergone positive selection. Equation 4 is representative of the naive empirical Bayes (NEB) approach under which MLEs ($\hat{\theta}_{\mathrm{M2a}}$) are used to compute posterior probabilities.

The NEB approach ignores potential errors in parameter estimates that can lead to false inference of positive selection at a site (i.e., a false positive). The resulting false positive rate can be especially high for alignments with low information content. An example setting with low information content arises when there are a substantial number of invariant sites, since these provide little information about the substitution process. The issue of low information content is well illustrated by the extreme case of the *tax* gene, HTLV-I [63]. The alignment consists of 20 sequences with 181 codon sites, 158 of which are invariant. The 23 variable sites have only one substitution each: 2 are synonymous and 21 are nonsynonymous. The high ratio of nonsynonymous-to-synonymous substitutions suggests that the gene underwent positive selection. This hypothesis was supported by analytic results: the LLR for the M1a vs M2a contrast was 6.96 corresponding to a p-value of approximately 0.03 [82]. The omnibus test therefore supported the conclusion that the gene underwent positive selection. However, the MLE for p_2 under M2a was $\hat{p}_2 = 1$. Using this value in Eq. 4 gives $\Pr(c_2 \mid x, \hat{\theta}_{\mathrm{M2a}}) = 1$ for all sites, including the 158 invariable sites. Such an unreasonable result can occur under NEB because, despite the possibility of large sampling errors in

MLEs due to low information, $\hat{\theta}_{M2a}$ is treated as a known value in Eq. 4.

Bayes empirical Bayes (BEB; [82]), a partial Bayesian approach under which rate ratios and their corresponding proportions are assigned discrete prior distributions (cf. [21]), was proposed as an alternative to NEB. Numerical integration over the assumed priors tends to provide better estimates of posterior probabilities, particularly in cases where information content is low. Using BEB in the analysis of the *tax* gene, for example, the posterior probability was $0.91 < \Pr(c_2 \mid x,\hat{\theta}_{M2a}) < 0.93$ for the 21 sites with a single nonsynonymous change and $0.55 < \Pr(c_2 \mid x,\hat{\theta}_{M2a}) < 0.61$ for the remaining sites [82]. Hence, the BEB approach mitigated the problem of low information content, as the posterior probability of positive selection at invariant sites was reduced. An alternative to BEB is called smoothed bootstrap aggregation (SBA) [38]. SBA entails drawing site patterns from X with replacement (i.e., bootstrap) to generate a set of alignments $\{X_1, \ldots, X_m\}$ with similar information content as X. The MLEs $\{\hat{\theta}_i\}_{i=1}^{m}$ for the vector of model parameters θ are then estimated by fitting the CSM to each $X_i \in \{X_1, \ldots, X_m\}$. A kernel smoother is applied to these values to reduce sampling errors. The mean value of the resulting smoothed $\{\hat{\theta}_i\}_{i=1}^{m}$ is then used in Eq. 4 in place of the MLE for θ obtained from the original alignment to estimate posterior probabilities. This approach was shown to balance power and accuracy at least as well as BEB. But, SBA has the advantage that it can accommodate the uncertainty of all parameter estimates (not just those of the ω distribution, as in BEB) and is much easier to implement. When SBA was applied to the *tax* gene, the posterior probabilities for positive selection were further reduced: $0.87 < \Pr(c_2 \mid x,\hat{\theta}_{M2a}) < 0.89$ for the 21 sites with a single nonsynonymous change, and $0.55 < \Pr(c_2 \mid x,\hat{\theta}_{M2a}) < 0.60$ for the remaining sites [38].

The problem of low information content was fairly obvious in the case of the *tax* gene, as 158 of the 181 codon sites within that dataset were invariant. However, it can sometimes be unclear whether there is enough variation in an alignment to ensure reliable inferences. It would be useful to have a method to determine whether a given data set might be problematic. An MLE $\hat{\theta}$ will always converge to a normal distribution centered at the true parameter value θ with variance proportional to $1/n$ as the sample size n (a proxy for information content) gets larger, provided that the CSM satisfies certain "regularity" conditions (a set of technical conditions that must hold to guarantee that MLEs will converge in distribution to a normal, and that the LLR for any pair of nested models will converge to its expected chi-squared distribution). This expectation makes it possible to assess whether an alignment is sufficiently informative to obtain the benefits of regularity. The

first step is to generate a set of bootstrap alignments $\{X_1, \ldots, X_m\}$. The CSM can then be fitted to these to produce a sample distribution $\{\hat{\theta}_i\}_{i=1}^{m}$ for the MLE of any model parameter θ. If the alignment is sufficiently informative with respect θ, then a histogram of $\{\hat{\theta}_i\}_{i=1}^{m}$ should be approximately normal in distribution. Serious departures from normality (e.g., a bimodal distribution) indicate unstable MLEs, which are a sign of insufficient information or an irregular modeling scenario. Mingrone et al. [38] recommend using this technique with real data as a means of gaining insight into potential difficulties of parameter estimation using a given CSM.

3.1.1 Irregularity and Penalized Likelihood

Issues associated with low information content can be made worse by violations of certain regularity conditions. For example, M2a is the same as M1a but for two extra parameters, p_2 and ω_2. Usual likelihood theory would therefore predict that the limiting distribution of the LLR is χ^2_2. However, this result is valid only if the regularity conditions hold. Among these conditions is that the null model is not obtained by placing parameters of the alternate model on the boundary of parameter space. Since M1a is the same as M2a but with $p_2 = 0$, this condition is violated. The same can be said for many nested pairs of Phase I CSMs, such as M7 vs M8 [81] or M1 vs branch-site Model A [79]. Although the theoretical limiting distribution of the LLR under some irregular conditions has been determined by Self and Liang [54], those results do not include cases where one of the model parameters is unidentifiable under the null [2]. Since M1a is M2a with $p_2 = 0$, the likelihood under M1a is the same for any value of ω_2. This makes ω_2 unidentifiable under the null. The limiting distribution for the M1a vs M2a contrast is therefore unknown [74].

A penalized likelihood ratio test (PLRT; [39]) has been proposed to mitigate problems associated with unidentifiable parameters. Under this method, the likelihood function for the alternate model (e.g., M2a) is modified so that values of p_2 closer to zero are penalized. This has the effect of drawing the MLE for p_2 away from the boundary, and can interpreted as a way to "regularize" the model. PLRT seems to be more useful in cases where the analysis of a real alignment produces a small value of $\hat{p}_2$ accompanied by an unrealistically large value of $\hat{\omega}_2$. This can happen because $\hat{\omega}_2$ is influenced by fewer and fewer site patterns as $\hat{p}_2$ approaches zero, and is therefore subject to larger and larger sampling errors. In addition, $\hat{\omega}_2$ and $\hat{p}_2$ tend to be negatively correlated, which further contributes to the large sampling errors. For example, Mingrone et al. [39] found that M2a fitted to a 5-taxon alignment with 198 codon sites without penalization gave $(\hat{p}_2, \hat{\omega}_2) = (0.01, 34.70)$. These MLEs, if taken at face value, suggest that a small number of sites in the gene underwent positive

selection. However, such a large rate ratio is difficult to believe given that its estimate is consistent with only approximately 2 codon sites (e.g., an estimated 1% of the 198 sites or ≈2 sites). Using the PLRT, the MLEs were $(\hat{p}_2,\hat{\omega}_2) = (0.09,1.00)$. These suggest that selection pressure was nearly neutral at a significant proportion of sites in the gene. In this case, the rate ratio is consistent with 9% of the 198 sites or ≈18 sites and is therefore less likely to be an artifact of sampling error. We expect this approach to be useful in a wide variety of evolutionary applications that rely on mixture models to make inferences (e.g., [13, 34, 47, 66]).

Other approaches for dealing with low information content in the data for an individual gene include the empirical Bayes approach of Kosiol et al. [33] and the parametric bootstrapping methods of Gibbs [14]. Both methods exploit the additional information content available from other genes. Kosiol et al. [33] adopted an empirical Bayes approach, where ω values varied over edges and genes according to a distribution. Because empirical posterior distributions are used, the approach is more akin to detecting sites under positive selection (e.g., using NEB) than formal testing. By contrast, Gibbs [14] adopted a test-based approach and utilized parametric bootstrapping [15] to approximate the distribution of the likelihood ratio statistic using data from other genes to obtain parameter sets to use in the bootstrap. Whereas this approach can attenuate issues associated with low information content, it can also be computationally expensive, especially when applied to large alignments.

3.2 Case Study B: Model Misspecification

The mechanisms that give rise to the diversity of site patterns in a set of homologous genes are highly complex and not fully understood. CSMs are therefore necessarily simplified representations of the true generating process, and are in this sense misspecified. The extent to which misspecification might cause an omnibus LRT to falsely detect positive selection was of primary concern during Phase I of model development. We use a particular form of the YN-BSM called Model A [79] to illustrate this issue. In its original form, the omnibus LRT assumes a null under which a proportion p_0 of sites evolved under stringent selection with $\omega_0 = 0$ and the remaining sites evolved under a neutral regime with $\omega_1 = 1$ on all branches of the tree (i.e., model M1 in [44]). This is contrasted with Model A, which is the same as M1 except that it assumes that some stringent sites and some neutral sites evolved under positive selection with $\omega_2 > 1$ on a prespecified branch called the foreground branch. The omnibus test contrasting M1 with Model A was therefore designed to detect a subset of sites that evolved adaptively on the same branch of the tree.

During this period of model development, the standard method to test the impact of misspecification on the reliability of

Table 1
Rate ratios (ω) for regimes X and Z taken from Zhang [85]

Sites	1–20	21–40	41–60	61–80	81–100	101–120	121–140	141–160	161–180	181–200
ω regime X	1.00	1.00	0.80	0.80	0.50	0.50	0.20	0.20	0.00	0.00
ω regime Z	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

an omnibus LRT was to generate alignments in silico using a more complex version of the CSM to be tested as the generating model. This usually involved adding more variability in ω across sites and/or branches than assumed by the fitted CSM while leaving all other aspects of the generating model the same. In Zhang [85], for example, alignments were generated using site-specific rate matrices, as in Eq. 1, with rate ratios ω specified by predetermined selection regimes, two of which are shown in Table 1. In one simulation, 200 alignments were generated using regime Z on a single foreground branch and regime X on all of the remaining branches of a 10 or 16 taxon tree. The gene therefore underwent a mixture of stringent selection and neutral evolution over most of the tree (regime X), but with complete relaxation of selection pressure on the foreground branch (regime Z). Positive selection did not occur at any of the sites. Nevertheless, the M1 vs Model A contrast inferred positive selection in 20–55% of the alignments, depending on the location of the foreground branch. Such a high rate of false positives was attributed to the mismatch between the process used to generate the data compared to the process assumed by the null model M1 [85].

The branch-site model was subsequently modified to allow $0 < \omega_0 < 1$ instead of $\omega_0 = 0$ (Modified Model A in [86]). Furthermore, the new null model is specified under the assumption that some proportion p_0 of sites (the stringent sites) evolved under stringent selection with $0 < \omega_0 < 1$ everywhere in the tree except on the foreground branch, where those same sites evolved neutrally with $\omega_2 = 1$. All other sites in the alignment (the neutral sites) are assumed to have evolved neutrally with $\omega_1 = 1$ everywhere in the tree. This is contrasted with the Modified Model A, which assumes that some of the stringent sites and some of the neutral sites evolved under positive selection with $\omega_2 > 1$ on the foreground. Hence, unlike the original omnibus test that contrasts M1 with Model A, the new test contrasts Modified Model A with $\omega_2 = 1$ against Modified Model A with $\omega_2 > 1$. These changes to the YN-BSM were shown to mitigate the problem of false inference. For example, using the same generating model with regimes X and Z, the modified omnibus test falsely inferred positive selection in only 1–7.5% of the alignments, consistent with the 5% level of significance of the test [86].

This case study demonstrates how problems associated with model misspecification were traditionally identified, and how they could be completely corrected through relatively minor changes to the model. However, the generating methods employed by studies such as Zhang [85] and Zhang et al. [86], although sophisticated for their time, produced alignments that were highly unrealistic compared to real data. For example, it was recently shown that a substantial proportion of variation in many real alignments might be due to selection effects associated with shifting balance over static site-specific fitness landscapes [25, 26]. This process results in random changes in site-specific rate ratios, or heterotachy, that cannot be replicated using traditional CSMs as the generating model. While the mitigation of statistical pathologies due to low information content (e.g., using BEB or SBA) or model misspecification (e.g., by altering the null and alternative hypotheses or the use of penalized likelihood) were critical advancements during Phase I of CSM development, other statistical pathologies went unrecognized due to reliance on unrealistic simulation methods. This issue is taken up in the next section.

4 Phase II: Advanced CSMs

A typical protein-coding gene evolves adaptively only episodically [59]. The evidence of adaptive evolution of this type can be very difficult to detect. For example, it is assumed under the YN-BSM that a random subset of sites switched from a stringent or neutral selection regime to positive selection together on the same set of foreground branches. The power to detect a signal of this kind can be very low when the proportion of sites that switched together is small [77]. Perhaps encouraged by the reliability of Phase I models demonstrated by extensive simulation studies [2, 3, 29, 31, 37, 70, 77, 82, 85, 86], combined with experimental validation of results obtained from their application to real data [1, 71, 76], investigators began to formulate increasingly complex and parameter-rich CSMs [31, 41, 48, 50, 55, 64, 65]. The hope was that carefully selected increases in model complexity would yield greater power to detect subtle signatures of positive selection overlooked by Phase I models. The introduction of such CSMs marks the beginning of Phase II of their historical development.

Phase II models fall into three broad categories:

1. The first consists of Phase I CSMs modified to account for more variability in selection effects across sites and branches than previously assumed, with the aim of increasing the power to detect subtle signatures of positive selection (e.g., the branch-site random effects likelihood model, BSREL; [31]).

2. The second category includes Phase I CSMs modified to contain parameters for mechanistic processes not directly associated with selection effects. Many such models have been motivated by a particular interest in the added mechanism (e.g., the fixation of double and triple mutations; [26, 40, 83]), or by the notion that increasing the mechanistic content of a CSM can only improve inferences about selection effects (e.g., by accounting for variations in the synonymous substitution rate; [30, 51]).
3. The third category of models abandons the traditional formulation of Eq. 1 in favor of a substitution process expressed in terms of explicit population genetic parameters, such as population size and selection coefficients [45, 48–50, 64, 65].

An example of the first category of models is BSREL, which accounts for variations in selection effects across sites and over branches by assuming a different rate-ratio distribution $\{(\omega_i^b, p_i^b) : i = 1, \ldots, k_b)\}$ for each branch b of a tree [31]. BSREL was later found to be more complex than necessary, so an adaptive version was formulated to allow the number of components k_b on a given branch to adjust to the apparent complexity of selection effects on that branch (aBSREL; [55]). A further reduction in model complexity led to the formulation of the test known as BUSTED (for branch-site unrestricted statistical test for episodic diversification; [41]), which we use to illustrate the problem of confounding in Case Study C. An example of the second category of models is the addition of parameters for the rate of double and triple mutations to traditional CSMs, the most sophisticated version of which is RaMoSSwDT (for Random Mixture of Static and Switching sites with fixation of Double and Triple mutations; [26]). This model is used in Case Study D to illustrate the problem of phenomenological load.

Models in the third category are the most ambitious CSMs currently in use, and are far more challenging to fit to real alignments than traditional models. One of the most impressive examples of their application is the site-wise mutation-selection model (swMutSel; [64, 65]) fitted to a concatenated alignment of 12 mitochondrial genes (3598 codon sites) from 244 mammalian species. Based on the mutation-selection framework of Halpern and Bruno [19], swMutSel estimates a vector of selection coefficients for each site in an alignment. This and similar models (e.g., [48–50]) appear to be reliable [58], but require a very large number of taxa (e.g., hundreds). Phase II models of this category are therefore impractical for the majority of empirical datasets. Here, we utilize MutSel as an effective means to generate realistic alignments with plausible

levels of variation in selection effects across sites and over time rather than as a tool of inference.

4.1 Case Study C: Confounding

By expressing the codon substitution process in terms of explicit population genetic parameters, the MutSel framework facilitates the investigation of complex evolutionary dynamics, such as shifting balance on a fixed fitness landscape or adaptation to a change in selective constraints (i.e., a peak shift; [6, 25]) that are missing from alignments generated using traditional methods. Specifically, by assigning a different vector of fitness coefficients for the 20 amino acids to each site, MutSel can generate more variation in rate ratio across sites and over time than has been realized in the past simulation studies (e.g., Table 1). In this way, MutSel provides the basis of a generating model that can be adjusted to produce alignments that closely mimic real data [26]. MutSel therefore serves to connect demonstrably plausible evolutionary dynamics to the pathology we refer to as confounding.

Under MutSel, the dynamic regime at the hth codon site (e.g., shifting balance, neutral, nearly neutral, or adaptive evolution) is uniquely specified by a vector of fitness coefficients $\mathbf{f}^h = f_1^h, \ldots, f_m^h$. It is generally assumed that mutation to any of the three stop codons is lethal, so $m = 61$ for nuclear genes and $m = 60$ for mitochondrial genes. And, although it is not a requirement, it is typical to assume that the f_j^h are constant across synonymous codons [25, 57]. Given $\mathbf{f}^h$, the elements of a site-specific instantaneous rate matrix A^h can be defined as follows for all $i \neq j$ (cf. Eq. 1):

$$A_{ij}^h \propto \begin{cases} \mu_{ij} & \text{if } s_{ij}^h = 0 \\ \mu_{ij} \dfrac{s_{ij}^h}{1 - \exp\left(-s_{ij}^h\right)} & \text{otherwise} \end{cases} \tag{5}$$

where μ_{ij} is the rate at which codon i mutates to codon j and $s_{ij}^h = 2N_e(f_j^h - f_i^h)$ is the scaled selection coefficient for a population of haploids with effective population size N_e. The probability that the new mutant j is fixed is approximated by $s_{ij}^h/\{1 - \exp(-s_{ij}^h)\}$ [9, 28].

The rate matrix A^h defines the dynamic regime for the site as illustrated in Fig. 3. The bar plot shows codon frequencies $\boldsymbol{\pi}^h = \pi_1^h, \ldots, \pi_m^h$ sorted in descending order. A site spends most of its time occupied by codons to the left or near the "peak" of its landscape. The codon-specific rate ratio for the site (dN_i^h/dS_i^h for codon i) is low near the peak (red line plot in Fig. 3) since mutations away from the peak are seldom fixed. However, if selection is not too stringent, the site will occasionally drift to the right into the "tail" of its landscape. When this occurs, the codon-specific rate

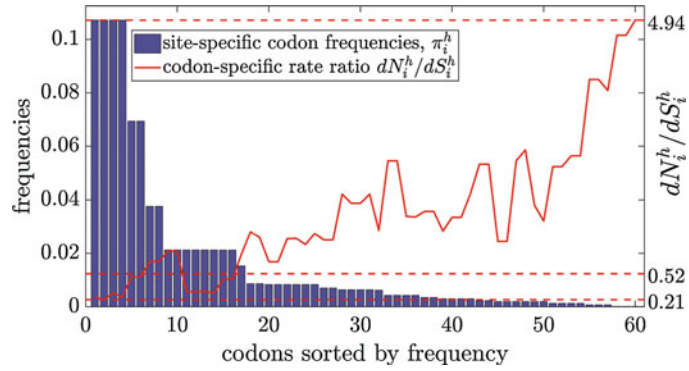

Fig. 3 Fitness coefficients for the 20 amino acids were drawn from a normal distribution centered at zero and with standard deviation $\sigma = 0.001$. Bars show the resulting stationary frequencies (a proxy for fitness) sorted from largest to smallest. They compose a metaphorical site-specific landscape over which the site is imagined to move. The solid red line shows the codon-specific rate ratio dN_i^h/dS_i^h for the sorted codons. This varies depending on the codon currently occupying the site, and can be greater than one following a chance substitution into the tail (to the right) of the landscape. In this case, the codon-specific rate ratio for the site ranged from 0.21 to 4.94 with a temporally averaged site-specific rate ratio of $dN^h/dS^h = 0.52$

ratio will be elevated for a time until a combination of drift and positive selection moves the site back to its peak. This dynamic between selection and drift is reminiscent of Wright's shifting balance. It implies that, when a population is evolving on a fixed fitness landscape (i.e., with no adaptive evolution), its gene sequences can nevertheless contain signatures of temporal changes in site-specific rate ratios (heterotachy), and that these might include evidence of transient elevation to values greater than one (i.e., positive selection). Such signatures of positive selection due to shifting balance can be detected by Phase II CSMs [25].

For example, BUSTED [41] was developed as an omnibus test for episodic adaptive evolution. The underlying CSM was formulated to account for variations in the intensity of selection over both sites and time modeled as a random effect. This is in contrast to the YN-BSM, which treats temporal changes in rate ratio as a fixed effect that occurs on a prespecified foreground branch (although the sites under positive selection are still a random effect). We therefore refer to the CSM underlying BUSTED as the random effects branch-site model (RE-BSM) to serve as a reminder of this important distinction. Under RE-BSM, the rate ratio at each site and branch combination is assumed to be an independent draw from the distribution $\{(\omega_0, p_0), (\omega_1, p_1), (\omega_2, p_2)\}$. In this way, the model accounts for variations in selection effects both across sites and over time. BUSTED contrasts the null hypothesis that $\omega_0 \leq \omega_1 \leq \omega_2 = 1$ with the alternative that $\omega_0 \leq \omega_1 \leq 1 \leq \omega_2$.

When applied to real data, rejection of the null is interpreted as evidence of episodic adaptive evolution.

Unlike the YN-BSM that aims to detect a subset of sites that underwent adaptive evolution together on the same foreground branches (i.e., coherently), BUSTED was designed to detect heterotachy similar to the type predicted by the mutation-selection framework: shifting balance on a static fitness landscape. Jones et al. [25] recently demonstrated that plausible levels of shifting balance can produce signatures of episodic positive selection that can be detected. BUSTED inferred episodic positive selection in as many as 40% of alignments generated using the MutSel framework. Significantly, BUSTED was correct to identify episodic positive selection in these trials. Even though the generating process assumed fixed site-specific landscapes (so there was no episodic adaptive evolution), and the long-run average rate ratio at each site was necessarily less than one [57], positive selection nevertheless did sometimes occur by shifting balance. This illustrates the general problem of confounding. Two processes are said to be confounded if they can produce the same or similar patterns in the data. In this case, episodic adaptive evolution (i.e., the evolutionary response to changes in site-specific landscapes) and shifting balance (i.e., evolution on a static fitness landscape) are confounded because they can both produce rate-ratio distributions that indicate episodic positive selection. The possibility of confounding underlines the fact that there are limitations in what can be inferred about evolutionary processes based on an alignment alone.

4.2 Case Study D: Phenomenological Load

Phenomenological load (PL) is a statistical pathology related to both model misspecification (Case Study B) and confounding (Case Study C) that was not recognized during Phase I of CSM development. When a model parameter that represents a process that played no role in the generation of an alignment (i.e., a misspecified process) nevertheless absorbs a significant amount of variation, its MLE is said to carry PL [26]. This is more likely to occur when the misspecified process is confounded with one or more other processes that did play a role in the generation of the data, and when a substantial proportion of the total variation in the data is unaccommodated by the null model [26]. PL increases the probability that a hypothesis test designed to detect the misspecified process will be statistically significant (as indicated by a large LLR) and can therefore lead to the incorrect conclusion that the misspecified process occurred. Critically, Jones et al. [26] showed that PL was only detected when model contrasts were fitted to data generated with realistic evolutionary dynamics using the MutSel model framework.

To illustrate the impact of PL, we consider the case of CSMs modified to detect the fixation of codons following simultaneous double and triple (DT) nucleotide mutations. The majority of

CSMs currently in use assume that codons evolve by a series of single-nucleotide substitutions, with the probability for DT changes set to zero. However, recent model-based analyses have uncovered evidence for DT mutations [32, 68, 83]. Early estimates of the percentage of fixed mutations that are DT were perhaps unrealistically high. Kosiol et al. [32], for example, estimated a value close to 25% in an analysis of over 7000 protein families from the Pandit database [69]. Alternatively, when estimates were derived from a more realistic site-wise mutation-selection model, DT changes comprised less than 1% of all fixed mutations [64]. More recent studies suggest modest rates of between 1% and 3% [5, 20, 27, 53]. Whatever the true rate, several authors have argued that it would be beneficial to introduce a few extra parameters into a standard CSM to account for DT mutations (e.g., [40, 83]). The problem with this suggestion is that episodic fixation of DT mutations can produce signatures of heterotachy consistent with shifting balance.

Recall the comparison of M1, a CSM containing parameters represented by the vector θ_1, and M2, the same model but for the inclusion of one additional parameter ψ, so that $\theta_2 = (\theta_1, \psi)$. The parameter ψ will reduce the deviance of M2 compared to M1 by some proportion of the baseline deviance between the simplest CSM (M0) and the saturated model $P_S(\hat{\theta}_S)$. We call this the percent reduction in deviance (PRD) attributed to $\hat{\psi}$:

$$\mathrm{PRD}(\hat{\psi}) = \frac{\Delta D(\hat{\theta}_{M1}, \hat{\theta}_{M2})}{\Delta D(\hat{\theta}_{M0}, \hat{\theta}_S)} \tag{6}$$

Suppose M1 and M2 were fitted to an alignment and that the $\mathrm{LLR} = \Delta D(\hat{\theta}_{M1}, \hat{\theta}_{M2})$ was found to be statistically significant. This would lead an analyst to attribute the $\mathrm{PRD}(\hat{\psi})$ to real signal for the process ψ was meant to represent, possibly combined with some PL and noise. Now, consider the case in which the process represented by ψ did not actually occur (i.e., it was not a component of the true generating process). Under this scenario, $\mathrm{PRD}(\hat{\psi})$ would contain no signal, but would be entirely due to PL plus noise. When this is known to be the case, we set $\mathrm{PRD}(\hat{\psi}) = \mathrm{PL}(\hat{\psi})$. As illustrated below, $\mathrm{PL}(\hat{\psi})$ can be large enough to result in rejection of the null, and therefore lead to a false conclusion about the data generating process.

We illustrate PL by contrasting the model RaMoSS with a companion model RaMoSSwDT that accounts for the fixation of DT mutations via two rate parameters, α (the double mutation rate) and β (the triple mutation rate) [26]. RaMoSS combines the standard M-series model M3 with the covarion-like model CLM3 (cf., [12, 18]). Specifically, RaMoSS mixes (with proportion p_{M3}) one model with two rate-ratio categories $\omega_0 < \omega_1$ that are constant over the entire tree with a second model (with proportion

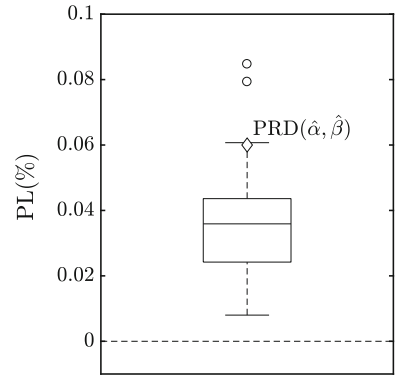

Fig. 4 The box plot depicts the distribution of the phenomenological load (PL) carried by $(\hat{\alpha}, \hat{\beta})$ produced by fitting the RaMoSS vs RaMoSSwDT contrast to 50 alignments generated under MutSel-mmtDNA: the circles represent outliers of this distribution. The diamond is the percent reduction in deviance for the same parameters estimated by fitting RaMoSS vs RaMoSSwDT to the real mtDNA alignment

$p_{\mathrm{CLM3}} = 1 - p_{\mathrm{M3}}$) under which sites switch randomly in time between $\omega_0' < \omega_1'$ at an average rate of δ switches per unit branch length. Fifty alignments were simulated to mimic a real alignment of 12 concatenated H-strand mitochondrial DNA sequences (3331 codon sites) from 20 mammalian species as distributed in the PAML package [73]. The generating model, MutSel-mmtDNA [26], was based on the mutation-selection framework and produced alignments with single-nucleotide mutations only. Since DT mutations are not fixed under MutSel-mmtDNA, the PRD carried by $(\hat{\alpha}, \hat{\beta})$ in each trial can be equated to PL (plus noise). The resulting distribution of PL$(\hat{\alpha}, \hat{\beta})$ is shown as a boxplot in Fig. 4.

Although DT mutations were not fixed when the data was generated, shifting balance on a static landscape can produce similar site patterns as a process that includes rare fixation of DT mutations (site patterns exhibiting both synonymous and nonsynonymous substitutions; [26]).[3] DT and shifting balance are therefore confounded. And since shifting balance tends to occur at a substantial proportion (approximately 20%) of sites when an alignment is generated under MutSel-mmtDNA, DT mutations were falsely inferred by the LRT in 48 of 50 trials at the 5% level of significance (assuming LLR $\approx \chi^2_2$ for the two extra parameters α and β in RaMoSSwDT compared to RaMoSS). The PRD $(\hat{\alpha}, \hat{\beta})$ when RaMoSS vs RaMoSSwDT was fitted to the real mmtDNA is

[3] It has previously been noted that the rapid fixation of compensatory mutations following substitution to an unstable base pair (e.g., AT→GT→GC) can also produce site patterns that suggest fixation of DT mutations [74, p. 46].

shown as a diamond in the same plot. Although $(\hat{\alpha}, \hat{\beta})$ estimated from the real mmtDNA were found to be highly significant (LLR = 84, p-value $<< 0.001$), the PRD$(\hat{\alpha}, \hat{\beta})$ was found to be just under the 95th percentile of PL$(\hat{\alpha}, \hat{\beta})$ (PRD = 0.060% compared to the 95th percentile of PL = 0.061). The evidence for DT mutations in the real data is therefore only marginal, and it is reasonable to suspect that its PRD$(\hat{\alpha}, \hat{\beta})$, if not entirely the result of PL, is at least partially caused by PL.

5 Discussion

CSMs have been subjected to a certain degree of censure, particularly during Phase I of their development [11, 22, 23, 46, 60–63, 85]. We maintain that it is not the model in and of itself, or the maximum likelihood framework it is based on, that gives rise to statistical pathologies, but the relationship between model and data. This principle was illustrated by our analysis of the history of CSM development, which we divided into two phases. Phase I was characterized by the formulation of models to account for differences in selection effects across sites and over time that comprise the major component of variation in an alignment. Starting with M0, such models represent large steps toward the fitted saturated model in Fig. 2, and also provide a better representation of the true generating process. The main criticism of Phase I models was the possibility of falsely inferring positive selection in a gene or at an individual codon site [62, 63, 85]. But, the most compelling empirical case of false positives was shown to be the result of inappropriate application of a complex model to a sparse alignment [63]. Methods for identifying (bootstrap) and dealing with (BEB, SBA, and PLRT) low information content were illustrated in Case Study A.

The other big concern that arose during Phase I development was the possibility of pathologies associated with model misspecification. The method used to identify such problems was to fit a model to alignments generated under a scenario contrived to be challenging, as illustrated in Case Study B. There, the omnibus test based on Model A of the YN-BSM was shown to result in an excess of false positives when fitted to alignments simulated using the implausible but difficult "XZ" generating scenario (e.g., with complete relaxation of selection pressure at all sites on one branch of the tree; Table 1). Subsequent modifications to the test reduced the false positive rate to acceptable values. Hence, Case Study B underlines the importance of the model–data relationship. However, it is not clear whether a model adjusted to suit an unrealistic data-generating process is necessarily more reliable when fitted to a real alignment. This difficulty highlights the need to find ways, for the

purpose of model testing and adjustment, to generate alignments that mimic real data as closely as possible.

Confidence in the CSM approach, combined with the exponential increase in the volume of genetic data and the growth of computational power, spurred the formulation CSMs of ever-increasing complexity during Phase II. The main issue with these models, which has not been widely appreciated, is confounding. Two processes are confounded if they can produce the same or similar patterns in the data. It is not possible to identify such processes when viewed through the narrow lens of an alignment (i.e., site patterns) alone. This was illustrated by Case Study C, where shifting balance on a static landscape was shown to be confounded with episodic adaptive evolution [7, 25]. Confounding can lead to what we call phenomenological load, as demonstrated in Case Study D. In that analysis, the parameters (α, β) were assigned a specific mechanistic interpretation, the rate at which double and triple mutations arise. It was shown that (α, β) can absorb variations in the data caused by shifting balance; hence, the MLEs $(\hat{\alpha}, \hat{\beta})$ resulted in a significant reduction in deviance in 48/50 trials (Fig. 4), and therefore improved the fit of the model to the data. However, the absence of DT mutations in the generating process invalidated the intended interpretation of $(\hat{\alpha}, \hat{\beta})$. This result underlines that a better fit does not imply a better mechanistic representation of the true generating process.

It is natural to assume that a better mechanistic representation of the true generating process can be achieved by adding parameters to our models to account for more of the processes believed to occur. The problem with this assumption is that the metric of model improvement under ML (reduction in deviance) is independent of mechanism. A parameter assigned a specific mechanistic interpretation is consequently vulnerable to confounding with other processes that can produce the same distribution of site patterns. As CSMs become more complex, it seems likely that the opportunity for confounding will only increase. It would therefore be desirable to assess each new model parameter for this possibility using something like the method shown in Fig. 4 whenever possible. The idea is to generate alignments using MutSel or some other plausible generating process in such a way as to mimic the real data as closely as possible, but with the new parameter set to its null value. To provide a second example, consider the test for changes in selection intensity in one clade compared to the remainder of the tree known as RELAX [67]. Under this model, it is assumed that each site evolved under a rate ratio randomly drawn from $\boldsymbol{\omega}_R = \{\omega_1, \ldots, \omega_k\}$ on a set of prespecified reference branches, and from a modified set of rate ratios $\boldsymbol{\omega}_T = \{\omega_1^m, \ldots, \omega_k^m\}$ on test branches, where m is an exponent. A value $0 < m < 1$ moves the rate ratios in $\boldsymbol{\omega}_T$ closer to one compared to their corresponding values in $\boldsymbol{\omega}_R$, consistent with relaxation of selection pressure at all sites on the test

branches. Relaxation is indicated when the contrast of the null hypothesis that $m = 1$ versus the alternative that $m < 1$ is statistically significant. The distribution of PL($\hat{m}$) can be estimated from alignments generated with $m = 1$. The PRD($\hat{m}$) estimated from the real data can then be compared to this to assess the impact of PL (cf. Fig. 4). This approach is predicated on the existence of a generating model that could have plausibly produced the site patterns in the real data. Jones et al. [26] present a variety of methods for assessing the realism of a simulated alignment, although further development of such methods is warranted. Software based on MutSel is currently available for generating data that mimic large alignments of 100-plus taxa (Pyvolve; [56]). Other methods have been developed to mimic smaller alignments of certain types of genes (e.g., MutSel-mmtDNA; [25]). It is only by the use of these or other realistic simulation methods that the relationship between a given model and an alignment can be properly understood.

References

1. Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26:255–271
2. Anisimova M, Bielawski JP, Yang ZH (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18:1585–1592
3. Anisimova M, Bielawski JP, Yang ZH (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958
4. Bielawski JP, Yang ZH (2004) A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol Evol 59:121–132
5. De Maio N, Holmes I, Schlötterer C, Kosiol C (2013) Estimating empirical codon hidden Markov models. Mol Biol Evol 30:725–736
6. dos Reis M (2013). http://arxiv:1311.6682v1. Last accessed 26 Nov 2013
7. dos Reis M (2015) How to calculate the non-synonymous to synonymous rate ratio protein-coding genes under the Fisher-Wright mutation-selection framework. Biol Lett 11:1–4.
8. Field SF, Bulina MY, Kelmanson IV, Bielawski JP, Matz MV (2006) Adaptive evolution of multicolored fluorescent proteins in reef-building corals. J Mol Evol 62:332–339
9. Fisher R (1930) The distribution of gene ratios for rare mutations. Proc R Soc Edinb 50:205–220
10. Forsberg R, Christiansen FB (2003) A codon-based model of host-specific selection in parasites, with an application to the influenza a virus. Mol Biol Evol 20:1252–1259
11. Friedman R, Hughes AL (2007) Likelihood-ratio tests for positive selection of human and mouse duplicate genes reveal nonconservative and anomalous properties of widely used methods. Mol Phylogenet Evol 542:388–393
12. Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866–873
13. Gaston D, Susko E, Roger AJ (2011) A phylogenetic mixture model for the identification of functionally divergent protein residues. Bioinformatics 27:2655–2663
14. Gibbs RA (2007) Evolutionary and biomedical insights from the Rhesus macaque genome. Science 316:222–234
15. Goldman N (1993) Statistical tests of models of DNA substitution. J Mol Evol 36:182–198
16. Goldman N, Yang ZH (1994) Codon-based model of nucleotide substitution for protein-coding DNA-sequences. Mol Biol Evol 11:725–736
17. Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 862–864

18. Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci USA 101:12957–12962
19. Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15:910–917
20. Harris K, Nielsen R (2014) Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Res 9:1445–1554
21. Huelsenbeck JP, Dyer KA (2004) Bayesian estimation of positively selected sites. J Mol Evol 58:661–672
22. Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99:364–373
23. Hughes AL, Friedman R (2008) Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. Immunogenetics 60:495–506
24. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class-1 loci reveals overdominant selection. Nature 335:167–170
25. Jones CT, Youssef N, Susko E, Bielawski JP (2017) Shifting balance on a static mutation-selection landscape: a novel scenario of positive selection. Mol Biol Evol 34:391–407
26. Jones CT, Youssef N, Susko E, Bielawski JP (2018) Phenomenological load on model parameters can lead to false biological conclusions. Mol Biol Evol 35:1473–1488
27. Keightley P, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter M (2009) Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. Genet Res 19:1195–1201
28. Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47:713–719
29. Kosakovsky Pond SL, Frost SDW (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208–1222
30. Kosakovsky Pond SL, Muse SV (2007) Site-to-site variations of synonymous substitution rates. Mol Biol Evol 22:2375–2385
31. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K (2011) A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28:3033–3043
32. Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24:1464–1479
33. Kosiol C, Vinař T, daFonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six mammalian genomes. PLoS Genet 4:1–17
34. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109
35. Liberles DA, Teufel AI, Liu L, Stadler T (2013) On the need for mechanistic models in computational genomics and metagenomics. Genome Biol Evol 5:2008–2018
36. Lopez P, Casane D, Phillipe H (2002) Heterotachy, and important process of protein evolution. Mol Biol Evol 19:1–7
37. Lu A, Guindon S (2013) Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. Mol Biol Evol 31:484–495
38. Mingrone J, Susko E, Bielwaski JP (2016) Smoothed bootstrap aggregation for assessing selection pressure at amino acid sites. Mol Biol Evol 33:2976–2989
39. Mingrone J, Susko E, Bielwaski JP (2018) Modified likelihood ratio tests for positive selection (submitted). Bioinformatics, Advance Access https://doi.org/10.1093/bioinformatics/bty1019
40. Miyazawa S (2011) Advantages of a mechanistic codon substitution model for evolutionary analysis of protein-coding sequences. PLoS ONE 6:20
41. Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, Scheffler K, Pond SLK (2015) Gene-wide identification of episodic selection. Mol Biol Evol 32:1365–1371
42. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. Mol Biol Evol 11:715–724
43. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426
44. Nielsen R, Yang ZH (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936
45. Nielsen R, Yang Z (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol 20:1231–1239
46. Nozawa M, Suzuki Y, Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci USA 106:6700–6705

47. Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571–581
48. Rodrigue N, Lartillot N (2014) Site-heterogeneous mutation-selection models with the PhyloBayes-MPI package. Bioinformatics 30:1020–1021
49. Rodrigue N, Lartillot N (2016) Detection of adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation-selection codon substitution model. Mol Biol Evol 34:204–214
50. Rodrigue N, Philippe H, Lartillot N (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc Natl Acad Sci USA 107:4629–4634
51. Rubinstein ND, Doron-Faigenboim A, Mayrose I, Pupko T (2011) Evolutionary model accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. Mol Biol Evol 28:3297–3308
52. Sawyer SL, Emerman M, Malik HS (2007) Discordant evolution of the adjacent antiretroviral genes trim22 and trim5 in mammals. PLoS Pathog 3:e197
53. Schrider D, Hourmozdi J, Hahn M (2014) Pervasive multinucleotide mutational events in eukaryotes. Curr Biol 21:1051–1054
54. Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions. J Am Stat Assoc 82:605–610
55. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Pond SLK (2015) Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol 32:1342–1353
56. Spielman S, Wilke CO (2015) Pyvolve: a flexible Python module for simulating sequences along phylogenies. PLoS ONE 10:1–7
57. Spielman S, Wilke CO (2015) The relationship between dN/dS and scaled selection coefficients. Mol Biol Evol 34:1097–1108
58. Spielman S, Wilke CO (2016) Extensively parameterized mutation-selection models reliably capture site-specific selective constraints. Mol Biol Evol 33:2990–3001
59. Struder RA, Robinson-Rechavi M (2009) Evidence for an episodic model of protein sequence evolution. Biochem Soc Trans 37:783–786
60. Suzuki Y (2008) False-positive results obtained from the branch-site test of positive selection. Genes Genet Syst 83:331–338
61. Suzuki Y, Nei M (2001) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol Biol Evol 18:2179–2185
62. Suzuki Y, Nei M (2002) Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol Biol Evol 19:1865–1869
63. Suzuki Y, Nei M (2004) False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom Thalassiosira weissflogii and the tax gene of the human T-cell lymphotropic virus. Mol Biol Evol 21:914–921
64. Tamuri AU, dos Reis M, Goldstein RA (2012) Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. Genetics 190:1101–1115
65. Tamuri AU, Goldman N, dos Reis M (2014) A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. Genetics 197:257–271
66. Wang H, Li K, Susko E, Rodger AJ (2008) A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol 8:1–13
67. Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K (2014) Relax: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol 32:820–832
68. Whelan S, Goldman N (2004) Estimating the frequency of events that cause multiple-nucleotide changes. Genetics 167:2027–2043
69. Whelan S, de Bakker PIW, Quevillon E, Rodriguez N, Goldman N (2006) Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Res 34(Database issue): D327–D331
70. Wong WSW, Yang ZH, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168:1041–1051
71. Yang ZH (2005) The power of phylogenetic comparison in revealing protein function. Proc Natl Acad Sci USA 102:3179–3180
72. Yang ZH (2006) On the varied pattern of evolution in 2 fungal genomes: a critique of Hughes and Friedman. Mol Biol Evol 23:2279–2282
73. Yang ZH (2007) PAML4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

74. Yang ZH (2014) Molecular evolution: a statistical approach. Oxford University Press, Oxford
75. Yang ZH (2017) PAML: phylogenetic analysis by maximum likelihood. http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf
76. Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503
77. Yang ZH, dos Reis M (2011) Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28:1217–1228
78. Yang ZH, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46: 409–418
79. Yang ZH, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917
80. Yang ZH, Nielsen R (2007) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25:568–579
81. Yang ZH, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449
82. Yang ZH, Wong SWS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118
83. Zaheri M, Dib L, Salamin N. (2014) A generalized mechanistic codon model. Mol Biol Evol 31:2528–2541
84. Zhai W, Nielsen R, Goldman N, Yang ZH (2012) Looking for Darwin in genomic sequences – validity and success of statistical methods. Mol Biol Evol 20:2889–2893
85. Zhang J (2004) Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol 21:1332–1339
86. Zhang J, Nielsen R, Yang ZH (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479