Detection of Regional Variation in Selection Intensity within Protein-Coding Genes Using DNA Sequence Polymorphism and Divergence

Zi-Ming Zhao,[†,1] Michael C. Campbell,[†,1,2] Ning Li,[3] Daniel S.W. Lee,[1] Zhang Zhang,[4] and Jeffrey P. Townsend*,[1,3,5]

[1]Department of Biostatistics, Yale University, New Haven, CT

[2]Department of Biology, Howard University, Washington, DC

[3]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT

[4]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[5]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: jeffrey.townsend@yale.edu.

Associate editor: John Parsch

Abstract

Numerous approaches have been developed to infer natural selection based on the comparison of polymorphism within species and divergence between species. These methods are especially powerful for the detection of uniform selection operating across a gene. However, empirical analyses have demonstrated that regions of protein-coding genes exhibiting clusters of amino acid substitutions are subject to different levels of selection relative to other regions of the same gene. To quantify this heterogeneity of selection within coding sequences, we developed Model Averaged Site Selection via Poisson Random Field (MASS-PRF). MASS-PRF identifies an ensemble of intragenic clustering models for polymorphic and divergent sites. This ensemble of models is used within the Poisson Random Field framework to estimate selection intensity on a site-by-site basis. Using simulations, we demonstrate that MASS-PRF has high power to detect clusters of amino acid variants in small genic regions, can reliably estimate the probability of a variant occurring at each nucleotide site in sequence data and is robust to historical demographic trends and recombination. We applied MASS-PRF to human gene polymorphism derived from the 1,000 Genomes Project and divergence data from the common chimpanzee. On the basis of this analysis, we discovered striking regional variation in selection intensity, indicative of positive or negative selection, in well-defined domains of genes that have previously been associated with neurological processing, immunity, and reproduction. We suggest that amino acid-altering substitutions within these regions likely are or have been selectively advantageous in the human lineage, playing important roles in protein function.

Key words: **model averaged site selection, Poisson Random Field, natural selection, polymorphism, divergence, human evolution.**

Introduction

One of the principle goals in evolutionary biology is to identify genetic variants under natural selection leading to adaptation. Many statistical tests have been developed to detect selection (Nei and Gojobori 1986; Hudson 1987; Tajima 1989; McDonald and Kreitman 1991; Sawyer and Hartl 1992; Fay and Wu 2000) based on divergent substitutions between species (Nei and Gojobori 1986; Nielsen and Yang 1998), polymorphisms within species (Tajima 1989; Fay and Wu 2000), or both (McDonald and Kreitman 1991; Sawyer and Hartl 1992). In particular, among-species analyses compare the ratio of substitution rates at replacement divergent (RD) and silent divergent (SD) sites (d_N/d_S), and typically detect selective events that occurred at a relatively deep time scale (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008). In contrast, polymorphism data within species reveal more recent episodes of selection based on the allele frequency spectrum or patterns of linkage disequilibrium (LD; Nurminsky et al. 1998; Parsch et al. 2005; Nielsen et al. 2007; Aguileta et al. 2009; Saminadin-Peter et al. 2012). Intermediate between these approaches, however, are methods that infer selection using both polymorphism and divergence data (McDonald and Kreitman 1991; Sawyer and Hartl 1992; Bustamante et al. 2001; Zhu and Bustamante 2005). A classic example of this latter approach is the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991) which compares the observed number of replacement polymorphic (RP) and silent polymorphic (SP) sites within species to the observed number of RD and SD sites between species in a 2 × 2 contingency table. An advantage of this statistical test is that the incorporation of polymorphism data increases its power to detect selection

doi:10.1093/molbev/msx213 Advance Access publication July 28, 2017

(Egea et al. 2008). Polymorphic and divergent sites also serve as input for the Poisson Random Field (PRF) model, which utilizes a theoretical framework based on strong selection, infrequent mutation, and consequent independence of sites to quantify selection intensity (Sawyer and Hartl 1992).

An inherent limitation of the MK test and other tests based on the PRF model, however, is that they will not be powerful when a large proportion of sites in the gene are not under selection. Moreover, current implementations based on Sawyer and Hartl (1992), Bustamante et al (2001), and Zhu and Bustamante (2005) typically provide a single estimate that represents a uniform selection intensity across the entire length of a gene, implying that all amino acid replacement sites experience the same selective pressure. While this assumption could be valid for some cases of selection (Schlenke and Begun 2003; Nielsen 2005; Nielsen et al. 2007; Sackton et al. 2007; Kerns et al. 2008), studies have demonstrated that selection can vary intragenically; specifically, positive or negative selection can operate on amino acid-altering changes in small defined regions of genes (Holmes et al. 1992; Hughes and Yeager 1998; Nielsen and Yang 1998; Yang and Swanson 2002; Wagner 2007; Kerns et al. 2008; Tamborero et al. 2013), resulting in intragenic heterogeneity of selection intensity.

To increase the power to identify heterogeneity of selection intensity and thus characterize selection in protein-coding genes, we have developed Model-Averaged Site Selection with Poisson Random Field (MASS-PRF). MASS-PRF identifies clusters of polymorphic and divergent variant sites within genes (i.e., clustering models), and calculates the probability of observing polymorphic and divergent variants at each site within a given sequence across clustering models. These probabilities are then used as entries in the PRF theory to estimate selection intensity (gamma, γ) on a site-by-site basis along the length of a given gene. Using simulations, we confirmed the reliability of MASS-PRF to identify highly localized groups of nucleotide changes in coding sequence, and to estimate the average probability of a variant appearing at each site. To evaluate the effects of demography and recombination on the selection inference by MASS-PRF, we used ms (Hudson 2002) and INDELible (Fletcher and Yang 2009) to perform coalescent simulations of neutrally evolving genes under different demographic scenarios and levels of recombination. These simulations demonstrated a modest impact of demography and recombination on MASS-PRF inference of selection intensity. To further illustrate the utility of MASS-PRF, we applied our method to a set of human protein-coding genes involved in neurological processing, immune response and reproduction, gathering human polymorphism data from the 1000 Genomes Project and divergence data from a comparison with *Pan troglodytes* (the common chimpanzee). These analyses identified signatures of positive selection ($\gamma > 4$) at replacement sites within well-defined regions of our genes, implying that these substitutions are or have been selectively advantageous in the human species. The results correlate well with prior studies that demonstrated that many of these genic regions play a role in gene function. The detection of genic regions under selection is highly informative for biomedical studies focused on the identification of functionally relevant sites involved in key biological processes, such as fertility, host–pathogen interactions, and drug resistance.

New Approaches

Overview of MASS-PRF Algorithm

MASS-PRF operates on polymorphism and divergence data in two steps: construction of clustering models (Step 1) and estimation of selection intensity (Step 2).

Step 1: Construction of Clustering Models

MASS-PRF examines aligned sequences of length *N*, scoring invariant sites as '0' and variable sites as '1'. MASS-PRF iteratively partitions the entire sequence into three regions: (1) a central region bounded by a start position (c_s) and end position (c_e), where ($0 \leq c_s < c_e \leq N - 1$); (2) a starting region flanking the central region; and (3) an ending region flanking the central region. For example, for a gene with a length of 100 base pairs (bp), the start (c_s) and end (c_e) positions of cluster model #1 could be nucleotide position 3 and nucleotide position 4, respectively; for cluster model #2 the start and end positions would be nucleotide position 3 and nucleotide position 5, respectively; for cluster model #3 the start and end positions would be nucleotide position 3 and nucleotide position 6, respectively. In this example, MASS-PRF hierarchically generates start and end positions for all possible clustering models (i.e., blocks of sequence) within a given gene, until the central model encompasses all but two sites, and the flanking models consist of sites 1 and 2 or sites 99 and 100. After constructing this first set of models, MASS-PRF will recursively partition the central and flanking sequences into all possible sets of three regions (bounded by all possible start and end positions). In summary, MASS-PRF will exhaustively construct all possible models (i.e., blocks of sequence) from the start of the gene to the end of the gene. MASS-PRF performs this clustering model step separately for each of four categories of mutations. These four categories are SP, SD, RP, and RD sites. Our method counts the number of variant sites in the starting (positions 1 to c_s), central (c_s to c_e), and ending (c_e to the end of the gene) regions.

Then, MASS-PRF calculates the maximum likelihood of observing SP, SD, RP, and RD within the central region (denoted by a probability p_c per site within the cluster; that is, the number of variable sites divided by the total number of sites within a given cluster; supplementary fig. S1, Supplementary Material online) and the maximum likelihood of observing SP, SD, RP, and RD in the neighboring regions outside of the central region (denoted by a probability p_0 per site outside the cluster; that is, the total number of variant sites outside the cluster divided by the total number of sites outside a given cluster; supplementary fig. S1, Supplementary Material online). Each clustering model (consisting of the central and flanking regions) thus has a binomial likelihood associated with it (fig. 1 and supplementary fig. S1, Supplementary Material online). Optionally, one may estimate model-averaged maximum likelihood probability for

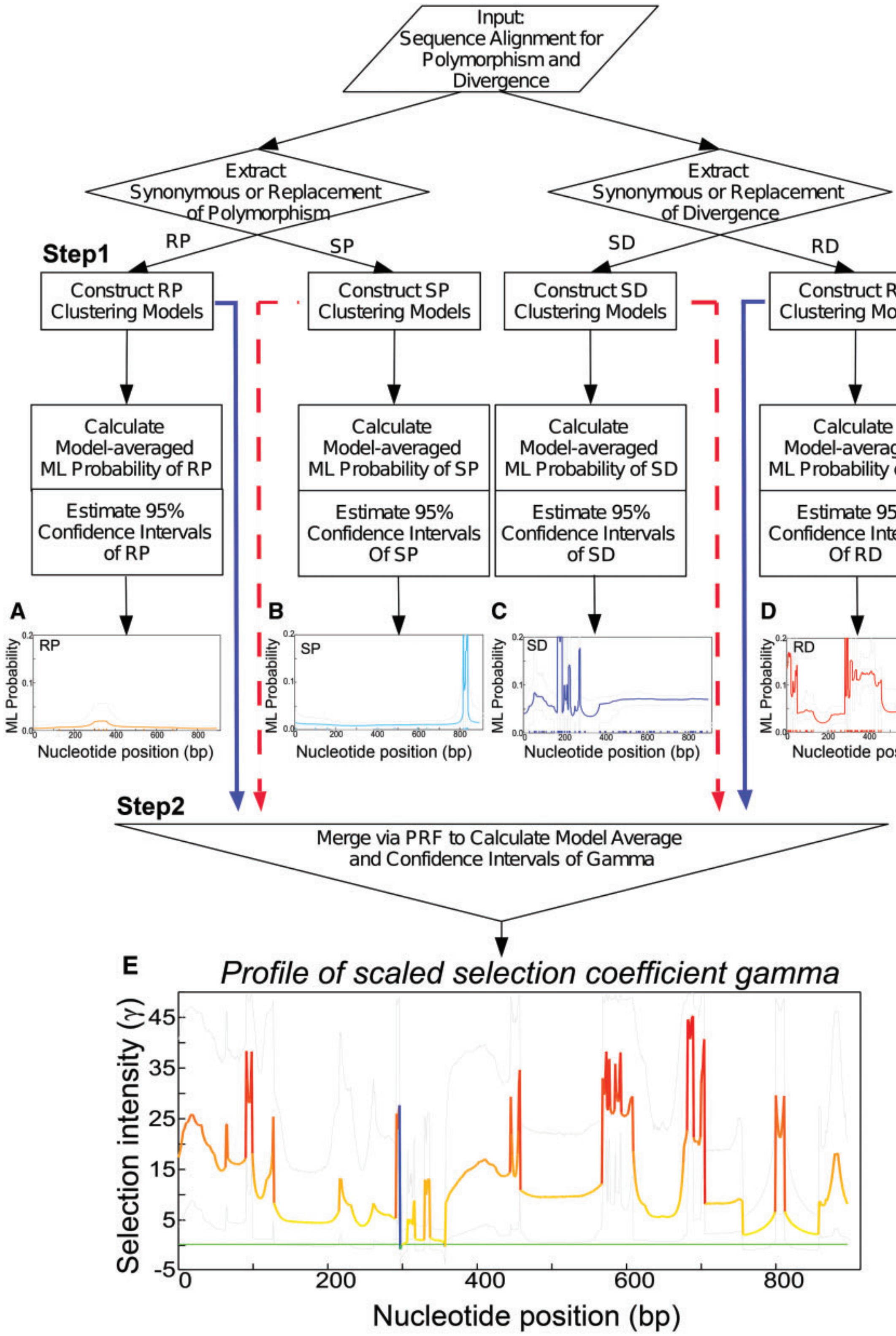
Input:
Sequence Alignment for Polymorphism and Divergence
Extract Synonymous or Replacement of Polymorphism
Extract Synonymous or Replacement of Divergence
Step1
RP
SP
SD
RD
Construct RP Clustering Models
Construct SP Clustering Models
Construct SD Clustering Models
Construct RD Clustering Models
Calculate Model-averaged ML Probability of RP
Calculate Model-averaged ML Probability of SP
Calculate Model-averaged ML Probability of SD
Calculate Model-averaged ML Probability of RD
Estimate 95% Confidence Intervals of RP
Estimate 95% Confidence Intervals Of SP
Estimate 95% Confidence Intervals of SD
Estimate 95% Confidence Intervals Of RD
A
B
C
D
ML Probability
0.2
0.1
0.0
0
200
400
600
800
Nucleotide position (bp)
Step2
Merge via PRF to Calculate Model Average and Confidence Intervals of Gamma
E
Profile of scaled selection coefficient gamma
Selection intensity (γ)
45
35
25
15
5
-5
0
200
400
600
800
Nucleotide position (bp)

Fig. 1. The workflow of the Model Averaged Site Selection via Poisson Random Field (MASS-PRF) approach. Step 1 consists of the construction of clustering models. Step 2 consists of the estimation of model-averaged selection intensity γ and its 95% model uncertainty intervals for each site. Step 1 can be applied separately to aligned sequences of (*A*) replacement polymorphism (RP), (*B*) synonymous (silent) polymorphism (SP), (*C*) synonymous divergence (SD), and (*D*) replacement divergence (RD). Step 2 uses observed probabilities of SP, SD, RP, and RD, and merges them into PRF theory to estimate (*E*) selection intensity. RP and RD are used to estimate model-averaged selection intensities and their 95% model uncertainty intervals (solid line in blue); SP and SD can be combined to represent intragenic inhomogeneity of mutation rate (dashed line in red) for calculations of site-specific divergence time by default, or assuming homogeneity of mutation rate, can be replaced by gene-level divergence time calculated from total counts of SP and SD, or species divergence time can be exogenously supplied by the user.

each site. MASS-PRF averages the probability of observing SP, SD, RP, and RD at each site within a given gene across models, yielding a site-specific probability of being a variant. Visual profiles of these probabilities over the gene sequence can be generated, including a 95% model uncertainty interval for these probabilities per site (fig. 1, insets *A*–*D*). Models are penalized for overparameterization using the Akaike or Bayesian Information Criteria (AIC or BIC).

Step 2: Estimation of Model-Averaged Selection Intensity and Model Uncertainty Intervals for Each Site

Our estimate of the selection intensity at site *i* is the model-averaged value of γ across all or across a stochastic sample of joint cluster models (for SP, SD, RP, and RD). The model uncertainty interval is calculated as the central 95% of joint model probability weight. The weight of each joint model is a product of the model weights for RP, RD, SP, and SD models. Model-averaged γ is then calculated by weighing every γ associated with each joint model. Models are penalized for overparameterization using the AIC or BIC. To estimate 95% model uncertainty intervals for selection intensity γ, we developed and implemented two algorithms: (1) an exhaustive option for sampling cluster models that should be selected for analysis of shorter genes where intensive computation is not prohibitive, or (2) a stochastic option for sampling cluster models that should be selected for analysis of longer genes in a relatively short time and with a high level of accuracy and precision over sufficient iterations.

MASS-PRF calculates the site-specific selection intensity (γ, fig. 1, inset *E*) with each set of four cluster models (SP, SD, RP, and RD) by providing the model-based variant probabilities at each site to the classic PRF model (Sawyer and Hartl 1992). More specifically, MASS-PRF solves for γ, setting the ratio of the expected number of RD and RP sites from Sawyer and Hartl (1992) equal to the ratio of the estimated probabilities of RD and RP appearing at each site that is calculated in “Step 1” for each model (fig. 1), and using the site-specific probabilities of SP and SD variants from the cluster models to estimate the expected neutral divergence parameter in the PRF model. The patterns of aggregation or grouping of variant sites provide the means for MASS-PRF to estimate regional selection intensity.

Results

Analysis of Clustering of Variant Sites with Polymorphism and Divergence Data Using MASS-PRF

Theoretically, variants in coding sequences can be distributed in multiple ways, ranging from a random distribution of variants across coding sequences to localized clustering of variants within well-defined regions of genes. To assess how well MASS-PRF detects clusters of nucleotide substitutions, we simulated protein-coding sequences of 1,200 bp length, for which the numbers of SP, RP, SD, and RD sites were assigned a priori (see Materials and Methods, fig. 2). Variants were distributed within clusters of 100–1,200 bp (cluster size increased in increments of 100 bp). We performed 100 replicates for SP, SD, RP, and RD. For each replicate incorporating all four variant types, we calculated the probability of observing SP, SD, RP, and RD per site across the gene estimated by MASS-PRF and using an unclustered model in which all sites had identical probabilities of being variant. We calculated the accuracy and precision of MASS-PRF and the unclustered model at estimating the actual probabilities of SP, SD, RP, and RD. The extent to which the estimated and expected probability distributions of variants differed was assessed using Kullback–Leibler (KL) divergence (0 indicates no difference, a positive value or negative value indicates poor fit). The average KL (*D*) and the inverse precision (σ) were close to zero (fig. 2), indicating minimal error between the estimated and expected probabilities of each site being a variant in the clustered model (MASS-PRF). In contrast, however, in the nonclustered model, *D* and σ diverged positively from zero (fig. 2). Overall, these results indicate that MASS-PRF is able to calculate the probability of variant sites clustered in relatively small genic regions with a high level of accuracy and precision.

Coalescent Simulations Assess Robustness to Demographic and Recombination History

To assess the effects of demography and recombination on the performance of MASS-PRF, we simulated six sets of 20 neutrally evolved genes (for a total of 120 genes) undergoing three demographic events (bottleneck, constant, and expansion) with and without genetic recombination. MK tests (McDonald and Kreitman 1991) of the simulated data did not reject the null hypothesis that most genes (98%; supplementary table S1, Supplementary Material online) were neutrally evolving. We then applied MASS-PRF on the simulated data specifying two distinct options: (1) a fixed divergence time of 6 Ma (million years ago), and (2) a site-specific divergence time calculated based on the clustering of silent sites, which is the default option in MASS-PRF. Using the simulated data analyzed by MASS-PRF, we estimated the false positive rate (FPR) associated with inference of selection ($\gamma < -1$ or $\gamma > 4$, as defined in Materials and Methods). The FPR was <2% for each of the six scenarios using the site-specific divergence time in MASS-PRF (supplementary table S1, Supplementary Material online), suggesting modest impacts of demography and recombination. In particular, the FPRs without recombination and with a fixed divergence time of 6 Ma were 12% (Bottleneck), 6% (Constant), and 1% (Expansion; supplementary table S1, Supplementary Material online). In contrast, for all three demographic scenarios without recombination and with a divergence time that was estimated using the site-specific divergence time, the FPR ranged from 1% to 2% (supplementary table S1, Supplementary Material online). Adding recombination when divergence time was estimated site-specifically yielded the same range of FPRs (1–2%; supplementary table S1, Supplementary Material online), whereas adding recombination when divergence time was fixed at 6 Ma led to FPRs that ranged from 1% to 8% (supplementary table S1, Supplementary Material online). These results indicate that the effects of demography and recombination are mitigated

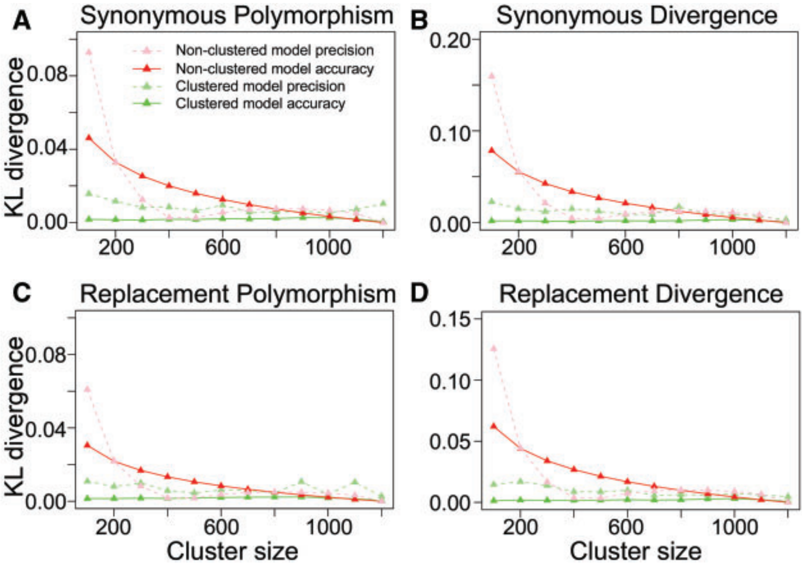

Fig. 2. Comparison of accuracy and precision based on the Kullback–Leibler (KL) divergence between simulated probabilities and estimated probabilities for the clustered and nonclustered model. The KL divergence quantifies the divergence of the distribution of expected (simulated) probabilities from the distribution of estimated probabilities. A value of KL divergence closer to zero indicates that the estimated probability from the clustered (green triangles) or nonclustered model (red triangles) match the expected probability imposed within the simulation. The simulated 1200 bp sequences featured variant sites restricted to cluster sizes (lengths of regions with all variants) ranging from 100 bp (a tight cluster) to 1,200 bp, incrementing cluster size by 100 bp. Analyses of accuracy (solid lines) and precision (dashed lines) are displayed for four classes of variants: (*A*) synonymous (silent) polymorphism, (*B*) replacement polymorphism, (*C*) synonymous divergence, and (*D*) replacement divergence.

by usage of the site-specific divergence time compared with specifying a fixed lineage divergence time of 6 Ma.

Application of MASS-PRF to Empirical Data

As an initial validation, we applied MASS-PRF to a set of genes (*SLC6A5*, *GRIN2C*, *RNASEL*, *IL18RAP*, *TGM4*, *WBP2NL*, and *SPAG5*) from the 1,000 Genomes Project (supplementary table S2, Supplementary Material online) that were previously inferred to be under selection based primarily on the ratio of divergence at replacement and silent sites between species (Bustamante et al. 2005; Clark and Swanson 2005; Kosiol et al. 2008). The tests for selection used in these prior studies yielded single gene-wide estimates for each gene. For example, estimates of γ for *RNASEL*, *IL18RAP*, *SPAG5*, and *GRIN2C* were 5.3 (C.I. 0.6–14.6), 8.3 (C.I. 1.2–19.4), 5.3 (C.I. 0.7–14.3), and 8.2 (C.I. 1.0–19.6), respectively, based on the mean of the posterior distribution of γ for each gene in Bustamante et al. (2005). Kosiol et al. (2008) did not estimate γ for *WBP2NL* and *SLC6A5* genes, but computed *P* values using a likelihood ratio test for selection (6×10^{-4} and 3×10^{-4}, respectively); the estimate of d_N/d_S for *TGM4* was 2.14 ($P = 2 \times 10^{-3}$; Clark and Swanson 2005). MASS-PRF analysis also indicated that these loci were under selection, and demonstrated that selection intensity was not constant across these genes. Instead, our data showed that selection varied across sites, revealing adaptive evolution within defined regions of protein-coding genes.

On the basis of our method, we detected weak ($\gamma > 4$; lower bound < 0), moderate ($\gamma > 4$; $0 <$ lower bound < 4), and strong ($\gamma > 4$; lower bound > 4) evidence for positive selection in specific genic regions. The lower bound refers to the lower 95% model uncertainty interval for the estimate of γ, and γ is the selection intensity ($2N_es$, where N_e is the effective population size and *s* is the canonical population genetic selection coefficient). In *SLC6A5*, for example, we found localized peaks of γ, indicative of positive selection ($\gamma > 4$; lower bound < 0), encompassing fixed replacement sites 316, 318, 329, 371, and 460 in exon 2 (fig. 3A, supplementary table S3, Supplementary Material online). In a prior study, the deletion of exons 2 and 3 resulted in the loss of part of the large cytoplasmic N-terminus of SLC6A5 protein, inhibiting synaptic transmission in mammals, and suggesting that these exons serve an important function (Gill et al. 2011). Our analysis of *GRIN2C* revealed two distinct peaks of γ; one peak ($\gamma > 4$; lower bound < 0) occurred in a region encompassing replacement substitutions at sites 67, 68, 212, 266, 272, and 299 in exon 1 and the other peak ($\gamma > 4$; $0 <$ lower bound < 4) at site 2551 in exon 12 (fig. 3B, supplementary table S3, Supplementary Material online). This inference of selection is supported by other studies reporting that exon 1, encoding amino acid residues in the N-terminal domain of GRIN2C, is responsible for the functional differences between members of the *GRIN2* gene family (Teng et al. 2010). The signal of positive selection present in exon 12 of *GRIN2C* suggests that replacement substitutions within this region have contributed to the functional divergence of this gene.

Our analysis also uncovered strong evidence for positive selection within specific regions of genes that are involved in immunity and inflammation (Fumagalli et al. 2011). In *RNASEL*, for example, we inferred positive selection ($\gamma > 4$; $0 <$ lower bound < 4) at replacement sites 8, 83, 107, 119, 140, 212, and 219 in exon 2, as well as 640, 648, and 667 in

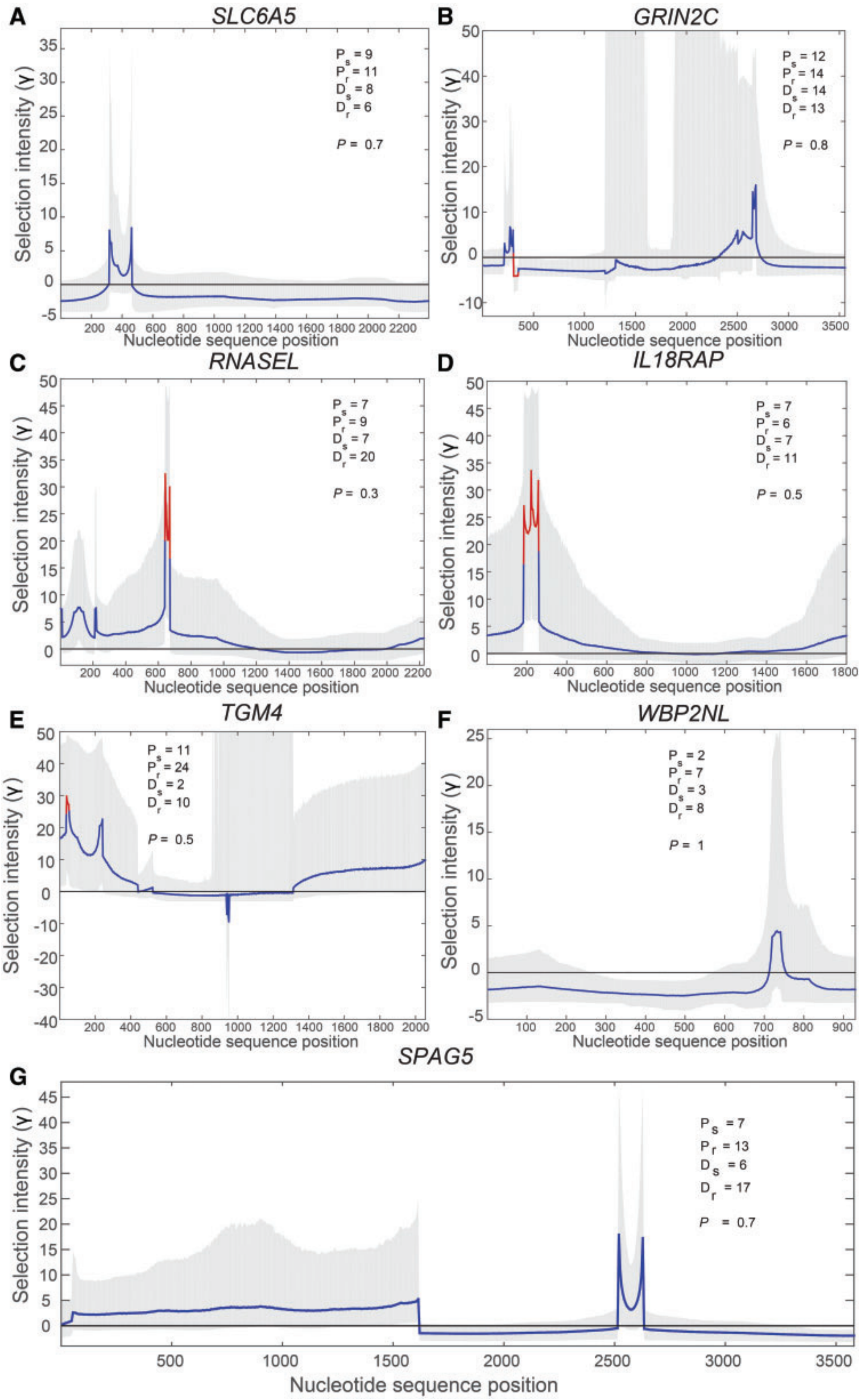

FIG. 3. Profiles of selection intensity (γ) across nucleotide positions for seven genes: (*A*) *SLC6A5*, (*B*) *GRIN2C*, (*C*) *RNASEL*, (*D*) *IL18RAP*, (*E*) *TGM4*, (*F*) *WBP2NL*, (*G*) *SPAG5*. A line indicates the model-averaged γ (red if the lower bound of $\gamma > 4$ or if the upper bound of $\gamma < -1$, otherwise blue), and a grey band indicates the 95% model uncertainty interval. The black horizontal line in each plot indicates $\gamma = 0$. Each figure reports synonymous polymorphic sites (P_s), replacement polymorphic sites (P_r), synonymous divergent sites (D_s), replacement divergent sites (D_r), and the Fisher Exact *P* value for the corresponding MK test.

exon 5 ($\gamma > 4$; lower bound > 4; fig. 3C, supplementary table S3, Supplementary Material online), which encode amino acids in regions of RNASEL that play a role in antiviral immune response. This localization of selection agrees with Wagner (2007), who identified a similar clustering of amino acid substitutions in exon 2, encoding amino acid residues that bind the activator molecule 2-5A, based on a comparative analysis of human and chimp anzee protein-coding sequences. These results imply strong positive selection for variants in this genic region. We suggest that the substitutions clustered in exon 5 could also encode functionally important protein domains. In *IL18RAP*, our results indicated that fixed replacement substitutions at positions 186, 222, 231, and 258 in exon 2 (fig. 3*D*, supplementary table S3, Supplementary Material online) were under strong positive selection ($\gamma > 4$; lower bound > 4), likely in response to pressure from pathogens.

We also found evidence for adaptive evolution at genes associated with human sexual reproduction. In *TGM4*, for example, we detected positive selection ($\gamma > 4$; lower bound > 4) at positions 39, 51, 225, and 240 in exon 2 (fig. 3*E*, supplementary table S3, Supplementary Material online). Indeed, a recent in vitro study demonstrated that exon 2 influences the expression of TGM4 isoforms (namely, 4-L, -M and -S; Choi et al. 2010), consistent with our inference that substitutions within this genic region are likely functionally important. We also estimated a peak of $\gamma > 4$ (but with lower bound < 0) encompassing fixed replacement sites 720, 729, 732, and 741 in exon 6 of *WBP2NL* (fig. 3*F*, supplementary table S3, Supplementary Material online). The corresponding amino acids 240, 243, 244, and 247 are located in functional repeat domains of the sperm-specific WBP2NL protein (Wu et al. 2007). Our analysis of *SPAG5* (fig. 3*G*, supplementary table S3, Supplementary Material online) revealed high peaks of γ ($\gamma > 4$; $0 <$ lower bound < 4) at sites 2520 and 2628 in exon 15, encoding divergent sites that bind the kinetochore-localized astrin (KNSTRN) protein which plays a role in chromosome alignment and normal cell division (Dunsch et al. 2011).

In addition to the seven genes previously inferred to be under positive selection in Bustamante et al. (2005), Clark and Swanson (2005), Kosiol et al. (2008), we applied MASS-PRF to an additional set of 51 protein-coding genes that had not been previously inferred to be adaptively evolving in the human lineage (Bakewell et al. 2007; Gaya-Vidal and Alba 2014). Selection intensities varied among and within genes (fig. 4). Graphical depictions of the levels of selection across domains of four exemplars arranged in a 2 × 2 grid of panels typify our findings across these 51 genes: (1) *SLC8A1*, inferred to be under selection both by MASS-PRF and the MK test (a Yes–Yes scenario); (2) *NT5C1B*, inferred to be under selection by MASS-PRF but not MK (a Yes–No scenario); (3) *MGAM*, not inferred to be under selection by MASS-PRF, but inferred to be under selection by the MK test (a No–Yes scenario); and (4) *TPH2*, not inferred to be under selection by either MASS-PRF or the MK test (a No–No scenario). More specifically, our analysis of *SLC8A1* showed a strong signature of positive selection spanning replacement sites in exon 2 (fig. 4*A*) that encodes amino acids in the calcium-binding domain of the SLC8A1 cell membrane protein. We also observed a high peak in γ at positions in exon 5 of *NT5C1B* (encompassing replacement sites 812, 813, 815, 819, 829, and 845; supplementary table S3, Supplementary Material online) consistent with a model of positive selection (fig. 4*B*). While the precise function of these substitutions at *NT5C1B* is unknown, our analysis suggests that these nucleotides are likely to be biologically relevant sites. Interestingly, we did not observe elevated peaks of γ at *MGAM* using MASS-PRF, in contrast to the findings from our MK analysis (fig. 4*C*). We also did not find evidence for positive selection at *TPH2* (fig. 4*D*), which is consistent with the current literature (McKinney et al. 2009; Chen and Miller 2012; Taub and Page 2016). Examples of other genes that fell within each of the above four categories can be found in supplementary figure S2 and Supplementary Material online. The family-wise error rates of MK tests can be found in supplementary table S4, Supplementary Material online.

Discussion

Here, we have shown that the MASS-PRF algorithm has power to detect localized groups of amino acid changes within genes using an unbiased approach that does not require a priori knowledge of cluster size or count. Our simulation results also demonstrated that MASS-PRF reliably estimates the maximum likelihood probability of each site being a variant with a high level of accuracy and precision. Furthermore, using this algorithm, we identified potential important targets of natural selection, within small genic regions in a set of protein-coding genes, illustrating that MASS-PRF can provide additional insights into the process of adaptive evolution. Indeed, because classical tests assume a uniform selection intensity across the entire gene, the phenomenon of clustered adaptive change in protein-coding sequence genes is a case of selection that has been understudied in evolutionary genetics. Most importantly, our coalescent simulations demonstrated that MASS-PRF is robust to demographic and recombination events.

MASS-PRF does not assume homogeneity of selection intensity across sites. Accordingly, it exhibited power to detect different regions in individual genes that are experiencing different intensities of selection. The level of selection inferred in the PRF framework is scaled to a model of neutral evolution typically based on the observed silent site polymorphism and divergence (Sawyer and Hartl 1992). MASS-PRF uses an unbiased approach to detect localized groups of amino acid changes within genes and does so without a priori knowledge of cluster size or count. Indeed, MK sliding windows have been used dating back to studies of mitochondrial genes (Rand et al. 1994). However, in addition to their lack of formal statistical tractability, results generated by sliding window analyses have been challenging to interpret and have not been widely accepted: the problem is that it is very hard to decide on an appropriate sliding window size a priori for the analysis, and that the decision on the size of the sliding window dramatically affects the outcome of the analysis, which significantly lessens the persuasiveness of findings based on this approach. Moreover, the “appropriate size” of sliding window will likely change across a given gene depending on the size of component functional domains (e.g., transmembrane domains, extracellular domains, DNA-binding domains, etc.) and the consequent clustering of selected sites. That is to

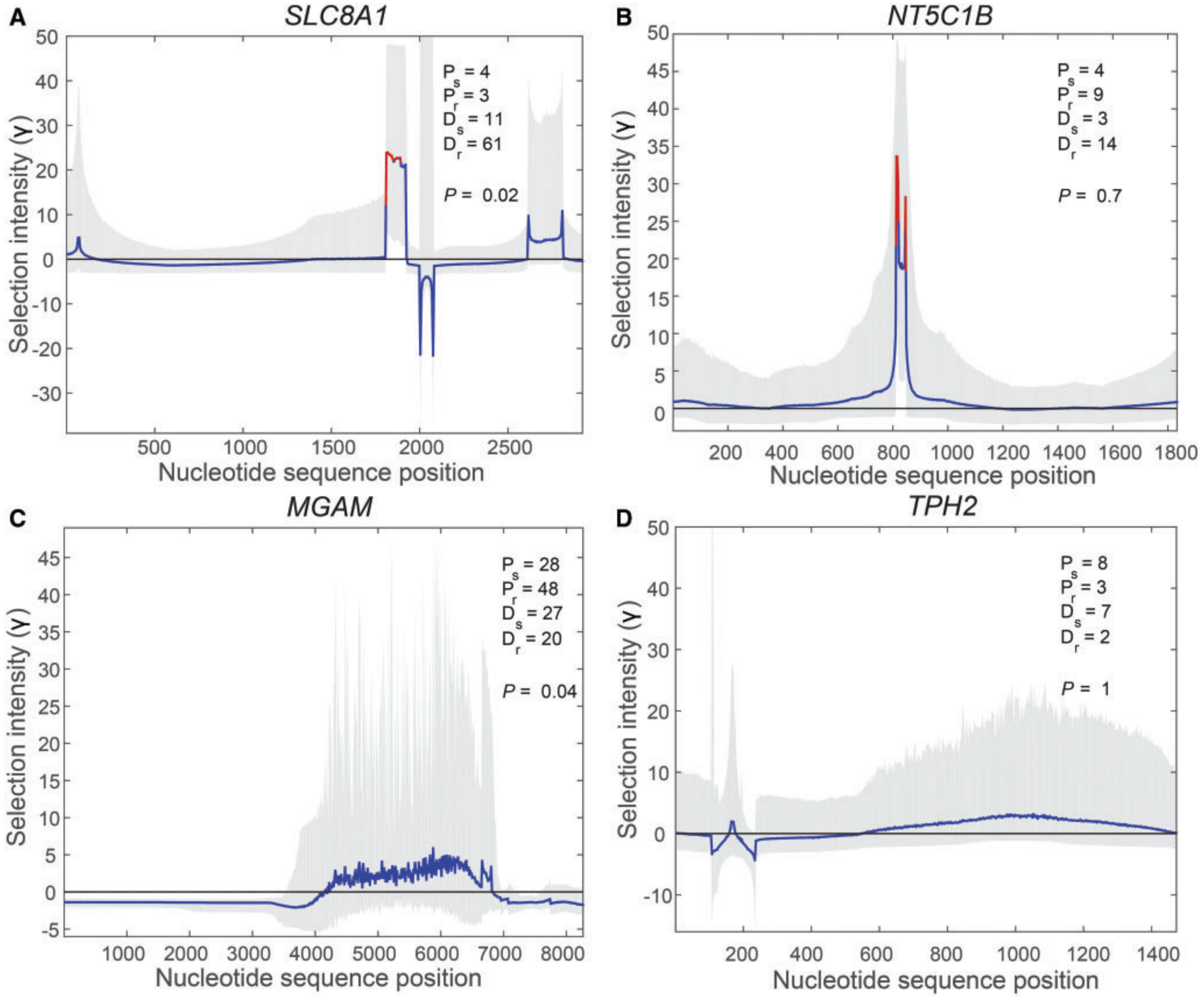

Fig. 4. Four scenarios for comparing selection inference by MASS-PRF and the MK test. (*A*) *SLC8A1* was inferred to be under selection both by MASS-PRF and the MK test ($P = 0.02$); (*B*) *NT5C1B* was inferred to be under selection by MASS-PRF, but not MK ($P = 0.67$); (*C*) *MGAM* was not inferred to be under selection by MASS-PRF, but was inferred to be under selection by the MK test ($P = 0.04$); (*D*) *TPH2* was not inferred to be under selection by either MASSPRF or the MK test ($P = 1$). A line indicates the model-averaged γ (red if the lower bound of $\gamma > 4$ or if the upper bound of $\gamma < -1$, otherwise blue), and a grey band indicates the 95% model uncertainty interval. The black horizontal line in each plot indicates $\gamma = 0$. Each figure reports synonymous polymorphic sites (P_s), replacement polymorphic sites (P_s), synonymous divergent sites (D_s), replacement divergent sites (D_r), and the Fisher Exact *P* value for the corresponding MK test.

say, if window size is too small or too large, it is possible to miss signals of selection operating within a particular genic region. For genome-wide analyses, this issue is compounded by the fact that genes will have widely varying counts of polymorphic and divergent sites, so determining an appropriate sliding window size automatically for each of the thousands of genes to be analyzed will be very difficult. So what would be better would be some approach that generates all possible sliding window sizes, including different sliding window sizes in different parts of the gene, and integrates over them all to give a single, statistically supported estimate for *each site* in a gene. That analysis is in essence what MASS-PRF does, but in a maximum likelihood model averaging framework linked to Sawyer and Hartl's PRF (Sawyer and Hartl 1992).

MASS-PRF shares some limitations in common with many existing methods for detecting positive selection. For example, MASS-PRF assumes that species have persisted at a constant population size. However, in general, biases in inferences of selection can arise when the populations have experienced recent bottlenecks or expansions, resulting in either an excess of intermediate frequency polymorphisms (indicative of balancing selection) or an excess of low-frequency polymorphisms (indicative of positive or weak purifying selection), respectively (Eyre-Walker 2002; Parsch et al. 2009). Though demographic events can mimic gene-wide signatures of selection, we do not expect that changes in population size would lead to localized clustering of multiple protein-altering sites in small genic regions, violating the predicted uniform distribution of mutations in gene space under neutral evolution (Wagner 2007). Our evaluation of demographic events, such as bottlenecks or exponential growth, revealed modest effects of demography on the estimation of selection intensity using simulated data. We argue that this modest impact of demography is likely due to the fact that MASS-PRF does not estimate selection using the site frequency spectrum, which is known to be highly sensitive to the effects of demographic history.

MASS-PRF, like other MK and PRF methods, assumes an infinite allele model and independence between sites. Whereas the infinite allele model is a reasonable approximation for most eukaryotic populations (Desai and Plotkin 2008), the assumption of free recombination between sites is less realistic. Typically, over time recombination breaks up genomic regions containing variants, leading to shorter blocks of LD, and thus less correlation, among nucleotide sites on the same chromosome. Within a gene, LD is expected to be present. However, the presence of replacement variants in close proximity—particularly fixed replacement substitutions—cannot be easily explained by LD alone. Under a scenario of frequent rapid adaptation, for instance, it is conceivable that neutral sites—including neutral replacement sites—could 'sweep' to fixation in a given species due to genetic hitchhiking with a selected mutation in the deep past, potentially leading to a clustering of sites and a subsequent bias in inferred selection intensity. Though more explicit modeling of the varying degrees of LD between sites, demography and different scenarios of selection is challenging, quantifying the effects of these factors on estimates of natural selection is an important topic that requires more in-depth study in evolutionary biology (Kryazhimskiy and Plotkin 2008; Zeng and Charlesworth 2010; Racimo and Schraiber 2014).

Our simulations demonstrated, however, modest effects of demography and levels of recombination on the estimation of selection intensity using MASS-PRF; these effects can be further mitigated by implementation of a site-specific divergence time based on model-averaged clustering of silent sites. Our analysis of coalescent simulations demonstrated that estimating silent site clustering helps to alleviate the effects of intragenic recombination and demographic changes. This improvement arises because the silent site clustering allows MASS-PRF to accommodate different coalescent histories (Arenas and Posada 2010) as well as different mutation rates (Barr et al. 2007) across a given gene. Assuming a constant mutation rate, detection of clustering of silent site polymorphism enables better quantification of the depth of the coalescent tree associated with intragenic sequence. Dense silent site polymorphism will tend to occur in regions of genes with deep coalescence, whereas sparse silent site polymorphism will tend to occur in regions of genes with recent coalescence (Arenas and Posada 2010; Ferretti et al. 2013). For example, if the first half of a given gene has a deep coalescence among alleles and the second half of the same gene has a shallow coalescence among alleles, the first half will have a greater number of silent site polymorphisms (and, under neutrality, RPs as well). The silent site clustering we perform will then detect this difference between the first and second half, attribute different probabilities of silent sites to the two halves of the gene, and thus—in principle—accurately estimate $2N_es$ despite the difference in coalescent history.

By default MASS-PRF uses this site-specific divergence time, quantifying intragenic silent site polymorphism which can reflect any intragenic recombination and demographic events that may have occurred, and incorporating those model-averaged site-specific divergence times in calculating the background silent mutation rate for each site. Incorporating the silent site clustering helps to alleviate error in the estimation of the level selection that arises as a consequence of recombination and demography. Our simulations show that this is true in practice as well as in principle. Overall, our simulation results with MASS-PRF are in broad agreement with methods, such as the MK test, that have demonstrated a modest sensitivity to the effects of demography (Nielsen 2001, 2005; Eilertson et al. 2012) and recombination (Kreitman 2000) on the detection of selection using polymorphism and divergence data.

The theoretical framework underlying MASS-PRF permits three ways of parameterizing the neutral model. First, silent sites can be analyzed for clustering which can arise due to differential mutation rate variation across sites or differential intragenic coalescence due to recombination (reflecting the baseline clustering of mutations under a neutral model). Clustering of silent site polymorphism and divergence enables the estimate of selection to incorporate heterogeneous rates of mutation across sites or heterogeneous histories due to differential intragenic coalescence. Alternatively, intragenic mutation rate can be specified as homogeneous and gene-wide coalescence can be assumed by calculating divergence time with gene-wide counts of polymorphic and divergent silent sites, as in the original PRF framework. Lastly, users of MASS-PRF can specify species divergence time as a parameter. When specifying a divergence time, selection estimates are driven only by the clustering of RDs relative to the clustering of RPs, avoiding bias caused by the effects of weak selection on silent sites. While in general silent sites are believed to be selectively neutral, a few studies have demonstrated that silent sites in coding sequences can be subject to weak selection (Akashi 1995; Ohta 2002; Singh et al. 2007) due to their functional roles in mRNA splicing (Chamary and Hurst 2005), RNA editing (Shabalina et al. 2006, 2013) and protein translation (Akashi 1994). A bias in estimates of selection based only on divergence data (d_N/d_S) can also arise as a consequence of relaxed functional constraint, in which the total number of replacement substitutions within genes is increased, mimicking genetic signatures of positive selection (Lazzaro 2005; Arbiza et al. 2006; Wagner 2007). However, simulations have been conducted to argue that clusters of amino acid changes in protein-coding genes cannot be attributed solely to relaxed functional constraint (Wagner 2007). Accordingly, estimates of selection by MASS-PRF would be less affected by relaxed functional constraint than other approaches that assume uniform selection intensity.

Unlike MASS-PRF, many extended PRF models (Bustamante et al. 2001; Zhu and Bustamante 2005) use the full allele-frequency spectrum of polymorphism data (i.e., the number of polymorphisms at a frequency 1 out of n, 2 out of n, . . . ($n-1$) out of n, where n is the number of sequences sampled) to infer selection intensity relative to neutrality. However, there are disadvantages to this approach in the MASS-PRF framework. Firstly, breaking down polymorphisms into the allele frequency spectrum can provide additional information regarding selection, but doing so also increases the sensitivity of inferred selection intensity to the effects of demographic events, leading to false signals of selection (Bustamante et al. 2001). Like the MK test itself (Nielsen 2001, 2005; Eilertson et al. 2012), MASS-PRF minimizes its sensitivity to demography by utilizing counts of SP and RP sites rather than site frequency data. Moreover, any gain in power that could be attained by differentiating along the frequency spectrum would be more than offset by the consequent requirement to cluster sites independently for each frequency category of the spectrum, which would both impose a considerable computational cost and weaken the precision with which MASS-PRF clusters polymorphic sites. Each additional allele frequency category induces smaller partitions of the polymorphic sites that can be analyzed for within-category clustering; for individual genes, obtaining sufficient polymorphism can sometimes be the limiting factor even for an approach that does not differentiate among allele frequencies. If it is not the case that there is heterogeneity of selection intensity across sites and/or the detection of selection is focusing on one or more whole genes, gene-level approaches using this information, such as SnIPRE (*S*electio*n* *I*nference using *P*oisson *R*andom *E*ffects; Eilertson et al. 2012), would be more appropriate methods to apply.

Other standard frequency-based approaches, such as Tajima's *D* and Fay and Wu's *H* (Tajima 1989; Fay 2000), and haplotype-based tests, such as EHH and *i*HS (Sabeti et al. 2002; Voight et al. 2006), have also been used to interrogate genomes for nucleotide positions that may have been targets of classical selective sweeps. However, these methods primarily provide insight into selection operating at a more recent time scale than do MK or PRF methods, which detect advantageous mutations that are fixed in one species relative to another species after their divergence at deeper time scales. Therefore, it would not be appropriate to use methods designed to detect recent selection at a microevolutionary level (i.e., within populations) to capture selective events that occurred over a deeper time scale at a species level. In particular, genetic signatures typically used by frequency- and haplotype-based methods to infer selection (such as reduced homozygosity and long-range haplotypes) persist only until recombination and mutation restore diversity at the selected locus over time (Vitti et al. 2013). Overall, MASS-PRF identifies genetic differences that have arisen due to selection between species (rather than within species; Vitti et al. 2013).

Application of MASS-PRF to Empirical Data

At a molecular level, selection can be inferred from nucleotide changes in coding sequence data from related lineages (Bakewell 2008). While standard tests of neutrality summarize patterns of diversity and are weakened by an assumption of homogeneity of selection pressure across genes, MASS-PRF is highly informative for inferring selection at coding genes, even within small genic regions in analyses of sequence data in related species. Thus, MASS-PRF is a viable method for detecting selection in genes that appear to have experienced local selection, complementing existing approaches aimed at identifying signatures of adaptive evolution.

In the present study, we detected moderate evidence for positive selection in well-defined regions of *GRIN2C* (fig. 3*B*) and *SLC6A5* (fig. 3*A*). These signals of positive and negative selection indicate regions of the genes that likely are, or have been, functionally important (Biswas and Akey 2006; Bakewell 2008). Amino acids in GRIN2C and SLC6A5 that are positively selected according to MASS-PRF interact with amino acids in other proteins; for example, amino acids in GRIN2C interact with proteins playing roles in learning, memory and synaptic communication (Teng et al. 2010), and amino acids in SLC6A5 interact with proteins playing roles in rapid sound localization (Lin 2011). We also presented strong evidence for selection at replacement sites in *SLC8A1* (fig. 4*A*), which encodes a membrane protein that regulates intracellular calcium concentrations in excitable cells, such as neurons, which is an important homeostatic function (Khananshvili 2013). Indeed, one of the key behavioral traits that defines modern *Homo sapiens* is complex cognition, and archaeological evidence has suggested that this modern behavioral trait arose at a relatively early stage of human evolution (Campbell et al. 2014). However, the development of knowledge (such as, technological advances in stone tools) and the transmission of this information within and between generations likely required changes in cognitive processes, such as memory and learning, in modern *H. sapiens* (Bakewell 2008; Heyes 2012). Moreover, changes in hearing sensitivity in humans might have been beneficial for understanding spoken language (Martinez et al. 2004). Therefore, it is not unexpected that selection for mutations at genes associated with these traits occurred in the human lineage after the divergence between the ancestors of modern humans and our closest living relative, the common chimpanzee.

In the immune system, well-known genes that regulate host defense against infection, such as the MHC (Hughes et al. 1990) and leukocyte antigens (Vallender and Lahn 2004), are characterized by a high level of replacement substitutions. In the present study, we identified signatures of selection at replacement substitutions in exon 2 of the *RNASEL* gene (fig. 3C), corresponding to amino acids that participate in the binding of the 2-5A protein, a key step in initiating the cleavage of viral RNA by the RNASEL enzyme (Tanaka et al. 2004). Furthermore, the strong signals of selection at replacement sites in the well-defined regions of *IL18RAP* (fig. 3D) suggest that these substitutions also play a role in disease resistance. Specifically, *IL18RAP* encode proteins that interact with other immune-related proteins to form complexes that trigger an antiviral response (Fink and Grandvaux 2013; Blaszczyk et al. 2015). Although the function of these replacement sites are currently unknown, these genic regions should be the focus of future studies aimed at understanding the evolution and genetic basis of human-specific biological traits. Lastly, results from our MASS-PRF analysis have implications for the design of interventions against infectious disease. Specifically, estimates of γ along DNA sequences can be informative for identifying functionally important regions of proteins (encoded by genes under selection), and this information may be of potential interest to scientists who develop protein-based therapeutics for the treatment and prevention of viral infections (Wagner 2007; Koellhoffer et al. 2014).

Like the genes associated with immune response, genes involved in sexual reproduction also evolved under strong positive selection (Torgerson et al. 2002). We identified positive selection at amino acid-altering nucleotides in exon 6 of *WBP2NL* (fig. 3F) corresponding to the repeat motif (YGAPPLG) in the WBP2NL protein. A recent in vitro study showed that the presence of YGXPPXG repeating motifs increases the binding specificity of WBP2NL to proteins in the oocyte leading to the activation of DNA and protein synthesis in the egg after fertilization (Wu et al. 2007). Other examples of functional mutations under positive selection include a replacement substitution at site 2628 in exon 15 of *SPAG5* (fig. 3G) that participates in the binding of KNSTRN, a key protein involved in chromosome segregation during mitosis. Additionally, sites under selection in exon 2 of *TGM4* (fig. 3E) encode amino acids that bind to the surface of sperm to minimize the activation of an antisperm immune response in the female reproductive tract (Clark and Swanson 2005). Thus, the replacement substitutions that we identified using MASS-PRF appear to play critical roles in species-specific fertilization, gamete recognition and human development. Lastly, replacement sites in exon 5 of *NT5C1B* (fig. 4B) also exhibited high levels of γ indicative of positive selection. Although the function of these substitutions remains unclear, our results suggest that these nucleotide substitutions in exon 5 likely represent targets of selection associated with gene function.

In addition, our analysis of *SLC8A1* and *TPH2* by MASS-PRF showed consistent selection results with MK tests (fig. 4*A* and *D*). Interestingly, we did not observe statistically significant peaks of γ within *MGAM* using MASS-PRF, in contrast to a statistically significant result for *MGAM* using the MK test ($P = 0.04$; fig. 4C). Examination of the distribution of the number of divergent and polymorphic sites ($P_s = 28$, $P_r = 48$, $D_s = 27$, and $D_r = 20$) reveals that significance of the MK test was driven by an excess of polymorphic replacement sites, which is consistent with a scenario of balancing selection. Because the PRF model is a model of directional selection, it is entirely appropriate that MASS-PRF would not yield a statistically significant peak of γ in *MGAM*: this gene does not appear to be under positive selection. In contrast, MASS-PRF analysis of *NT5C1B* revealed strong selection in a restricted region close to nucleotide position 800. *NT5C1B* does not demonstrate statistically significant selection across the entire gene ($P = 0.67$) using the MK test (fig. 4*B*). MASS-PRF indicates that the action of selection on this gene is limited to a small region—therefore it would be unlikely that any gene-wide methodology would detect this selection. While the genes analyzed in this study may have evolved in direct response to environmental pressures, it is possible that replacement substitutions in these genes could have occurred due to co-evolution. Specifically, co-evolution of genes could arise through protein/protein interaction in which compensatory changes in one protein occurs in response to changes in a partner protein directly under positive selection. That is to say, amino acid changes in one protein under selection could exert pressure on another protein to maintain particular amino acid changes that facilitate continued protein/protein interactions (Qian et al. 2015). Intriguingly, a recent genome-wide analysis of 1000 Genomes data reported strong signatures of recent selection among interacting proteins involved in signal transduction, neurogenesis and immune function, suggesting that the process of co-evolution has influenced patterns of variation in human genes (Wyckoff et al. 2000; Qian et al. 2015). Similar results have been described for genes associated with immunity and reproduction in *Drosophila* (Obbard et al. 2006, 2009). Thus, MASS-PRF can provide additional insights into the evolution of gene–gene (or protein–protein) interactions.

In general, because RD substitutions are fixed in the human lineage, it can be argued that these changes are likely ancient, predating both the separation of humans into different subpopulations beginning ~90,000–100,000 years ago in Africa and the migration of humans across the globe from Africa beginning ~80,000 years ago (Campbell et al. 2014). Given the recent increase in genome sequence data from Neanderthal remains, MASS-PRF could be applied in future studies to investigate changes that evolved uniquely in *H. sapiens* since the divergence of modern humans and Neanderthals from a common ancestor ~400,000–500,000 years ago. Overall, MASS-PRF is an informative tool for detecting functional targets of selection at relatively deep evolutionary time scales and for understanding the origins of species-specific traits.

Materials and Methods

Identifying Clustering Models Using Observed Polymorphism and Divergence Data

To analyze patterns of variation within and between species, we aligned homologous sequences, scored each site as either 0 (representing no variant) or 1 (representing a variant site), and used model-averaged clustering of discrete linear sequences (Zhang and Townsend 2009). This method calculates all likely models of linear clustering by partitioning the entire coding sequence into all possible sets of three regions, and the multiple-Bernoulli likelihood was estimated for each clustering model. Models were penalized for overparameterization via the AIC, BIC, or AIC (corrected; AICc). We then calculated the weighted average probability of a variant appearing at each site across all models (as in MACML; Zhang and Townsend 2009), using the weights

$$w_i = \frac{e^{-\frac{1}{2}\left(a_i - \breve{a}\right)}}{\sum_{j=1}^{m} e^{-\frac{1}{2}\left(a_j - \breve{a}\right)}}, \tag{1}$$

where a_i is the AIC of model i, $\breve{a}$ is the smallest AIC in all models, and m is the number of all candidate models.

To perform the model averaging, we defined the probability of a variant at each site i as the weighted average probability across all models for each site i,

$$p(i) = \sum_{j=1}^{m} w_j \times p(i|j), \tag{2}$$

where $p(i|j)$ is $p(i)$ given model j. To calculate the central 95% of weighted models for the probabilities at each site, we first sorted all the models by $p(i)$, then summed the weight of each model from low to high. When the cumulative weight reached 0.025 and 0.975, the $p(i)$ associated with that model was the lower and upper 95% bound, respectively. This clustering can be performed on all four categories of variants: SP, SD, RP, and RD. Model averaging of the probability of each site being a variant is optional fig. 1, insets A–D.

Calculating the Selection Intensity on Each Site

The estimated maximum likelihood probabilities of polymorphism and divergence data computed by the MACML algorithm above were used to parameterize the PRF model (Sawyer and Hartl 1992). Theoretical expectations for polymorphism and divergence in the PRF model are calculated as follows:

$$E(SD) = 2\mu_s \times \left(t + \frac{1}{m} + \frac{1}{n}\right), \tag{3}$$

$$E(SP) = 2\mu_s \times [L(m) + L(n)], \tag{4}$$

$$E(RD) = 2\mu_r \times \left(\frac{2\gamma}{1 - e^{-2\gamma}}\right) \times [t + G(m) + G(n)], \tag{5}$$

and

$$E(RP) = 2\mu_r \times \left(\frac{2\gamma}{1 - e^{-2\gamma}}\right) \times [F(m) + F(n)], \tag{6}$$

where μ_s is the silent mutation rate per N_e generations, t is the species divergence, m and n are the corresponding sample sizes of sequences in the gene alignment from each species (the $1/m$ and $1/n$ terms account for actually polymorphic sites that may be observed as monomorphic in each sample), μ_r is the replacement mutation rate per N_e generations. In these equations, the scaled selection coefficient is $\gamma = 2N_e s$, where N_e is the effective population size and s is the canonical population genetic selection coefficient. One explicit assumption of this method is that N_e is the same between two species being compared. Under PRF theory, a scaled selection coefficient $\gamma > 0$ corresponds to patterns of substitution driven by positive selection, $\gamma < 0$ corresponds to negatively selected substitutions, and $\gamma = 0$ corresponds to a pattern of substitution driven by neutral evolution (Sawyer and Hartl 1992). Ohta (2002) suggested that weak selection or near neutrality corresponds to $0.5 < |N_e s| < 3$ in *Drosophila* (Ohta 1992, 2002), and that estimates of γ above this range should be strong enough so that the change in frequency of a mutation mainly depends on selection (Sawyer et al. 2007). Using simulation data, we defined nearly neutral as $-1 < \gamma < 4$, strong positive selection as $\gamma > 4$ and strong negative selection as $\gamma < -1$.

Specifically, for each demographic and recombination scenario, we plotted the percentage of statistically significant sites over all the simulated genes (the FPR) as a function of the γ threshold, testing all threshold values from -10 to 10 with an interval of 0.1. As we expected, the FPR for positive selection falls sharply with increasing γ thresholds, whereas the FPR for negative selection increases rapidly with increasing thresholds for γ (supplementary fig. S3, Supplementary Material online). On the basis of this analysis, we found that γ thresholds associated with low FPRs (<0.1) were close to our specified thresholds of neutrality at –1 and 4. In addition, we defined levels of positive selection based on estimates of γ as follows: weak ($\gamma > 4$; lower bound < 0), moderate ($\gamma > 4$; $0 <$ lower bound < 4), and strong ($\gamma > 4$; lower bound > 4) evidence for positive selection in specific genic regions. The lower bound in these inequalities refers to the lower 95% model uncertainty interval for the estimate of γ.

The functions L, F, and G appearing in equations (4–6) all account for sites that are polymorphic in the population that may be observed as monomorphic within finite sample sizes m and n, and are defined in Sawyer and Hartl (1992) as

$$L(n) = \sum_{i=1}^{n-1} \frac{1}{i}, \tag{7}$$

$$F(n) = \int_0^1 \frac{1 - x^n - (1-x)^n}{1 - x} \cdot \frac{1 - e^{-2\gamma x}}{2\gamma x} dx, \text{ and} \tag{8}$$

$$G(n) = \int_0^1 (1-x)^{n-1} \frac{1 - e^{-2\gamma x}}{2\gamma x} dx. \tag{9}$$

To parameterize t in equations (3) and (5), divergence estimates can be derived from other data, or divergence between the two species can be inferred from presumably neutral silent substitutions (Sawyer and Hartl 1992). In the latter case, the observed numbers of SPs within species and SD between species are tallied from all aligned DNA sequences. A nominal divergence t can then be estimated at a gene level by setting the ratio of the expectations in equations (4) and (3) equal to the ratio of observations,

$$\frac{L(m) + L(n)}{t + \frac{1}{m} + \frac{1}{n}} = \frac{N_{sp}}{N_{sd}}, \tag{10}$$

where N_{sp} is the number of SP and N_{sd} is the number of SD across a given coding gene, while and $L(m)$ and $L(n)$ are defined by equation (7).

We modified the PRF method (Sawyer and Hartl 1992), which typically calculates a single γ value for all divergent sites along a given sequence. In our modified method, we parameterized the PRF model with the average maximum likelihood probability of observing RP and RD at each site *i*. We then estimated γ_i by setting the ratio of the expectations in equations (6) and (5) equal to the ratio of estimated probabilities of RP and RD at site *i*,

$$\frac{F(m)+F(n)}{t+G(m)+G(n)}=\frac{p_{rp}(i)}{p_{rd}(i)}. \tag{11}$$

$F(m)$, $F(n)$, $G(m)$, and $G(n)$ are defined by equations (8) and (9). A computationally tractable simplification of equation (11) arises when $m=1$ (i.e., we are using just one divergent sequence). In that case,

$$\begin{aligned} f(\gamma_i) = t + \int_0^1 & \frac{1-e^{-2\gamma_i x}}{2\gamma_i x} \\ & \times\left[1+(1-x)^{n-1}-\frac{p_{rd}(i)}{p_{rp}(i)}\frac{1}{1-x}\right. \\ & \left.\times\left(1-x^n-(1-x)^n\right)\right]dx, \end{aligned} \tag{12}$$

$$\begin{aligned} \text{and } f^{'}(\gamma_i) = \int_0^1 & \left[1+(1-x)^{n-1}-\frac{p_{rd}(i)}{p_{rp}(i)}\frac{1}{1-x}\right. \\ & \left.\times\left(1-x^n-(1-x)^n\right)\right] \\ & \times\frac{e^{-2\gamma_i x}\left(1+2\gamma_i x-e^{2\gamma_i x}\right)}{2{\gamma_i}^2 x}dx. \end{aligned} \tag{13}$$

We estimated γ_i for a single model by implementing the Newton–Raphson method (Lange 1999) using equation (12) and its derivative equation (13).

In equations (10–12), parameter *t* is assumed to be constant across sites within a gene. Alternatively, to incorporate potential heterogeneity of mutation rate across sites when estimating γ_i, it is possible to subscript *t* by site *i* in equation (10), in which case, the site-specific divergence time t_i can stand as a proxy for variation in silent mutation rate across sites:

$$t(i)=\frac{p_{sd}(i)}{p_{sp}(i)}\left[L(m)+L(n)\right]-\frac{1}{m}-\frac{1}{n}. \tag{14}$$

Estimating the Model-Averaged Selection Intensity and Its Model Uncertainty Intervals for Each Site

Our estimate of the selection intensity at site *i* is the model-averaged value across all or a stochastic sample of joint models, and the model uncertainty interval is calculated as the central 95% of joint model probability weight. The weight of each joint model is a product of the model weights (eq. 1) for RP and RD models or alternatively a product of the model weights for RP, RD, SP, and SD models (if silent site clustering is implemented to account for mutation rate heterogeneity within genes). Model-averaged γ is then calculated by weighing every γ associated with each joint model. To estimate 95% model uncertainty intervals for selection intensity, we developed and implemented two algorithms that can be selected based on user needs: (1) an exhaustive option that should be selected for analysis of shorter genes where intensive computation is not prohibitive, or (2) a stochastic option that should be selected for analysis of longer genes in a relatively short time and with a high level of accuracy and precision over sufficient iterations.

For the exhaustive algorithm, weights for all possible joint models are calculated as above. The model-based selection intensity at site *i* for each joint model is calculated using equations (12) and (13). These model-based selection intensities are then averaged by their joint weights to obtain the model-averaged selection intensity γ_i. To calculate the 95% model uncertainty interval of selection intensity at each site, all joint models are sorted by their selection intensity at site *i*, then the weight of each joint model is summed from low to high. When the cumulative weight reaches 0.025 and 0.975, the selection intensity associated with those joint models represents the lower and upper 95% bound, respectively.

In contrast, the stochastic algorithm samples a subset of the possible joint models, requiring markedly less computation in comparison with the exhaustive algorithm. The sum of all weighted joint models equals 1. To sample stochastically, random variables between 0 and 1 are generated to select a large number of joint models proportionately to their joint model weight (30,000 by default). Once the subset of joint models is selected, model-averaged selection intensity is calculated as the average value of model-based selection intensities sampled at site *i*, and the 95% model uncertainty interval at each site is bounded by the 0.025 and 0.975 quantiles of the sampled ranked model-based selection intensities.

Implementation of MASS-PRF

MASS-PRF was written in the standard C++ programming language. The software package is accompanied by a manual, example data, source codes, and compiled executable commands for Windows/Linux/Mac. Source codes are released to GPLv3, and can be downloaded from https://github.com/Townsend-Lab-Yale/MASSPRF (last accessed August 2, 2017).

Simulations of Clustering of Variant Sites with Polymorphism and Divergence Data

To demonstrate the robustness of MASS-PRF, we simulated SP, SD, RP, and RD sites distributed within clusters ranging from 100 bp to 1,200 bp in length, with cluster size increasing by 100 bp up to a maximum cluster size of 1,200 bp (the entire length of the gene).

Our simulations were conducted as follows:

(1) Each simulated sequence represents protein-coding sequence with a total length of 1,200 bp (corresponding to a protein 400 amino acids in length).

(2) The numbers of SP, RP, SD, and RD were specified as 15, 10, 25, and 20 respectively in a 1,200 bp long nucleotide coding sequence. These numbers for SP, RP, SD, and RD are within the ranges of organismal polymorphism and divergence data.

(3) SP, RP, SD, and RD sites were randomly and uniformly distributed within the cluster length.

We used the expected probabilities derived from our simulation parameters for SP, SD, RP, and RD, using count data for each variant type and information about the width of the region in which these sites occur (i.e., either the size of a cluster ranging from 100 bp to 1,200 bp assessed in increments of 100 bp. Sequences were generated by custom Perl scripts and we performed $M = 100$ replicates for each simulation scenario. The maximum likelihood probability of a variant site for both the clustered and non-clustered model was compared with the expected (simulated) probabilities using the Kullback–Liebler distance to determine the accuracy and precision of MASS-PRF in detecting clusters of nucleotide substitutions.

Accuracy and Precision of MASS-PRF in Detecting Clusters of Substitutions Using Simulated Data

To assess the performance of MASS-PRF and an unclustered approach in detecting clusters of substitutions, we quantified the divergence between the multiple Bernoulli probability distribution p corresponding to the expected probabilities across sites and the multiple Bernoulli probability distribution p^* corresponding to the estimated probabilities across sites by calculating the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951). In the unclustered approach, the probability at each site p was constrained to be identical across sites. The KL divergence measures the difference between two probability distributions p and p^*, and is defined as

$$D(p||p^*) = p \times \log_2 \frac{p}{p^*} + (1-p) \times \log_2 \frac{(1-p)}{(1-p^*)}. \quad (15)$$

With M replicates for each variable combination, and N sites for each sequence, the average KL divergence over $M \times N$ was calculated as

$$\bar{D} = \frac{1}{M \times N} \sum_{j=1}^{M \times N} D(p||p^*). \quad (16)$$

Equation (16) measures the average closeness of our estimated probabilities to expected probabilities, quantifying the accuracy of the estimation. The precision σ was calculated as

$$\sigma = \sqrt{\frac{1}{M \times N} \sum_{j=1}^{M \times N} [D(p||p^*) - \bar{D}]^2}. \quad (17)$$

Because KL divergence measures the difference between the two distributions, an average KL divergence $\bar{D}$ and a σ approaching 0 indicate a good match between the estimated and expected distributions. Thus, accuracy and precision based on KL divergence are interpreted to be high when $\bar{D}$ and σ are low. The KL divergence was performed on all four categories of variants: SP, SD, RP, and RD (fig. 2).

Assessment of Robustness of MASS-PRF to Demographic and Recombination History Using Coalescent Simulations

To evaluate the impacts of demography and recombination on the estimation of selection intensity using MASS-PRF, we simulated six sets of neutrally evolved genes under three demographic events: bottleneck, constant and expansion scenarios with and without recombination. Firstly, we used Hudson's ms (Hudson 2002) to simulate 20 coalescent graphs, each containing 101 samples, for six conditions incorporating three demographic events and recombination. Within each iteration of generating each coalescent graph, we specified a diploid population size of $N_0 = 10^4$, a mutation rate of 10^{-8}, and a recombination rate of $\rho = 0.36$ (as suggested in Hudson's ms application) with a gene length of 900 bp. Within each iteration, the 101 samples include one divergent sequence and 100 polymorphic sequences, which diverged 6 Ma (by eliminating migration from that time forward).

We specified three demographic histories with and without recombination: an instantaneous 5-fold bottleneck at 100 ka (thousand years ago), a constant population, and a 10-fold exponential growth beginning 100 ka. After generating 20 trees under each of the six scenarios, we used Fletcher and Yang's INDELible (Fletcher and Yang 2009) to evolve sequences consistent with the coalescent trees under a neutral codon substitution model (the two-ratio model M1 with $\omega_0 = 0$, $\omega_1 = 1$, and $p_0 = 0.5$) as demonstrated by Yang and Nielsen (2002). In the recombination scenarios, the multiple trees for each coalescent simulation that were output by ms were independently evolved in NDELible and then reconcatenated to produce full 900 bp sequences.

MK tests were then applied to the simulated data to test the null hypothesis that genes were neutrally evolving. Finally, we estimated selection intensity for the six sets of simulated sequences using MASS-PRF with two divergence time strategies: (1) the site-specific divergence time calculated using silent sites, and (2) a fixed 6 Ma divergence time corresponding to the human–chimpanzee species divergence time (of the human and chimpanzee lineages) specified in our ms coalescent simulations. FPRs of selection intensities inferred by MASS-PRF on the simulated data were calculated.

Application of MASS-PRF to Empirical Data

We evaluated selection intensities by MASS-PRF in seven genes: *SLC6A5*, *GRIN2C*, *RNASEL*, *IL18RAP*, *TGM4*, *WBP2NL*, and *SPAG5*, which were detected to be under positive selection by Bustamante et al. (2005), Kosiol et al. (2008), or Clark et al. (2005). We extracted phased polymorphisms in these seven genes from the 1000 Genomes whole exome sequence data (supplementary table S2, Supplementary Material online) based on the genomic coordinates given in the National Center for Biotechnology Information (NCBI) database (Build GRCh37). We then converted the genomic coordinates of polymorphisms in the above genes to their corresponding transcript coordinates using the Variant Annotation Integrator software on the UCSC

Genome Browser website. We downloaded the transcripts for each gene (which served as a reference) from the UCSC Genome Browser and inserted the phased polymorphic alleles into the reference transcript using their transcript coordinates, creating two copies of the protein-coding gene. We performed this step for each of the 297 individuals in the 1000 Genomes whole exome data set.

Once we reconstructed the diploid sequences for each individual based on their polymorphisms and the reference transcript, we applied MASS-PRF to these data and the aligned orthologous gene transcripts from *P. troglodytes* (common chimpanzee). For this analysis, we specified the default option of using clustering of silent sites to calculate the time of divergence (t) between the ancestors of humans and chimpanzees. We specified the BIC criterion for calculating weights of cluster models and weights of joint models, using the stochastic algorithm (sampling 30,000 models) to calculate the model-averaged selection intensity and its 95% model uncertainty intervals. For selection intensities indicative of strong positive or negative selection, we examined our sequence data to confirm the presence of clusters of divergent replacement sites in our human samples compared with chimpanzee sequence.

To distinguish nucleotide substitutions that occurred in the human lineage from those that occurred in the chimpanzee lineage, we used *Gorilla gorilla* as an outgroup (supplementary table S5, Supplementary Material online). The chimpanzee and gorilla sequences were downloaded from Ensembl. In addition to the above genes previously inferred to be under positive selection in Bustamante et al. (2005), Clark and Swanson (2005), Kosiol et al. (2008), we also analyzed 51 genes that were not previously identified to be under selection based on polymorphic and divergent data (Bakewell et al. 2007; Gaya-Vidal and Alba 2014) using MASS-PRF.

To detect departures from neutral evolution in real data, we applied the MK tests to the above 58 genes (figs. 3, 4 and supplementary fig. S2, Supplementary Material online). We also compared the counts of replacement and silent sites within and between species (human and chimp anzee) for all of our genes using the MK test (McDonald and Kreitman 1991). In the MK test, under neutrality, the within- to between-species ratio for silent variant counts is expected to be the same as the within- to between-species ratio of replacement variant counts. For this between-species comparison, we used sequences from *P. troglodytes* downloaded from NCBI. Significance for the MK statistic was determined using Fisher's Exact Test. We calculated Q values to control for possible type I errors when conducting multiple tests for MK tests. The Q values were computed by inputting the P values from the MK tests into the MATLAB Bioinformatics Toolbox command mafdr (www.mathworks.com/help/bioinfo/ref/mafdr.html), which implements a false discovery rate estimation following (Storey 2002).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Angelika Hofmann for reading and making comments on the manuscript, and many colleagues too numerous to name who have shared their thoughts. We also thank the Yale High Performance Computing Center and NIH grant S10RR029676-01 for providing computational resources, and Robert Bjornson and Nicholas Carriero for their assistance with cluster computation for this project. This work was supported by a Yale Endowed Postdoctoral Fellowship to N.L. and by funds from the Notsew Orm Sands Foundation to J.P.T.

References

Aguileta G, Refregier G, Yockteng R, Fournier E, Giraud T. 2009. Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infect Genet Evol.* 9(4):656–670.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136(3): 927–935.

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics* 139(2):1067–1076.

Arbiza L, Dopazo J, Dopazo H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput Biol.* 2(4):288–300.

Arenas M, Posada D. 2010. Coalescent simulation of intracodon recombination. *Genetics* 184(2):429–437.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci U S A.* 104(18):7489–7494.

Bakewell MA, Zhang J. 2008. Positive selection on genes in humans as compared to chimpanzees. In: Encyclopedia of life sciences (ELS). Chichester (United Kingdom): John Wiley & Sons, Ltd.

Barr CM, Keller SR, Ingvarsson PK, Sloan DB, Taylor DR. 2007. Variation in mutation rate and polymorphism among mitochondrial genes of Silene vulgaris. *Mol Biol Evol.* 24(8):1783–1791.

Biswas S, Akey JM. 2006. Genomic insights into positive selection. *Trends Genet.* 22(8):437–446.

Blaszczyk K, Olejnik A, Nowicka H, Ozgyin L, Chen YL, Chmielewski S, Kostyrko K, Wesoly J, Balint BL, Lee CK, et al. 2015. STAT2/IRF9 directs a prolonged ISGF3-like transcriptional response and antiviral activity in the absence of STAT1. *Biochem J.* 466(3):511–524.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153–1157.

Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159(4):1779–1788.

Campbell MC, Hirbo JB, Townsend JP, Tishkoff SA. 2014. The peopling of the African continent and the diaspora into the new world. *Curr Opin Genet Dev.* 29:120–132.

Chamary JV, Hurst LD. 2005. Biased codon usage near intron–exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet.* 21(5):256–259.

Chen GL, Miller GM. 2012. Advances in tryptophan hydroxylase-2 gene expression regulation: new insights into serotonin–stress interaction and clinical implications. *Am J Med Genet B Neuropsychiatr Genet.* 159b(2):152–171.

Choi SY, Choi K, Jeon JH, Kim CW, Shin DM, Lee JB, Lee SE, Kim CS, Park JS, Jeong EM, et al. 2010. Differential alternative splicing of human transglutaminase 4 in benign prostate hyperplasia and prostate cancer. *Exp Molec Med.* 42(4):310–318.

Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. *PLOS Genet.* 1(3):e35.

Desai MM, Plotkin JB. 2008. The polymorphism frequency spectrum of finitely many sites under selection. *Genetics* 180(4):2175–2191.

Dunsch AK, Linnane E, Barr FA, Gruneberg U. 2011. The astrin–kinastrin/SKAP complex localizes to microtubule plus ends and facilitates chromosome alignment. *J Cell Biol.* 192(6):959–968.

Egea J, Erlacher C, Montanez E, Burtscher I, Yamagishi S, Hess M, Hampel F, Sanchez R, Rodriguez-Manzaneque MT, Bösl MR, et al. 2008. Genetic ablation of FLRT3 reveals a novel morphogenetic function for the anterior visceral endoderm in suppressing mesoderm differentiation. *Genes Dev.* 22(23):3349–3362.

Eilertson KE, Booth JG, Bustamante CD. 2012. SnIPRE: selection inference using a Poisson random effects model. *PLoS Comput Biol.* 8(12):e1002806.

Eyre-Walker A. 2002. Changing effective population size and the McDonald–Kreitman test. *Genetics* 162(4):2017–2024.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.

Ferretti L, Disanto F, Wiehe T. 2013. The effect of single recombination events on coalescent tree height and shape. *PLoS ONE.* 8(4):e60123.

Fink K, Grandvaux N. 2013. STAT2 and IRF9: beyond ISGF3. *JAKSTAT* 2(4):e27521.

Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. *Molec Biol Evol.* 26(8):1879–1888.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R, Akey JM. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLOS Genet.* 7(11):e1002355.

Gaya-Vidal M, Alba MM. 2014. Uncovering adaptive evolution in the human lineage. *BMC Genomics.* 15:599.

Gill JL, Capper D, Vanbellinghen JF, Chung SK, Higgins RJ, Rees MI, Shelton GD, Harvey RJ. 2011. Startle disease in Irish wolfhounds associated with a microdeletion in the glycine transporter GlyT2 gene. *Neurobiol Dis.* 43(1):184–189.

Heyes C. 2012. New thinking: the evolution of human cognition. *Philos Trans R Soc B-Biol Sci.* 367(1599):2091–2096.

Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Brown AJL. 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human-immunodeficiency-virus type-1 within a single infected patient. *Proc Natl Acad Sci U S A.* 89(11):4835–4839.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.

Hudson RR, Martin K, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.

Hughes AL, Ota T, Nei M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class-I major-histocompatibility-complex molecules. *Molec Biol Evol.* 7(6):515–524.

Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Ann Rev Genet.* 32:415–435.

Kerns JA, Emerman M, Malik HS. 2008. Positive selection and increased antiviral activity associated with the PARP-containing isoform of human zinc-finger antiviral protein. *PLOS Genet.* 4(1):e21.

Khananshvili D. 2013. The SLC8 gene family of sodium-calcium exchangers (NCX) – structure, function, and regulation in health and disease. *Mol Aspects Med.* 34(2–3):220–235.

Koellhoffer JF, Higgins CD, Lai JR. 2014. Protein engineering strategies for the development of viral vaccines and immunotherapeutics. *FEBS Lett.* 588(2):298–307.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* 4(8):e1000144.

Kreitman M. 2000. Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet.* 1:539–559.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLOS Genet.* 4(12):e1000304.

Kullback S, Leibler RA. 1951. On information and sufficiency. *Ann Math Stat.* 22(1):79–86.

Lange K. 1999. Numerical Analysis for Statisticians. New York: Springer-Verlag.

Lazzaro BP. 2005. Elevated polymorphism and divergence in the class C scavenger receptors of *Drosophila melanogaster* and *D. simulans. Genetics* 169(4):2023–2034.

Lin X. 2011. Perception of sound and gravity by TMC1 and TMC2. *J Clin Investig.* 121(12):4633–4636.

Martinez I, Rosa M, Arsuaga JL, Jarabo P, Quam R, Lorenzo C, Gracia A, Carretero JM, de Castro JMB, Carbonell E. 2004. Auditory capacities in Middle Pleistocene humans from the Sierra de Atapuerca in Spain. *Proc Natl Acad Sci U S A.* 101(27):9976–9981.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351(6328):652–654.

McKinney JA, Turel B, Winge I, Knappskog PM, Haavik J. 2009. Functional properties of missense variants of human tryptophan hydroxylase 2. *Hum Mutat.* 30(5):787–794.

Nei MG, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molec Biol Evol.* 3:418–426.

Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* 86(Pt 6):641–647.

Nielsen R. 2005. Molecular signatures of natural selection. *Ann Rev Genet.* 39:197–218.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet.* 8(11):857–868.

Nielsen R, Yang ZH. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3):929–936.

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. *Nature* 396(6711):572–575.

Obbard DJ, Jiggins FM, Halligan DL, Little TJ. 2006. Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Curr Biol.* 16(6):580–585.

Obbard DJ, Welch JJ, Kim KW, Jiggins FM. 2009. Quantifying adaptive evolution in the Drosophila immune system. *Plos Genet.* 5(10):e1000698.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Ann Rev Ecol Syst.* 23(1):263–286.

Ohta T. 2002. Near-neutrality in evolution of genes and gene regulation. *Proc Natl Acad Sci U S A.* 99(25):16134–16137.

Parsch J, Meiklejohn CD, Hartl DL. 2005. Inferring evolutionary history through inter- and intraspecific DNA sequence comparison: the *Drosophila janus* and ocnus genes. Selective Sweep. *Molec Biol Intell Unit.* 1–12. See https://link.springer.com/book/10.1007/0-387-27651-3

Parsch J, Zhang Z, Baines JF. 2009. The influence of demography and weak selection on the McDonald–Kreitman test: an empirical study in Drosophila. *Molec Biol Evol.* 26(3):691–698.

Qian W, Zhou H, Tang K. 2015. Recent coselection in human populations revealed by protein–protein interaction network. *Genome Biol Evol.* 7:136–153.

Racimo F, Schraiber JG. 2014. Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. *PLOS Genet.* 10(11):e1004697.

Rand DM, Dorfsman M, Kann LM. 1994. Neutral and non-neutral evolution of Drosophila mitochondrial DNA. *Genetics* 138(3):741–756.

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol.* 239(2):226–235.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419(6909):832–837.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in Drosophila. *Nat Genet.* 39(12):1461–1468.

Saminadin-Peter SS, Kemkemer C, Pavlidis P, Parsch J. 2012. Selective sweep of a cis-regulatory sequence in a non-African population of *Drosophila melanogaster*. *Molec Biol Evol.* 29(4):1167–1174.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics* 132(4):1161–1176.

Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. *Proc Natl Acad Sci U S A.* 104(16):6504–6510.

Schlenke TA, Begun DJ. 2003. Natural selection drives drosophila immune system evolution. *Genetics* 164(4):1471–1480.

Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* 34(8):2428–2437.

Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res.* 41(4):2073–2094.

Singh ND, DuMont VLB, Hubisz MJ, Nielsen R, Aquadro CF. 2007. Patterns of mutation and selection at synonymous sites in Drosophila. *Molec Biol Evol.* 24(12):2687–2697.

Storey JD. 2002. A direct approach to false discovery rates. *J R Stat Soc: Ser B (Stat Methodol).* 64(3):479–498.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. 2013. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* 29(18):2238–2244.

Tanaka N, Nakanishi M, Kusakabe Y, Goto Y, Kitade Y, Nakamura KT. 2004. Structural basis for recognition of 2′, 5′-linked oligoadenylates by human ribonuclease L. *EMBO J.* 23(20):3929–3938.

Taub DR, Page J. 2016. Molecular signatures of natural selection for polymorphic genes of the human dopaminergic and serotonergic systems: a review. *Front Psychol.* 7:857.

Teng H, Cai W, Zhou L, Zhang J, Liu Q, Wang Y, Dai W, Zhao M, Sun Z. 2010. Evolutionary mode and functional divergence of vertebrate NMDA receptor subunit 2 genes. *PLoS ONE.* 5(10):e13342.

Torgerson DG, Kulathinal RJ, Singh RS. 2002. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Molec Biol Evol.* 19(11):1973–1980.

Vallender EJ, Lahn BT. 2004. Positive selection on the human genome. *Hum Molec Genet.* 13(Suppl 2):R245–R254.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annu Rev Genet.* 47:97–120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4(3):e72.

Wagner A. 2007. Rapid detection of positive selection in genes and genomes through variation clusters. *Genetics* 176(4):2451–2463.

Wu ATH, Sutovsky P, Manandhar G, Xu W, Katayama M, Day BN, Park K-W, Yi Y-J, Xi YW, Prather RS, et al. 2007. PAWP, a sperm-specific WW domain-binding protein, promotes meiotic resumption and pronuclear development during fertilization. *J Biol Chem.* 282(16):12164–12175.

Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403(6767):304–309.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molec Biol Evol.* 19(6):908–917.

Yang ZH, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molec Biol Evol.* 19(1):49–57.

Zeng K, Charlesworth B. 2010. The effects of demography and linkage on the estimation of selection and mutation parameters. *Genetics* 186(4):1411–U1537.

Zhang Z, Townsend JP. 2009. Maximum-likelihood model averaging to profile clustering of site types across discrete linear sequences. *Plos Comput Biol.* 5(6):e1000421.

Zhu L, Bustamante CD. 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* 170(3):1411–1421.