Natural Selection in Population and Phylogenetics: a Combined Approach

Running head

Natural Selection in Population and Phylogenetics

Daniel J. Wilson* and Molly Przeworski

Department of Human Genetics, University of Chicago, 920 East 58th Street, Chicago, IL 60637, United States of America.

* Corresponding author. E-mail daniel.wilson@ndm.ox.ac.uk. Present address: Nuffield Department of Clinical Medicine, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, United Kingdom. Telephone 01865 287500. Fax 01865 287501.

Last saved 9/4/10 15:53

Abstract

Patterns of polymorphism and divergence across multiple species are highly informative about the selective forces that worked to shape those populations. Through an analysis of polymorphism within and divergence between species we can hope to learn about the distribution of selection coefficients across the genome, the changes in adaptive landscape that occur over time, and the specific sites involved in key adaptations that distinguish modern day species. Yet the cross-fertilization of ideas from population genetics and phylogenetics in this area has been limited. Methods designed to detect fine-scale variation in selection pressures such as codeml and omegaMap exploit respectively divergence or polymorphism data alone. The Poisson random field approach utilizes both, but under a restrictive set of modeling conditions. Here we propose a combined population genetics and phylogenetics approach that brings together the advantages of all these techniques. To demonstrate its flexibility, and connection to existing methods, we illustrate our approach with example analyses of human toll-like receptors, the meningococcal *porB3* antigen, and CTL escape and reversion in an HLA-discordant HIV-1 transmission pair.

Author summary

Introduction

The ultimate questions in biology [1] concern the role of adaptation versus alternative, non-adaptive forces [2] in shaping the diversity of life within and between species. Consequently, detecting the genetic signature of natural selection in patterns of

polymorphism and divergence across multiple species has become a major goal of evolutionary biology [3-5]. From analyses of polymorphism within and divergence between species, we can hope to learn about the distribution of selection coefficients acting on the genome [6], changes in the adaptive landscape over time [7], and the specific sites in the genome that underlie adaptive phenotypes [8].

A variety of molecular patterns are predicted by natural selection. Within species, selection may skew allele frequencies towards favored forms [9], propel advantageous derived mutants to high frequency [10], drive population differentiation [11], and leave behind extended regions of linkage disequilibrium [12]. Between species, selection may conserve favorable genotypes over long periods of evolutionary time [13], or accelerate the substitution of novel forms in response to disruptive fitness regimes [14]. Traditionally, in protein-coding regions, the rate of substitution of synonymous mutations has been used as a neutral yardstick against which to compare the rate of amino acid-changing substitutions [15], but this approach has been broadened to include comparisons with flanking untranslated regions, introns, putative regulatory regions and non-genic regions [16, 17].

By and large, the models used to understand how selection influences gene frequencies in polymorphism data can be traced back to the diffusion theory of population genetics pioneers such as Sewall Wright [18] and Motoo Kimura [19]. Diffusion theory [20] underpins a range of neutrality tests [9, 10] and methods for estimating selection coefficients from allele frequency spectra, with or without ancestral information [21, 22], but its mathematical elegance owes much to simplifying conditions such as biallelic loci (*i.e.* no more than two alleles co-segregate at a locus), infinite sites

(*i.e.* no repeat mutation, back mutation, or saturation of sites), or parent-independent mutation (in which the rate of mutation to an allele is independent of the mutating allele). As a result these methods often require compromises when analyzing data, such as pooling alleles or discarding sites where there are more than two alleles, and reconstructing ancestral states by parsimony, provided the ancestral allele can be unambiguously identified from an outgroup.

The point of departure for analyses of divergence data is usually the reconstruction of a tree relating the molecular sequences in question. The estimation of substitution rates along the branches of the tree allows the detection of regions of accelerated evolution [23], or an excess rate of non-synonymous relative to synonymous substitution, usually expressed as the *DN*/*DS* ratio [24] at particular sites [25], along particular branches [26], or both [27]. Modern likelihood-based phylogenetic methods [28, 29] are based on Markov chain models [20]. They allow for the joint estimation of the tree and a sophisticated substitution model that may incorporate multiple alleles (usually 4 nucleotides, 20 amino acids or the 61 non stop codons), highly parent-dependent rates (*e.g.* different rates for synonymous versus non-synonymous substitutions) and probabilistic inference of ancestral states [30]. However, the application of phylogenetic methods directly to polymorphism data can have pathological consequences. Recombination between sites within a population can resemble repeat mutation when rigidly imposing a tree structure [31], and this in turn may cause a false signal of adaptive evolution [32, 33]. Population samples are likely to contain weakly deleterious polymorphisms that are segregating in the population but ultimately destined for loss. This can give the misleading impression of a relaxation of selective constraint

near the tips of the tree [34]. These issues are especially pertinent in the context of the growth of high-throughput technologies aimed at deeply sequencing populations [5].

Polymorphism and divergence offer complementary angles on the evolutionary process. The snapshots afforded by polymorphism data are especially informative about deleterious mutation because mildly advantageous and deleterious variants alike can segregate within populations, albeit with different expected frequencies. Over time, selection acts on these differences so that substitutions are enriched for advantageous mutations, and impoverished of deleterious mutations. The McDonald-Kreitman (MK) test [35] exploits this contrast to detect adaptation where divergence or polymorphism data alone might not, owing to variation in selection coefficients within a gene. A preponderance of deleterious mutants might limit the *DN*/*DS* ratio to a value much less than unity, and thereby swamp the signal of adaptive substitution at other sites. Yet an excess *DN*/*DS* ratio compared to *PN*/*PS*, its population analog, could reveal a surplus of non-synonymous substitution compared to polymorphism, indicative of adaptive change.

Several model-based interpretations of the MK test have been proposed [36-38], of which the Poisson random field (PRF) approach is most widely used [36, 39, 40]. Except for a class of unviable mutants, PRF does not model variation in selection coefficients within a gene. Arguably, this sets a high threshold for detecting adaptation, as the net effect of selection within the gene must be adaptive change, whereas one might conservatively expect the evolution of a protein-coding gene often to involve a small number of adaptive changes against a backdrop of selective constraint. Perhaps this explains why scans of the human genome have found fewer genes under adaptive change than expected from the false positive rate [40, 41]. Methods to detect fine-scale variation

in selection pressures such as codeml [25, 28] and omegaMap [42] exist but exploit respectively divergence and polymorphism data alone. The mathematical conveniences of diffusion theory, particularly the infinite alleles model, make PRF simple and attractive to use. But they also make it difficult to extend to scenarios requiring multiple alleles, multiple species, sophisticated mutation models, probabilistic inference of ancestral states, variable selection pressures and localizing the signal of selection.

The aim of this paper is to overcome these limitations by combining population genetic and phylogenetic approaches to the analysis of polymorphism and divergence. We propose a likelihood-based framework that integrates the underlying diffusion and Markov chain models to facilitate the analysis of natural selection in multiple populations, using multiallelic, parent-dependent models of mutation with variable selection pressure. We illustrate the principles of the approach, and its connection to PRF, through an example biallelic analysis of human toll-like receptors. We use Wright-Fisher simulations to test the multiallelic parent-dependent model by employing the idea of calibration. Through an analysis of polymorphism data in the meningococcal *porB3* antigen locus, we characterize intragenic variation in selection pressure, and compare our results to those of omegaMap. Finally, we perform an analysis of polymorphism and divergence data from multiple HIV-1 populations sampled from a transmission pair, and identify specific sites subject to adaptive change in response to the host genotype.

Results

Population genetics and phylogenetics are concerned with the same evolutionary processes, but at different timescales. The diffusion processes of population genetics consider changes in population allele frequency over time (Figure 1A), while the Markov

chains of phylogenetics consider the succession of allelic substitutions over time. The latter may be viewed as a summary or simplification of the former (Figure 1B), appropriate at longer timescales where the fixation and loss of novel mutations occurs essentially instantaneously. Here we attempt to reconcile the approaches, because the patterns of polymorphism that represent snapshots of the evolutionary process, and which we are able to observe directly, can complement our understanding of selection gleaned from patterns of divergence between species.

Working with diffusion processes is difficult because they are mathematically unwieldy [20]. Models of any practical value generally yield a probability distribution for the frequency of alleles in a population at equilibrium. Such results come at a cost, usually requiring simplifying assumptions such as constant population size, no population structure and parent-independent mutation [18]. We inherit these assumptions, but return to the issue of parent-independent mutation later. A stationary distribution of gene frequencies is useful for the analysis of polymorphism from a single population. To make it useful for the analysis of multiple populations, we have to introduce a connection to the phylogenetic model.

At a snapshot in time, the phylogenetic model reports the ancestral allele in the population. Our approach hinges on modifying the stationary distribution of allele frequencies in the population so that it depends on the ancestral allele (Figure 1C). Intuitively, one expects the ancestral allele to be at high frequency, quite possibly fixed, in the population, and the idea is to modify the stationary distribution to reflect this. We call this modification to the stationary distribution *phylogenetic conditioning* because by

conditioning on the ancestral allele, which we do by Bayes rule (see Methods), we introduce a phylogenetic dependency to the model of population allele frequencies.

In the analysis of polymorphism data, we cannot know for certain the identity of the ancestral allele in a population. Thus for statistical inference, which requires calculation of the likelihood of the observed data, we must sum over all possible ancestral alleles, not just at the internal nodes in the phylogeny relating our populations, but also at the tips of the phylogeny (Figure 1D). This is an extension to the usual pruning algorithm [30], which we call *population pruning* (see Table 1 for a glossary of terms). Population pruning is an important conceptual development because it implies that sequences sampled from extant (or indeed extinct) species should not be analyzed directly by the phylogenetic model, even in samples of size one, because of the inevitable presence of derived alleles in those sequences. This may help explain the apparent relaxation of selective constraint often seen at the tips of a phylogeny [34].

Biallelic Model Applied to Toll-Like Receptors

As a proof-of-principle, or sanity check, and to demonstrate the close connection to the Poisson random field (PRF) approach [36, 39], our first application is to the analysis of data in the form of McDonald-Kreitman (MK) tables [35]. The MK table is a two-by-two contingency table that classifies, for a pair of species, the number of sites in a gene that are synonymous (*S*) or non-synonymous (*N*), polymorphic (*P*) or divergent (*D*). The MK test is a test of the null hypothesis, under the neutral theory [15], that the odds ratio $(DN\ PS)/(DS\ PN)$ equals one. A DN/DS significantly greater than PN/PS is indicative of adaptive evolution between the two species.

PRF offers a parametric interpretation of the MK table based on diffusion theory that we parallel using our approach. In the infinite sites setting of the PRF, non-synonymous mutations with selective advantage s arise at rate μ_R per locus per generation. Synonymous mutations, assumed neutral, arise at rate μ_S. Because of the nuances of population genetics, the parameters estimable from the data are the population-scaled counterparts $\theta_S = 2PN_e\mu_S$, $\theta_R = 2PN_e\mu_R$ and $\gamma = 2PN_e s$, where P is the ploidy and N_e the effective population size; $1/PN_e$ can be thought of as the strength of genetic drift. The fourth parameter is τ, the divergence time for the two species, measured in units of PN_e generations.

In our finite-sites version we assume a rate of back mutation equal to the forward mutation rate. We arbitrarily label the two alleles at a locus 1 and 2. To assign selection coefficients to alleles, we mirror the PRF approach by assuming that the derived mutation has population-scaled selective advantage γ, regardless of which allele that may be. Thus, when $\gamma > 0$, selection favors new phenotypes: this is known as positive selection. Conversely, when $\gamma < 0$, selection favors the incumbent phenotype: this is negative selection. Measuring fitness relative to the ancestral allele implies that were a back mutation to arise, it would have advantage γ. In other words, the model implies that the selection regime changes upon substitution, and this is a property shared by the phylogenetic *DN*/*DS* model of Nielsen and Yang [25]. We call this recurrent directional selection, and while it is convenient for inference, in the Discussion we return to a consideration of its merits and demerits.

To evaluate the performance of our method and compare it to PRF, we estimated by maximum likelihood the evolutionary parameters for a family of genes known as toll-like receptors [43]. In primates, toll-like receptors (TLRs) are involved in innate immunity to intra- and extra-cellular pathogens. TLRs 3, 7, 8 and 9 confer immunity to viruses by recognizing foreign, intracellular nucleic acids. The remaining six TLRs confer resistance to bacteria and macroparasites through recognition of a variety of foreign extracellular molecules.

Barreiro and colleagues sequenced all ten TLRs in several hundred humans and constructed MK tables using the chimpanzee reference sequence. Interestingly, they found that the four TLRs involved in viral recognition exhibited patterns of positive selection while the remaining six exhibited patterns of negative selection, which may reflect the different selection pressures exerted by viral versus cellular pathogens on the host immune system [43]. In a reanalysis of their data (Table 2), we confirm this observation, obtaining very similar estimates of the synonymous and non-synonymous mutation rates, divergence times and selection coefficients for all ten genes by PRF and our method. This is reassuring, and indeed expected since PRF can be considered a special case of our method when the mutation rate is very low.

A Multiallelic Hot-or-Not Model

Summarizing patterns of polymorphism and divergence using MK tables is convenient for a pair of species when diversity is relatively low. But for multiple species or data that exhibit moderate to high diversity, multiple nucleotides might be polymorphic, divergent or both at the same position. When multiple nucleotide positions within the same codon are coincidentally polymorphic or divergent, it becomes difficult

to classify sites as synonymous or non-synonymous. Furthermore, polymorphism at the same site in multiple species, or multiple substitutions at the same position, make it difficult to infer ancestral states by parsimony. In these situations it would be preferable to analyze the molecular sequences directly, using a probabilistic model, rather than to summarize them.

In addition, a codon-based sitewise model of molecular evolution has advantages from the modeling perspective. The relationship between any pair of codons can be defined unambiguously as synonymous or non-synonymous. For explicit finite-sites models, standard phylogenetic tools are available for dealing with uncertainty in ancestral states, branch lengths, tree topology and evolutionary rates [28, 29, 30]. And it is well understood how to model variation in evolutionary rates between sites, or along specific branches [25, 26], allowing us to localize the signal of selection.

To enjoy these advantages, we constructed a multiallelic model of codon evolution. Our starting point was the stationary distribution of allele frequencies under Wright's [18] multiallelic diffusion process with parent-independent mutation and selection (PIMS) that we would modify using phylogenetic conditioning. Wright allows for quite general models of selection. However, to be practically useful the parameters must be estimable from data, and calculation of the likelihood must be computationally efficient. With these considerations in mind, we developed a *hot-or-not* model in which there are just two fitness classes. Measuring fitness relative to the ancestral state, we define γ as the population-scaled selective advantage of codons encoding an amino acid different to that encoded by the ancestral codon. As in the biallelic example, this is a recurrent selection model so when $\gamma > 0$, amino acids different to the ancestor are favored

(positive selection), and when $\gamma < 0$, the ancestral amino acid type is favored (negative selection).

In a codon-based framework, a parent-independent mutation model is of dubious provenance, because one expects the rate of mutation to be much higher to codons that differ by one, rather than two or three nucleotides. To address this we developed an approximation to parent-dependent mutation described in the Methods. In summary, we exploit the method of phylogenetic conditioning to force the mutation rate to depend on the identity of the ancestral allele. Intuitively, the ancestral allele is likely to be the most frequent, so the majority of mutations arise on the ancestral background. Therefore we expect the approximation to perform well.

We used this approximation to parent-dependent mutation to implement a codon-based version of Hasegawa, Kishino and Yano's model (HKY85 [44]) in which π is the stationary frequency of codon usage and κ is the rate of transitions relative to transversions. Within a codon, only one nucleotide is allowed to mutate at a time, and the expected population-scaled mutation rate is θ. Under this HKY85 codon model with recurrent selection, the phylogenetic substitution rate is identical to that of Nielsen and Yang (NY98 [25]), where their parameter representing the *DN*/*DS* ratio, ω, equals $\gamma/\left(1-\mathrm{e}^{-\gamma}\right)$. However, in the next section we examine why our approach is not equivalent to applying the phylogenetic NY98 model directly to polymorphism data.

Testing the Method by Simulation

Owing to a number of approximations made in the construction of the model, we wished to test the performance of the likelihood in an inference setting. Two scenarios

were of particular interest: in the first, we wished to assess the performance of parameter estimation from polymorphism data alone, when the ancestor is known at each site. This serves to test the method of phylogenetic conditioning, and the approximate model of parent-dependent mutation. In the second, we wished to assess the performance of parameter estimation from polymorphism data when there is some divergence information. In this scenario, the ancestor is assumed known 10 PN_e generations prior to sampling. This serves as an additional test of the phylogenetic model and the population pruning algorithm.

We used as the basis of our test the Bayesian idea of calibration [45]. In simulations, if our model works correctly, then the 95% credible interval (CI) for a parameter should include the truth 95% of the time. Thus we performed 200 Wright-Fisher simulations of 30 sequences length 250 codons under each scenario, drawing the evolutionary parameters θ, κ and γ from a prior distribution (see Methods). For each scenario, we expected that in 190 simulations the 95% CI would include the truth, with an acceptable range of 184-196.

For all parameters in both scenarios, the actual number of datasets in which the 95% CI enveloped the truth lay within the acceptable range (Figure 2). Broadly speaking, there is greater statistical uncertainty in the parameter estimates (represented by the width of the credible intervals) in the first scenario in which inference was based on polymorphism data alone. In the second scenario, one might have expected a loss of information because the identity of the ancestor is provided 10 PN_e generations prior to sampling. However, this seems to have been outweighed by the additional information gained from the signal of divergence.

All three parameters, including the transition:transversion ratio κ, were well-estimated from the data, indicating that the approximation to parent-dependent mutation works well. The uncertainty in estimating θ is greater for lower values, presumably because there are fewer informative sites. In estimating γ, there appears to be good power to distinguish positive from negative selection. However, the credible intervals get wider, representing greater uncertainty, as one moves away from $\gamma = 0$. This pattern is asymmetrical in the second scenario, in which the uncertainty in estimating γ decreases more slowly for positive than negative selection. This can also be explained by the number of informative sites: positive selection promotes diversity while negative selection suppresses it.

Colored vertical lines in Figure 2 highlight 95% CIs that did not include the true value. This allows one to visualize any systematic patterns that may be present in the simulations. For κ and γ there does not appear to be a consistent pattern in the datasets for which the 95% CI excluded the truth. However for θ it appears that in most of these cases, the mutation rate was under-estimated. This suggests a slight downward bias, although the scatter of point estimates (solid circles) around the truth indicates that it is not severe. The overall picture, therefore, is encouraging and consistent with the method being well-calibrated.

Analysis of Polymorphism in a Meningococcal Antigen

As noted earlier, the HKY85 codon model with recurrent selection that we have developed has the same substitution rate matrix as the NY98 model, which is frequently used for phylogenetic analyses of selection. Nielsen and Yang [46] originally investigated

the link between the population-scaled selection coefficient γ and the *DN*/*DS* ratio (which they call ω) in a model of recurrent selection. By equating the substitution rate in the two cases, they obtained the relationship $\omega = \gamma / \left(1 - e^{-\gamma}\right)$. It has become popular to apply phylogenetic models of selection such as NY98 to polymorphism data, and here we compare methods that would estimate γ by inverting this relationship to the *DN*/*DS* ratio against our new approach.

For some time it has been known that intragenic recombination within populations can cause problems for phylogenetic analyses of selection because recombination can resemble repeat mutation when rigidly imposing a tree structure, and this in turn can generate false positives in tests for selection [32, 33]. To counter this problem, Wilson and McVean [42] proposed a method, omegaMap, which estimates the *DN*/*DS* ratio within recombining populations by approximating the coalescent with recombination. OmegaMap estimates the *DN*/*DS* ratio ω by treating selection within populations as a form of mutational bias: when $\omega > 1$ (positive selection) non-synonymous mutations are more likely to arise than neutral mutations, and when $\omega < 1$ (negative selection), non-synonymous mutations are less likely to arise than neutral mutations. All mutations, having arisen, behave neutrally.

We refer to this treatment of selection at the population level as *aristogenetic*, in honor of Osborn's discredited theory of adaptive evolution [47] in which adaptive mutations have an inherent propensity to arise in the genome. We can emulate the aristogenetic model of omegaMap using our approach by formulating a neutral model with mutation rate given by the NY98 model. Whereas omegaMap models linkage

disequilibrium between sites, and estimates the population-scaled recombination rate ρ, we assume independence between sites, equivalent to free recombination. We compare omegaMap and this aristogenetic NY98 model to our HKY85 model with recurrent selection. This comparison serves two purposes. Firstly, in comparing omegaMap to an aristogenetic NY98 model, we can learn what effect our assumption of independence between sites has on the analysis, and further test the adequacy of our approximation to parent-dependent mutation. Secondly, in comparing the aristogenetic models to the HKY85 model with recurrent selection, we can quantify what difference, if any, this makes to inference.

Our frame of reference is a collection of 23 meningococcal *porB3* sequences isolated from carriers in England and Wales [48]. The *porB3* locus encodes an antigenic cell surface porin protein important for proper cell growth and pathogenesis. The protein consists of eight loop regions that span the cell membrane. Figure 3A shows the point estimate and 95% CI for ω using the adaptive sliding window model of omegaMap [42]. The majority of the gene appears subject to negative selection, with evidence for positive selection in four of the extracellular loop regions. The point estimate of ρ was 48.9 per kilobase (95% CI 33.3 – 68.4), which is considerable, but not what you would call free recombination.

In Figure 3B, the point estimate and 95% CI are shown as estimated from an aristogenetic NY98 model assuming the same model of variation in ω and no linkage disequilibrium (LD) between sites. Both point estimate and credible interval are nigh-on identical between omegaMap and the aristogenetic NY98 model. This is reassuring on at least two levels. Firstly, it is further evidence that our approximation to parent-dependent

mutation works well. Secondly, to find that the estimates of ω should be so similar despite the very different approaches to dealing with recombination suggests that, for aristogenetic models at least, inference of selection is robust to LD. As a disclaimer, however, when the recombination rate is low it would still be prudent to test robustness by simulation. It is also worth emphasizing that the assumption of free recombination is quite different to the assumption of no recombination (complete LD) when directly applying phylogenetic analyses of selection [25], which are vulnerable to false positives. By a very informal argument, one might hope that imposing less structure on a problem (*i.e.* by assuming sites are unlinked within a population) would be less pathological when the model assumptions are violated than imposing more structure on a problem (by assuming sites are completely linked).

In contrast, the differences in inference between the aristogenetic models and the HKY85 model with recurrent selection are striking; Figure 3C illustrates the transformed parameter $\omega = \gamma / \left(1 - e^{-\gamma}\right)$ for the purposes of comparison. Principally, there is much greater uncertainty in parameter estimation, as witnessed by the wider 95% CI. As a consequence, the point estimate changes more smoothly over the gene, and the distinction between sites estimated to be under positive versus negative selection is less pronounced. Figure 3D shows the posterior probability of positive selection ($\omega > 1$) under the three models. OmegaMap and the aristogenetic NY98 model (red and yellow lines) are nearly indistinguishable from one another. Compared to the HKY85 model with recurrent selection (green line), inference under the aristogenetic models produces a sharper transition from areas of low to high probability of positive selection, and absolute

probabilities are more extreme (closer to 0 or 1). The locations of sites under positive or negative selection, however, are consistent.

Estimating the strength of selection within a population is clearly quite a different undertaking to estimating the strength of selection at the phylogenetic level. On a phylogenetic timescale, the *DN*/*DS* ratio relates directly to the selection coefficient γ in a model of recurrent selection. However, the examination here of aristogenetic models suggests that while it may perfectly possible to obtain a good estimate of *DN*/*DS* within a population, the relationship to the selection coefficient is less immediate. In other words, there is more stochasticity in the relationship between selection and the *DN*/*DS* ratio within a population; thus the latter predicts the former less efficaciously. This should not come as a surprise [49]; indeed it forms the basis for the MK test [35]. Nevertheless, our combined population genetics-phylogenetics approach lets that additional stochasticity to be properly taken into account.

Differential Selection in an HIV-1 Transmission Pair

Finally we apply our population genetics/phylogenetics approach to the analysis of selection acting on multiple populations. Human immunodeficiency virus type 1 (HIV-1) is a rapidly evolving retrovirus that chronically infects patients by continually adapting to evade the host immune response. The human immune system mounts a range of responses including the cytotoxic T lymphocyte (CTL) response, which is mediated by class I human leukocyte antigen (HLA) molecules. Human cells expressing HLA-I recognize and present virus molecules, initiating a CTL response that kills infected cells to halt viral replication. The success of the CTL response depends on the recognition of specific viral antigens, or epitopes, by the HLA-I repertoire. The virus can evade the CTL

response by mutating to epitopes not recognized by the host's particular HLA-I repertoire. Viral proteins carrying such mutations may suffer a reduction in functional efficacy, and thereby carry a fitness cost in a different HLA-I environment. Reversions to wild type are therefore often seen following infection of a new host [50].

Here we characterize natural selection in the *gag* gene of three HIV-1 populations sampled from a donor-recipient transmission pair: one from the donor (D0) and recipient (R0) shortly following transmission, and one from the recipient four years later (R1). The donor had been infected for eight years previously, and the recipient received no antiretroviral treatment over the course of the study [50]. The analysis serves to test the performance of our methods because the evolution of HIV-1 in these patients was to some extent predicted by their HLA types, and recognition of virus epitopes by the host can be confirmed in the laboratory. We do not pretend that the HIV-1 population *in vivo* meets all the assumptions made by the model – in particular one might expect demographic growth following transmission. The ability of the method to perform despite this is one of factors of interest.

Following the omegaMap approach [42], we modeled variation in the selection coefficient γ, which we estimated separately for the three populations, using an adaptive sliding window model, with an average window length of five codons. Figure 5 shows the posterior probability of positive selection ($\gamma > 0$) for each population. The plots are annotated with colored vertical bars indicating the sample frequency of amino acid polymorphisms relative to the sequence inferred close to the point of transmission. Black lines above depict known *gag* epitopes that are restricted to the donor (GL8) and recipient

(QW11, VL8, EW10, FK10) HLA types. Functional activity towards known epitopes by the host immune system was confirmed by enzyme-linked immunospot assay [50].

Overall, the signal of selection was strongest in population R1, probably because it has the largest sample size and longest branch leading to it. A stronger signal confers tighter CIs around γ (not shown) and greater perturbation in the probability of positive selection away from the prior expectation of 0.5 (Figure 5). Informally then, we consider sites to be subject to positive selection when they have a posterior probability exceeding 0.5 and a visually discernible increase over the background probability for that population.

Positive selection was detected at an Ile to Leu mutation within the QW11 epitope and a Val to Leu mutation flanking the EW10 epitope in the recipient. The transmitted forms of QW11 and EW10 are both HLA-restricted to (*i.e.* recognized by) recipient HLA allele A*2501. Functional assays confirmed a decrease in CTL recognition of the mutant form for both epitopes [50]. These mutations probably represent CTL escape, evolved under selection pressure imposed by the new HLA-I regime of the recipient. We also detected positive selection in a Phe to Leu mutation in the GL8 epitope in both donor and recipient. The Leu variant, which is HLA-B8-restricted, was absent from the R0 population, implying that Phe was the transmitted form. The donor and recipient were HLA-B8 positive and negative respectively. Therefore, we interpret the detection of positive selection in D0 and R1 to reveal an incomplete sweep in the donor of the Phe epitope, which was transmitted, followed by a reversion to the Leu wild type in the recipient.

The ability to experimentally verify allelic differences in one aspect of viral fitness (strength of the elicited immune response) supports the population and phylogenetic evidence for positive selection at some sites. This is reassuring, and it leads one to speculate that the detection of positive selection at various other sites for which there is currently no functional explanation may indeed be real. However, what of the known epitopes at which there was no evidence for positive selection? In general, it does not follow that the existence of a selective pressure will necessarily lead to an adaptive change, for reasons including lack of mutational opportunity. However, the mutation rate in HIV is high, and low-frequency amino acid variants were observed in epitopes VL8 and FK10. But for CTL escape to occur, a mutation must not only confer evasion of the immune response, but also avoid too onerous a penalty in the normal functioning of the protein. In other words, the existence of a known vulnerability to an HLA-I allele does not imply the existence of alternative epitopes able to confer a net fitness advantage in that environment. The absence of a signal of positive selection in epitopes VL8 and FK10 is not, therefore, an obvious problem.

Discussion

In this paper we have introduced a combined population genetics/phylogenetics approach to the analysis of selection in patterns of polymorphism and divergence. Our goal was to apply flexible models of selection that incorporate multiple alleles, multiple species and variation in selection pressures spatially and over time, in a likelihood-based framework. To illustrate our approach and its connection to related methods such as PRF [36] and omegaMap [42], we performed analyses of human toll-like receptors [43], the

meningococcal *porB3* antigen [48] and CTL escape and reversion in an HIV-1 transmission pair [50].

Although the motivation for developing a combined population genetics/phylogenetics approach was the analysis of selection, it can be thought of as a more general method for integrating these two groups of techniques that will become increasingly important with the advent of deep population sequencing [5]. We introduced the idea of phylogenetic conditioning that allows the stationary distribution of allele frequencies from classical population genetics models grounded in diffusion theory to be integrated with modern phylogenetic models based on Markov chains, and we described the population pruning algorithm which takes into account uncertainty in the ancestral state of populations at the tips of the tree. Our approach may help understand the apparent relaxation of functional constraint observed towards the tips of phylogenies [34], which we explain in terms of sampling low-frequency derived mutations that are destined for loss. This method of combining population genetics and phylogenetics models could also be used for multi-species or multi-population analyses of neutral evolution.

The diffusion models underpinning our work make a variety of assumptions, including constant population size, no population structure and parent-independent mutation. In response to the last of these, we developed an approximation to parent-dependent mutation to allow for inference under very general mutation models. Simulations showed the approximation to perform very well. However, we inherited the remaining assumptions of these methods, and this is a clear avenue for further investigation. The extension to models incorporating demographic change and migration are obvious starting points.

Although we showed in one of our analyses of the meningococcal *porB3* antigen that inferences about selection were robust to our assumption of free recombination between sites within a population, this is an area for improvement. In the neutral case, our model suggests that conditioning on the ancestral allele within a population is analogous to sampling one individual with that allele. This suggests an obvious way to extend the PAC approximation [56] to the coalescent with recombination to multiple populations. It may be that PAC can also be adapted to model selection within populations as a way of taking this forward.

In this paper we concentrated almost exclusively on models of recurrent selection, in which the fitness of an allele is measured relative to the ancestral allele [25, 36]. This is a mathematical convenience that has peculiar consequences. Under positive selection ($\gamma > 0$), the ancestral allele is disfavored, creating a continual drive for innovation. However, under negative selection ($\gamma < 0$), derived alleles are disfavored. It follows that were a mildly deleterious allele to fix by drift, upon fixation the selection regime would change and the back mutation would not restore fitness as one might expect, but in fact erodes it further. While it is difficult to make biological sense of this behavior, it will occur rarely for moderate to highly deleterious selection coefficients. When the substitution rate is low enough that one expects no more than a single substitution per branch per site, the problem basically does not arise. In any case, models of recurrent selection have a hold on phylogenetic analyses of selection, and thus it makes sense to use this as a starting point. Nonetheless, our work provides a platform for developing non-recurrent models of selection in which the substitution process might be modeled separately from changes in the selection regime. We express optimism that such models

might pave the way for the sitewise analysis of non-coding regions, opening up an area of investigation that has heretofore received much less attention.

Methods

Overview

At a fundamental level, population genetics and phylogenetics differ in that the former models changes in allele frequency (a continuous quantity) within species while the latter models changes in the ancestral allele (a discrete quantity) between species. To reintegrate the approaches, we introduce phylogenetic dependency to the stationary distribution of allele frequencies within a population by conditioning on the identity of the ancestral allele. To calculate the likelihood of an observed sample of alleles from a population, conditional on the ancestral allele, we integrate over the population allele frequencies. To calculate the joint likelihood of samples from several populations we extend the usual pruning algorithm [30] by summing over the identity of ancestral alleles for each of the populations, in addition to the internal nodes in the phylogeny, weighted by the phylogenetic substitution probabilities that arise from the population genetic model.

To illustrate, we construct a biallelic model that includes, as a special case, PRF, and adapt it to the analysis of MK tables [35]. For the multiallelic treatment that follows, we use as our starting point the stationary distribution of allele frequencies in a parent-independent model with selection [18]. We develop a codon-based model of recurrent directional selection with two fitness classes, by adopting the standard technique of measuring fitness relative to the ancestral allele [25, 36]. To introduce parent-dependent

mutation, we adapt the likelihood by conditioning the mutation rate on the ancestral allele, and employing a coalescent approximation; through Wright-Fisher simulations we test the adequacy of the approximation. We use an adaptive sliding window model for intragenic variation in selection pressure [42], and apply the new approach to Bayesian analyses of polymorphism data in a meningococcal antigen [48] and polymorphism and divergence data in three HIV-1 populations [50].

Phylogenetic Conditioning and Population Pruning

At a single locus, let **f** be a vector of the frequencies of K alleles in a population (typically, $K = 4$ nucleotides, 20 amino acids or 61 non stop codons), where $\sum_{i=1}^{K} f_i = 1$. We introduce *phylogenetic conditioning* to the stationary distribution of allele frequencies in the population, $p(\mathbf{f})$, using Bayes rule. Suppose A is the identity of the ancestral allele, then the phylogenetically conditioned distribution is

$$p(\mathbf{f} \mid A) = \frac{\Pr(A \mid \mathbf{f})\, p(\mathbf{f})}{\Pr(A)}, \tag{1}$$

where $\Pr(A \mid \mathbf{f})$ is the probability that allele A is ancestral given **f**, and $\Pr(A)$ is a normalizing constant. Suppose **x** is a vector of the number of times each allele was counted in a sample at a single locus, where $\sum_{i=1}^{K} x_i = n$. To calculate the likelihood $\Pr(\mathbf{x} \mid A)$, conditional on the ancestral allele, we integrate over the population allele frequencies

$$\Pr(\mathbf{x} \mid A) = \int_{\mathbf{f}} \Pr(\mathbf{x} \mid \mathbf{f})\, p(\mathbf{f} \mid A), \tag{2}$$

where $\Pr(\mathbf{x} \mid \mathbf{f})$ is an appropriate sampling distribution such as the multinomial.

Inherent to phylogenetic conditioning is the idea that the ancestral allele cannot be directly observed for a population. Therefore we have to account for uncertainty in the ancestral allele by summing over all possibilities when calculating the likelihood. This extends the standard phylogenetic approach of summing over the allelic state at internal nodes on a tree, most efficiently done by Felsenstein's pruning algorithm [30], by requiring that the ancestral states at the tips of the tree are also summed over. We define the *population pruning algorithm* as follows:

1. Label nodes 1…r in order of greater tree depth, where r is the root.
2. For i in 1…(r – 1) and j in 1…K calculate α_{ij}, where

$$\alpha_{ij} = \begin{cases} \sum_{k=1}^{K} p_{jk}^{(t_i)} \alpha_{D_{i1}k} \alpha_{D_{i2}k} & \text{if } i \text{ is internal} \\ \sum_{k=1}^{K} p_{jk}^{(t_i)} \varepsilon_{ik} & \text{if } i \text{ is a tip.} \end{cases} \quad (3)$$

3. Calculate the full likelihood as $\sum_{k=1}^{K} \pi_k \alpha_{D_{r1}k} \alpha_{D_{r2}k}$.

In the above, D_{i1} and D_{i2} are the descendants of node i, t_i the branch length above node i, , π_k and $p_{jk}^{(t)}$ the stationary distribution and transition probabilities of the phylogenetic Markov chain and $\varepsilon_{ik} = \Pr(\mathbf{x}_i \mid A = k)$ is the phylogenetically conditioned likelihood for the population at node i.

Biallelic Case

Consider a two-allele model of directional selection where s is the selective advantage of allele 1 over allele 2 and μ_1 and μ_2 are the per-generation rates of mutation to alleles 1 and 2 respectively. The stationary distribution of allele frequencies is [18]

$$p(\mathbf{f}) \propto \mathrm{e}^{\gamma f_1} f_1^{\theta_1 - 1} f_2^{\theta_2 - 1}, \tag{4}$$

where $\gamma = 2PN_e s$ and $\theta_i = 2PN_e \mu_i$ are the population-scaled selection coefficient and mutation rates, P the ploidy and N_e the effective population size. The model is time-reversible [51], so $\Pr(A \mid \mathbf{f})$ equals the forward-in-time fixation probability. This is known exactly in the biallelic case, but for analytic tractability and for the benefit of the multiallelic treatment that follows, we take the low-mutation limit [19]

$$\Pr(A \mid \mathbf{f}) = \begin{cases} \dfrac{1 - \mathrm{e}^{-\gamma f_1}}{1 - \mathrm{e}^{-\gamma}} & \text{if } A = 1 \\ \dfrac{\mathrm{e}^{-\gamma f_1} - \mathrm{e}^{-\gamma}}{1 - \mathrm{e}^{-\gamma}} & \text{if } A = 2. \end{cases} \tag{5}$$

A difficulty in analyses of selection is the question of how to assign selection coefficients to individual alleles. The conventional solution [25, 36] is to measure selective advantage relative to the ancestral allele. When $\gamma > 0$, derived types are favored over the ancestral type; this is known as positive selection. Conversely, when $\gamma < 0$, the ancestral type is favored over derived types; this is known as negative selection. We adopt this approach, which we refer to as recurrent directional selection. Thus

$$p(\mathbf{f} \mid A) = \frac{\left(1 - \mathrm{e}^{-\gamma f_A}\right) f_1^{\theta_1 - 1} f_2^{\theta_2 - 1}}{\mathrm{B}(\theta_1, \theta_2)\left[1 - {}_1\mathrm{F}_1(\theta_A, \theta_1 + \theta_2, -\gamma)\right]} \tag{6}$$

where $\mathrm{B}(a,b)$ is the beta function and ${}_1\mathrm{F}_1(a,b,c)$ is the confluent hypergeometric function. Integrating over a binomial sampling distribution,

$$\begin{aligned} p(\mathbf{x} \mid A) &= \int_0^1 \binom{n}{x_1} f_1^{x_1} (1 - f_1)^{x_2} \Pr(\mathbf{f} \mid A) \mathrm{d} f_1 \\ &= \binom{n}{x_1} \frac{\mathrm{B}(x_1 + \theta_1, x_2 + \theta_2)}{\mathrm{B}(\theta_1, \theta_2)} \frac{\left[1 - {}_1\mathrm{F}_1(x_A + \theta_A, n + \theta_1 + \theta_2, -\gamma)\right]}{\left[1 - {}_1\mathrm{F}_1(\theta_A, \theta_1 + \theta_2, -\gamma)\right]}. \end{aligned} \tag{7}$$

For the population pruning algorithm we also need the phylogenetic substitution rate per PN_e generations, approximated by

$$q_{ij} = \lim_{f \to 0} \frac{\theta_j}{2} \frac{1 - e^{-\gamma f}}{f\left(1 - e^{-\gamma}\right)} = \frac{\theta_j}{2} \frac{\gamma}{1 - e^{-\gamma}} \tag{8}$$

[Nie03], and the resulting stationary allele frequencies $\pi_i = \theta_i / \left(\theta_1 + \theta_2\right)$.

This model includes PRF [36] as a special case in the infinite sites limit. To emphasize the close connection between the methods, we apply our biallelic model to the analysis of data in the form of MK tables, in which the number of synonymous and non-synonymous sites that are polymorphic or divergent between a pair of species are counted. Following the MK-PRF parameterization [36, 39], θ_S and θ_R are the population-scaled synonymous and non-synonymous mutation rates respectively, τ is the divergence time between the species (in units of PN_e generations) and γ is the selection coefficient, shared by all sites. We assume independence between sites, which is equivalent to assuming linkage equilibrium between sites within each population. Using Equations 6-8 and the population pruning algorithm, we calculate the probabilities (φ_{PS}, φ_{DS}, φ_{PN}, φ_{DN}) that a site is polymorphic (*P*) or divergent (*D*) by summing over possible configurations in the two species that lead to these patterns for synonymous (*S*; $\gamma = 0$) and non-synonymous (*N*; $\gamma \neq 0$) sites separately. We take $1 - \sum \varphi$ to be the probability a site is neither polymorphic nor divergent, and estimate the parameters (θ_S, θ_R, τ, γ) by maximum likelihood using a multinomial likelihood with five outcomes.

Multiallelic Hot-or-Not Model

For multiallelic models, our starting point is the Wright-Dirichlet distribution, which is the solution to the stationary distribution of allele frequencies in a diffusion model with parent-independent mutation and selection (PIMS). Conjectured by Wright [18], it is a variant on the Dirichlet distribution in which gene frequency combinations conferring higher population fitness are up-weighted. In our notation,

$$p(\mathbf{f}) \propto \mathrm{e}^{W(\mathbf{f})} \prod_{i=1}^{K} f_i^{\theta_i - 1}, \tag{9}$$

where $w(\mathbf{f})$ is the population fitness as a function of $\mathbf{f}$ and $W(\mathbf{f}) = 2PN_e w(\mathbf{f})$ is its population-scaled counterpart.

We concentrate on models with two fitness classes, which we refer to as hot-or-not models. In the hot-or-not model, alleles encoding the favored (hot) phenotype have selective advantage γ over alleles encoding other phenotypes; in a codon model, these phenotypes correspond to the 20 amino acids. In the hot-or-not model, the Wright-Dirichlet distribution simplifies to

$$p(\mathbf{f}) \propto \mathrm{e}^{\gamma F_H} \prod_{i=1}^{K} f_i^{\theta_i - 1}, \tag{10}$$

where F_i is the total frequency of alleles with the same phenotype as allele i, and H represents an allele encoding the hot phenotype.

To proceed, we make two assertions informed by the biallelic case. First, that this multiallelic model is time-reversible, which implies that the probability allele A is ancestral given $\mathbf{f}$ equals the fixation probability. Second, that in the low-mutation limit, the fixation probability is

$$\Pr(A \mid \mathbf{f}) = \begin{cases} \dfrac{1-\mathrm{e}^{-\gamma F_H}}{1-\mathrm{e}^{-\gamma}} \dfrac{f_A}{F_H} & \text{if } A \text{ is hot} \\ \dfrac{\mathrm{e}^{-\gamma F_H}-\mathrm{e}^{-\gamma}}{1-\mathrm{e}^{-\gamma}} \dfrac{f_A}{1-F_H} & \text{if } A \text{ is not.} \end{cases} \tag{11}$$

Based on these assertions and assuming recurrent selection, in Appendix A we derive the phylogenetically conditioned distribution

$$p(\mathbf{f} \mid A) = \frac{f_A}{F_A} \frac{\Theta_A}{\theta_A} \frac{\left(1-\mathrm{e}^{-\gamma F_A}\right) \prod_{i=1}^{K} f_i^{\theta_i - 1}}{\mathrm{B}(\boldsymbol{\theta}) \left[1 - {}_1\mathrm{F}_1\left(\Theta_A, \Theta, -\gamma\right)\right]}, \tag{12}$$

and corresponding likelihood

$$p(\mathbf{x} \mid A) = \binom{n}{\mathbf{x}} \frac{\left(x_A + \theta_A\right)\Theta_A}{\theta_A \left(X_A + \Theta_A\right)} \frac{\mathrm{B}(\mathbf{x}+\boldsymbol{\theta})}{\mathrm{B}(\boldsymbol{\theta})} \frac{\left[1 - {}_1\mathrm{F}_1\left(X_A + \Theta_A, n + \Theta, -\gamma\right)\right]}{\left[1 - {}_1\mathrm{F}_1\left(\Theta_A, \Theta, -\gamma\right)\right]}, \tag{13}$$

where X_A and Θ_A are the total number of copies and total mutation rate for alleles with the ancestral phenotype, and Θ is the total mutation rate across all alleles.

Parent-Dependent Models

Phylogenetic conditioning lends itself naturally to approximating parent-dependent mutation because the ancestral allele will often be the genetic background upon which a mutation arises. Therefore we modify the mutation rate vector $\boldsymbol{\theta}$ depending on the ancestral allele. In Appendix B we detail the rationale for our choice of ancestor-dependent $\boldsymbol{\theta}$. Briefly, we match the rates for a parent-independent and a parent-dependent model by using coalescent-averaged mutation probabilities, in which we calculate the expected probability of mutation from the ancestral allele A to every other allele, averaging over the neutral coalescent time.

We use our parent-dependent approximation to implement a codon-based analog to Hasegawa, Kishino and Yano's [44] mutation model (henceforth HKY85). In the codon-based HKY85 model the alleles are the $K = 61$ non stop codons, and the mutation rate is

$$\theta_{ij} = \frac{\pi_j}{C}\begin{cases} 1 & \text{if } i \text{ and } j \text{ differ by 1 transversion} \\ \kappa & \text{if } i \text{ and } j \text{ differ by 1 transition} \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where C normalizes the rate matrix so that the expected mutation rate is $\theta/2$ per PN_e generations. This allows us to produce a population genetic formulation of the phylogenetic Nielsen and Yang model [25] commonly used for analyses of selection. We call this model HKY85 with recurrent selection; it has stationary distribution π and (following Equation 8) phylogenetic substitution rate

$$q_{ij} = \frac{\theta_{ij}}{2}\begin{cases} 1 & \text{if codons } i \text{ and } j \text{ are synonymous} \\ \dfrac{\gamma}{1-\mathrm{e}^{-\gamma}} & \text{if codons } i \text{ and } j \text{ are non-synonymous.} \end{cases} \tag{15}$$

Nielsen and Yang [46] also formulated a population genetic interpretation of their phylogenetic model [25] (henceforth NY98). In their formulation, which is followed by omegaMap [42], selection is treated as a form of mutational bias, so that advantageous alleles are more likely to occur by mutation than neutral alleles, and deleterious alleles less likely. Thereafter, the population dynamics of all alleles are treated as neutral. We can mimic this treatment of selection as mutational bias, which we refer to as the aristogenetic NY98 model, by taking the neutral limit $\gamma \to 0$ of the likelihood (Equation 13) and equating the mutation rate to the phylogenetic substitution rate (Equation 15).

Testing the Model by Simulation

Since our likelihood function relies on several approximations, we wished to evaluate its performance in inference. To do so, we employed the Bayesian idea of calibration [45]; a 95% credible interval is said to be well-calibrated if it contains the true parameter value with probability 0.95. In simulations this can be tested directly. For the HKY85 model with recurrent selection, we simulated parameters 200 times as follows: θ from a log-uniform distribution with range (0.02, 0.2), κ from a log-uniform distribution with range $(0.05, 20)$, γ from a normal distribution with mean 0 and standard deviation 10. For each of the 200 draws of (θ,κ,γ) we used Wright-Fisher simulations to generate data comprising $n = 30$ sequences of length $L = 250$ codons; sites were simulated independently. We then performed inference using Markov chain Monte Carlo (MCMC) to obtain a posterior distribution on the parameters (θ,κ,γ) for each dataset using the same priors that simulated the parameters. Under these conditions, if our likelihood approximation is adequate, we expect that the 95% credible intervals for each parameter will contain the truth in 95% of simulations, with an acceptance range of 184-196 datasets.

We performed this procedure for two scenarios: one in which we assumed the ancestral allele is known for each site in the population. This serves as a test of the phylogenetically conditioned likelihood. In the second, we assumed the ancestral allele is known 10 PN_e generations prior to sampling. This serves as an additional test of the phylogenetic substitution probabilities and the population pruning algorithm. Preliminary simulations revealed a sensitivity to the precise definition of which allele is ancestral. In

Appendix C we provide details of our investigation into the definition of the ancestral allele. The outcome of these deliberations was to define the ancestral allele as the oldest allele currently segregating in the population.

Inferring Variation in Selection in a Meningococcal Antigen

For the analysis of polymorphism data in a sample of 23 meningococcal *porB3* sequences [48] we used an adaptive sliding window model for variation in selection pressure [42]: variation in the *DN*/*DS* ratio $\omega = \gamma / \left(1 - e^{-\gamma}\right)$ along the gene is treated as a piecewise constant function, such that adjacent sites share the same ω with probability $1 - p$ and have different values of ω with probability p. Reversible jump MCMC is used to integrate over the number and position of transitions in ω, and the values of ω [42], resulting in a smoothly varying point estimate and credible intervals for ω along the gene.

Three population genetic formulations were investigated: omegaMap, the NY98 aristogenetic model, and the HKY85 recurrent selection model. We assumed improper log-uniform priors for θ, κ, and (in the case of omegaMap) ρ, an exponential prior with mean 1 for ω, and an average window length of $1/p = 30$ codons. Sites containing indels were removed from the analysis. Regular MCMC moves were used to integrate over θ, κ and ρ [42]. For each analysis two chains of length 500,000 iterations were run and, after removing a burn-in of 10,000 iterations, compared for convergence and merged to obtain the posterior distribution.

Analysis of Multiple Populations of HIV-1

We analyzed polymorphism and divergence in the *gag* gene of three HIV-1 populations [50] sampled from a donor-recipient transmission pair: one from the donor (D0) and recipient (R0) shortly following transmission, and one from the recipient four years later (R1). Owing to partially overlapping PCR products and the presence of indels, the number of sequences per site was not constant in *gag*. The mean sample size was 8.8, 8.8 and 26.2 sequences in D0, R0 and R1 respectively. We treated indels as missing data, rather than discarding the entire site in the alignment.

We separately parameterized the three branches of the tree relating populations D0, R0 and R1. Applying the HKY85 recurrent selection model, we estimated θ, κ, γ and the branch length T for each. We assumed improper log-uniform priors on θ, κ and T. We employed the same model of variation in γ along the gene as for ω above, assuming a mean window length of 5 codons and a normal prior distribution for γ with mean 0 and standard deviation 10. We used MCMC to obtain the posterior distribution of the parameters, modifying the reversible jump moves because γ can be positive or negative. Two chains of 1,000,000 iterations were run, with a burn-in of 10,000 iterations. These were compared for convergence and merged to obtain final results.

Acknowledgements

We would like to thank Graham Coop, Peter Donnelly, Bob Griffiths, Dick Hudson, Marty Kreitman, Guy Sella and the members of the Pritchard, Przeworski and Stephens labs for fruitful discussions and useful insight and the city of Chicago for the backdrop. DJW was funded by XXX. MP was funded by XXX.

Last saved 9/4/10 15:53

References

1. Tinbergen N (1963) On aims and methods of ethology. Zeitschrift für Tierpsychologie 20: 410-433.
2. Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm. Proc Roy Soc Lond B 205: 581-598.
3. Kelley JL, Swanson W (2008) Positive selection in the human genome: from genome scans to biological significance. Ann Rev Genom Hum Genet 9: 143-160.
4. Sella G, Petrov D, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the Drosophila genome? PLoS Genet 5: e1000495.
5. Pool JE, Hellman I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. Genom Res 20: 291-300.
6. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4: e1000083.
7. Barrerio LB, Quintana-Murci L (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11: 17-30.
8. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, et al. (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 6: e90.
9. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.
10. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

11. Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. Genetics 74: 175-195.
12. McVean G (2007) The structure of linkage disequilibrium around a selective sweep. Genetics 175: 1395-1406.
13. McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. PLoS Genetic 5: e1000471.
14. Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, et al. (2009) Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet 5: e1000753.
15. Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275-276.
16. Barrerio LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340-345.
17. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, et al. (2009) Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet 5: 1000592.
18. Wright S (1949) Adaptation and selection. In: Jepson GL, Simpson GG, Mayr E, editors. Genetics, Paleontology and Evolution. Princeton University Press. pp. 365-389.
19. Kimura M (1955) Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symp Quant Biol 20: 33-55.

20. Karlin S, Taylor HM (1981) A Second Course in Stochastic Processes. New York: Academic Press.
21. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159: 1779-1788.
22. Donnelly P, Nordborg M, Joyce P (2001) Likelihoods and simulation methods for a class of nonneutral population genetics models. Genetics 159: 853-867.
23. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, et al. (2006) Forces shaping the fastest evolving regions in the human genome. PLoS Genet 2: e168.
24. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418-426.
25. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929-936.
26. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908-917.
27. Zhang J, Nielsen R, Yang Z (2005) An improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472-2479.
28. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.
29. Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

30. Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22: 240-249.
31. Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156: 879-891.
32. Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164: 1229-1236.
33. Shriner D, Nickle DC, Jensen MA, Mullins JI (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet Res 81: 115-121.
34. Ho SY, Shapiro B, Phillips MJ, Cooper A, Drummond AJ (2007) Evidence for time dependency of molecular rate estimates. Syst Biol 56: 515-522.
35. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.
36. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161-1176.
37. Charlesworth B (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res 63: 213-227.
38. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415: 1022-1024.
39. Bustamante CD, Nielsen R, Sawyer SA, Olsen KA, Purugganan MD, et al. (2002) The cost of inbreeding in Arabidopsis. Nature 416: 531-534.

40. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.

41. Nielsen R, Bustamante CD, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3: e170.

42. Wilson DJ, McVean G (2006) Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172: 1411-1425.

43. Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, et al. (2009) Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. PLoS Genet 5: e100562.

44. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160-174.

45. Dawid AP (1982) The well-calibrated Bayesian. Journal of the American Statistical Association. 77: 605-610.

46. 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.

47. Osborn HF (1934) Aristogenesis, the creative principle in the origin of species. Am Nat 716: 193-235.

48. Urwin R, Holmes EC, Fox AJ, Derrick JP, Maiden MCJ (2002) Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen PorB. Mol Biol Evol 19: 1686-1694.

49. Kryazhimskiy S, Plotkin J (2008) The population genetics of dN/dS. PLoS Genet 4: e1000304.

50. Liu Y, McNevin J, Cao J, Zhao H, Genowati I, et al. (2006) Selection on the human immunodeficiency virus type 1 proteome following primary infection. J Virol 80: 9519-9529.

51. Griffiths RC (2003) The frequency spectrum of a mutation, and its age, in a general diffusion model. Theor Popul Biol 64: 241-251.

52. Erdéyli A (1939) Integration of certain systems of linear partial differential equations of hypergeometric type. Proc Roy Soc Edin A 59: 224-241.

53. Phillips PCB (1988) The characteristic function of the Dirichlet and multivariate F distributions. Cowles Foundation Discussion Paper No. 865.

54. Kingman JFC (1982) On the genealogy of large populations. J Appl Prob 19A: 27-43.

55. Griffiths RC, Lessard S (2005) Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. Theor Popul Biol 68: 167-177.

56. Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165: 2213-2233.

Figure Legends

Figure 1 Synthesizing population and phylogenetic components of an evolutionary model. At the phylogenetic timescale, fluctuations (A) in gene frequency over time are conceptually reduced (B) to a consideration of the substitution process alone. When considering a snapshot of the population (C), we employ a population genetics model of gene frequencies conditioned on the ancestral allele, whose identity is governed by the phylogenetic substitution process. To calculate the likelihood of a sample of sequences from several populations (D), we can use Felsenstein's pruning algorithm to sum over the ancestral alleles at internal nodes (d,e), as usual, and additionally at the tips (a-c). This approach accounts for the presence of derived alleles in molecular sequences.

Figure 2 Testing calibration by simulation. The posterior mean (circles) and 95% credible interval (vertical lines) of the mutation rate (θ), transition:transversion ratio (κ) and strength of selection (γ) are plotted against their true values for 200 simulated datasets under two scenarios. To test phylogenetic conditioning, inference was performed with known ancestral states. To test phylogenetic conditioning and population pruning, the ancestral state was recorded 10 PN_e generations prior to sampling. Colored lines draw attention to datasets for which the truth lies outside the 95% credible interval. The top left number in each graph reports the number of simulations for which the 95% credible interval enveloped the truth (a range of 184–196 is acceptable). In all cases 30 sequences of length 250 codons were simulated per dataset.

Figure 3 Inference of selection from polymorphism data in meningococcal *porB3*: comparing aristogenetic to recurrent selection models. The inferred posterior mean (solid line) and 95% credible interval (shaded region) of the *dN*/*dS* ratio (ω) is plotted for (A) omegaMap (B) an aristogenetic Wright-Dirichlet model designed to emulate omegaMap (C) a Wright-Dirichlet model of recurrent directional selection. In (D) the posterior probability of positive selection ($\omega > 1$) is compared between the three approaches.

Figure 4 Immune escape and reversion in an HIV-1 transmission pair. (A) The posterior probability of positive selection in gag p17 and p24 is shown for three population samples: the donor shortly after transmission (D0), the recipient shortly after transmission (R0) and the recipient 4 years later (R1). The height of colored vertical bars indicates the sample frequency of amino acid variants relative to the inferred sequence close to transmission. The locations of known epitopes that are HLA-restricted in the recipient (4 epitopes) or donor (GL8) are marked by horizontal solid black bars. (B) A schematic of the sampling points in the donor-recipient transmission chain.

Figure S1 The operational definition of ancestral identity affects the accuracy of phylogenetic conditioning. (A) When the operational definition of ancestral identity is the last allele to have fixed or – as here – the state of the population MRCA, there is a discrepancy between theory (purple bars) and simulations (green bars). Simulations, which were conducted for a codon model with $\theta = 0.3$, $\kappa = 1$, $\gamma = 0$, report a non-negligible probability of failing to sample the ancestral allele, which is not predicted from theory, and could be erroneously attributed to positive selection (red bars). (B) When the

operational definition of ancestral identity is the oldest allele segregating in the population, the differences are resolved. (C) The cause of the problem: an ancestral allele (cyan) is lost from the population long before any other allele fixes, generating appreciable periods of time when the ancestral allele is no longer segregating in the population.

Table 1 Glossary of terms

Aristogenetic mutation: the treatment of selection within populations merely as a form of mutational bias in which better (fitter) forms are more likely to be generated.

Coalescent averaged: the method of taking the neutral expectation of a quantity such as a mutation rate.

Hot-or-not model: an adaptive landscape in which there are just two fitness forms, one that is hot (*i.e.* fit) and one that is not (*i.e.* less fit).

Phylogenetic conditioning: the introduction of phylogenetic dependency into a distribution of gene frequencies by conditioning on the identity of the ancestral allele.

PIMS: a model of parent-independent mutation and selection.

Population pruning: an extension of Felsenstein's pruning algorithm to include uncertainty in the identity of ancestral alleles at the tips of the phylogeny.

Recurrent selection: a model of directional selection in which the process of substitution and shifts in the adaptive landscape are deliberately confounded by assuming that fitness is determined by the ancestral allele. Positive selection arises when the ancestral type is less fit than derived types, and the converse is referred to as negative selection.

Wright-Dirichlet distribution: a variant of the Dirichlet distribution in which gene frequencies that confer higher population fitness are up-weighted. Conjectured by Wright to be the solution to the stationary distribution of gene frequencies under PIMS.

Table 2 Emulating Sawyer and Hartl

θ_S	θ_R	τ	γ
1.11 / 1.11	2.86 / 2.87	3.52 / 3.54	-0.774 / -0.764
1.11 / 1.11	1.11 / 1.11	2.61 / 2.62	0.000 / 0.001
1.42 / 1.42	1.21 / 1.21	5.33 / 5.36	0.614 / 0.615
0.79 / 0.79	2.19 / 2.19	1.53 / 1.53	-0.775 / -0.768
1.11 / 1.11	2.84 / 2.85	7.14 / 7.18	-1.44 / -1.43
0.79 / 0.79	2.84 / 2.85	15.5 / 15.6	-2.27 / -2.27
0.998 / 0.999	0.607 / 0.607	5.01 / 5.02	1.19 / 1.19
2.00 / 2.00	0.123 / 0.123	5.01 / 5.04	6.02 / 5.98
1.26 / 1.27	0.251 / 0.251	8.49 / 8.54	3.90 / 3.89
1.58 / 1.58	3.41 / 3.43	4.70 / 4.73	-2.27 / -2.27

A comparison of maximum likelihood estimates of the synonymous (θ_S) and non-synonymous (θ_R) mutation rates, the divergence time (τ) and strength of selection (γ) using the Poisson random field model of Sawyer and Hartl (left values) and our finite sites version based on a Wright-Dirichlet model (right values) for Toll Like Receptors 1-10 in humans.

Appendix A: Deriving the Multiallelic Hot-or-Not Model

The greatest difficulty in applying the Wright-Dirichlet distribution to data analysis is computing the normalization constant. Various methods are available, including importance sampling and Taylor series expansion [22]. In a multiallelic PIMS model of directional selection with no dominance, the Wright-Dirichlet distribution simplifies to

$$p(\mathbf{f}) \propto \prod_{i=1}^{K} \mathrm{e}^{\gamma_i f_i} f_i^{\theta_i - 1}, \tag{A1}$$

where γ_i is the population-scaled selection coefficient for allele i. The normalizing constant for this distribution is equal to the characteristic function for a Dirichlet distribution with parameter $\boldsymbol{\theta}$. In addition to the methods mentioned above, it can be computed numerically using contour integration in the complex plane [52, 53]

$$\int_{\Phi_K} \prod_{i=1}^{K} e^{\gamma_i f_i} f_i^{\theta_i - 1} \, \mathrm{d}\mathbf{f} = \frac{\prod_{i=1}^{K} \Gamma(\theta_i)}{2\pi I} \int_L e^t \prod_{i=1}^{K} (t - \gamma_i)^{-\theta_i} \, \mathrm{d}t, \tag{A2}$$

where I represents the imaginary unit, Φ_K is a $(K-1)$-simplex. The contour L is a loop beginning and ending at $-\infty$, and encircling in the positive direction (anticlockwise when the x and y axes correspond to the real and imaginary lines respectively) all the finite singularities of the integrand (*i.e.* the values of γ_i). We have *Mathematica* code that implements this integral (available on request) but found it numerically unstable.

However, in the hot-or-not model we can use Equation A2 to show that

$$p(\mathbf{f}) = \frac{\mathrm{e}^{\gamma F} \prod_{i=1}^{K} f_i^{\theta_i - 1}}{\mathrm{B}(\boldsymbol{\theta}) \, {}_1\mathrm{F}_1(\Theta_H, \Theta, \gamma)}, \tag{A3}$$

where γ is the population-scaled selective difference between fitness classes, F is the total frequency of alleles in the favored (hot) fitness class, Θ_H is the total mutation rate to alleles of that class and Θ is the total mutation rate for all alleles. $\mathrm{B}(\boldsymbol{\theta})$ is the beta function with vector argument, so that

$$\mathrm{B}(\boldsymbol{\theta}) = \prod_{i=1}^{K} \Gamma(\theta_i) \Big/ \Gamma\left(\sum_{i=1}^{K} \theta_i\right), \tag{A4}$$

and ${}_1\mathrm{F}_1(a,b,c)$ is the confluent hypergeometric function, which is the solution to the confluent hypergeometric equation. The biallelic directional selection model can be posed in terms of this equation. Hypergeometric equations appear frequently in theoretical physics, and we speculate that the solutions to related hypergeometric equations may yield the solutions to other problems of interest in population genetics.

Let $\mathcal{H}$ be the set of favored alleles, and $\mathcal{N}$ the set of disfavored alleles. The numbers of alleles in the two classes are K_H and K_N. Define $\mathbf{h} = \{f_i / F : i \in \mathcal{H}\}$, $\mathbf{g} = \{f_i / (1-F) : i \in \mathcal{N}\}$, $\boldsymbol{\theta}^{(H)} = \{\theta_i : i \in \mathcal{H}\}$ and $\boldsymbol{\theta}^{(N)} = \{\theta_i : i \in \mathcal{N}\}$. By making the change of variables $(\mathbf{f}) \to (F, \mathbf{g}, \mathbf{h})$, Equation A3 can be factorized so that

$$\begin{aligned} p(F,\mathbf{g},\mathbf{h}) = &\left[\frac{\mathrm{e}^{\gamma F} F^{\Theta_H - 1}(1-F)^{\Theta_N - 1}}{\mathrm{B}(\Theta_H,\Theta_N)\,{}_1\mathrm{F}_1(\Theta_H,\Theta,\gamma)}\right] \\ &\times \left[\frac{1}{\mathrm{B}(\boldsymbol{\theta}^{(N)})}\prod_{i=1}^{K_N} g_i^{\theta_i^{(N)}-1}\right]\left[\frac{1}{\mathrm{B}(\boldsymbol{\theta}^{(H)})}\prod_{i=1}^{K_H} h_i^{\theta_i^{(H)}-1}\right], \end{aligned} \tag{A5}$$

where $\Theta_N = \Theta - \Theta_H$. This demonstrates that F follows a biallelic Wright-Dirichlet distribution with parameters (Θ_H, Θ_N, γ), and $\mathbf{g}$ and $\mathbf{h}$ follow independent Dirichlet distributions with parameters $\boldsymbol{\theta}^{(N)}$ and $\boldsymbol{\theta}^{(H)}$ respectively. The Dirichlet distribution arises

from the neutral case, and this factorization suggests that in a hot-or-not model, evolution within class $\mathcal{H}$ or $\mathcal{N}$ can be characterized as neutral. This in turn supports our assertion (see main text) that, in the low-mutation limit, the probability of fixation of allele *A* equals the biallelic fixation probability for the whole class multiplied by the neutral fixation probability for allele *A* within its class.

To introduce phylogenetic conditioning, suppose allele *A*, which is a member of the favored class, is ancestral. From Equations 11 and A5,

$$\begin{aligned} p(F,\mathbf{g},\mathbf{h}\mid A) &= \frac{\left(\mathrm{e}^{\gamma F}-1\right)F^{\Theta_H-1}(1-F)^{\Theta_N-1}\prod_{i=1}^{K_N} g_i^{\theta_i^{(N)}-1} h_A \prod_{i=1}^{K_H} h_i^{\theta_i^{(H)}-1}}{\int_0^1 \left(\mathrm{e}^{\gamma\varphi}-1\right)\varphi^{\Theta_H-1}(1-\varphi)^{\Theta_N-1}d\varphi \int_{\Phi_{K_N}} \prod_{i=1}^{K_N}\zeta_i^{\theta_i^{(N)}-1}d\zeta \int_{\Phi_{K_H}} \eta_A \prod_{i=1}^{K_H}\eta_i^{\theta_i^{(H)}-1}d\eta} \\ &= \frac{\left(\mathrm{e}^{\gamma F}-1\right)F^{\Theta_H-1}(1-F)^{\Theta_N-1}\prod_{i=1}^{K_N} g_i^{\theta_i^{(N)}-1} h_A \prod_{i=1}^{K_H} h_i^{\theta_i^{(H)}-1}}{\mathrm{B}(\Theta_H,\Theta_N)\left[{}_1\mathrm{F}_1(\Theta_H,\Theta,\gamma)-1\right]\mathrm{B}\left(\boldsymbol{\theta}^{(N)}\right)\frac{\theta_A}{\Theta_H}\mathrm{B}\left(\boldsymbol{\theta}^{(H)}\right)}. \end{aligned} \tag{A6}$$

Reversing the change-of-variables,

$$p(\mathbf{f}\mid A) = \frac{f_A}{F}\frac{\Theta_H}{\theta_A}\frac{\left(\mathrm{e}^{\gamma F}-1\right)\prod_{i=1}^{K} f_i^{\theta_i-1}}{\mathrm{B}(\boldsymbol{\theta})\left[{}_1\mathrm{F}_1(\Theta_H,\Theta,\gamma)-1\right]}. \tag{A7}$$

Assuming a multinomial sampling distribution, the likelihood for a sample of size *n* comprising x_i copies of allele *i* equals

$$\begin{aligned} p(\mathbf{x}\mid A) &= \int_{\Phi_K}\binom{n}{\mathbf{x}}\prod_{i=1}^{K} f_i^{x_i}\, p(\mathbf{f}\mid A)\,\mathrm{d}\mathbf{f} \\ &= \binom{n}{\mathbf{x}}\frac{(x_A+\theta_A)\Theta_H}{\theta_A(X_H+\Theta_H)}\frac{\mathrm{B}(\mathbf{x}+\boldsymbol{\theta})}{\mathrm{B}(\boldsymbol{\theta})}\frac{\left[{}_1\mathrm{F}_1(X_H+\Theta_H,n+\Theta,\gamma)-1\right]}{\left[{}_1\mathrm{F}_1(\Theta_H,\Theta,\gamma)-1\right]}, \end{aligned} \tag{A8}$$

where X_H is the number of alleles sampled in class $\mathcal{H}$ and $\binom{n}{\mathbf{x}} = n!\prod_{i=1}^{K}\frac{1}{x_i!}$. However, in a recurrent selection model, the ancestral allele is actually *disfavored* when $\gamma > 0$, leading to the forms for Equations 12 and 13 in the main text.

Appendix B: Approximating Parent-Dependent Mutation

In developing an approximation to parent-dependent mutation we use a coalescent-averaged probability of mutation from the ancestral allele, A, to every other allele. The coalescent [54] is a statistical distribution for the topology and branch lengths of the genealogy of a random sample of sequences that arises in a selectively neutral model. It is convenient to employ the neutral case as the basis for our approximation, because there is an explicit separation between the genealogical process and the mutational process. Our approach is to find a parent-independent mutation rate vector $\boldsymbol{\theta}'$ that matches the parent-dependent probabilities of observing an individual with allele i given the ancestral allele A in a neutral population.

Taking the neutral limit of the phylogenetically conditioned Wright-Dirichlet distribution (Equation 13) yields

$$p(\mathbf{f} \mid A) = \frac{\prod_{i=1}^{K} f_i^{\theta_i + e_{Ai} - 1}}{\mathrm{B}(\boldsymbol{\theta} + \mathbf{e}_A)}, \qquad \text{(B1)}$$

i.e. a Dirichlet distribution with parameter $\boldsymbol{\theta} + \mathbf{e}_A$, where e_{Ai} equals 1 if $i = A$ and 0 otherwise. This suggests that conditioning on the ancestral allele is probabilistically equivalent to conditioning on having observed allele A in a sample of size 1. Using the

coalescent, we can calculate the probability, m_{AB}, of observing an allele of type B having already observed an allele of type A, for general mutation models,

$$m_{AB} = \int_0^\infty p_{AB}^{(2t)} \, \mathrm{e}^{-t} \, \mathrm{d}t \,, \tag{B2}$$

since t follows an exponential distribution with rate 1 when time is measured in units of PN_e generations. For a parent-independent model where the rate of mutation to allele i is $\theta_i/2$, the transition probability from allele A to B in time t is

$$p_{AB}^{(t)} = \begin{cases} \mathrm{e}^{-\Theta t/2} + \dfrac{\theta_B}{\Theta}\left(1 - \mathrm{e}^{-\Theta t/2}\right) & \text{if } A = B \\ \dfrac{\theta_B}{\Theta}\left(1 - \mathrm{e}^{-\Theta t/2}\right) & \text{if } A \neq B, \end{cases} \tag{B3}$$

where $\Theta = \sum_{i=1}^{K} \theta_i$, therefore

$$m_{AB} = \begin{cases} \dfrac{\theta_B + 1}{\Theta + 1} & \text{if } A = B \\ \dfrac{\theta_B}{\Theta + 1} & \text{if } A \neq B. \end{cases} \tag{B4}$$

For a parent-dependent, time-reversible model,

$$p_{AB}^{(t)} = \sum_{i=1}^{K} v_{Ai} \, \mathrm{e}^{d_i t} \, v_{iB}^{(-1)}, \tag{B5}$$

where $\mathbf{v}$ is a matrix of eigenvectors of the mutation rate matrix, $\mathbf{v}^{(-1)}$ is its inverse, and $\mathbf{d}$ is a vector of the corresponding eigenvalues, so

$$m_{AB} = \sum_{i=1}^{K} \frac{v_{Ai} v_{iB}^{(-1)}}{1 - 2d_i} \,. \tag{B6}$$

For non time-reversible models, the calculation is a little more involved.

For a given ancestral allele A, we wish to find $\boldsymbol{\theta}'$ so that m_{AB} ($B = 1 \ldots K$) is matched for the parent-independent and parent-dependent models. The solution to this problem satisfies

$$\theta'_B = \begin{cases} m_{AB}(\Theta' + 1) - 1 & \text{if } A = B \\ m_{AB}(\Theta' + 1) & \text{if } A \neq B, \end{cases} \tag{B7}$$

where the vector $\mathbf{m}_A$ is calculated from Equation B6 for a time-reversible model. However, there are only $K - 1$ degrees of freedom because $\mathbf{m}_A$ is constrained to sum to 1; therefore Θ' is not identifiable. We circumvent the problem by imposing the additional constraint that $\theta'_A = 0$, interpretable as no back mutation to the ancestral allele, which is consistent with the low-mutation limit for the fixation probability used to motivate our implementation of phylogenetic conditioning. This implies that

$$\Theta' = \frac{1 - m_{AA}}{m_{AA}}. \tag{B8}$$

Appendix C: On the Definition of Allelic Ancestry

We used Wright-Fisher simulations to test the adequacy of our approximate likelihood, as described in the Methods. During simulation we kept a record of the ancestral allele at every generation, which we initially defined as the last allele to have reached fixation. Preliminary results revealed a discrepancy between our theory and the simulations. The major departure from theoretical expectations was the frequency with which we drew samples in which the ancestral allele was completely absent. Figure S1A shows the distribution of the number of derived alleles in a sample of 30 in a neutrally evolving population from theory (purple bars) and simulation (green bars). For the parameters used, simulations suggested there should be a 5% probability of failing to

sample the ancestral type. Troublingly, only in the case of positive selection (red bars) did theory predict an appreciable probability of not sampling the ancestral allele. That neutral simulations should generate a pattern that the method confuses with positive selection was potentially of great concern.

Visualization of simulations over time revealed the cause of the problem (Figure S1C). When there are multiple alleles, the ancestral allele can be lost many generations before any other allele fixes. During these periods samples cannot contain the ancestral allele. By contrast, in the low-mutation limit, no more than two alleles ever co-segregate so the problem does not arise. Hence the discrepancy stems from utilizing the low-mutation limit for the probability of fixation during phylogenetic conditioning. The problems gets worse as the mutation rate increases, to the point where fixation rarely occurs, and the identity of the last allele to have fixed has little relevance to the current state of the population. In fact, in a PIMS model when $\theta > 1$, there is a non-zero probability that no allele will fix in a finite number of generations [20].

The fact that fixation is not even theoretically well-defined for some values of θ, and the observation that theory and simulations diverge even for modest mutation rates, led us to re-evaluate what the appropriate definition of allelic ancestry might in fact be when mutation is non negligible. We considered three alternatives for the operational definition of allelic ancestry:

1. The last allele to have fixed.
2. The allelic identity of the population MRCA.
3. The oldest allele still segregating in the population.

Note that all definitions are equivalent in the low-mutation limit. We repeated simulations using the alternative definitions of allelic ancestry, keeping track of the allelic state of the population most recent common ancestor (MRCA) or oldest allele. This is more difficult to implement because the population genealogy must be recorded and updated (by branching and pruning) from one generation to the next during Wright-Fisher simulation.

Using the population MRCA guarantees that the ancestral allele is well-defined, and quantitatively alleviated the discrepancy between theory and simulations, but did not eliminate it. Only by defining the ancestral allele as the oldest allele still segregating in the population, a definition inspired by a consideration of the age-ordered Ewens sampling formula [55], did we reconcile the simulations with the results expected from theory (Figure S1B). This leads us to reason that firstly, the low-mutation limit of the probability that an allele was the last to fix (equivalently, is the next to fix) is a good approximation to the probability that an allele is the oldest (will persist the longest) when mutation is non negligible. Secondly, the identity of the oldest allele characterizes the state of the population better than other definitions of allelic identity when mutation is non negligible. Therefore we might re-interpret the phylogenetic substitution process in terms of a change in the oldest allele, rather than a change in the fixed allele.

It is worth noting that in recurrent selection models, the definition of allelic ancestry affects the evolutionary dynamics, because the selection regime changes upon a change in the ancestral allele. We do not seek to defend this quirk of the model here except to say that it is consistent with the general approach of deliberately confounding the selection regime with the substitution process by measuring fitness relative to the

ancestral allele, which we prefer to interpret as the oldest allele segregating in the population.