Differential Effect of Selection against LINE Retrotransposons among Vertebrates Inferred from Whole-Genome Data and Demographic Modeling

Alexander T. Xue[1,2,*], Robert P. Ruggiero[3], Michael J. Hickerson[1,4], and Stéphane Boissinot[3,*]

[1]Department of Biology: Subprogram in Ecology, Evolutionary Biology, and Behavior, City College and Graduate Center of City University of New York

[2]Human Genetics Institute of New Jersey and Department of Genetics, Rutgers University, Piscataway

[3]New York University Abu Dhabi, Saadiyat Island Campus, United Arab Emirates

[4]Division of Invertebrate Zoology, American Museum of Natural History, New York, New York

*Corresponding authors: E-mails: xanderxue@gmail.com; sb5272@nyu.edu.

Accepted: April 20, 2018

Data deposition: This project has been deposited at Dryad under the accession http://dx.doi.org/10.5061/dryad.qd300cb.

Abstract

Variation in LINE composition is one of the major determinants for the substantial size and structural differences among vertebrate genomes. In particular, the larger genomes of mammals are characterized by hundreds of thousands of copies from a single LINE clade, L1, whereas nonmammalian vertebrates possess a much greater diversity of LINEs, yet with orders of magnitude less in copy number. It has been proposed that such variation in copy number among vertebrates is due to differential effect of LINE insertions on host fitness. To investigate LINE selection, we deployed a framework of demographic modeling, coalescent simulations, and probabilistic inference against population-level whole-genome data sets for four model species: one population each of threespine stickleback, green anole, and house mouse, as well as three human populations. Specifically, we inferred a null demographic background utilizing SNP data, which was then exploited to simulate a putative null distribution of summary statistics that was compared with LINE data. Subsequently, we applied the inferred null demographic model with an additional exponential size change parameter, coupled with model selection, to test for neutrality as well as estimate the strength of either negative or positive selection. We found a robust signal for purifying selection in anole and mouse, but a lack of clear evidence for selection in stickleback and human. Overall, we demonstrated LINE insertion dynamics that are not in accordance to a mammalian versus nonmammalian dichotomy, and instead the degree of existing LINE activity together with host-specific demographic history may be the main determinants of LINE abundance.

Key words: retrotransposons, transposable elements, purifying selection, comparative population genomics, composite likelihood optimization, approximate Bayesian computation.

Introduction

Vertebrate genomes differ considerably in size and structure, with transposable element (TE) abundance and diversity among the genomic features that show the most variation between species (Tollis and Boissinot 2012). Specifically, copy number for Long Interspersed Nuclear Elements (LINEs) largely accounts for the generally greater genome sizes of mammals relative to nonmammalian vertebrates. LINEs constitute a diverse and ancient group of mobile DNA, consisting of 28 clades defined by specific functional features and supported by phylogenetic analysis (Kapitonov et al. 2009). Each of these clades can then be represented within an organism by multiple discrete groups of elements, denoted as families. As non-long terminal repeat retrotransposons, LINEs replicate by reverse transcription of their RNA at the site of insertion by a process called target primed reverse transcription (Luan et al. 1993; Cost et al. 2002). This is an inefficient process since the majority of novel insertions are truncated at the 5′ end and

thus no longer capable of further transposition, that is, dead-on-arrival. LINE insertions can cause genetic defects and genomic instability (Ostertag and Kazazian 2001; Burns and Boeke 2012), yet they contribute a substantial source of evolutionary novelties that can affect genes and genomic regulation (Rebollo et al. 2012; Warren et al. 2015; Mita and Boeke 2016).

Mammalian genomes contain hundreds of thousands of LINE copies generated by a single clade, L1, which accounts for at least 18% of genome size (International Human Genome Sequencing et al. 2001; Mouse Genome Sequencing Consortium et al. 2002). L1 elements have accumulated from a single lineage of families since the origin of mammals and are mostly ancient remnants of past activity, with only the most recently evolved family active at a time (Smit et al. 1995; Furano 2000; Khan et al. 2005). There is however a number of differences in L1 content between mammalian species (supplementary table S1, Supplementary Material online). For example, the copy number of potentially active L1 progenitors is far larger in mouse ($>$ 3,000) than in human ($<$ 100), resulting in a higher rate of L1 transposition in the former (Ostertag and Kazazian 2001). Additionally, three distinct L1 subfamilies are concurrently active in the mouse genome (L1Md_A, Tf, and Gf), all of which entail further subsets. Some of these subsets have evolved in a strictly vertical manner (e.g., L1Md_A, Tf_III, and Gf_II), while others have invaded the genome following hybridization with a sister species (e.g., Tf_I, Tf_II, and Gf_I) (Rikke et al. 1995; Hardies et al. 2000; Goodier et al. 2001; Sookdeo et al. 2013). Conversely, the human genome has a single active family, Ta, comprised of two closely related subsets: Ta-0, which was mostly active $>$ 1.5 Ma; and Ta-1, which is responsible for the bulk of novel insertions in modern human (Boissinot et al. 2000, 2004; Sheen et al. 2000).

In contrast, nonmammalian vertebrates possess a much larger diversity of active LINEs represented by divergent families distributed across multiple clades (supplementary table S1, Supplementary Material online). For instance, the genome of the lizard *Anolis carolinensis* has five active clades (L1, L2, CR1, RTE, and R4), with both L1 and L2 containing $>$20 active and highly divergent families (Novick et al. 2009; Alföldi 2011). However, each specific LINE family in nonmammalian vertebrates tends to be represented by a small copy number ($<$ 100) of recently inserted elements (Furano et al. 2004; Novick et al. 2009; Chalopin et al. 2015). As in mammals, there are substantial differences in the diversity and abundance patterns among species. For example, the zebrafish (Furano et al. 2004) and *Xenopus tropicalis* (Hellsten et al. 2010) genomes harbor a great diversity of LINE elements similar to *A. carolinensis*, whereas the genomes of stickleback (Blass et al. 2012) and fugu contain much smaller numbers of families, though with numbers of recent clades similar to other fish (Duvernell et al. 2004; Chalopin et al. 2015).

It remains unclear why the abundance of LINE insertions differs so greatly among vertebrates. It has been proposed that this pattern reflects variation in fixation rates of polymorphic LINE insertions caused by differential fitness effects on the host organism (Furano et al. 2004; Novick et al. 2009). A presumed fitness dichotomy between mammals and nonmammal vertebrates has been hypothesized to result from differing rates of ectopic recombination among lineages, where crossing over of nonhomologous loci occurs and produces deleterious chromosomal rearrangements. Specifically, it has been proposed that mammalian genomes intrinsically have a reduced rate of ectopic recombination as an adaptation to an extreme accumulation of L1 elements (Furano et al. 2004). Indeed, the majority of L1 insertions appear to behave as neutral alleles in mammals (Boissinot et al. 2006), such that population-level frequency and chance of fixation depend solely on the stochastic process of genetic drift that is shaped by the host demographic history. However, L1 elements may not be fully neutral in mammals since longer elements are found at lower allele frequencies than shorter ones within human populations, suggesting an appreciable fitness cost related to TE length (Boissinot et al. 2006). This is consistent with the observation that longer L1 elements occupy genomic compartments with a lower recombination rate, which would be less likely to mediate ectopic recombination events, than shorter L1 elements (Boissinot et al. 2001; Song and Boissinot 2007). Conversely, the young age and low genome-wide abundance of LINEs in fish and reptiles has been interpreted as evidence for rapid turnover (Duvernell et al. 2004; Furano et al. 2004), similar to that observed in *Drosophila* (Charlesworth and Charlesworth 1983; Kaplan and Brookfield 1983; Langley et al. 1983). Under this scenario, most LINEs remain at low allele frequencies and rarely reach fixation due to the deleterious effect of novel insertions, which may be due to the ectopic recombination rate being intrinsically higher in these organisms than mammals.

Here, we performed a model-based demographic analysis for several vertebrate species to explore the relative effect of selection versus genetic drift on LINE polymorphisms (Ewing and Jensen 2016). Previous studies investigating this question used a limited number of TE insertions, mostly derived from the published reference genomes (Boissinot et al. 2006; Blass et al. 2012; Tollis and Boissinot 2013). However, the recent availability of resequenced whole-genome data at population-level sampling in threespine stickleback fish (*Gasterosteus aculeatus*), green anole (*A. carolinensis*), house mouse (*Mus musculus*), and human (*Homo sapiens*) now permits access to a more complete collection of polymorphic LINE insertions as well as genome-wide SNP data (1000 Genomes Project Consortium 2010; Chain et al. 2014; Harr et al. 2016; Ruggiero et al. 2017). We exploited these SNP resources to inform null demographic backgrounds that were then leveraged against the LINE data to infer the presence, direction, and magnitude of selection (fig. 1) (Williamson et al. 2005;

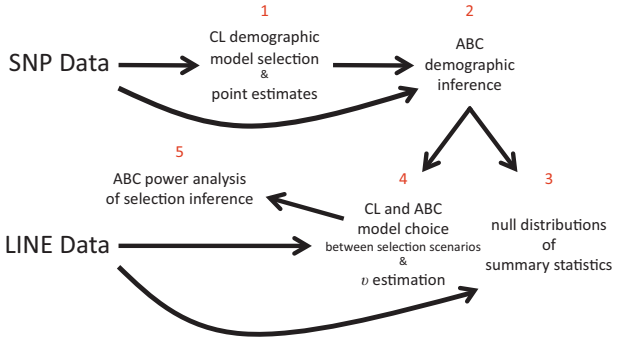

Fig. 1.—Flowchart of analyses. 1) Null demographic background is inferred from the whole-genome SNP data using CL optimization. This includes model selection between two-epoch expansion and three-epoch bottleneck-expansion via Akaike's relative weight. 2) ABC posterior distributions are estimated against the SNP data under the CL-chosen demographic model and with prior distributions informed by the CL point estimates. 3) Null distributions of summary statistics are simulated from the ABC posterior distributions and compared with the empirical whole-genome LINE data. 4) Selection is inferred from the LINE data using both of the statistical frameworks CL and ABC. This involves choosing between models of no selection, negative selection, and positive selection; demographic parameters are informed by the ABC posterior distributions and selection is approximated via the parameter υ, which converts to the standard exponential size change parameter r. 5) ABC leave-one-out cross-validation is performed to examine accuracy and bias in selection model choice and parameter estimation of υ.

Jensen et al. 2007). In particular, following recent efforts to elucidate selective effects by accounting for genome-wide heterogeneity in demographic parameter values (Aeschbacher et al. 2016; Roux et al. 2016; Rougemont et al. 2017; Rougeux et al. 2017), our methodology included a separate analysis that incorporates an additional demographic parameter to mimic the effect of selection within the context of demographic history.

Materials and Methods

Resequenced Whole-Genome Data

We analyzed one population each of stickleback, anole, and mouse, as well as three populations of human. For anole, mouse, and human, we focused on L1 elements because L1 is the clade that accounts for most of the genome size variation among vertebrates (Furano et al. 2004; Tollis and Boissinot 2012). However, this clade has not been recently active in stickleback, thus we instead collected stickleback LINE data for the Maui element, which belongs to the L2 clade. The published genome contains ~2,400 Maui copies and this family was shown to be polymorphic within stickleback populations, suggesting that it is currently amplifying in this species (Blass et al. 2012). Our total data set consisted of: six stickleback individuals from a river population in Europe (Chain et al. 2014); seven anole individuals (Ruggiero et al. 2017) assigned to the Gulf-Atlantic population as defined by Tollis and Boissinot (2012), which is the most widespread of the five distinct North American populations; ten mouse individuals belonging to a Northern Indian population, obtained from http://www.ebi.ac.uk/ena/data/view/PRJEB2176, last accessed April 28, 2018 (Harr et al. 2016); and 179 individuals across the human populations of Yoruba from Ibadan (YRI; N = 59), Han Chinese from Beijing combined with Japanese from Tokyo (CHB + JPT, henceforth CHJ; N = 60), and Utah residents of Central European ancestry (CEU; N = 60) (Stewart et al. 2011).

For stickleback, anole, and mouse, whole-genome data were processed following the procedure described in Ruggiero et al. (2017). To detect LINEs, we employed the program MELT (Mobile Element Detector Tool), which had been used to discover TE polymorphisms within the human genome (Sudmant et al. 2015). This program requires a library of TE consensus sequences to identify split reads, indicative of polymorphic insertions. We searched the stickleback genome utilizing a single Maui consensus since this family is highly homogenous in sequence (Blass et al. 2012), anole genome with the 20 consensus sequences described in Novick et al. (2009), and the mouse genome given the consensus of the three active families (Tf, Gf, L1Md_A). For human, data were retrieved from the 1000 Genomes Project pilot phase, with SNPs in VCF format at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/low_coverage/snps/, last accessed April 28, 2018, L1 insertion data listed in supplementary Table S1, Supplementary Material online of Stewart et al. (2011), and recombination hotspot location information in the file "hotspot_positions_b36.txt" located within the archive "1000G_LC_Pilot_genetic_map_b36_genotypes_10_2010.tar.gz" at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/supporting_data/recombination_hotspots/, last accessed April 28, 2018.

Processing Data into SFS Format

Polymorphic SNP and LINE insertion data for stickleback, anole, mouse, YRI, CHJ, and CEU were converted to the single-population folded site frequency spectrum (SFS), that is, not polarized against an ancestral outgroup for derived frequencies and instead based on minor allele frequencies. Importantly, information about monomorphic sites was not included as it is challenging to reliably collect such data for TEs, thus only the relative shape of the SFS (i.e., each allele frequency class bin was scaled to proportion of total SNPs rather than SNP count) was considered. As a result, SNP mutation rate as well as LINE transposition rate are largely irrelevant to downstream simulation and modeling efforts since only markers that are already polymorphic are included. For the SNP data, only biallelic positions were considered, with a minimum coverage of 10× read depth and minimum threshold of 10 haploid individuals, except for the human

populations where the minimum coverage was disregarded since these data sets had site frequencies inferred using a validated probabilistic framework on the low coverage (1×–3×) raw reads (Stewart et al. 2011). The number of SNPs was further reduced to minimize linkage disequilibrium. For the stickleback, anole, and mouse populations, SNPs were chosen sequentially along each chromosome and contig at a distance >10,000 sites from the preceding retained SNP. In the human data sets, since a linkage map was available, the first SNP was selected per length of sequence between HapMap-defined recombination hotspots; if no valid SNP was available within a particular sequence block, then the most adjacent SNP in the preceding hotspot zone was considered, if available. For LINE data, three independent SFS were calculated per population based on TE length: short LINEs were at most 1,200 bp; long LINEs were at most 6,000 bp and longer than short; and FL (full-length) LINEs were over 6,000 bp. Notably, some L1 families in anoles generally considered to be FL are shorter than 6 Kb and are thus included in the long category. For stickleback, since FL Maui elements are substantially shorter than 6 Kb, only short and long LINE SFS were constructed.

SFS files were created with the *Python* module *dadi* (Gutenkunst et al. 2009), which utilizes a sampling projection technique that entails considering all possible combinations of subsampling haploid individuals in order to accommodate missing data. A data projection of 10 haploid samples was deployed among data sets for comparative purposes. Additionally, a sampling projection that permits a more complete data matrix was applied as well: 12 haploid samples for stickleback; 14 haploid samples for anole; 20 haploid samples for mouse; and 50 haploid samples for each human population. In total, there were 12 observed SNP SFS, at two sampling projections per population of stickleback, mouse, anole, YRI, CHJ, and CEU, as well as 34 empirical LINE SFS, with three length conditions imposed for all combinations of sampling projections and populations (less two for no FL data in stickleback).

Inferring Null Demographic Model from SNP Data

Demographic inference was conducted under single-population instantaneous size change models to fit a null demographic background to the whole-genome SNP data using the program *fastsimcoal2.5221* (Excoffier et al. 2013) (fig. 1). Importantly, the results for this model fitting exercise are not the focus of this study, but rather the aim is to construct a valid null model that can reproduce the genome-wide SNP data. Hence, assumption violations arising from *a priori* misspecifications are tolerable as long as the genomic signal underlying the empirical SNPs is reasonably captured (i.e., the SNP-based SFS can be attained via simulation). To achieve probabilistic inference, *fastsimcoal2.5221* employs coalescent simulations to approximate the expected SFS given a set of parameter values under a demographic model, and Brent's conditional maximization algorithm on each parameter iteratively (ECM) to optimize the composite likelihood (CL) of the expected SFS against the observed data assuming a multinomial distribution. Every expected SFS was built from 100,000 coalescent simulations of genealogies, and a total of 10–40 ECM cycles were performed, with a single cycle comprising of all parameters individually optimized once and the number of cycles determined from a stopping criterion of 0.01 (i.e., the minimum relative difference in parameters between two cycles). The relative likelihoods of two demographic models were compared per observed SNP SFS: two-epoch expansion and three-epoch bottleneck-expansion. These models represent simplified yet commonly tested size change scenarios that encompass a variety of single-population histories due to the wide search ranges utilized (supplementary table S2, Supplementary Material online). Each model inference entailed 50 total independent optimization iterations, thus involving different initial parameter draws, to avoid local optima and in turn approach the global CL. In sum, across six populations, two sampling projections, and two models, there were $(6\times2\times2\times50)=1{,}200$ total executions of *fastsimcoal2.5221* here. Following optimization, model selection was conducted for every data set via Akaike's relative weight of evidence, which utilizes the individual run with the highest CL per model and is based on Akaike Information Criterion (AIC) scores (Excoffier et al. 2013). In cases where model selection contradicted between the two sampling projections, the favored model at the higher data projection was assumed for both in the interest of consistency.

Furthermore, we obtained additional point estimates as well as uncertainty measures from approximate Bayesian computation (ABC) posterior distributions, which were derived under the CL chosen model given prior distributions centered around the corresponding CL point estimates so as to optimize intensive ABC efforts (fig. 1 and supplementary table S3, Supplementary Material online). This ABC approach supplements the CL framework by offering an alternative model of statistical inference that operates under a different procedure and associated set of assumptions. One major benefit of ABC is that credibility intervals are easily procured simultaneously with point estimates, while CL bootstrapping, an analogous measure of uncertainty, would be prohibitively expensive computationally, requiring an additional $(100\times1{,}200)=120{,}000$ *fastsimcoal2.5221* replicates (assuming 100 bootstraps). Conversely, ABC posterior distributions are sensitive to prior distributions, whereas multiple independent replicates of the Markovian process elicited in *fastsimcoal2* for CL optimization should be more robust to user specification of search ranges. ABC simulations were facilitated with the *R* package *Multi-DICE* (Xue and Hickerson 2017), which allowed straightforward co-opting of *fastsimcoal2.5221* under its "FREQ" mode to populate multiple independent ABC reference tables for the separate populations and two data projections. To clarify, each reference table

consists of a set of simulations generated under the same model as well as data specification, and can be applied to an empirical data set for ABC inference. A total of 500,000 simulations were generated per reference table, the SFS were converted to relative frequencies (i.e., frequency classes in units of proportion from total SNPs rather than SNP count), and demographic inference was executed using the *abc R* package and eponymous function (Csilléry et al. 2012) under the simple rejection algorithm at a 0.003 tolerance threshold (leading to 1,500 accepted simulations).

Notably, although a three-population model exploiting information from joint site frequencies (i.e., shared polymorphisms, private alleles, and fixed differences) could have been employed for the human populations, the results would not have been relevant here within a comparative context since the other three species comprise of only a single population. Furthermore, a three-population model would require sites to be present across all three populations to construct the multipopulation SFS, which is problematic for the LINE data given allelic dropout and nonhomology coupled with a very low quantity of loci at several dozen per length category. Moreover, the additional parameterization and topological complexity involved would be a nontrivial increase in required computational resources. Although disregarding shared ancestry could possibly result in biased size change estimates, migration rates between these human populations have previously been established to be low (Gutenkunst et al. 2009; Gronau et al. 2011). Therefore, a single-population model is an appropriate simplifying assumption for constructing a valid null model to be utilized downstream for selection inference. Although incorporating more complex multipopulation models is worth exploring in future applications, it is beyond the scope of this study.

Testing Null Hypothesis on LINE Data with Simulated Summary Statistics

To evaluate deviation in the LINE data from the null demographic background, which would suggest the presence of selection in addendum to the whole-genome signal, we resimulated from each of the 12 empirical ABC posterior distributions to construct null distributions for a battery of summary statistics (fig. 1). To achieve this, the sets of parameter draws from each of the 1,500 ABC accepted simulations were exploited under the inferred demographic model to generate 1,500 new SFS simulations. These new SFS were simulated under the "SNP" simulation model in *fastsimcoal2.5221* according to the exact sampling specifications of each LINE SFS. This resulted in slower runs than the "FREQ" setting but better accounted for variance from number of LINE loci sampled, which was lower than the SNP sampling by orders of magnitude. The simulations were subsequently converted into separate summary statistics, including the standard population genetic summary statistic Tajima's *D* for the total set of LINE loci, multinomial-distribution CL scores, individual principal components (PCs), and individual SFS allele frequency classes scaled to relative frequency. To clarify, CL scores were calculated from the total product of the SFS bins, which is a reduction (given no monomorphic sites) of the CL equation deployed by *fastsimcoal2.5221* that assumes these represent independent probabilities. To obtain PCs, the 1,500 simulated SFS were entered into a principal component analysis (PCA), which was then leveraged against the simulations for transformation into PC vectors of the same size but with variation maximized. For each summary statistic, the 1,500 simulated values constituted an individual null distribution to which the corresponding observed data value was compared. Empirical values outside of the central 95% density were considered rejection of the null model (Bustamante et al. 2001; Thornton and Andolfatto 2006), thus implicitly supporting selection (with values departing from the central 50% density also highlighted). Importantly, while this utilizes a traditional null model test, it does not clearly indicate selection presence due to the multitude of summary statistics, nor explicitly qualify the type or quantify the magnitude of selection.

Demographic Models Testing Selection on LINE Data

As a complementary approach to detect putative selection within LINE data sets as well as infer directionality (i.e., purifying or positive) and magnitude, we tested an expanded set of three discrete demographic models against each LINE SFS: 1) the null model inferred from the according SNP data set without modification; 2) the null model inferred from the according SNP data set with an additional exponential growth parameter; and 3) the null model inferred from the according SNP data set with an additional exponential contraction parameter. To clarify, the first model represents a null hypothesis of no selection that is consistent with the genome-wide SNP background, whereas the exponential size change parameters act as a proxy for selective dynamics and are independent from the instantaneous population size change(s) in the null demographic background (Wright 2005). Specifically, exponential growth (i.e., model 2) mimics weak purifying selection given that both similarly affect the SFS with an increase in rare alleles (Nielsen 2005; Williamson et al. 2005), and likewise exponential contraction (i.e., model 3) approximates positive directional selection as both shape the SFS toward intermediate frequency bins (Nielsen 2005; Gattepaille et al. 2013). Exponential size change is parameterized by υ, the ratio of total effective population exponential size change scaled from the bigger to smaller size. To clarify, the υ CL search range/ABC prior distribution is identical between the growth and contraction models, yet its subsequent conversion to the standard exponential rate of size change parameter *r* reflects either growth or contraction with a negative or positive sign, respectively. To accomplish this conversion, exponential size change begun at an arbitrary time of 125,000 generations

into the past, which assures that the initiation of exponential size change occurred as the earliest historical event across all inferred null demographic backgrounds. Values for υ close to 1.000 converge to zero size change beyond the genome-wide null, consistent with LINE neutrality, whereas a large selective effect is expected to confer values substantially >1.000.

To perform this task, we again applied the CL framework in *fastsimcoal2.5221* following the same software settings as for the SNP data, as well as ABC inference using the *fastsimcoal2.5221* simulation machinery under the "SNP" setting to match empirical LINE sampling along with the *Multi-DICE* and *abc R* packages (fig. 1). Both approaches used the distribution $\upsilon \sim \ln U(1, 1000)$, as a search range and prior, respectively. For the CL framework, which included (6 populations$\times$2 data samplings$\times$3 selection models$\times$50 replicates) = 1,800 total optimization replicates, bounded search ranges were set for the null background demographic parameters (i.e., CL optimization may not surpass bounds, unlike for υ) by the corresponding SNP ABC posterior distribution 50% credibility intervals (CIs), that is, 25% and 75% quantile values, permitting uncertainty in these demographic parameters to inform selection parameter space optimization. As with the null demographic models, Akaike's relative weights were calculated from the highest AIC score per set of 50 replicates to choose among the three selection models. For ABC, 500,000 simulations were generated under every selection model per population/sampling combination, with uncertainty in the null demographic inference similarly integrated by exploiting the entire corresponding SNP-based posterior distributions as the prior. To clarify, the prior for each population/sampling combination was composed of a discrete uniform distribution over 1,500 ABC-retained vectors of parameter draws. The simple rejection algorithm assuming 1,500 accepted simulations was utilized to choose between the three selection models by combining the 1,500,000 total simulations into a single reference table (0.001 tolerance threshold), as well as estimate υ posterior distributions given solely the negative and positive selection model simulations, respectively (0.003 tolerance threshold).

Additionally, we performed ABC leave-one-out cross-validation, which involves iteratively extracting a single simulation with known true values to infer as a pseudoobserved data set (POD) against remaining simulations, to assess statistical power (fig. 1). This was conducted against every reference table for both model choice and υ parameter estimation under the identical protocol described for ABC inference, on 150 PODs consisting of 50 PODs per model for the former, and 100 PODs each for median and mode point estimates, respectively, in the latter. For model choice cross-validation, a confusion matrix, which describes accuracy and bias through counts of correct model identifications as well as misclassifications, and mean posterior probabilities, which entail calculating the mean model posterior distribution among all 50 PODs per true model, were produced. For cross-validation of υ parameter estimation, true values were compared against estimates through calculating Pearson's *r* correlation and root mean squared error (RMSE) across each set of 100 PODs. Notably, a comparable CL power analysis would be prohibitively expensive computationally in the same manner as aforementioned for CL bootstrapping.

Importantly, unlike the null distribution simulations, this strategy allows explicit testing between neutrality, purifying selection, and positive selection through model choice within a probabilistic framework, as well as quantification of selective intensity through υ estimates, for comparison between populations and LINE sizes.

Results

SFS Data

Resequencing data provided a genome-wide distribution of putatively unlinked SNPs (on the order of tens of thousands) and LINEs among populations, with the number of LINEs widely varying between data sets (table 1). Specifically, LINEs were at least an order of magnitude more numerous in mouse (53,674 in total across all length categories) than any other species ($\sim$1,000 insertions in stickleback and anole), whereas human populations contained only a few dozen LINEs per size type. Overlaying the SFS of SNPs with those of the LINEs, scaled to relative frequencies, suggests that LINEs of different populations and size specifications are subject to varying levels of selective pressure (fig. 2). However, further analysis beyond these mere observations is needed to statistically test the presence, direction, and intensity of selection.

Null Demographic Background Based on SNP Data

All populations except for stickleback supported the three-epoch bottleneck-expansion model based on Akaike's relative weights across CL inferences (fig. 3 and table 2). In mouse and YRI, this determination was more ambiguous due to a conflict of favored models between sampling projections, with the 10 haploids data set favoring the two-epoch expansion model in these cases. Nonetheless, the more complex three-epoch model was designated as the inferred demographic model in both populations given its Akaike's relative weight at the higher sampling projection, especially considering that support for the simpler two-epoch model was not overwhelming for either mouse or YRI ($\sim$ 0.6 higher in relative weight for both). Among populations, time point estimates consistently coincided either with postglacial or at least later Pleistocene activity, and size change magnitudes were generally fairly moderate, with most $<10\times$ for either expansion or contraction (fig. 3 and supplementary table S4, Supplementary Material online). Notably, stickleback, CHJ, and CEU experienced minor expansions of $\sim3\times$ or less, whereas larger expansions of at least $5\times$ were inferred in anole and mouse.

Table 1
Number of Whole-Genome SNPs and LINEs for SFS Construction

Population	Sampling Projection of Haploids	SNPs	Short LINEs	Long LINEs	FL LINEs
Stickleback	10	38,439	667	502	N/A
Stickleback	12	11,636	716	543	N/A
Anole	10	86,599	479	348	120
Anole	14	8,658	567	407	134
Mouse	10	192,291	23,138	6,536	7,861
Mouse	20	155,158	32,002	9,240	12,432
Human (YRI)	10	18,501	65	34	37
Human (YRI)	50	28,769	86	48	58
Human (CHJ)	10	22,579	43	20	23
Human (CHJ)	50	29,935	57	33	30
Human (CEU)	10	19,854	60	49	37
Human (CEU)	50	28,378	60	40	46

Across CL point estimates, most fell within the ABC central 95% density CIs, and many of these further were contained by the central 50% density CIs.

Null Distributions of Summary Statistics against LINE Data

Comparing null distributions, generated from the ABC posterior distributions of demographic parameters, to the mouse LINE data sets almost universally displayed signals of selection across the various summary statistic measures, three LINE lengths, and two sampling projections (fig. 4, table 3, and supplementary table S5, Supplementary Material online). In fact, the observed summary statistic calculated from the LINE data was well beyond the entire null distribution in many cases. Furthermore, there is a clear trend of increasing number of outlier empirical summary statistics as well as greater distance from the null distributions positively correlating with LINE size. In anole, there was apparent selection on short and long LINEs across both haploid samplings, but seemingly not so for the FL LINEs. Along with the fact that there is a higher number of outlier summary statistics and more distance from the null distributions for short LINEs versus long LINEs, there surprisingly seems to be an inverse relationship between selective effect and LINE length for anole. The stickleback results are more ambiguous, with selection seemingly present for the short LINEs but a lack of convincing evidence to reject the null hypothesis for the long LINEs. Conversely, the human populations were almost wholly compatible with the null demographic background, with the CEU long LINEs the only convincing case of selection (and mostly at the 10 haploids sampling projection). Notably, all outlier empirical values were more negative for Tajima's *D* and more positive for the singleton allele frequency class with respect to the according null distributions, except for the stickleback short LINE data sets where the opposite was true, suggesting purifying selection in the former cases and positive selection on the stickleback short LINEs. Importantly, there were instances when results varied nonnegligibly between the two sampling projections, especially in human populations where the disparity in number of haploids was greatest, hence the importance of having a common sampling of individuals for comparative purposes. Additionally, when the simulated distributions, which include PC values, were compared with corresponding SNP data sets, the empirical datapoint near-universally fell within the central 95% density and in fact was usually contained within the central 50% density as well (supplementary table S5, Supplementary Material online). This demonstrates that the null demographic model, based on simulations from the posterior, has a reasonable predictive fit to the empirical data.

Selection Modeling on LINE Data

Model-based inference of selection directionality and magnitude, based on model choice and parameter estimation, respectively, was overall largely compatible and complementary with the summary statistic null distribution results (fig. 5, table 4, and supplementary table S6, Supplementary Material online). Specifically, each of the mouse LINE data sets heavily supported the negative selection model based on Akaike's relative weights across CL inferences as well as ABC model posterior probabilities, with purifying selection exclusively favored (i.e., Akaike's relative weight or ABC model posterior probability of 1.000) in most cases. Additionally, υ estimates demonstrate a distinct positive correlation with LINE size across both sampling projections, indicating increasing deleterious effect with longer length. For anole, the surprising negative correlation between selection signal and LINE length implied by the null distributions is unequivocal here. The short LINE data sets highly favored negative selection, the long LINE data sets weakly supported negative selection, and the FL LINE data sets converged to the null demographic background of the SNP data. Estimates for υ further corroborate this relationship. Conversely, no selection was generally inferred for stickleback and human populations among all size filters and sampling projections based on Akaike's relative weights across CL

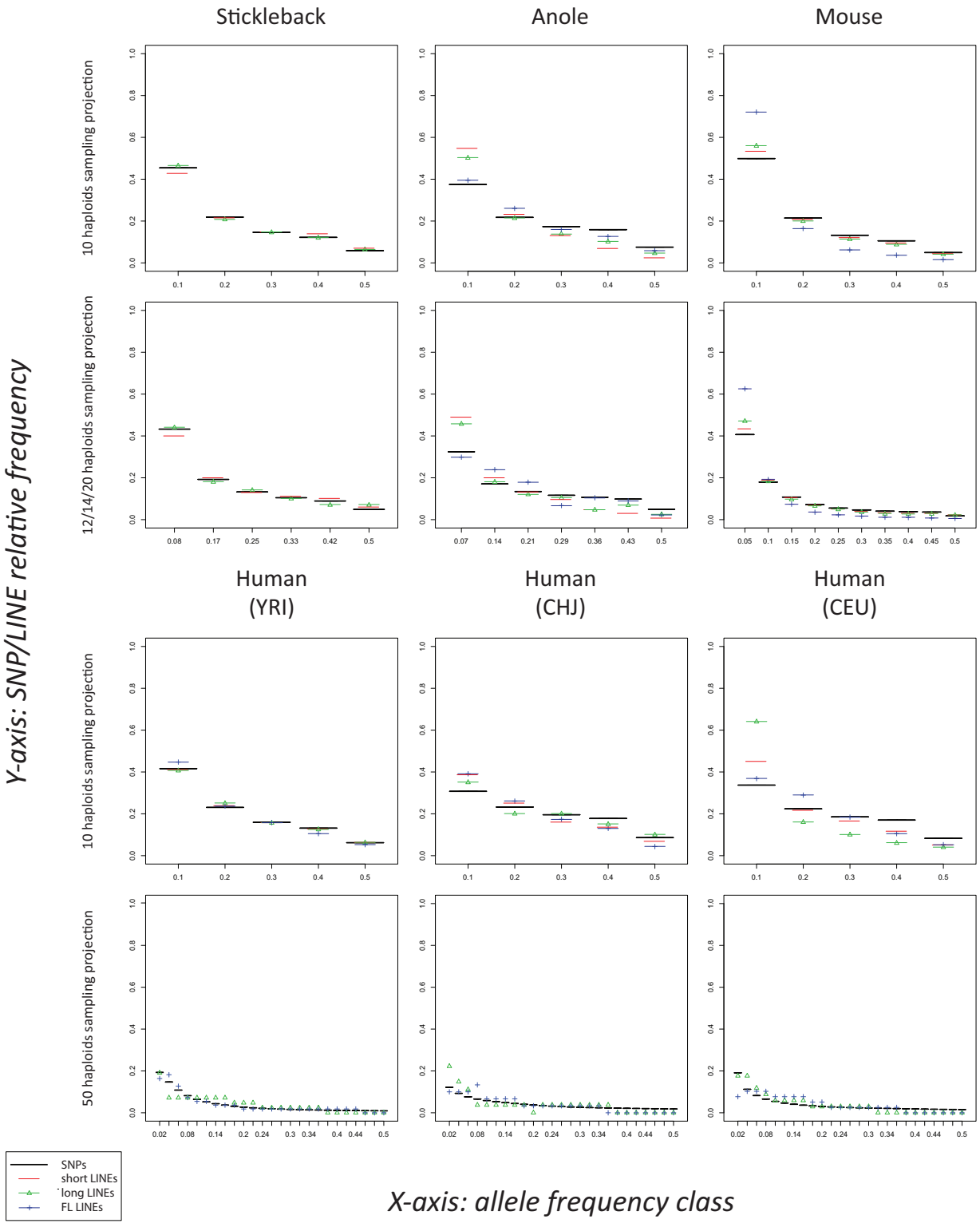

Fig. 2.—Observed SFS for SNP and LINE data. In the second row, 12 haploids refer to stickleback, 14 haploids refer to anole, and 20 haploids refer to mouse.

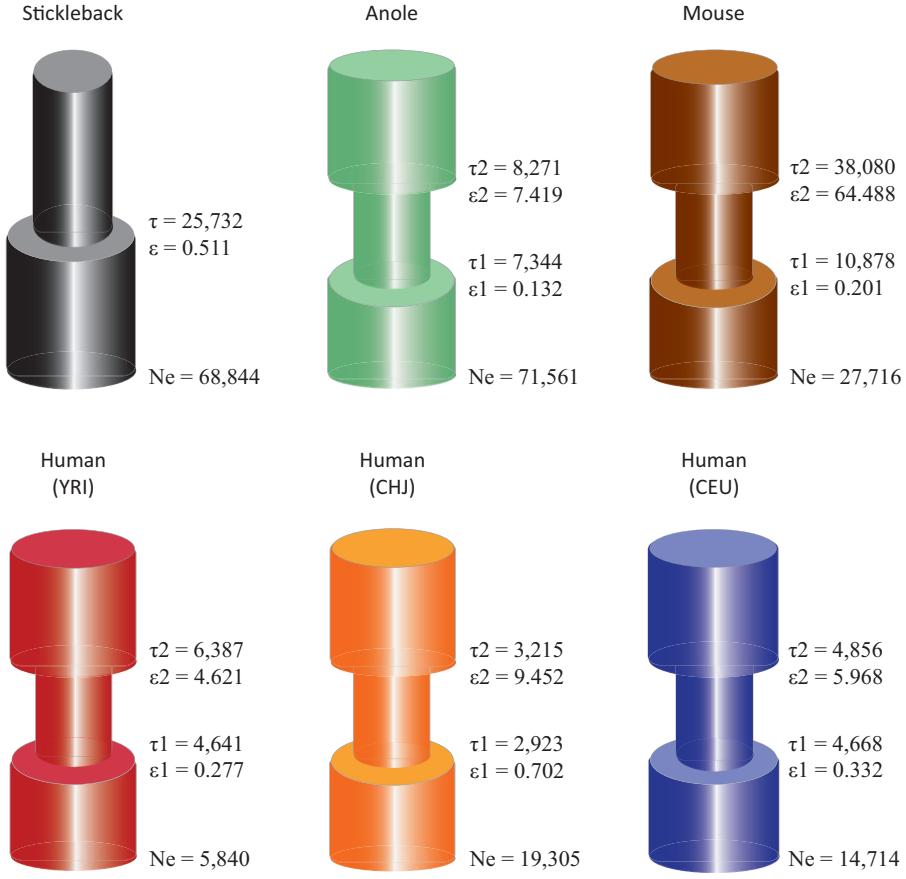

Fig. 3.—Null demographic background CL inference. The three-epoch bottleneck-expansion model was selected for all populations except for stickleback, which favored the two-epoch expansion model across both sampling projections. For comparative purposes, parameter estimates here are from the 10 haploids sampling projection. The symbols τ, ε, and Ne represent time in number of generations, proportion of instantaneous size change, and current-day effective haploid population size, respectively. The numerical suffixes delineate the demographic event (i.e., expansion or bottleneck) and are in reverse chronological order (e.g., $\tau 1 =$ time of the more recent expansion event). Note that ε is scaled to the effective population size immediately following the instantaneous change, such that $\varepsilon 1 \leq 1.000$ (i.e., expansion) and $\varepsilon 2 \geq 1.000$ (i.e., bottleneck) must be true.

Table 2
Null Demographic Background CL Model Selection via Akaike's Relative Weight

	Stickleback	Anole	Mouse
Selected Model	Two-epoch expansion	Three-epoch bottleneck-expansion	Three-epoch bottleneck-expansion
Akaike's relative weight (10 haploids)	0.852	1.000	0.214
Akaike's relative weight (higher sampling)	0.641 (12 haploids)	1.000 (14 haploids)	0.701 (20 haploids)
	Human (YRI)	**Human (CHJ)**	**Human (CEU)**
Selected Model	Three-epoch bottleneck-expansion	Three-epoch bottleneck-expansion	Three-epoch bottleneck-expansion
Akaike's relative weight (10 haploids)	0.182	1.000	1.000
Akaike's relative weight (50 haploids)	1.000	1.000	1.000

inferences, with three exceptions where purifying selection was supported: YRI short and FL LINEs at 50 haploids, though with low corresponding υ estimates; and CEU long LINEs at 10 haploids, which was near-exclusively in favor of negative selection with accordingly high υ estimates and corroborated by the ABC model posterior probability. There were also several

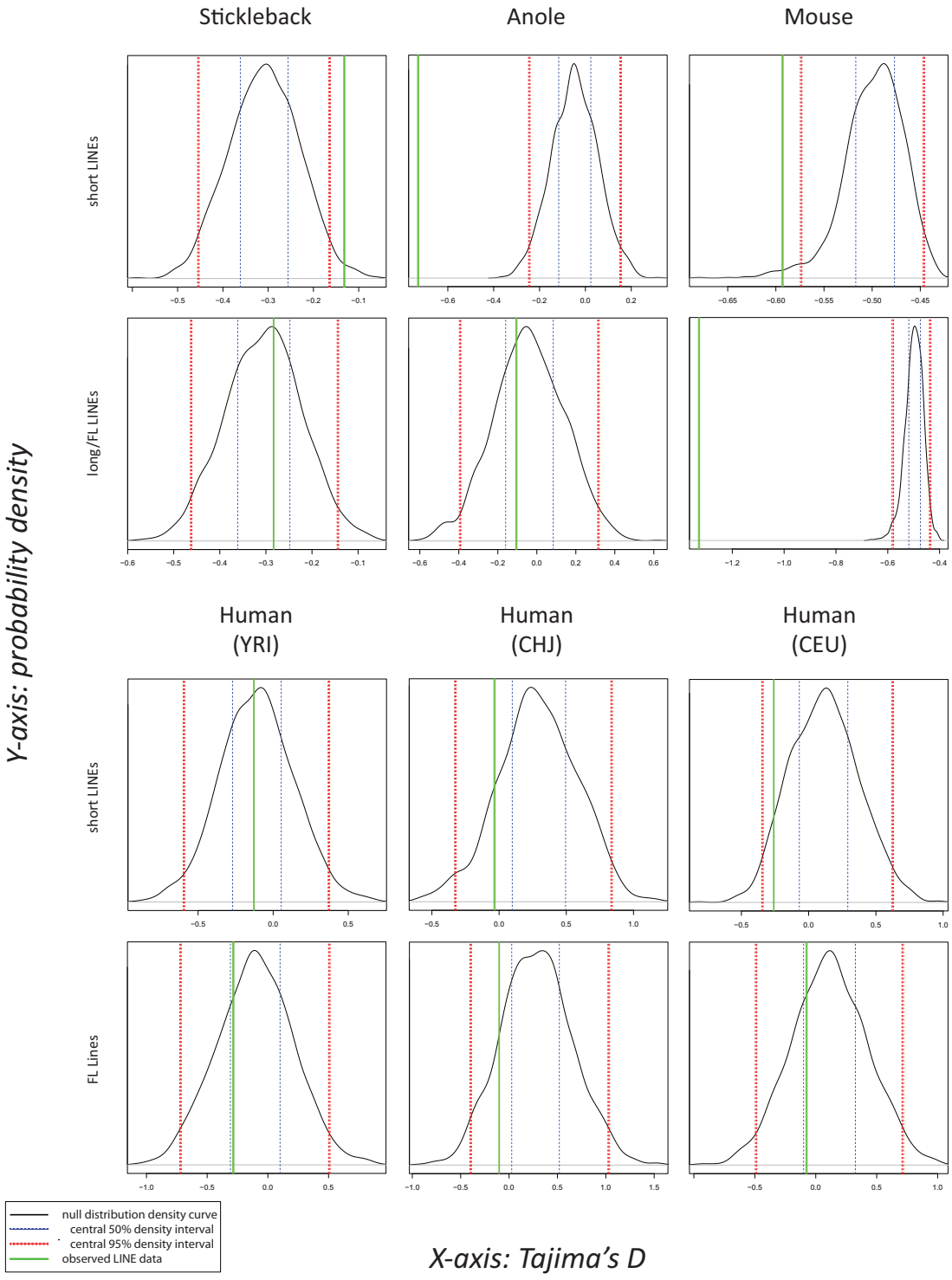

Fig. 4.—Simulated null distribution against empirical LINE data set for Tajima's *D*. Based on the 10 haploids sampling projection. In the second row, long LINEs refer to stickleback (since FL Maui elements fall under our size specification for long LINEs; see Materials and Methods) and FL LINEs refer to anole and mouse.

Table 3
Simulated Null Distributions against Empirical LINE Data for Select Summary Statistics

	Short LINEs			Long LINEs			FL LINEs		
	TE value	Null 2.5%	Null 97.5%	TE value	Null 2.5%	Null 97.5%	TE value	Null 2.5%	Null 97.5%
					Stickleback				
Tajima's *D*	**−0.132**	−0.454	−0.164	−0.283	−0.462	−0.144	N/A	N/A	N/A
CL score	**1.33E-04**	6.60E-05	1.23E-04	1.07E-04	6.46E-05	1.27E-04	N/A	N/A	N/A
PC1	−0.034	−0.043	0.046	0.007	−0.050	0.051	N/A	N/A	N/A
AFC1	0.427	0.426	0.502	0.464	0.419	0.506	N/A	N/A	N/A
					Anole				
Tajima's *D*	**−0.729**	−0.245	0.154	**−0.473**	−0.260	0.174	−0.105	−0.392	0.315
CL score	**2.61E-05**	9.93E-05	1.85E-04	**6.88E-05**	9.42E-05	1.90E-04	1.22E-04	5.95E-05	2.13E-04
PC1	**0.147**	−0.055	0.056	**0.100**	−0.060	0.061	−0.023	−0.100	0.099
AFC1	**0.546**	0.358	0.455	**0.501**	0.354	0.458	0.395	0.319	0.487
					Mouse				
Tajima's *D*	**−0.593**	−0.574	−0.447	**−0.691**	−0.584	−0.432	**−1.328**	−0.579	−0.436
CL score	5.62E-05	5.60E-05	7.46E-05	**4.53E-05**	5.42E-05	7.68E-05	**4.28E-06**	5.49E-05	7.62E-05
PC1	**0.029**	−0.013	0.020	**0.058**	−0.018	0.023	**0.241**	−0.017	0.021
AFC1	**0.533**	0.495	0.524	**0.559**	0.491	0.527	**0.721**	0.492	0.525
					Human (YRI)				
Tajima's *D*	−0.128	−0.594	0.370	−0.123	−0.762	0.601	−0.284	−0.717	0.506
CL score	1.26E-04	2.50E-05	2.15E-04	1.24E-04	0.00E+00	2.38E-04	9.27E-05	0.00E+00	2.29E-04
PC1	−0.009	−0.142	0.140	−0.015	−0.200	0.201	0.027	−0.179	0.185
AFC1	0.413	0.286	0.540	0.406	0.250	0.594	0.447	0.263	0.579
					Human (CHJ)				
Tajima's *D*	−0.034	−0.326	0.836	0.220	−0.504	1.077	−0.102	−0.394	1.025
CL score	1.43E-04	4.08E-05	2.80E-04	2.10E-04	0.00E+00	2.81E-04	1.01E-04	0.00E+00	2.80E-04
PC1	0.066	−0.154	0.173	0.043	−0.211	0.229	0.053	−0.206	0.214
AFC1	0.386	0.182	0.477	0.350	0.150	0.500	0.391	0.130	0.522
					Human (CEU)				
Tajima's *D*	−0.259	−0.344	0.623	**−0.936**	−0.420	0.662	−0.076	−0.489	0.710
CL score	9.48E-05	4.28E-05	2.53E-04	**2.46E-05**	3.49E-05	2.58E-04	1.09E-04	0.00E+00	2.62E-04
PC1	0.090	−0.143	0.134	**0.296**	−0.154	0.152	−0.014	−0.169	0.186
AFC1	0.450	0.241	0.483	**0.640**	0.240	0.500	0.368	0.211	0.526

NOTE.—All for 10 haploids sampling projection. Bold values are outside the 95% null distribution intervals. Underlined values are outside the 50% null distribution intervals but within the 95% intervals.

PC1, principal component 1; AFC1, allele frequency class 1 (i.e., singletons).

contradictory ABC results, most notably stickleback short LINEs supported positive selection at both samplings, and all but one of the CHJ data sets were consistent with selection.

ABC cross-validation confusion matrices and mean posterior probabilities suggest adequate power to differentiate the selection models for the nonhuman populations (supplementary table S7, Supplementary Material online). Specifically, the confusion matrices establish that false-positive rates are lower than false-negative rates, such that accuracy was highest when the true model was no selection coupled with near-absent misclassification between the negative and positive selection models. Model choice is much more problematic for the human populations though, where LINE sampling was drastically decreased (table 1). High error is especially apparent in the confusion matrices for the YRI and CHJ data sets under the 10 haploids sampling projection, where there was minimal inference of no selection. Moreover, bias toward purifying selection is showcased in the confusion matrices for the YRI and CHJ data sets at 50 haploids. Across confusion matrices for all organisms, samplings, and LINE sizes, negative and positive selection models were more likely to be misclassified under the no selection model than each other (except for YRI and CHJ at 10 haploids), while biased error toward purifying selection is exhibited when the true model included no selection (except for the human data sets with 10 haploids). Additionally, the correct mean posterior probability is almost always higher for the negative and positive selection models than the no selection model. Furthermore, there is evident power for υ parameter estimation in stickleback, mouse, several data sets of anole, and YRI given 50 haploids (though seemingly biased toward underestimation), whereas accurate υ estimation was lacking in the

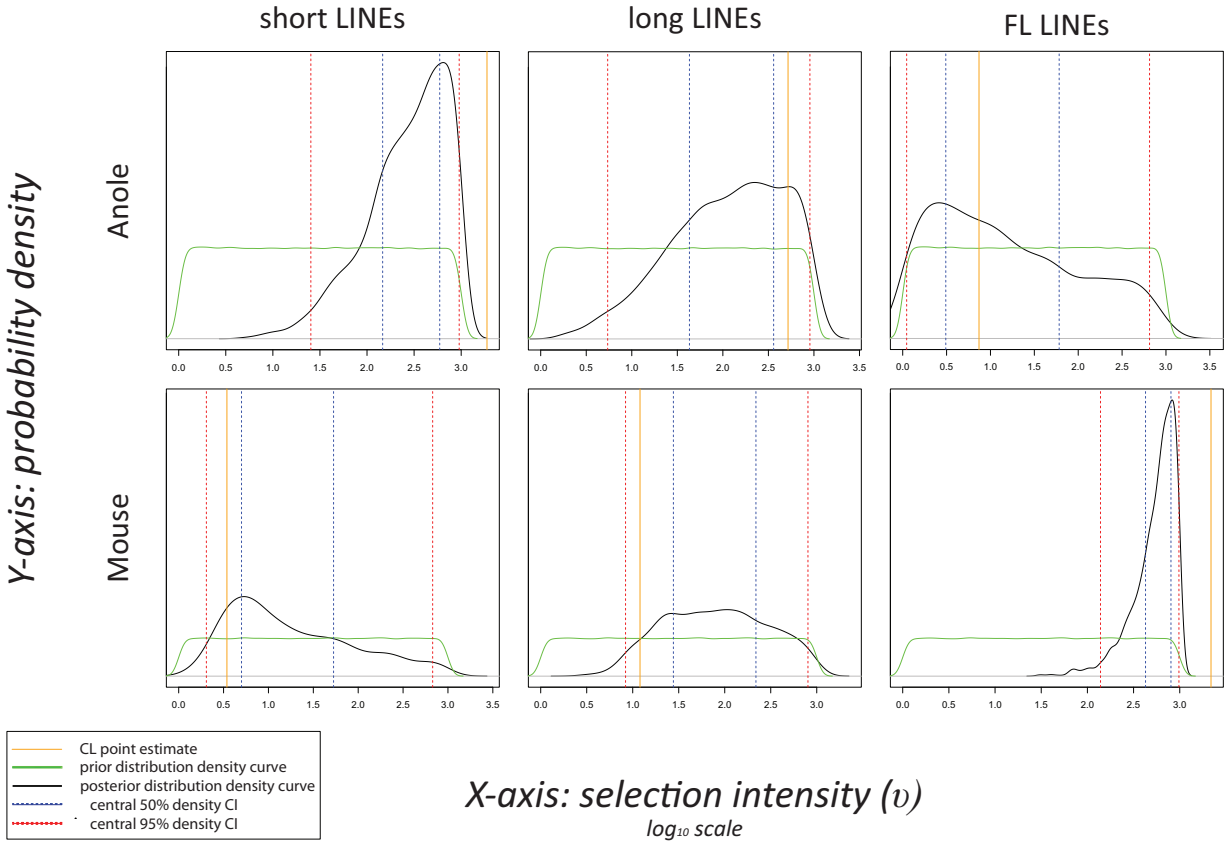

Fig. 5.—CL point estimate with ABC prior and posterior distributions of υ for anole and mouse. Given the negative selection model and 10 haploids sampling projection. Only anole and mouse are displayed here since these are the populations that demonstrated a clear signal of selection.

remaining cases (supplementary fig. S1 and table S7, Supplementary Material online). Overall, inferring υ is usually more difficult under purifying selection than the positive selection model.

Discussion

Using population resequencing data and demographic modeling of selection processes, we discovered LINE selection dynamics to vary greatly among four model vertebrate species that differ in total abundance of LINE insertions across the genome as well as number of active families. However, the comparative pattern of selection signal does not seem to be related to these differences nor evolutionary proximity between the species. For example, purifying selection against LINE polymorphisms is very strong within anole whereas there was no such evidence of negative selection for stickleback, despite their genomes likewise containing several sets of a few highly related copies. Similarly, LINEs were estimated to be much more deleterious in mouse than in human, which are both mammals with genomes dominated by an extremely large number of L1 elements.

Strong Purifying Selection against L1 Inserts in Mouse and Anole Lizard

We identified a strong signature of purifying selection against L1 in mouse, with stronger selection against long elements than short ones and FL insertions the most deleterious. This is consistent with previous analyses in human (Boissinot et al. 2006; Song and Boissinot 2007) and fruit fly (Petrov et al. 2003). Such a strong bias against longer elements may have three nonexclusive explanations. First, it could be due to the increased ability of longer elements to mediate ectopic recombination and thus cause deleterious chromosomal breaks (Dray and Gloor 1997), which is supported by experimental evidence that showed elements in mouse shorter than 1.2 Kb are unlikely to mediate ectopic recombination events (Cooper et al. 1998). Second, FL insertions are the source of further transposition, which in turn can be deleterious to the host. Third, a toxic effect of FL active insertions producing RNA or proteins deleterious to the host may be a contributing factor (Nuzhdin 1999). Recent analyses demonstrating that L1 activity in somatic tissue could have a deleterious effect with respect to aging or oncogenesis provide a mechanism for selection to act specifically against FL insertions (Faulkner and Garcia-Perez 2017).

Table 4
Selection Directionality CL Model Choice via Akaike's Relative Weight

	Short LINEs	Long LINEs	FL LINEs	Short LINEs	Long LINEs	FL LINEs	Short LINEs	Long LINEs	FL LINEs
		Stickleback			Anole			Mouse	
Positive Selection	0.288	0.218	N/A	0.000	0.140	0.207	0.000	0.000	0.000
No Selection	**0.522**	**0.576**	N/A	0.000	0.370	**0.535**	0.001	0.000	0.000
Negative Selection	0.190	0.206	N/A	**1.000**	**0.490**	0.257	**0.999**	**1.000**	**1.000**
		Human (YRI)			Human (CHJ)			Human (CEU)	
Positive Selection	0.219	0.223	0.213	0.210	0.213	0.193	0.182	0.000	0.196
No Selection	**0.571**	**0.568**	**0.576**	**0.572**	**0.575**	**0.526**	**0.498**	0.001	**0.535**
Negative Selection	0.210	0.209	0.212	0.219	0.212	0.281	0.320	**0.998**	0.270

NOTE.—All for 10 haploids sampling projection. Bold values are for the favored models.

Strong purifying selection on LINEs in anole is consistent with previous studies (Tollis and Boissinot 2013; Ruggiero et al. 2017), though the detection of stronger selection against short LINEs compared with longer ones as well as FL elements appearing neutral is surprising. This result may be due to anole FL LINEs being so deleterious that most are quickly eliminated from the population, and thus only very weakly deleterious or neutral FL LINEs segregate in populations (Nielsen 2005). Importantly, exponential growth as deployed in our model would not reproduce very strong negative selection signatures (Williamson et al. 2005), and thus would not account for an extremely deleterious effect. Counter-intuitively, a signature of purifying selection would thus remain readily detectable for short LINEs because of a lessened deleterious effect of those insertions compared with longer elements.

LINE Insertions Possibly Neutral in Human and Stickleback

Most of the analyses for human, across populations and L1 of different length, indicated neutral dynamics, which is surprising since it had been previously demonstrated that FL L1 are selected against, though truncated insertions behave like neutral alleles (Boissinot et al. 2006). A possible explanation for this discrepancy may reside in the type of insertions examined. The Boissinot et al. (2006) study focused on Ta-1, which is the most recently active subset of L1 and accounts for the vast majority of novel insertions (Boissinot et al. 2000; Brouha et al. 2003). Conversely, the present study applied to all types of L1 polymorphisms such as Ta-0, which amplified long before Ta-1 and thus has low present activity, yet still has segregating alleles that tend to be at higher population frequencies than Ta-1 polymorphisms (Boissinot et al. 2000). It is likely that Ta-0 deleterious insertions have already been eliminated and only nearly or fully neutral elements remain. Hence, the absence of a clear selection signature here is plausibly due to heterogeneity of insertion ages, with Ta-1 negatively selected in present populations and the more ancient Ta-0 having already gone through the sieve of purifying selection. Unfortunately, we are unable to distinguish Ta-1 from Ta-0 here, but this hypothesis is consistent with the results for YRI under CL inference and CHJ under ABC inference, which differed across the widely disparate sampling projections and accordingly temporal resolutions (Keinan and Clark 2012; Robinson et al. 2014; Xue and Hickerson 2015). Specifically, the older and more neutral Ta-0 polymorphisms may be predominant in the data sets for 10 haploids, whereas the newer and more deleterious Ta-1 insertions could have higher representation at 50 haploids. Surprisingly, the opposite pattern is observed for CEU, where there was strong evidence of purifying selection on long LINEs, yet only at the lower level sampling. In this case, the negative selection model was favored with high υ point estimates under both CL and ABC inference, including a near-exclusive Akaike's relative weight (i.e., approaching 1.0) and very high model posterior probability that far surpasses the according cross-validation mean posterior probability, as well as additional support from several null distribution outliers. This perhaps indicates idiosyncratic dynamics occurring specifically within the CEU population, on L1 of this particular length range, and/or at a given period of time, which may be a consequence of biological, environmental, and/or stochastic conditions. However, interpretation should be tempered since discriminatory power is limited here given the low sampling of LINEs. Nonetheless, this interesting result warrants further investigation.

A similar process to that proposed in human could explain the apparent neutrality of LINEs in stickleback. It is well known that TE amplification tends to occur in waves, where periods of intense amplification alternate with periods of low activity (Pascale et al. 1990; Furano 2000; Khan et al. 2005; Sookdeo et al. 2013). The fraction of segregating insertions under negative selection then depends on which part of the amplification wave a species is experiencing. For instance, a species experiencing a high rate of transposition (i.e., top of the wave) will incur many deleterious insertions, generating a stronger signal of negative selection, as is likely the case in anole and mouse. In contrast, a species presently undergoing a low rate

of transposition (i.e., bottom of the wave) could have the majority of segregating elements be older, generated when transposition was stronger and now having already passed prolonged purifying selection. Interestingly, our ABC inferences coupled with the null distribution outlier results support low to moderate positive selection on short Maui elements, which may further suggest that many of these remaining polymorphic insertions have been co-opted for functional benefit. As in human, stickleback may have these higher frequency insertions masking the effect of negative selection acting on a smaller number of recently inserted LINEs. This process would result in an underestimate of the fitness cost imposed by LINE activity in current populations since our model does not consider such temporal change in transposition rate, thus obscuring comparative analysis. This hypothesis is supported by the fact that the average divergence among Maui insertions in stickleback is substantially higher ($\sim$ 2.2%) than those for L1 elements in anole ($<$ 1% for 16 of the 20 L1 families) or mouse ($\sim$ 0.3–1.3%), indicating most insertions in stickleback are older (Blass et al. 2012). Another important factor is the use of Maui, which is a family representative of the L2 clade and was used instead of L1 because of the extremely small numbers of L1 insertions in the stickleback genome (Blass et al. 2012). Although L1 and L2 are believed to replicate by similar biochemical processes (Sugano et al. 2006), L2 seems less efficient in producing FL progenitors and thus could be intrinsically less deleterious than L1 (Ruggiero et al. 2017). However, it has also been demonstrated in lizard that the frequency distributions of L1 and L2 polymorphic insertions were virtually indistinguishable, suggesting a similar fitness cost (Ruggiero et al. 2017). The effect of LINEs in fish therefore requires further investigation, particularly in a model species exhibiting a large number of recently active families, such as the zebrafish.

Importantly, our modeling effort may be considered more conservative in estimating the presence of selection since the uncertainty in demographic inference from the SNP data was incorporated into the simulations testing for selection. Therefore, demography alone was allowed to explain a greater variance in the observed LINE SFS, possibly obfuscating an effect from selection. Indeed, this is illustrated in our ABC cross-validation of model selection, where false-negative rates are consistently greater than false-positive rates. Relatedly, the number of LINEs was very low in the human populations, as well as generally so across many of the LINE data sets in contrast to the corresponding SNP data sets, which may likewise result in the modeling being too permissive of genetic drift solely effecting SFS differences between the SNP and LINE data. The detriment of low LINE sampling on accurate detection of selection is likewise showcased in the ABC cross-validation, specifically dramatically decreased disambiguation for the human data sets, which was even further displayed among human data sets with decreasing numbers of LINEs. As a result, our model as well as the data are likely to be less sensitive to capturing weaker selective scenarios, and thus cannot definitively reject such occurrences. Hence, further exploration with more comprehensive TE data sets would help elucidate these complex dynamics within human and stickleback.

Why Is Selection against LINE Insertions Stronger in Certain Organisms over Others?

We found no evidence for purifying selection against LINEs in stickleback despite support for selection in anole, the other nonmammalian vertebrate; truncated elements in anole impose a high fitness cost on their host even though short insertions have been found to be neutral in other species; and purifying selection against L1 is much stronger in mouse than in human populations, where we did not detect a clear signature for selection in most cases. We propose three broad categories of explanations to generally account for these species-specific observations.

First, the deleterious effect of LINE insertions may be intrinsically stronger in some species than others. This hypothesis was originally proposed to account for the higher diversity and lower abundance of L1 in zebrafish compared with mammals (Furano et al. 2004), and was eventually extended to other nonmammalian vertebrates such as the green anole (Novick et al. 2009). Furano et al. (2004) further proposed that regulatory mechanisms evolved in mammals to prevent ectopic recombination, thereby reducing the deleterious impact of L1 insertions. This would have rendered mammalian genomes more tolerant to the accumulation of L1 insertions, resulting in the extremely high copy numbers for L1 that are found in those genomes. However, studies have shown that even within mammals, there is wide variation in recombination rates (Jensen-Seaman et al. 2004; Winckler 2005). In contrast, with ectopic recombination being more frequent in zebrafish and anoles, LINE insertions would be more deleterious and rarely reach fixation, hence the large number of insertions segregating in populations yet small number of insertions accumulating in those genomes (Ruggiero et al. 2017). Consistent with this notion, we detected strong purifying selection in anole and especially so for truncated elements, implying that all elements regardless of size could be capable of mediating chromosomal rearrangements through ectopic recombination within this lizard species. In contrast, short L1 insertions as well as Alu elements, which are also shorter in length, have been shown to behave like neutral alleles in human (Boissinot et al. 2006; Cordaux et al. 2006) and not be restricted to low recombining regions of the genomes to the same extent as long elements (Boissinot et al. 2001; Myers et al. 2005; Song and Boissinot 2007).

Second, the number of insertions present in a genome can be a significant factor affecting selection intensity against novel insertions. Specifically, a TE family becomes deleterious when it reaches a certain copy number threshold, creating a

positive feedback loop where the deleterious impact of a family escalates alongside its copy number (Montgomery et al. 1987; Petrov et al. 2003, 2011). This scenario could contribute to the differences reported here between mouse and human given that the number of polymorphic insertions is much larger in the former. A possible explanation for this frequency-dependence could be that a larger number of segregating insertions throughout the genome increases the probability of ectopic recombination, which is also believed to be further exacerbated when insertions are in the heterozygous state (Montgomery et al. 1991). Therefore, when insertions are numerous across the genome but at low allele frequencies (i.e., greater chance to be heterozygous), they are more likely to be deleterious. Moreover, the total number of insertions in a population is directly related to the rate of transposition as well, which may be intrinsically higher in some organisms than others. Periods of high activity often correlate with the acquisition of novel features by LINE elements, allowing these elements to bypass the host repressive machinery and thereby reducing host fitness. This process is well documented in human where the acquisition of novel promoter sequences allows L1 to evade the transcriptional repression of KRAB zinc-finger proteins (Khan et al. 2005; Jacobs et al. 2014). It is likely that different rates of transposition occur in the model systems analyzed here, as implied by the observation that the relative proportion of genetic defect caused by retrotransposons is larger in mouse than human (Ostertag and Kazazian 2001).

Third, differences in demographic history could possibly account for host-specific LINE characteristics. Notably, anole and mouse experienced expansions of much greater magnitude than stickleback and human, and the signature of population expansion is an excess of low frequency SNPs, thus an increase in heterozygosity (Nielsen 2005). Assuming heterozygosity increases the rate of ectopic recombination (Montgomery et al. 1991), then higher ectopic recombination rates are expected in populations with large expansions, thus yielding more negative selection on TEs. Furthermore, the greater expansions may imply larger recent population sizes, where selection is more effective at eliminating deleterious insertions, whereas genetic drift is more prominent in smaller populations and thus may permit slightly deleterious insertions to persist at higher allele frequencies (Lynch and Conery 2003; Blass et al. 2012; Tollis and Boissinot 2013; Ruggiero et al. 2017). Importantly, although it has been reported from a number of models that smaller populations yield a lower efficacy for selection, the more subtle effect of population expansion has not received the same scrutiny. An effect of complex demography on the intensity of selection against LINEs is an intriguing hypothesis and will require further experiments, such as comparing the fitness cost of LINEs in multiple populations of anole and mouse with different demographic histories.

Supplementary Material

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

Acknowledgments

We would like to thank Emmeline Ma for assistance in designing the model diagram figures; Andrew D. Kern for thoughtful comments on the manuscript and lab support; Edward A. Myers and Ivan Prates for providing computational resources; and Ana C. Carnaval and Brenna M. Henn for thoughtful comments on the manuscript. We also thank three anonymous reviewers for their excellent comments and suggestions. This work was supported by grants from National Institutes of Health (1R15GM096267-01 to S.B. and M.J.H.); National Science Foundation (DEB-1253710 to M.J.H., DEB-1343578 to A.C.C., M.J.H., and Kyle C. McDonald); FAPESP (BIOTA, 2013/50297-0 to A.C.C., M.J.H., and K.C.M.); and NASA through the Dimensions of Biodiversity Program. This work would not have been possible without help from the City University of New York High Performance Computing Center, with support from the National Science Foundation (CNS-0855217 and CNS-0958379).

Literature Cited

1000 Genomes Project Consortium. 2010. A map of human genome variation from population scale sequencing. Nature 467:1061–1073.
Aeschbacher S, Selby JP, Willis JH, Coop G. 2017. Population-genomic inference of the strength and timing of selection against gene flow. Proc Natl Acad Sci U S A. 114(27):7061–7066.
Alföldi J. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477(7366):587–591.
Blass E, Bell M, Boissinot S. 2012. Accumulation and rapid decay of non-LTR retrotransposons in the genome of the three-spine stickleback. Genome Biol Evol. 4(5):687–702.
Boissinot S, Chevret P, Furano AV. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol. 17(6):915–928.
Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. 2006. Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci U S A. 103(25):9590–9594.
Boissinot S, Entezam A, Furano AV. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol. 18(6):926–935.
Boissinot S, Entezam A, Young L, Munson PJ, Furano AV. 2004. The insertional history of an active family of L1 retrotransposons in humans. Genome Res. 14(7):1221–1231.
Brouha B, et al. 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A. 100(9):5280–5285.
Burns KH, Boeke JD. 2012. Human transposon tectonics. Cell 149(4):740–752.
Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. Genetics 159(4):1779–1788.
Chain FJJ, et al. 2014. Extensive copy-number variation of young genes across stickleback populations. PLoS Genet. 10(12):e1004830.
Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol. 7(2):567–580.

Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. Genet Res. (Camb) 42(01):1–27.
Cooper DM, Schimenti KJ, Schimenti JC. 1998. Factors affecting ectopic gene conversion in mice. Mamm Genome 9(5):355–360.
Cordaux R, Lee J, Dinoso L, Batzer MA. 2006. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. Gene 373:138–144.
Cost GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. EMBO J. 21(21):5899–5910.
Csilléry K, François O, Blum MGB. 2012. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 3(3):475–479.
Dray T, Gloor GB. 1997. Homology requirements for targeting heterologous sequences during P-induced gap repair in *Drosophila melanogaster*. Genetics 147(2):689–699.
Duvernell DD, Pryor SR, Adams SM. 2004. Teleost fish genomes contain a diverse array of L1 retrotransposon lineages that exhibit a low copy number and high rate of turnover. J Mol Evol. 59(3):298–308.
Ewing GB, Jensen JD. 2016. The consequences of not accounting for background selection in demographic inference. Mol Ecol. 25(1):135–141.
Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9(10):e1003905.
Faulkner GJ, Garcia-Perez JL. 2017. L1 mosaicism in mammals: extent, effects, and evolution. Trends Genet. 33(11):802–816.
Furano AV. 2000. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog Nucleic Acid Res Mol Biol. 64:255–294.
Furano AV, Duvernell DD, Boissinot S. 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet. 20(1):9–14.
Gattepaille LM, Jakobsson M, Blum MGB. 2013. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. Heredity (Edinb) 110(5):409–419.
Goodier JL, Ostertag EM, Du K, Kazazian HHJr. 2001. A Novel Active L1 Retrotransposon Subfamily in the Mouse. Genome Res 11(10):1677–1685.
Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 43(10):1031–1034.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5(10):e1000695.
Hardies SC, et al. 2000. LINE-1 (L1) lineages in the mouse. Mol Biol Evol. 17(4):616–628.
Harr B, et al. 2016. Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. Sci Data 3:160075.
Hellsten U, et al. 2010. The genome of the Western clawed frog *Xenopus tropicalis*. Science 328(5978):633–636.
International Human Genome Sequencing C, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Jacobs FMJ, et al. 2014. An evolutionary arms race between KRAB zinc finger genes 91/93 and SVA/L1 retrotransposons. Nature 516(7530):242–245.
Jensen JD, Thornton KR, Bustamante CD, Aquadro CF. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics 176(4):2371–2379.
Jensen-Seaman MI, et al. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. 14(4):528–538.
Kapitonov VV, Tempel S, Jurka J. 2009. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene 448(2):207–213.
Kaplan NL, Brookfield JFY. 1983. Transposable elements in Mendelian populations. III. Statistical results. Genetics 104:485–495. http://www.genetics.org/cgi/content/abstract/104/3/485.
Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336(6082):740–743.
Khan H, Smit A, Boissinot S. 2005. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16(1):78–87.
Langley CH, Brookfield JFY, Kaplan N. 1983. Transposable elements in Mendelian populations. I. A theory. Genetics 104(3):457–471.
Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72(4):595–605.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302(5649):1401–1404.
Mita P, Boeke JD. 2016. How retrotransposons shape genome regulation. Curr Opin Genet Dev. 37:90–100.
Montgomery E, Charlesworth B, Langley CH. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. Genet Res. (Camb) 49(01):31–41.
Montgomery EA, Huang SM, Langley CH, Judd BH. 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. Genetics 129(4):1085–1098.
Mouse Genome Sequencing Consortium, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310(5746):321–324.
Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39(1):197–218.
Novick PA, Basta H, Floumanhaft M, McClure MA, Boissinot S. 2009. The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard *Anolis carolinensis* shows more similarity to fish than mammals. Mol Biol Evol. 26(8):1811–1822.
Nuzhdin SV. 1999. Sure facts, speculations, and open questions about the evolution of transposable element copy number. Genetica 107:129–137.
Ostertag EM, Kazazian HH Jr. 2001. Biology of Mammalian L1 retrotransposons. Annu Rev Genet. 35(1):501–538.
Pascale E, Valle E, Furano AV. 1990. Amplification of an ancestral mammalian L1 family of long interspersed repeated DNA occurred just before the murine radiation. Proc Natl Acad Sci U S A. 87(23):9481–9485.
Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol. 20(6):880–892.
Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, Gonzalez J. 2011. Population genomics of transposable elements in *Drosophila melanogaster*. Mol Biol Evol. 28(5):1633–1644.
Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 46(1):21–42.
Rikke BA, Zhao Y, Daggett LP, Reyes R, Hardies SC. 1995. Mus spretus LINE-1 sequences detected in the Mus musculus inbred strain C57BL/6J using LINE-1 DNA probes. Genetics 139(2):901–906.
Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN. 2014. Sampling strategies for frequency spectrum-based population genomic inference. BMC Evol Biol. 14(1):254.
Rougemont Q, et al. 2017. Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and non-parasitic lamprey ecotypes. Mol Ecol. 26(1):142–162.

Rougeux C, Bernatchez L, Gagnaire P-A. 2017. Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish species pairs (*Coregonus clupeaformis*). Genome Biol Evol. 9(8):2057–2074.
Roux C, et al. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS Biol. 14(12):e2000234.
Ruggiero RP, Bourgeois Y, Boissinot S. 2017. LINE insertion polymorphisms are abundant but at low frequencies across populations of *Anolis carolinensis*. Front Genet. 8:44.
Sheen F, et al. 2000. Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res. 10(10):1496–1508.
Smit AFA, Tóth G, Riggs AD, Jurka J. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol. 246(3):401–417.
Song M, Boissinot S. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene 390(1-2):206–213.
Sookdeo A, Hepp CM, McClure MA, Boissinot S. 2013. Revisiting the evolution of mouse LINE-1 in the genomic era. Mob DNA 4(1):3.
Stewart C, et al. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. 7(8):e1002236.
Sudmant PH, et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526(7571):75–81.
Sugano T, Kajikawa M, Okada N. 2006. Isolation and characterization of retrotransposition-competent LINEs from zebrafish. Gene 365:74–82.
Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172(3):1607–1619.
Tollis M, Boissinot S. 2012. The evolutionary dynamics of transposable elements in eukaryote genomes. In: Garrido-Ramos MA editor, Repetitive DNA. Basel, Switzerland: Karger Publishers. p. 68–91.
Tollis M, Boissinot S. 2013. Lizards and LINEs: selection and demography affect the fate of L1 retrotransposons in the genome of the green anole (*Anolis carolinensis*). Genome Biol Evol. 5(9):1754–1768.
Warren IA, et al. 2015. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. Chromosom Res. 23(3):505–531.
Williamson SH, et al. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 102(22):7882–7887.
Winckler W. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308(5718):107–111.
Wright SI. 2005. The effects of artificial selection on the maize genome. Science 308(5726):1310–1314.
Xue AT, Hickerson MJ. 2015. The aggregate site frequency spectrum for comparative population genomic inference. Mol Ecol. 24(24):6223–6240.
Xue AT, Hickerson MJ. 2017. Multi-DICE: r package for comparative population genomic inference under hierarchical co-demographic models of independent single-population size changes. Mol Ecol Resour. 17(6):e212.

Associate editor: Josefa Gonzalez