Trends in Genetics

CellPress REVIEWS

Review

Studying Natural Selection in the Era of Ubiquitous Genomes

Timothy B. Sackton[1,*]

A major goal of comparative genomics research is modeling changes in DNA sequences between species to understand the evolutionary forces acting on species differences. Application of these models to a number of species over the past decade has revealed some commonalities across organisms, most notably a consistent role of positive selection in shaping the molecular evolution of the immune system. However, models of DNA sequence evolution also have important limitations that are increasingly being recognized, including issues with data quality and biases caused by simplifying assumptions. While new approaches have begun to address these challenges, ultimately, additional data, such as resequencing data from populations, will provide more power to fully understand the unique evolutionary forces acting on different species. In this review, I summarize the conclusions of recent genome-wide studies of selection, highlight some important challenges to applying these methods to large data sets, and discuss ways forward for the field.

Highlights

Studies of natural selection across many available genomes consistently identify genes in the immune system as frequently enriched for positive selection at the molecular level.

Attempts to identify specific targets of positive selection in particular lineages have been less successful, in part due to the lack of power of comparative genomic methods to detect lineage-specific selection.

As data sets grow in size, the challenge of ensuring high-quality alignments for molecular evolution models (necessary to avoid false positives) is growing, although new computational tools for the quality control of sequence alignments will help.

Larger data sets also reveal that some key assumptions of codon-based models are unrealistic in real data, with potentially substantial impacts on the robustness of conclusions from these methods.

Genomics and the Study of Adaptive Protein Evolution

A major concern in biology is understanding the evolutionary forces that are responsible for genetic differences between species. While there is considerable debate about the relative proportion of amino acid substitutions fixed between species by **positive selection** (see Glossary) and **genetic drift** [1,2], it is undeniable that at least a subset of proteins in most genomes, studied to-date, experience rapid adaptive evolution [3–7]. The identity of these proteins, the similarities (and differences) between species and clades, and the specific regions in particular proteins subject to positive selection can be informative about biological interactions shaping both clade-wide and lineage-specific evolution. While early studies were limited by the small number of available genomes, the explosive growth in assembled genomes in recent years [8–10] motivates a re-evaluation of biological conclusions from studies of positive selection in protein-coding genes across species and a look forward at outstanding challenges and issues in the field.

There are many approaches to detect positive selection in protein-coding genes, which focus on different timescales and requiring different types of data (Box 1). In this review, I primarily focus on phylogenetic methods from comparative genomics that can detect positive selection occurring over medium (between related species) and long (across an entire clade) timescales. These approaches, such as are implemented in the software **Phylogenetic Analysis by Maximum Likelihood** (PAML) [11] and **Hypothesis Testing using Phylogenies** (HyPhy) [12–14], use **codon-based models** of molecular evolution [15] to infer particular genes, codons, or lineages on a phylogeny under selection. These models use the number of amino acid (or **nonsynonymous**) substitutions (**dN**) relative to the number of neutral substitutions (typically approximated by the number of substitutions at **synonymous** sites, **dS**). The dN/dS ratio (often referred to as ω) is used as a measure of selection, where $\omega > 1$ is characteristic of an evolutionary model where certain amino acid substitutions are beneficial and, thus, have a higher

[1] Informatics Group, Harvard University, Cambridge, MA 02138, USA

*Correspondence: tsackton@g.harvard.edu (T.B. Sackton).

https://doi.org/10.1016/j.tig.2020.07.008

Box 1. Detecting Selection in Protein Coding Genes

Methods to detect the action of natural selection in protein-coding genes can be distinguished by the types of data they rely on and the timescale over which they operate. One major approach, first proposed by McDonald and Kreitman [108], relies on a comparison of polymorphism and divergence data to detect an excess of nonsynonymous substitutions relative to the expectation under neutrality, typically on short to moderate timescales (e.g., between closely related species). This approach and its extensions (reviewed in [113]) rely on the idea that, within populations, most or all segregating mutations will be neutral, while both neutral and adaptive mutations can fix between species. While these assumptions are not perfectly met, modifications to the test framework can account for this and give high power to detect lineage-specific adaptation [117].

An alternate approach is based on comparing rates of protein evolution to a neutral standard, such as rates of synonymous site evolution [83]. In this approach, a potentially large alignment of protein-coding genes is compared, allowing these methods to identify common patterns of natural selection across large phylogenies and relatively deep (clade-wide) timescales. An excess of nonsynonymous substitutions (as indicated by a dN/dS ratio >1) is indicative of some evolutionary process favoring fixations of nonsynonymous mutations, typically adaptive protein evolution. Several extensions to these models, allowing variation in dN/dS either among branches, across sites in a protein, or both, can improve power to detect positive selection that acts only on subsets of species or codons (reviewed in more detail in [118]).

probability of fixation compared with neutral mutations. Notably, these methods require only a single representative genome from each species in a phylogeny and, thus, have been particularly popular in the era of comparative genomics.

Codon-based models have been commonly used for two related, but distinct goals. By comparing protein sequences across many different species from particular groups, it is possible to ask questions about common threads of adaptive evolution. Typically, this is done using models that allow dN/dS to vary across sites in a protein, but not necessarily among lineages; these are referred to as **site models** (Box 2) [12,16,17]. Site models typically are best at finding **episodic selection**, where the same proteins and amino acid positions are subject to repeated adaptive evolution in different species, as is seen, for example, in the evolution of immune system genes in many species [3,7,18].

Box 2. Site-Based versus Lineage-Based Models of Molecular Evolution

Standard codon-based models can be divided into different classes, depending on the assumptions they make about variation in ω (dN/dS) among sites and branches.

Site Models

Selection in real proteins typically operates only on a subset of codons; therefore, models that require a single fixed ω for all positions within a gene are extremely conservative. Site models capture this variation by allowing each codon in the gene to have a different ω, typically subject to some modeling constraints. In PAML, this is done by assuming that ω values come from a distribution across the gene, detecting positive selection by comparing a model that allows an extra class with ω >1, to a model that does not [17]. In HyPhy, FUBAR implements a Bayesian model that similarly allows variation in ω across sites, albeit with different assumptions [119].

Branch Models

In some circumstances (e.g., when investigating molecular adaptations that may be responsible for unique lineage-specific biology), a major goal is to estimate lineage-specific rates of protein evolution, while still constraining each gene to a single ω value. These models can either allow two or more dN/dS ratios for a set of prespecified target lineages, or allow each branch to have its own ω.

Branch-Site Models

Branch-site models allow ω to vary among both codons and lineages. In principle, these methods have the most power to detect positive selection, which may only be present in a subset of lineages. In PAML, these methods are typically implemented by specifying a set of foreground and background lineages, and fitting models that allow different distributions of ω among these sets (only allowing a class of sites with ω >1 in the foreground lineages). In HyPhy, several different methods with varying assumptions allow detection of a class of codons in one or more lineages with ω >1, including MEME [16], BUSTED [12], and aBS-REL [13].

Glossary

Adaptive introgression: process whereby a variant from a different population or species increases fitness in the recipient population.

Adaptive radiation: process whereby organisms rapidly diversify from an ancestral species into new descent forms.

Ancestral polymorphisms: segregating variants retained from an ancestral population, typically in the context of shared variant between two descent species or populations.

Annotation errors: problems in alignments of genes that result from errors in identifying the correct gene model in each species, such as by misannotating some intronic sequence as coding sequence.

Branch models: models of molecular evolution that assume the same parameters for all sites in an alignment, but allow these parameters to vary among branches on a tree.

Branch-site models: models of molecular evolution that allow parameters to vary both among sites in an alignment and among branches on a tree.

Codon-based models: models of DNA sequence evolution that parameterize changes among codons, instead of single nucleotides.

dN: number of nonsynonymous changes per nonsynonymous site.

dS: number of synonymous changes per synonymous site.

Episodic selection: selection that occurs repeatedly over time on the same sites and genes.

Genetic drift: process of random sampling of individuals without respective to fitness, possibly leading to substitutions between species that have no fitness consequences.

Gene tree discordance: process of allele sorting during population divergence that produces gene trees that differ from the species tree.

Hidden Markov Model: statistical model where unknown (hidden) states can be described by a stochastic model (Markov chain) where the probability of each hidden state in sequence depends only on the immediately prior state.

Homologous: regions of the genome (nucleotides or genes) that share a common ancestor and, thus, are related by descent.

Hypothesis Testing using Phylogenies (HyPhy): a software

Alternatively, there is considerable interest in the role of natural selection in the evolution of lineage-specific traits that define unique species biology. In some cases, regulatory changes underlie lineage-specific biology (e.g., coat color patterns of rodents [19]), which will be invisible to models of protein evolution. However, in many other cases, protein-coding changes appear important, such as the evolution of gigantism and adaptation to marine life in cetaceans [20], or the adaptation to fresh water of typically marine organisms [21]. To detect lineage-specific protein evolution, where the rate of protein evolution is elevated above the neutral expectation only in one or a few lineages, requires models that allow dN/dS to vary across lineages (**branch models;** Box 2) [22] or that allow dN/dS to vary among sites and lineages (**branch-site models;** Box 2) [13,23,24]. In this review, I first discuss the emerging consensus regarding the types of protein experiencing natural selection in different clades and the properties they share. I then highlight emerging concerns about the reliability of these methods, along with potential solutions.

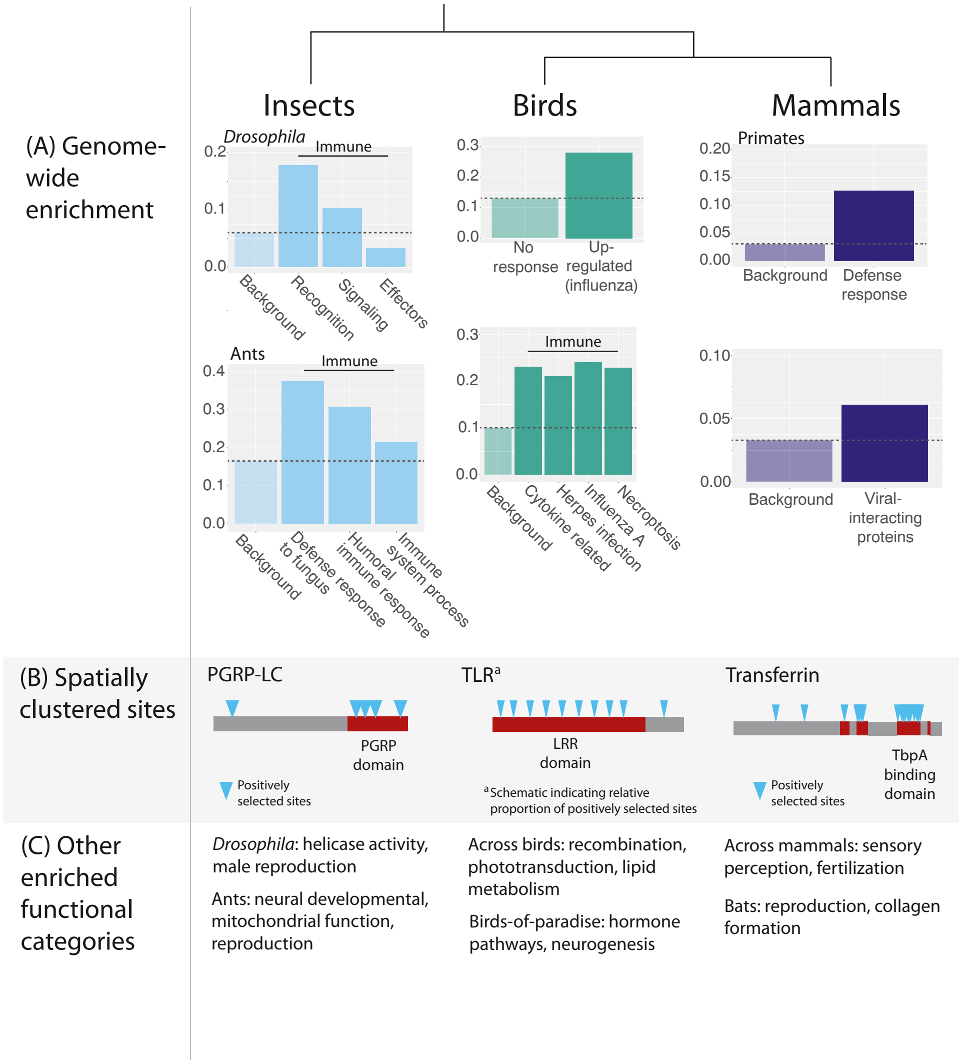
What Have We Learned about Natural Selection from Clade-Wide Studies?

With the first publications of multiple genomes of primates and fruit flies 15 years ago [6,25], an obvious question was whether there were any clear functional categories of genes showing an excess of positive selection across the genome. Genome-wide comparisons using site models of molecular evolution from these early studies showed that both in *Drosophila* [26,27] and mammals [7,28], genes related to immunity and defense, reproduction, and genomic conflict were among the most frequent targets of positive selection in the genome. These early results confirmed candidate gene studies highlighting the importance of intermolecular conflict in driving the rapid evolution of proteins, especially those involved in the interactions between hosts and pathogens [29,30], in interactions between sperm and egg in broadcast-spawning marine organisms [31,32], and in intragenomic conflict between transposable elements and genomic defenses [33].

More recently, a flood of high-quality mammalian genomes has resulted in several studies adding significant nuance and detail to these early conclusions. In a seminal paper using a data set spanning 24 mammalian genomes, Enard and coauthors [3] used multiple methods (including BUSTED and BS-REL from the HyPhy package) to identify signatures of episodic selection across the clade, and showed that proteins that interact with viruses experience at least twice as many adaptive amino acid substitutions as proteins that do not interact with viruses. This signal is pervasive across the genome, found both in components of the host immune response, and also in cell-surface receptors and other proteins used by viruses to enter host cells and proliferate [3]. This pattern is also apparent even in the most conserved subset of the human genome [34]. Signals of positive selection are not limited to viral-interacting proteins: a similar analysis of manually curated proteins interacting with *Plasmodium* (the causative agent of malaria) and other parasites in the order Piroplasmida showed that these proteins are also approximately twice as likely to experience adaptive evolution as matched controls [35]. These studies largely focused on broad mammalian comparisons. However, another recent study with a particular focus on primates revealed largely consistent patterns: genes positively selected across primates are strongly over-represented for both innate and adaptive immunity [36]. Similarly, a study of 18 bat genomes, using PAML site models, found that nearly 20% of the 181 positively selected genes across bats are associated with immune functions, far more than any other functional category [37]. Taken together, these studies point to a dominant role of host–pathogen interactions in driving protein adaptation in mammals (Figure 1).

Candidate gene studies of molecular evolution, largely in mammals, have been important for advancing our understanding of the mechanisms by which these arms race dynamics drive

package for testing models of phylogenetics and molecular evolution.
Nonsynonymous: substitution or mutation that changes the amino acid sequence of a protein-coding gene.
Phylogenetic Analysis by Maximum Likelihood (PAML): software package implementing many commonly used codon-based models of molecular evolution.
Positive selection: process driving the rapid fixation of mutations that confer a fitness advantage to the individual that carries them.
Primary sequencing errors: errors (base-level inaccuracies or other assembly errors) in the underlying assembled DNA sequence, which become alignment errors when these sequences are used to construct an alignment.
Residue: single amino acid in a protein.
Site models: models of molecular evolution that allow parameters to vary among sites in an alignment, but not among branches on a tree.
Synonymous: substitution or mutation that does not change the amino acid sequence of a protein-coding gene.



Trends in Genetics

Figure 1. Clade-Wide Patterns of Positive Selection from Genome-Wide Studies. A summary of studies of genome-wide patterns of positive selection from representative clades of mammals, insects, and birds where sufficient high-quality and genome-wide studies are available. (A) Evidence for enrichment of positive selection in genes with immune function. Each plot shows the proportion of genes in the genome with evidence for positive selection. In all cases, immune categories are significantly higher than non-immune categories. Data plotted are: manually annotated immune genes from *Drosophila* [6]; Gene Ontology categories associated with ant genes [48]; genes upregulated by influenza infection in birds [18]; specific immune-related Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways from birds [18]; Gene Ontology categories associated with primate genes [36]; and inferred excess of positive selection in viral interacting proteins [3]. (B) Examples of spatially clustered positively selected sites in immune proteins in insects, birds, and mammals. Data plotted are: positively selected residues in the pathogen recognition protein PGRP-LC in *Drosophila* [6]; schematic of the relative proportion of positively selected residues in the LRR domain compared with the rest of the protein in Toll-like receptors in birds [120]; and positive selection in the binding domain of transferrin in mammals [43]. (C) Other enriched categories of genes from genome-wide studies. Results compiled from: *Drosophila* [26], ants [48], birds clade-wide [18], birds-of-paradise [121], mammals clade-wide [7,28], and bats [37].

repeated rapid evolution across clades [38,39]. Detailed dissections of diverse proteins, including the complement system [40], protein kinase alpha [41], TRIM5alpha [42], and transferrin [43,44], show that positively selected amino acid **residues** are often directly involved in mediating interactions with pathogens (Figure 1B). These candidate gene studies have the advantage of being able to examine detailed localization of positively selected residues, often combining structural and experimental evidence to demonstrate a clear function for positively selected sites. For example, in a classic study of protein kinase alpha [41], Elde and coauthors showed experimentally that specific residues under positive selection in protein kinase alpha are associated with reduced susceptibility to a model virus that inhibits protein kinase alpha activity by encoding a mimic of its normal substrate.

Recently, another study [45] extended the logic of this kind of structural inference genome-wide by looking for positive selection on a site-by-site basis in a data set of >3000 structure-mapped alignments of mammalian proteins, to identify clusters of positivity selected sites. Given that these clusters are identified in folded protein space rather than in linear sequence space, they are not likely to be the result of alignment artifacts, as has been suggested for previous claims of clustering of selected residues [46]. Strikingly, nearly 50% of proteins identified by this study, with the strongest evidence for structurally clustered positive selection, were involved in the immune response; the remaining proteins were largely enzymes associated with the metabolism of xenobiotics [45]. Overall, positively selected clusters are significantly enriched near functionally important regions, such as catalytic sites or binding domains. Although the absolute numbers of robustly identified clusters of selected residues in this study were small, this work provides another line of evidence for the importance of positive selection to the evolution of the immune system.

While mammals remain the group with the best functional annotations of genes and the largest number of high-quality publicly available genomes, many other groups of organism now have sufficiently diverse assembled genomes to allow comparative studies of positive selection. While the lack of high-quality annotation of gene function can be a challenge, and the number of high-quality genome-wide studies outside of mammals is small, several important commonalities can be seen in these studies. In birds, recent work [18] demonstrates that not only are pathways involved in host–pathogen interactions enriched for positive selection (Figure 1), but orthologs of many of the genes experiencing positive selection in mammals are also similarly subject to positive selection in birds, hinting at deeply conserved loci of host–pathogen conflicts [18]. Older work in *Drosophila* [26,47] and other insects [48,49] also hints at frequent positive selection among genes encoding immune defense proteins (Figure 1), although, with many newly sequenced genomes among insects now available, these studies are ripe for an update. Overall, genes involved in immunity, as well those encoding proteins used by pathogens to facilitate their life cycle (such as cell-surface receptors required for viral entry), may be among the most consistent targets of natural selection in many lineages (Figure 1).

While genes encoding proteins with immune function may be the most frequently detected targets of positive selection in genome-wide scans, it is clear that these genes are not the only targets of positive selection. All of the genome-wide studies discussed earlier identified strong signals of positive selection in genes encoding proteins involved in other functions (Figure 1C). Reproductive proteins, in particular, appear to be common targets of positive selection: genes encoding proteins involved in spermatogenesis and fertilization are enriched in positive selection across mammals [7,28,37] and reproductive proteins are common targets of selection in *Drosophila* [50]. In birds, there is also evidence for enrichment of positive selection among genes mediating recombination (which is likely related to genomic defense against transposable elements) and phototransduction [18], and metabolic enzymes are enriched for clustered selected sites in mammals [45]. With the rapidly expanding number of available genomes to analyze, across a wider range of organisms than ever before, a more detailed and nuanced understanding of clade-wide selection and the role of arms race dynamics in driving positive selection is on the horizon.

Studying Lineage-Specific Selection with Comparative Genomics

The vast diversity of life on Earth, and the 'endless forms most beautiful' of living things, strongly imply that there must be more to natural selection than just host–pathogen arms races. In particular, the newly sequenced genomes of many species are motivated, in part, by a desire to understand the genomic changes associated with fascinating lineage-specific biology. Thus, it has become fairly standard practice to test for lineage-specific selection using various branch-site tests of positive selection (e.g., [51–54]).

However, interpretation of these analyses can be challenging, given the potentially high false positive and false negative rates of branch-site methods. A demonstration of this challenge can be seen by comparing multiple studies that have looked for lineage-specific patterns of positive selection in mole-rats [55–59]. African mole-rats (family Bathyergidae) are subterranean rodents that have attracted considerable interest due to their subterranean lifestyle and extreme longevity (relative to body size); the phenotypically convergent blind mole rats in the family Spalacidae have also attracted study [56]*.* While all studies reveal biologically plausible genes under selection, there is remarkably little consistency among them [55,59]. All of these studies use some variation on branch-site models to identify positive selection specific to mole-rat lineages, albeit with varying outgroups, filtering methods, and taxon sampling. Given the high false positive and false negative rates of branch-site models, it is likely that some of the failure to find consistent results across studies is due to methodological limitations of the approach. Notably, an analysis looking purely at rates of amino acid substitution in multiple subterranean mammals found strong evidence for convergent loss of genes involved in vision and adaptation for tunneling behavior [60], suggesting that a common signal of protein evolution is detectable across mole-rats.

Rather than attempt to infer specific genes that are rapidly evolving and associated with the unique phenotypic traits of a particular organism, which may be difficult with these methods, some studies have turned to the use of branch-site tests in a comparative context, examining the impact of positive selection on different lineages [61,62]. If, for example, a particular trait of a species or clade is associated with rapid adaptation to a new environment, it is plausible that more genes in the genome of the rapidly evolving lineage will show evidence for positive selection. Using evidence from multiple **adaptive radiations** in angiosperms, two recent studies showed that a higher proportion of the protein-coding genome had evidence for positive selection in adaptive radiations than in background lineages [61,62]. These studies densely sampled both rapid radiations and background lineages in multiple systems, allowing comparisons of independent site models on each focal clade, although requiring some care to correct for possible power differences due to variation in total tree length across clades [63]. However, the interpretation of these results as indicating a link between positive selection and adaptive radiations requires caution, because the rapid diversification associated with adaptive radiations will result in more **gene tree discordance** than in background lineages, which can increase the apparent substitution rate and, thus, lead to false inferences of positive selection [64]. Nonetheless, studies of cichlid radiations [65] and diversification of wild tomatoes [66] also support a role for positive selection in adaptative radiations, and additionally indicate an important role for sorting of **ancestral polymorphisms** and **adaptive introgression**. As more and larger genomic data sets become available, the comparative analysis of lineage-specific positive selection will become increasingly powerful, and will likely reveal many important insights.

Challenges and Solutions: The Problem of Data Quality

Despite the increasing ubiquity of sequenced genomes and the increasing quality and completeness of those genomes, data quality remains a serious hurdle to detecting selection using codon-based models of molecular evolution. These models depend on the assumption that, in an alignment, all columns truly represent **homologous** positions, so that differences between species reflect substitutions, not errors. Errors in alignment (arising from the complexity of the alignment problem itself, **primary sequencing errors**, or **annotation errors**) can easily result in spurious signals of positive selection [46,67,68], motivating the use of alignment filtering tools (Table 1).

The simplest solution is to remove alignment columns that include nonhomologous bases; the challenge is that defining what is nonhomologous can be surprisingly tricky. Given that regions with substantial alignment gaps are more prone to alignment error, a simple and popular approach is to focus on removing gap and gap-adjacent columns [69,70]. Alternatively, residue-based methods [71–73] attempt to generate reliability scores for each individual residue in an alignment, allowing filtering of unreliable residues, even if they are limited to specific species. While residue-based filtering appears more effective than column-based filtering [74], studies using simulations of rapidly evolving proteins call into question the need to filter at all, reporting little or no improvement in false positive or true positive rate on filtered alignments, at least for some aligners [75,76]. Nonetheless, many large-scale comparative genomics studies have used heuristic methods of residue filtering to dramatically reduce the presence of false positive inferences of positive selection [9,77].

There are also several other major sources of error. In particular, mistakes in the annotation of gene models (annotation errors) can lead to nonhomologous or nonprotein-coding sequences being included in an alignment. Additionally, base-level inaccuracies in a genome assembly (primary sequence errors) can result in elevated false substitution rates [78]. Simulations that include primary sequence error lead to very high false positive rates for tests of selection [79], implying that filtering is still likely to be critical.

Several new methods attempt to remove nonhomologous sequences from alignments to address this problem. HmmCleaner [79] starts by creating a profile **hidden Markov model** from the input multiple sequence alignment, which it uses to represent the consensus model for the alignment. Each individual sequence is then aligned to the profile model and scored to identify regions of high divergence from the consensus profile. To demonstrate the power of

Table 1. Methods for Alignment Filtering

Software name	Filtering type	Method	Software	Refs
GBlocks	Column	Heuristic removal of divergent or gappy blocks	http://molevol.cmima.csic.es/castresana/Gblocks.html	[69]
trimAl	Column	Heuristic removal of gappy columns	http://trimal.cgenomics.org/	[70]
GUIDANCE2	Column and residue	Alignment consistency	http://guidance.tau.ac.il/ver2/	[71,72,74]
TCS	Column and residue	Alignment consistency	http://www.tcoffee.org/Projects/tcs/	[73]
HmmCleaner	Residue	Profile HMMs	https://metacpan.org/release/Bio-MUST-Apps-HmmCleaner	[79]
SWAMP	Residue	Heuristic removal of segments of excess divergence	https://github.com/peterwharrison/SWAMP	[80]
PREQUAL	Residue	Prealignment removal of nonhomologous sequences	https://github.com/simonwhelan/prequal	[81]

this method, the authors [79] introduced simulated primary sequencing errors into protein-coding alignments. Tests for lineage-specific selection on these simulated data sets had extremely high false positive rates, which could be corrected by prefiltering with HmmCleaner. A similar, but more heuristic approach is implemented in the program SWAMP [80], which uses a sliding window to find alignment segments with unreasonably high nonsynonymous substitutions. An alternate approach, PREQUAL [81], attempts to identify nonhomologous stretches of sequence before alignment, so that they can be masked and removed. These tools appear to perform well, and further work should focus on improving residue-filtering methods to account for errors in primary sequence and annotation, which may be expected to increase in frequency with the move towards relying on long, error-prone sequencing technologies for genome assembly [82].

Challenges and Solutions: The Limits and Possible Extensions of Codon Models

Codon models have been extremely popular and successful for detecting natural selection, but, like all models, they remain an imperfect representation of real biology. While codon models capture some mutational variation (e.g., transition/transversion differences) and allow for variation in codon usage [83], there are several aspects of sequence evolution that these models do not currently account for, which are increasingly recognized to be a potential concern for detecting selection. While I focus on mutational biases for brevity, several other data complexities remain significant issues for models of molecular evolution. These include complexities arising from variation in the underlying topology of gene tress (e.g., gene tree discordance, incomplete lineage sorting, and other sources of phylogenetic incongruity [64,84]), and complexities arising from a mismatch between modeling assumptions and the behavior of real data (e.g., the importance of accounting for synonymous rate variation within genes [85–87]).

One major concern, particularly in vertebrate genomes, is the role of GC-biased gene conversion [88]. Gene conversion describes the unidirectional transfer of sequence information from a donor sequence to a highly similar target. A significant amount of evidence suggests that this process is biased, such that, in AT/GC heterozygotes, the G or C allele preferentially replaces the A or T allele (reviewed in [89]). GC-biased gene conversion is a particularly severe problem for methods that assume a single neutral model across the genome, because, in these cases, elevated mutational-driven substitution rates in GC-rich regions create strong local departures from the genome-wide neutral model [90–92]. However, GC-biased gene conversion can also impact estimates of positive selection from codon-based models that use synonymous sites to infer neutral rates for individual genes [93–95]. While some models to account for GC-biased gene conversion exist [96,97], these approaches are yet to be incorporated into popular software for estimating positive selection using codon-based models. However, new methods that incorporate within-gene variation in synonymous substitution rates are a promising development to accommodate processes such as GC-biased gene conversion [87].

Codon-based models also assume that all substitutions occur individually and independently. Work over the past decade has clearly shown that this assumption is unrealistic for real data, because 1–5% of all substitutions likely arise from multinucleotide or other complex mutation events, in which a single mutational event causes changes in several bases in close proximity [98,99]. Given that multinucleotide mutations increase the inferred number of nonsynonymous substitutions, these can have serious consequences for detecting selection [100]. Branch-site tests are particularly prone to false positives, because a single multinucleotide mutation can produce a signal of multiple nonsynonymous changes in a particular lineage [101]. Indeed, a recent analysis of typical data incorporating multinucleotide mutations showed that a very high proportion of presumed signals of lineage-specific selection using branch-site tests disappeared when multinucleotide mutations were allowed in the codon model [101]. Several updates to existing models have been proposed to help account for the issues arising from complex mutation events, including changes to underlying codon models to allow multiple substitutions [100–102] and changes to likelihood calculations to improve stability under cases of model misspecification [103]. While these new models are not yet routinely available in standard analysis packages, there is good potential to address concerns about multinucleotide mutations and other modeling challenges with improved methods.

Transition from One Genome per Species to Many Genomes per Species

Comparative genomics is in the midst of a transition from a focus on sampling one reference sequence per species, to sampling multiple individuals per species. Cheap and widely available resequencing data across diverse clades [104], and related advances in genome assembly that start to capture population variation as part of a reference sequence (such as diploid assembly [105] and genome graphs [106]) all point towards a future where reference variation, instead of a single reference sequence, will be common. Resequencing studies, especially in non-model species, are increasingly limited by the challenges of sample collection and storage (e.g., [107]), instead of the costs of DNA sequencing. This has important implications for tests for positive selection.

Fundamentally, multi-individual data sets contain vastly more power to make inferences about lineage specific selection than even large trees of single representative haplotypes per species [1]. Tests that incorporate polymorphism and divergence data, such as the McDonald–Kreitman test [108] and its numerous extensions [109–112], use synonymous and nonsynonymous variation within a population (both of which are expected to be largely neutral, because neither deleterious nor beneficial mutations will remain segregating for long) to calibrate the expected amount of amino acid divergence between species (reviewed in [113]). Thus, these methods allow more precise estimates of critical properties, such as the distribution of fitness effects on protein-altering mutations and the fraction of amino acids fixed by selection in specific lineages [113,114]. Methods are already being developed that incorporate polymorphism data to better estimate ω and related parameters [115,116] and, in the future, these methods will only become more powerful and more feasible to implement.

Concluding Remarks and Future Directions

Comparing protein alignments across species has a long history as an approach to detect the action of natural selection. With the advent of large comparative genomics data sets, important insights into frequent adaptation have emerged from comparative projects examining large numbers of species. However, these methods are optimized to detect certain kinds of selection, in particular when common sites are repeated targets of adaptation across many lineages. This model of selection is probably a good description of many molecular arms races, especially immune system proteins and, indeed, has been validated with functional work in some classic host–pathogen arms races. However, this model of selection is probably not a good description for unique, lineage-specific adaptations in particular species; while lineage-specific models, such as branch-site tests, exist, they are prone to error, have low power, and a high false positive rate. Careful data preprocessing and new modeling approaches incorporating more complex mutational dynamics can help overcome the high false positive rate of these methods. However, despite some clear success stories, it is likely that only the increasing availability of population resequencing data will begin to reveal the full richness of positive selection to shape the diversity of life on Earth (see Outstanding Questions).

Outstanding Questions

Given that vertebrates, and mammals in particular, have been the focus of most comparative genomics studies of selection (with the exception of a few model systems, such as *Drosophila*), how will conclusions about the ubiquity of selection on genes interacting with pathogens hold up as more diverse lineages across the tree of life are studied?

With increasing amounts of data and potentially new sources of error, what methods can be devised to allow simple and robust prefiltering and quality control of sequence alignments used as input to codon models?

How will more realistic models of sequence evolution change our understanding of which genes in which species are targets of positive selection?

As within species variation data become common, how can these best be leveraged to improve our understanding of both lineage-specific and clade-wide selection?

Acknowledgments

I thank Allison Shultz, Sara Wuitchik, Nadia Singh, Dakota McCoy, Chris Balakrishnan, Matthew Hahn, and two anonymous reviewers for helpful comments and discussion. This work was conducted on the traditional territory of the Wampanoag and Massachusett peoples.

References

1. Kern, A.D. and Hahn, M.W. (2018) The neutral theory in light of natural selection. *Mol. Biol. Evol.* 35, 1366–1371
2. Jensen, J.D. *et al.* (2018) The importance of the neutral theory in 1968 and 50 years on: a response to Kern & Hahn 2018. *Evolution* 73, 111–114
3. Enard, D. *et al.* (2016) Viruses are a dominant driver of protein adaptation in mammals. *Elife* 5, e12469
4. Galtier, N. (2016) Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet.* 12, e1005774
5. Langley, C.H. *et al.* (2012) Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192, 533–598
6. Drosophila 12 Genomes Consortium *et al.* (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–218
7. Kosiol, C. *et al.* (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genet.* 4, e1000144
8. Lewin, H.A. *et al.* (2018) Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333
9. Zhang, G. *et al.* (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346, 1311–1320
10. Koepfli, K.-P. *et al.* (2015) The Genome 10K Project: a way forward. *Annu. Rev. Anim. Biosci.* 3, 57–111
11. Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591
12. Murrell, B. *et al.* (2015) Gene-wide identification of episodic selection. *Mol. Biol. Evol.* 32, 1365–1371
13. Smith, M.D. *et al.* (2015) Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32, 1342–1353
14. Pond, S.L.K. *et al.* (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679
15. Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–736
16. Murrell, B. *et al.* (2012) Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 8, e1002764–10
17. Yang, Z. *et al.* (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431–449
18. Shultz, A.J. and Sackton, T.B. (2019) Immune genes are hotspots of shared positive selection across birds and mammals. *Elife* 8, 398362
19. Mallarino, R. *et al.* (2016) Developmental mechanisms of stripe patterns in rodents. *Nature* 539, 518–523
20. Tollis, M. *et al.* (2019) Return to the sea, get huge, beat cancer: an analysis of cetacean genomes including an assembly for the humpback whale (*Megaptera novaeangliae*). *Mol. Biol. Evol.* 36, 1746–1763
21. Kenny, N.J. *et al.* (2019) Symbiosis, selection and novelty: freshwater adaptation in the unique sponges of Lake Baikal. *Mol. Biol. Evol.* 36, 2462–2480
22. Yang, Z. and Nielsen, R. (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46, 409–418
23. Yang, Z. and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908–917
24. Zhang, J. *et al.* (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22, 2472–2479
25. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87
26. Sackton, T.B. *et al.* (2007) Dynamic evolution of the innate immune system in *Drosophila*. *Nat. Genet.* 39, 1461–1468
27. Heger, A. and Ponting, C.P. (2007) Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. *Genome Res.* 17, 1837–1849
28. Nielsen, R. *et al.* (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3, e170
29. Hughes, A.L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335, 167–170
30. Radwan, J. *et al.* (2020) Advances in the evolutionary understanding of MHC polymorphism. *Trends Genet.* 36, 298–311
31. Wilburn, D.B. *et al.* (2019) Indirect sexual selection drives rapid sperm protein evolution in abalone. *Elife* 8, e52628
32. Lee, Y.H. *et al.* (1995) Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* 12, 231–238
33. Cosby, R.L. *et al.* (2019) Host-transposon interactions: conflict, cooperation, and cooption. *Genes Dev.* 33, 1098–1116
34. Castellano, D. *et al.* (2019) Viruses rule over adaptation in conserved human proteins. *bioRxiv* Published online February 19, 2020. https://doi.org/10.1101/555060
35. Ebel, E.R. *et al.* (2017) High rate of adaptation of mammalian proteins that interact with Plasmodium and related parasites. *PLoS Genet.* 13, e1007023
36. van der Lee, R. *et al.* (2017) Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. *Nucleic Acids Res.* 45, 10634–10648
37. Hawkins, J.A. *et al.* (2019) A metaanalysis of bat phylogenetics and positive selection based on genomes and transcriptomes from 18 species. *Proc. Natl. Acad. Sci. U. S. A.* 116, 11351–11360
38. Daugherty, M.D. and Malik, H.S. (2012) Rules of engagement: molecular insights from host-virus arms races. *Annu. Rev. Genet.* 46, 677–700
39. Sironi, M. *et al.* (2015) Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* 16, 224–236
40. Cagliani, R. *et al.* (2016) The mammalian complement system as an epitome of host-pathogen genetic conflicts. *Mol. Ecol.* 25, 1324–1339
41. Elde, N.C. *et al.* (2009) Protein kinase R reveals an evolutionary model for defeating viral mimicry. *Nature* 457, 485–489
42. Sawyer, S.L. *et al.* (2005) Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proc. Natl. Acad. Sci. U. S. A.* 102, 2832–2837
43. Barber, M.F. and Elde, N.C. (2014) Escape from bacterial iron piracy through rapid evolution of transferrin. *Science* 346, 1362–1366
44. Demogines, A. *et al.* (2013) Dual host-virus arms races shape an essential housekeeping protein. *PLoS Biol.* 11, e1001571
45. Slodkowicz, G. and Goldman, N. (2020) Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. *Proc. Natl. Acad. Sci. U. S. A.* 117, 5977–5986
46. Markova-Raina, P. and Petrov, D. (2011) High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. *Genome Res.* 21, 863–874
47. Obbard, D.J. *et al.* (2009) Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genet.* 5, e1000698
48. Roux, J. *et al.* (2014) Patterns of positive selection in seven ant genomes. *Mol. Biol. Evol.* 31, 1661–1685
49. Barribeau, S.M. *et al.* (2015) A depauperate immune repertoire precedes evolution of sociality in bees. *Genome Biol.* 16, 83

50. Haerty, W. *et al.* (2007) Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. *Genetics* 177, 1321–1335
51. Lind, A.L. *et al.* (2019) Genome of the Komodo dragon reveals adaptations in the cardiovascular and chemosensory systems of monitor lizards. *Nat. Ecol. Evol.* 3, 1241–1252
52. Rane, R.V. *et al.* (2019) Genomic changes associated with adaptation to arid environments in cactophilic *Drosophila* species. *BMC Genomics* 20, 52
53. Ghosh, A. *et al.* (2020) A high-quality reference genome assembly of the saltwater crocodile, *Crocodylus porosus*, reveals patterns of selection in Crocodylidae. *Genome Biol. Evol.* 12, 3635–3646
54. Gloss, A.D. *et al.* (2019) Evolution of herbivory remodels a *Drosophila* genome. *bioRxiv* Published online September 17, 2020. https://doi.org/10.1101/767160
55. Davies, K.T.J. *et al.* (2015) Family wide molecular adaptations to underground life in African mole-rats revealed by phylogenomic analysis. *Mol. Biol. Evol.* 32, 3089–3107
56. Fang, X. *et al.* (2014) Genome-wide adaptive complexes to underground stresses in blind mole rats *Spalax*. *Nat. Commun.* 5, 3966
57. Kim, E.B. *et al.* (2011) Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* 479, 223–227
58. Fang, X. *et al.* (2014) Adaptations to a subterranean environment and longevity revealed by the analysis of mole rat genomes. *Cell Rep.* 8, 1354–1364
59. Sahm, A. *et al.* (2018) Long-lived rodents reveal signatures of positive selection in genes associated with lifespan. *PLoS Genet.* 14, e1007272
60. Partha, R. *et al.* (2017) Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *Elife* 6, e25884
61. Nevado, B. *et al.* (2019) Adaptive evolution is common in rapid evolutionary radiations. *Curr. Biol.* 29, 3081–3086
62. Nevado, B. *et al.* (2016) Widespread adaptive evolution during repeated evolutionary radiations in New World lupins. *Nat. Commun.* 7, 12384
63. Anisimova, M. *et al.* (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18, 1585–1592
64. Mendes, F.K. and Hahn, M.W. (2016) Gene tree discordance causes apparent substitution rate variation. *Syst. Biol.* 65, 711–721
65. Brawand, D. *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature* 513, 375–381
66. Pease, J.B. *et al.* (2016) Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.* 14, e1002379–24
67. Fletcher, W. and Yang, Z. (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* 27, 2257–2267
68. Mallick, S. *et al.* (2009) The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* 19, 922–933
69. Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552
70. Capella-Gutiérrez, S. *et al.* (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973
71. Sela, I. *et al.* (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* 43, W7–W14
72. Penn, O. *et al.* (2010) An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.* 27, 1759–1767
73. Chang, J.-M. *et al.* (2014) TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol. Biol. Evol.* 31, 1625–1637
74. Privman, E. *et al.* (2012) Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol. Biol. Evol.* 29, 1–5
75. Jordan, G. and Goldman, N. (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* 29, 1125–1139
76. Spielman, S.J. *et al.* (2014) Limited utility of residue masking for positive-selection inference. *Mol. Biol. Evol.* 31, 2496–2500
77. Larracuente, A.M. *et al.* (2008) Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* 24, 114–123
78. Thomas, G.W.C. and Hahn, M.W. (2019) Referee: reference assembly quality scores. *Genome Biol. Evol.* 11, 1483–1486
79. Di Franco, A. *et al.* (2019) Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.* 19, 21
80. Harrison, P.W. *et al.* (2014) SWAMP: sliding window alignment masker for PAML. *Evol. Bioinform. Online* 10, 197–204
81. Whelan, S. *et al.* (2018) PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* 34, 3929–3930
82. Watson, M. and Warr, A. (2019) Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* 37, 124–126
83. Yang, Z. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503
84. Diekmann, Y. and Pereira-Leal, J.B. (2015) Gene tree affects inference of sites under selection by the branch-site test of positive selection. *Evol. Bioinform. Online* 11, 11–17
85. Davydov, I.I. *et al.* (2019) Large-scale comparative analysis of codon models accounting for protein and nucleotide selection. *Mol. Biol. Evol.* 36, 1316–1332
86. Pond, S.K. and Muse, S.V. (2005) Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* 22, 2375–2385
87. Wisotsky, S.R. *et al.* (2020) Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: ignore at your own peril. *Mol. Biol. Evol.* Published online February 18, 2020. https://doi.org/10.1093/molbev/msaa037
88. Galtier, N. and Duret, L. (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23, 273–277
89. Duret, L. and Galtier, N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* 10, 285–311
90. Katzman, S. *et al.* (2010) GC-biased evolution near human accelerated regions. *PLoS Genet.* 6, e1000960
91. Kostka, D. *et al.* (2012) The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol. Biol. Evol.* 29, 1047–1057
92. Borges, R. *et al.* (2019) Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics* 212, 1321–1336
93. Bolívar, P. *et al.* (2018) Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. *Mol. Biol. Evol.* 35, 2475–2486
94. Bolívar, P. *et al.* (2019) GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol.* 20, 5
95. Ratnakumar, A. *et al.* (2010) Detecting positive selection within genomes: the problem of biased gene conversion. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 365, 2571–2580
96. Capra, J.A. *et al.* (2013) A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet.* 9, e1003684
97. Hu, Z. *et al.* (2019) Bayesian detection of convergent rate changes of conserved noncoding elements on phylogenetic trees. *Mol. Biol. Evol.* 36, 1086–1100
98. Schrider, D.R. *et al.* (2011) Pervasive multinucleotide mutational events in eukaryotes. *Curr. Biol.* 21, 1051–1054
99. Averof, M. *et al.* (2000) Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 287, 1283–1286
100. De Maio, N. *et al.* (2013) Estimating empirical codon hidden Markov models. *Mol. Biol. Evol.* 30, 725–736
101. Venkat, A. *et al.* (2018) Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat. Ecol. Evol.* 2, 1280–1288
102. Dunn, K.A. *et al.* (2019) Improved inference of site-specific positive selection under a generalized parametric codon model when there are multinucleotide mutations and multiple nonsynonymous rates. *BMC Evol. Biol.* 19, 22

103. Mingrone, J. *et al.* (2019) ModL: exploring and restoring regularity when testing for positive selection. *Bioinformatics* 35, 2545–2554

104. Corbett-Detig, R.B. *et al.* (2015) Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol.* 13, e1002112

105. Chin, C.-S. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054

106. Paten, B. *et al.* (2017) Genome graphs and the evolution of genome inference. *Genome Res.* 27, 665–676

107. Bakker, F.T. *et al.* (2020) The Global Museum: natural history collections and the future of evolutionary science and public education. *PeerJ* 8, e8225

108. McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351, 652–654

109. Bustamante, C.D. *et al.* (2002) The cost of inbreeding in *Arabidopsis*. *Nature* 416, 531–534

110. Sawyer, S.A. and Hartl, D.L. (1992) Population genetics of polymorphism and divergence. *Genetics* 132, 1161–1176

111. Eilertson, K.E. *et al.* (2012) SnIPRE: selection inference using a poisson random effects model. *PLoS Comput. Biol.* 8, e1002806–e1002814

112. Zhao, Z.-M. *et al.* (2017) Detection of regional variation in selection intensity within protein-coding genes using DNA sequence polymorphism and divergence. *Mol. Biol. Evol.* 34, 3006–3022

113. Eyre-Walker, A. (2006) The genomic rate of adaptive evolution. *Trends Ecol. Evol.* 21, 569–575

114. Smith, N.G.C. and Eyre-Walker, A. (2002) Adaptive protein evolution in *Drosophila*. *Nature* 415, 1022–1024

115. Mugal, C.F. *et al.* (2020) Polymorphism data assist estimation of the nonsynonymous over synonymous fixation rate ratio ω for closely related species. *Mol. Biol. Evol.* 37, 260–279

116. De Maio, N. *et al.* (2013) Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol. Biol. Evol.* 30, 2249–2262

117. Messer, P.W. and Petrov, D.A. (2013) Frequent adaptation and the McDonald-Kreitman test. *Proc. Natl. Acad. Sci. U. S. A.* 110, 8615–8620

118. Kosiol, C. and Anisimova, M. (2019) Selection acting on genomes. In *Evolutionary Genomics: Statistical and Computational Methods* (Anisimova, M., ed.), pp. 373–397, Springer

119. Murrell, B. *et al.* (2013) FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. *Mol. Biol. Evol.* 30, 1196–1205

120. Grueber, C.E. *et al.* (2014) Episodic positive selection in the evolution of avian toll-like receptor innate immunity genes. *PLoS One* 9, e89632

121. Prost, S. *et al.* (2019) Comparative analyses identify genomic features potentially involved in the evolution of birds-of-paradise. *Gigascience* 8, giz003