The Impact of Linked Selection in Chimpanzees: A Comparative Study

Susanne P. Pfeifer[1,2,3,*] and Jeffrey D. Jensen[1,2,3]

[1]School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

[3]School of Life Sciences, Arizona State University (ASU), Tempe, Arizona

*Corresponding author: E-mail: susanne.pfeifer@epfl.ch.

Accepted: September 24, 2016

Abstract

Levels of nucleotide diversity vary greatly across the genomes of most species owing to multiple factors. These include variation in the underlying mutation rates, as well as the effects of both direct and linked selection. Fundamental to interpreting the relative importance of these forces is the common observation of a strong positive correlation between nucleotide diversity and recombination rate. While indeed observed in humans, the interpretation of this pattern has been difficult in the absence of high-quality polymorphism data and recombination maps in closely related species. Here, we characterize genetic features driving nucleotide diversity in Western chimpanzees using a recently generated whole genome polymorphism data set. Our results suggest that recombination rate is the primary predictor of nucleotide variation with a strongly positive correlation. In addition, telomeric distance, regional GC-content, and regional CpG-island content are strongly negatively correlated with variation. These results are compared with humans, with both similarities and differences interpreted in the light of the estimated effective population sizes of the two species as well as their strongly differing recent demographic histories.

Key words: nucleotide diversity, selection, chimpanzee.

The Pervasive Relationship between Recombination and Variation

The correlation between nucleotide diversity and recombination rate is one of the most prevalent patterns in population genetics, and has been broadly interpreted as evidence for the strong effects of linked selection (see review of Cutter and Payseur 2013). The impact on linked neutral variation with negatively selected sites (termed Background Selection (BGS)), and with positively selected sites (termed Recurrent Hitchhiking (RHH)), is expected to be stronger in regions of low recombination (Begun and Aquadro 1992; Charlesworth et al. 1993). As such, both recurrent positive and recurrent purifying selection will serve to produce this observed relationship. Although both processes are likely at play (see Hudson 1994), and strong arguments have been made for the relative importance of one over the other in particular organisms, the pattern itself has been demonstrated to be remarkably pervasive – having been observed across mammals (e.g., Nachman 1997; Lohmueller et al. 2011), birds (e.g., Rao et al. 2011), insects (e.g., Begun and Aquadro 1992; Stump et al. 2005), fungi (e.g., Cutter and Moses 2011), plants (e.g., Dvorák et al. 1998), and viruses (e.g., Renzette et al. 2016). Although the observation is open to interpretation, an undeniable strength of BGS-based arguments is the fact that there is a far greater proportion of newly arising deleterious mutations compared to newly arising beneficial mutations across the genome, a notion already well appreciated in the early literature of the field (Timofeeff-Ressovsky 1940; Muller 1949, 1950; and see review of Bank et al. 2014). Thus, the selective removal of such mutations is likely a very common process.

Given the wealth of genomic data and an inherent interest in the potential effects of segregating deleterious mutations, humans have been a centrally important organism of study in this area. Perhaps most relatedly, Hellmann et al. (2005) found recombination rate to be the best predictor of human diversity levels, as well as a strong predictor of human-chimpanzee divergence. Exploring scenarios of RHH, BGS, as well as the possibility of a mutagenic effect of recombination, the authors were unable to well-discern between these models, largely

owing to insufficient polymorphism data and recombination rate estimates in their outgroup (chimpanzees). Further, though simulations suggested a slightly better fit under models of background selection, no predictive correlation was observed in humans between diversity and gene content, an observation somewhat at odds with a BGS-based explanation.

The subsequent decade has witnessed important advances in our understanding of chimpanzee genomics, making it possible to revisit these important results in greater depth. Most previous sub-genomic comparative work has suggested that humans harbor two- to four-fold lower levels of intra-species diversity (Kaessmann et al. 1999; Deinard and Kidd 1999; Jensen-Seaman et al. 2001; Satta 2001; Stone et al. 2002; Chimpanzee Sequencing and Analysis Consortium 2005)—with the observation generally being interpreted as a somewhat larger long-term effective population size in chimpanzees (Deinard and Kidd 1999; Kaessmann et al. 2001). However, these analyses have been based on a specific set of genes or genomic locations, resulting in ascertainment concerns.

With the availability of whole genome polymorphism data from the PanMap project, and the development of a fine-scale recombination map (Auton et al. 2012), we here visit the question of the pervasiveness of linked selection in the chimpanzee genome on a large scale while avoiding such ascertainment issues. Through a whole-genome analysis of 10 Western chimpanzees, a number of notable observations emerge. First and foremost, recombination rate is a strong predictor of variation at putatively neutral sites across the genome as is exon content.

In order to characterize this pattern and describe underlying predictors of variation in chimpanzees, each chromosome was divided into continuous windows of 1Mb (fig. 1), and the nucleotide diversity π within a window was compared against multiple genomic features, many of which are hypothesized to be either directly or indirectly linked with DNA damage or repair mechanisms (i.e., recombination rate, regional GC-content, gene density, CpG-island density, simple repeats content, distance to the centromere/telomere (Fryxell and Moon 2005; Elango et al. 2008; Tyekucheva et al. 2008; Chen et al. 2010)). Further, to avoid variance generated by differing levels of selective constraint, analyses were focused on putatively neutral regions.

Nucleotide Diversity between Autosomes in Chimpanzees

The autosomes in chimpanzees exhibit similar levels of nucleotide diversity (fig. 2)—with all autosomes being in the range of one standard deviation of the genome-wide average ($\pi = 6.9 \times 10^{-4}$). The contribution of different genomic factors influencing levels of variation was assessed using multiple linear regression as well as multiple logarithmic regression (see

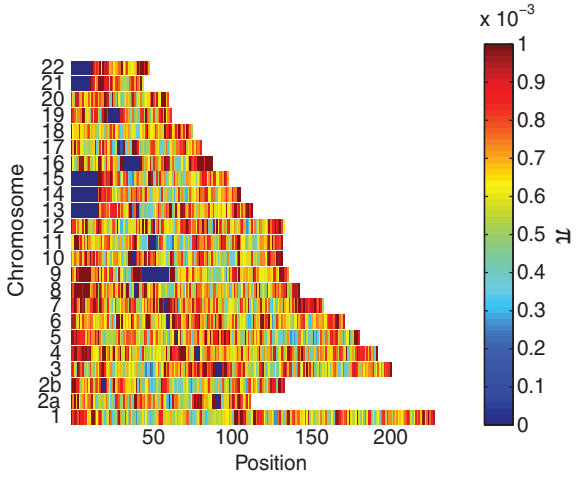

Fig. 1.—Nucleotide diversity levels π in Western chimpanzees across chromosomes (estimated in 1 Mb windows with at least 80% accessibility after filtering). $\pi = 0$ (dark blue): no data available.

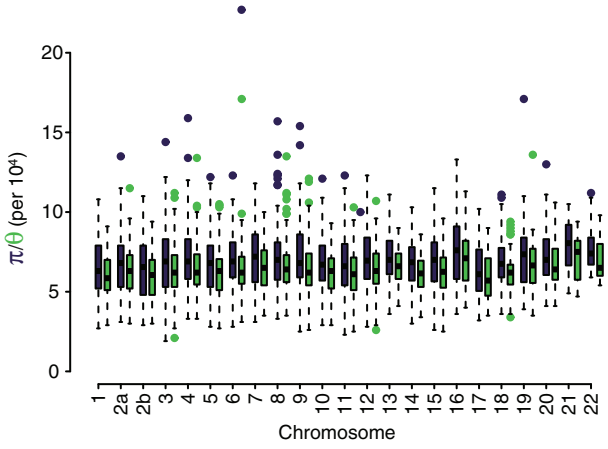

Fig. 2.—Distribution of estimates for nucleotide diversity π (blue) and Watterson's estimate of θ (green) by chromosomes. Calculated using 1 Mb windows with at least 80% accessibility after filtering.

Materials and Methods section). All parameters were scaled to have variance 1 and mean 0 in order to enable an easy comparison, and the autosomes were analyzed in 1 Mb segments with at least 80% accessibility (i.e., at least 80% of sites in the 1 MB segment passed the filter criteria), utilizing the genetic map from Auton et al. (2012). Across all sites, recombination rate was the primary predictor of nucleotide variation (fig. 3), with a strongly positive correlation (table 1), indicative of an important genome-wide role for linked selection. In addition, regional CpG-island and GC-content are strongly negatively correlated with variation (table 1), consistent with their associated effects on mutation rate. CpG dinucleotides, generally exhibiting high mutation rates (i.e., transition mutation rates

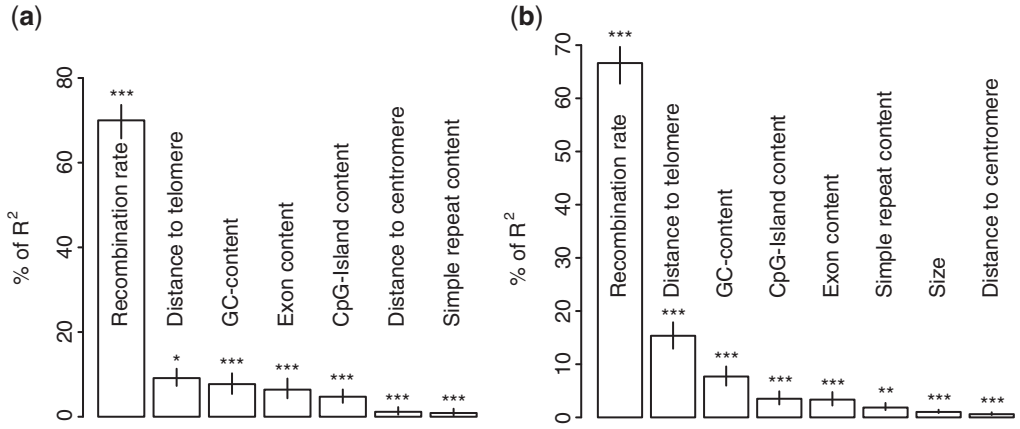

Fig. 3.—Relative importance of each significant regressor in the model for the nucleotide diversity estimate π in (a) Western chimpanzees ($R^2 = 45.85\%$) and (b) humans ($R^2 = 55.69\%$). The LMG metric (Lindemann et al. 1980) was used to calculate the relative importance of each predictor by partitioning R^2 by averaging over orders (computed using Gnu R's "rela.impo" package). Metrics were normalized to sum to 100%. Significance levels: *** 0; ** 0.001; * 0.01.

Table 1

Pairwise Correlation Between Different Predictors Used in the Regression Model for Western Chimpanzees (Significant Correlations with $P < 0.01$ are Highlighted in Bold)

	Recombination rate	GC content	Distance to centromere	Distance to telomere	Size	Exon content	CpG-Island content	Simple repeat content
π	**0.618**	**−0.207**	−0.032	**−0.239**	**−0.085**	**−0.256**	**−0.222**	−0.049
Recombination rate		**0.059**	**0.055**	**−0.509**	**−0.172**	**−0.142**	**−0.092**	**0.096**
GC content			0.046	**−0.389**	**−0.244**	**0.376**	**0.455**	**0.230**
Distance to centromere				**−0.312**	**0.251**	−0.021	0.002	**−0.056**
Distance to telomere					**0.332**	−0.009	**−0.087**	**−0.147**
Size						**−0.066**	−0.027	**−0.271**
Exon content							**0.290**	**0.084**
CpG-Island content								0.037

are elevated by ~30-fold in great apes (Hwang and Green 2004; Siepel and Haussler 2004; Keightley et al. 2011)) owing to spontaneous methylation-dependent deamination (e.g., Cooper and Youssoufian 1988; Duret 2009), are more stable within CpG-islands than in the rest of the genome—an observation that has often been attributed to the fact that CpG dinucleotides are usually unmethylated within CpG-islands (Polak and Arndt 2008; Cohen et al. 2011). In addition, CpG-islands can play a role in gene regulation, leading to more stable CpG-sites when they are under selection (Hodgkinson and Eyre-Walker 2011). Mutation rates of methylated CpG dinucleotides are also decreased in regions of high GC-content compared to other regions in the genome, possibly due to a lower melting of the DNA duplex as methylcytosine deamination preferentially occurs on single-stranded DNA (Fryxell and Moon 2005; Elango et al. 2008). Furthermore, exon content remains a fourth but significant predictor of variation in chimpanzees, presenting a stronger correlation than has been observed in humans.

Interpreting Similarities and Differences between Humans and Chimpanzees

Mutation-associated recombination has been suggested as a potential explanation for the observed correlation between recombination and diversity (see Kimura and Crow 1964). Indeed, there is some evidence of this in yeast (e.g., Strathern et al. 1995). However, sequence divergence in other organisms, ranging from Drosophila to humans, does not provide support for this explanation – an observation additionally supported by the data presented here. This result is particularly strong given the recent divergence of human and chimpanzee, and the broad-scale similarity of their

recombinational landscapes (McVicker et al. 2009; Lohmueller et al. 2011; Auton et al. 2012).

Thus, the predictive relationship between recombination and diversity in chimpanzees may be interpreted as evidence for the effects of linked selection. Importantly, RHH will have greater effects when positive selection is strong, with simulations suggesting an expected logarithmic relationship between recombination and diversity (Hellmann et al. 2005). Conversely, BGS will have the greatest effect when purifying selection is relatively weak (see Charlesworth 2012), with simulations suggesting an expected linear relationship between recombination and diversity (Hellmann et al. 2005). Interestingly, a logarithmic model is a better fit to both human and chimpanzee data, suggesting a potentially important role for positive selection in shaping genomic variation in both species (AIC_{log}(chimp) = 5,130; AIC_{linear}(chimp) = 5,369; AIC_{log}(human) = 4,683; AIC_{linear}(human) = 4,958).

Naturally, evidence is also observed for the important role of purifying selection. Note that in the weak-selection regime, a simple re-scaling of effective population size is no longer sufficient to account for BGS effects, and the site frequency spectrum may become strongly left-skewed (Ewing and Jensen 2016). Differences in this regard between these two species may be expected given their differing recent demographic histories, with human populations undergoing rapid growth (Coventry et al. 2010; Tennessen et al. 2012). More specifically, under this human demographic model, the increasing effective population size is expected to better prevent the fixation of strongly deleterious mutations, while the extreme population growth is expected to result in a larger proportion of segregating weakly deleterious mutations (Lohmueller et al. 2008; Keinan and Clark 2012; Gazave et al. 2013; Ewing and Jensen 2014; Lohmueller et al. 2014).

In order to investigate this expected difference with the genome-wide data used here, we compared the ratio of non-synonymous with synonymous polymorphic and divergent sites between chimpanzees and humans, utilizing a reconstructed ancestor as an outgroup. The mean genome-wide p*N*/p*S* in humans is 1.27 and 1.26 in chimpanzees (Wilcoxon rank sum test: P value < 0.03). Conversely, mean genome-wide d*N*/d*S* in humans is 1.04 and 1.23 in chimpanzees (Wilcoxon rank sum test: P value $< 2.5 \times 10^{-4}$). Thus, consistent with the above expectation owing to differences in recent demographic histories, a greater ratio of non-synonymous to synonymous variants is observed to be segregating in the human population, with a lesser ratio of non-synonymous to synonymous fixations on the human branch. Relatedly, there is a stronger relationship between exon content and diversity in chimpanzees than in humans. Consistent with this observation, Bataillon et al. (2015) recently argued for high rates of purifying selection in the coding regions of chimpanzee utilizing the inference framework of Eyre-Walker and Keightley (2009).

Thus, these different predictions developed in the theory literature are increasingly valuable for interpreting the influx of genomic polymorphism data from closely related species, and clearly suggest that the powerful and widely invoked correlation between recombination rate and diversity, combined with other genomic information, will be of great value in inferring differing strengths and rates of selection between species.

Materials and Methods

Diversity

A set of 5,323,301 autosomal chimpanzee single nucleotide polymorphisms (SNPs) from medium-coverage (average 9.1X) Illumina GAII sequencing of 10 Western chimpanzees (nine females and one male) was obtained from the PanMap project (Auton et al. 2012). PanMap SNP calls were subject to several filter criteria in order to minimize genotype errors (see Auton et al. 2012, SOM pp. 4–6). As the applied filter metrics can lead to the exclusion of a substantial fraction of sites in the genome, mask files, defining which nucleotides were accessible to the variant discovery in the study, were necessary to enable population genetic analysis. Mask files were generated using GATK's "UnifiedGenotyper" and "VariantFiltration" (Version 1.0.4705) (McKenna et al. 2010; DePristo et al. 2011) using the same filter criteria.

Each chromosome was divided into continuous windows of 1Mb size and the number of nucleotide differences per site between two randomly chosen sequences (π) (Nei and Li 1979) as well as Watterson's estimate of θ (Watterson 1975) were estimated for all windows with at least 80% accessibility (i.e., at least 80% of sites in the 1Mb segment passed the filter criteria).

In order to enable a comparison with humans, autosomal genotype data, consisting of 7,906,281 SNPs from a sample of 10 Yoruban (YRI) individuals (NA18522, NA19116, NA18912, NA19093, NA18516, NA18501, NA18870, NA18498, NA18510, and NA18499) was obtained from the 1000 Genomes Low Coverage Pilot Project SNP release (1000 Genomes Project Consortium 2010). These individuals exhibit a similar sequence coverage (~7.08 X on an average) as well as an equivalent concentration of recombination rate in the fine-scale genetic map than the chimpanzee sample (Auton et al. 2012).

All chimpanzee and human SNPs were annotated using ANNOVAR (Wang et al. 2010) with the information of the chimpanzee genome build panTro2 and the human genome build hg18, respectively (extracted from the "refGene" data set of the UCSC Genome table browser (Karolchik et al. 2004)), enabling the identification of synonymous (S) and non-synonymous (N) coding variants. Thereby, the panTro2 annotation contained 2,699 transcripts for 2,574 unique genes (including 653 without coding sequence annotation),

whereas the hg18 annotation contained 52,204 transcripts for 26,452 genes (including 11,833 without coding sequence annotation). Annotations were used to calculate the ratio of non-synonymous to synonymous polymorphisms (p*N*/p*S*) in both species. Mean genome-wide p*N*/p*S* were calculated in continuous windows of 100 kb size, excluding information from chimpanzee chromosome 2a and 2b and the orthologous regions in human chromosome 2 due to the different histories of the chromosomes (i.e., human chromosome 2 originated from a telomeric fusion event in the human ancestral lineage (IJdo et al. 1991)) as well as from chromosome 13 for which there were an insufficient number of annotated genes with polymorphisms available in the chimpanzee data set. The Wilcoxon rank sum test statistic was calculated using Gnu R's inbuilt functions.

Annotation of Genome Features

The relationship of regional nucleotide diversity levels in Western chimpanzees as well as in humans with specific genome features (namely recombination rate, GC-content, exon content, simple repeat content, CpG-island content, distance to the centromere and telomeres, as well as chromosome size) was studied to reveal whether there were any correlations between nucleotide diversity at large scales and these sequence features in either of the two species.

The length of each chromosome was determined and GC-content was measured using information obtained from the chimpanzee genome build panTro2.1 and the human genome build hg18, respectively, as downloaded from the UCSC genome browser. The distance from the middle of a given window to the centromere (typically consisting of large arrays of repetitive DNA) as well as to the closest telomere (often containing high GC content as well as high rates of recombination) was calculated to obtain information about the influence of large-scale chromosomal structure on nucleotide diversity rates. The distance was set to 0 if the entire window fell within a centromere/telomere region. Locations of centromeres and telomeres were obtained from the UCSC "Gap" database table (whereby chromosomal start and end coordinates were used to fill in missing telomere data). Genes were annotated using ANNOVAR (Wang et al. 2010) with the information of the chimpanzee genome build panTro2 and the human genome build hg18, respectively. Exon content was estimated as the percentage of sequence within exons. Similarly, the CpG-island content and the simple repeat content were estimated as the percentage of sequence within CpG-islands and simple repeats as obtained from the UCSC "CpG Island" and "Simple Repeats" tracks, respectively. Population recombination rates were calculated as the slope of a regression of genetic distances of markers within a 1Mb-window using the PanMap genetic map for chimpanzees and the genetic map build from the 10 YRI individuals for humans (Auton et al. 2012).

Regression Analyses

The contribution of the different genomic factors influencing the observed 1 Mb-scale variation in diversity levels across the genome of Western chimpanzees and humans was assessed using multiple linear regression as well as multiple logarithmic regression. Thereby, analyses were limited to intergenic 1 Mb windows with at least 80% accessibility. Estimates for diversity levels were log-transformed to be roughly normally distributed. All parameters were scaled to have variance 1 and mean 0 in order to enable an easier comparison of the different parameters (after the transformation, the slopes directly measure the strength of the relationship between the explanatory and the response variable). All possible models were analyzed and standard regression diagnostics were used to evaluate the validity of the model as well as to identify and remove outliers. The best model was chosen according to Akaike's Information Criterion (AIC) (as implemented in Gnu R). The LMG metric (Lindemann et al. 1980) was used to calculate the relative importance of each predictor by partitioning R^2 by averaging over orders (computed using Gnu R's "rela.impo" package).

Divergence

In order to identify fixed differences between the 10 Western chimpanzee individuals and humans, PanMap reads were additionally aligned against the human reference genome build hg18, using the same quality criteria than for the alignment against the chimpanzee reference genome (Auton et al. 2012). The ancestral allele for each site was determined using the four-way EPO alignments (downloaded from Ensemble). Analogous to the polymorphism data, divergent sites were annotated using ANNOVAR (Wang et al. 2010) and the ratio of non-synonymous to synonymous divergent sites (d*N*/d*S*) was calculated in continuous windows of 100 kb size in both species.

Acknowledgments

We would like to thank Adam Auton, Gil McVean, and two anonymous reviewers for helpful comments.

Literature Cited

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073.

Auton A, et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. Science 336(6078):193–198.

Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD. 2014. Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet. 30(12):540–546.

Bataillon T, et al. 2015. Inference of purifying and positive selection in three subspecies of chimpanzees (Pan troglodytes) from exome sequencing. Genome Biol Evol. 7(4):1122–1132.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356(6369):519–520.

Charlesworth B. 2012. The effects of deleterious mutations on evolution at linked sites. Genetics 190(1):5–22.
Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134(4):1289–1303.
Chen CL, et al. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res. 20(4):447–457.
Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055):69–87.
Cohen NM, Kenigsberg E, Tanay A. 2011. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell 145(5):773–786.
Cooper DN, Youssoufian H. 1988. The CpG dinucleotide and human genetic disease. Hum Genet. 78(2):151–155.
Coventry A, et al. 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun. 1:131.
Cutter AD, Moses AM. 2011. Polymorphism, divergence, and the role of recombination in *Saccharomyces cerevisiae* genome evolution. Mol Biol Evol. 28(5):1745–1754.
Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 14(4):262–274.
Deinard A, Kidd K. 1999. Evolution of a HOXB6 intergenic region within the great apes and humans. J Hum Evol. 36(6):687–703.
DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491–498.
Duret L. 2009. Mutation patterns in the human genome: more variable than expected. PLoS Biol. 7(2):e1000028.
Dvorák J, Luo MC, Yang ZL. 1998. Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing aegilops species. Genetics 148(1):423–434.
Elango N, Kim SH, Vigoda E, Yi SV. 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput Biol. 4(2):e1000015.
Ewing GB, Jensen JD. 2014. Distinguishing neutral from deleterious mutations in growing populations. Front Genet. 5:7.
Ewing GB, Jensen JD. 2016. The consequences of not accounting for background selection in demographic inference. Mol Ecol. 25(1):135–141.
Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 26(9):2097–2108.
Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol. 22(3):650–658.
Gazave E, Chang D, Clark AG, Keinan A. 2013. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. Genetics 195(3):969–978.
Hellmann I, et al. 2005. Why do human diversity levels vary at a megabase scale? Genome Res. 15(9):1222–1231.
Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. Nat Rev Genet. 12(11):756–766.
Hudson RR. 1994. How can the low levels of DNA sequence variation in regions of the drosophila genome with low recombination rates be explained? Proc Natl Acad Sci U S A. 91(15):6815–6818.
Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A. 101(39):13994–14001.
IJdo JW, Baldini A, Ward DC, Reeders ST, Wells RA. 1991. Origin of human chromosome 2: an ancestral telomere-telomere fusion. Proc Natl Acad Sci U S A. 88(20):9051–9055.
Jensen-Seaman MI, Deinard AS, Kidd KK. 2001. Modern African ape populations as genetic and demographic models of the last common ancestor of humans, chimpanzees, and gorillas. J Hered. 92(6):475–480.
Kaessmann H, Wiebe V, Pääbo S. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. Science 286(5442):1159–1162.
Kaessmann H, Wiebe V, Weiss G, Pääbo S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet. 27(2):155–156.
Karolchik D, et al. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32 (Database issue):D493–D496.
Keightley PD, Eöry L, Halligan DL, Kirkpatrick M. 2011. Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. Genetics 187(4):1153–1161.
Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336:740–743.
Kimura M, Crow JF. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725–738.
Lindemann RH, Merenda PF, Gold RZ. 1980. In: Foresman S, editor. Introduction to bivariate and multivariate analysis. Glenview (IL): Scott, Foresman & Co. p. 119ff.
Lohmueller KE, et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. Nature 451(7181):994–997.
Lohmueller KE. 2014. The distribution of deleterious genetic variation in human populations. Curr Opin Genet Dev. 29:139–146.
Lohmueller KE, et al. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 7(10):e1002326.
Nachman MW. 1997. Patterns of DNA variability at X-linked loci in *Mus domesticus*. Genetics 147(3):1303–1316.
Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 76(10):5269–5273.
McKenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.
McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5:e1000471.
Muller HJ. 1949. Redintegration of the symposium on genetics, paleontology, and evolution. In: Jepsen GL, Simpson GG, Mayr E, editors. Genetics, paleontology and evolution. Princeton (NH): Princeton University Press. p. 421–445.
Muller HJ. 1950. Our load of mutations. Am J Hum Genet. 2:111–176.
Polak P, Arndt PF. 2008. Transcription induces strand-specific mutations at the 5′ end of human genes. Genome Res. 18(8):1216–1223.
Rao Y, Sun L, Nie Q, Zhang X. 2011. The influence of recombination on SNP diversity in chickens. Hereditas 148(2):63–69.
Renzette N, Kowalik TF, Jensen JD. 2016. On the relative roles of background selection and genetic hitchhiking in shaping human cytomegalovirus genetic diversity. Mol Ecol. 25(1):403–413.
Satta Y. 2001. Comparison of DNA and protein polymorphisms between humans and chimpanzees. Genes Genet Syst. 76(3):159–168.
Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol. 21(3):468–488.
Stone AC, Griffiths RC, Zegura SL, Hammer MF. 2002. High levels of Y-chromosome nucleotide diversity in the genus Pan. Proc Natl Acad Sci U S A. 99(1):43–48.

Strathern JN, Shafer BK, McGill CB. 1995. DNA synthesis errors associated with double-strand-break repair. Genetics 140(3):965–972.
Stump AD, Shoener JA, Costantini C, Sagnon N, Besansky NJ. 2005. Sex-linked differentiation between incipient species of Anopheles gambiae. Genetics 169(3):1509–1519.
Tennessen JA, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090):64–69.
Timofeeff-Ressovsky NW. 1940. Mutations and geographical variation. In: Huxley JS, editor. The new systematics. London: The Systematics Association. p. 73–136.
Tyekucheva S, et al. 2008. Human-macaque comparisons illuminate variation in neutral substitution rates. Genome Biol. 9(4): R76.
Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Res. 38:e164.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7(2):256–276.

Associate editor: Naruya Saitou