Published online 25 November 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D903–D909***
doi:10.1093/nar/gkt1188

1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans

Marc Pybus[1], Giovanni M. Dall'Olio[1], Pierre Luisi[1], Manu Uzkudun[1], Angel Carreño-Torres[2], Pavlos Pavlidis[3], Hafid Laayouni[1], Jaume Bertranpetit[1,*] and Johannes Engelken[1,4,*]

[1]Program for Population Genetics, Institute of Evolutionary Biology (CSIC—Universitat Pompeu Fabra), 08003 Barcelona, Spain, [2]Population Genomics Node, National Institute for Bioinformatics (INB), Universitat Pompeu Fabra, 08003 Barcelona, Spain, [3]Institute of Molecular Biology and Biotechnology-FORTH, Heraklion, Crete GR 700 13, Greece and [4]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

Received July 31, 2013; Revised October 31, 2013; Accepted November 1, 2013

ABSTRACT

Searching for Darwinian selection in natural populations has been the focus of a multitude of studies over the last decades. Here we present the 1000 Genomes Selection Browser 1.0 (http://hsb.upf.edu) as a resource for signatures of recent natural selection in modern humans. We have implemented and applied a large number of neutrality tests as well as summary statistics informative for the action of selection such as Tajima's D, CLR, Fay and Wu's H, Fu and Li's F* and D*, XPEHH, ΔiHH, iHS, F_{ST}, ΔDAF and XPCLR among others to low coverage sequencing data from the 1000 genomes project (Phase 1; release April 2012). We have implemented a publicly available genome-wide browser to communicate the results from three different populations of West African, Northern European and East Asian ancestry (YRI, CEU, CHB). Information is provided in UCSC-style format to facilitate the integration with the rich UCSC browser tracks and an access page is provided with instructions and for convenient visualization. We believe that this expandable resource will facilitate the interpretation of signals of selection on different temporal, geographical and genomic scales.

INTRODUCTION

Initiatives such as the 1000 Genomes Project (1,2) are generating resequencing data from world-wide human populations on a genome-wide scale. Resequencing data constitutes a major leap for population genomic analysis due to its higher information density and limited SNP ascertainment bias compared to genotyping data. Therefore such data is appropriate to calculate summary statistics that are based on the site frequency spectrum like CLR or Tajima's D. Using the neutral evolutionary model as a null hypothesis, diverse statistics can be applied to genetic data to identify deviations from neutrality (Table 1). These statistical tests show varying degrees of robustness to demographic events (e.g. population bottlenecks and expansions) and sensitivity to different types of selection (e.g. positive, purifying or balancing). For instance, population bottlenecks, can lead to footprints that are similar to those caused by positive selection (21). Therefore, outlier approaches, which are commonly used to identify non-neutral loci in the extremes of a genome-wide distribution, are likely to contain a number of false positives in their extremes. Likewise, a number of false negatives, hence misidentified truly selected loci, are expected in a grey zone near the (arbitrary) outlier threshold (22). Outlier approaches in genome scans have proven powerful, but certainly they should be interpreted carefully in order to avoid storytelling (23). Even more, a profound understanding of adaptive evolution requires the integration of biological function (24) and if possible, validation on an experimental basis (25). Molecular network approaches can also give a functional context to the specific genes under adaptive selection (26,27). In all studies, care should be taken in communicating putative loci under selection to the public in order to avoid racist misinterpretation (28).

*To whom correspondence should be addressed. Tel: +34 933 160 840; Fax: +34 935 422 802; Email: johannesengelken@yahoo.com
Correspondence may also be addressed to Jaume Bertranpetit. Email: Jaume.Bertranpetit@upf.edu

The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

Table 1. List of available summary statistics

Method family	Method	Reference	Window size	Rank scores tail
Allele frequency spectrum	Tajima's D	Tajima (3)	30 kb	Lower
	CLR	Nielsen *et al.* (4)	Variable size	Upper
	Fay and Wu's H	Fay and Wu (5)	30 kb	Lower
	Fu and Li's F*	Fu and Li (6)	30 kb	Lower
	Fu and Li's D*	Fu and Li (6)	30 kb	Lower
	R^2	Ramos-Onsins and Rozas (7)	30 kb	Lower
Linkage disequilibrium structure	XP-EHH	modified from Sabeti *et al.* (8)	SNP-specific	Upper
	AiHH	modified from Voight *et al.* (9)	SNP-specific	Upper
	his	modified from Voight *et al.* (9)	SNP-specific	Upper
	EHH_average	modified from Sabeti *et al.* (10)	30 kb	Upper
	EHH_max	modified from Sabeti *et al.* (10)	30 kb	Upper
	Wall's B	Wall (11)	30 kb	Upper
	Wall's Q	Wall (12)	30 kb	Upper
	Fu's F	Fu (13)	30 kb	Lower
	Dh	Nei (14)	30 kb	Upper
	Za	Rozas *et al.* (15)	30 kb	Upper
	ZnS	Kelly (16)	30 kb	Upper
	ZZ	Rozas *et al.* (15)	30 kb	Upper
Population differentiation	Fst (global and pairwise)	Weir and Cockerham (17)	SNP-specific	Upper
	ΔDAF (standard and absolute)	Hofer *et al.* (18)	SNP-specific	Upper
	XP-CLR	Chen *et al.* (19)	0.1 cM (maximum window)	Upper
Descriptive statistics	Segregating sites		30 kb	NA
	Singletons		30 kb	NA
	pi (nucleotide diversity)	Nei and Li (20)	30 kb	NA
	DAF (derived allele frequency)		SNP-specific	NA
	MAF (minor allele frequency)		SNP-specific	NA

Despite of these limitations and the fact that complete selective sweeps may not be extremely widespread in humans (29), a large number of regions under strong positive selection can be expected in the genome (30).

DESCRIPTION OF APPLIED STATISTICAL TESTS

Due to linkage, neutral alleles in the surrounding region hitchhike with the selected allele. Maynard Smith and Haigh (31) described this process of genetic hitchhiking and the so-called selective sweep. More recent studies showed that genetic hitchhiking generates distinct polymorphism signatures on the genome such as: (i) reduction of polymorphism level and excess of low- and high-frequency derived variants (32), (ii) spatial patterns of linkage-disequilibrium (33) and (iii) increased genetic differentiation among populations (34). Taking advantage of these three theoretical expectations, several methods to detect positive selection have been developed in the last two decades. This makes reference to the fact that no single statistic is enough to describe selection under various demographic models and modes of selection (22).

Here, we implemented a large number of statistical tests (Table 1) in order to allow for a more comprehensive analysis of natural selection, especially, positive selection. In brief, we have assigned the statistical tests to different method families (Table 1). Within the first family which is based on the allele frequency spectrum, Tajima's D (3) is a classical neutrality test that compares estimates of the number of segregating sites and the mean pair-wise difference between sequences. CLR is a multi-locus, composite likelihood ratio test (4,35). Fay and Wu's H (5) uses another facet of the site-frequency spectrum, by comparing the number of derived segregating sites at high frequencies to the number of variants at intermediate frequencies. Fu and Li's F* compares the number of singletons to the mean pair-wise difference between sequences and Fu and Li's D* compares it to the total number of nucleotide variants in a genomic region (6). R_2 (7) is a statistical test for detecting population growth based on the comparison of the difference between the number of singletons per sequence and the average number of nucleotide differences.

Among the linkage disequilibrium structure methods, XP-EHH (8) is a cross-population test based on extended haplotype homozygosity (EHH). ΔiHH considers the difference between the integrated haplotype homozygosity scores for each allele in a single population while iHS (9) is defined as their log ratio. EHH average and EHH maximum (36); modified from (10) are based on the extended haplotype homozygosity. Wall's B (11) counts the number of pairs of adjacent segregating sites that are congruent (if the subset of the data consisting of the two sites contains only two different haplotypes), while Wall's Q (12) adds the number of partitions (two disjoint subsets whose union is the set of individuals in the sample) induced by congruent pairs to Wall's B. Fu's F (13) takes into account the haplotype diversity in the sample. Dh (14) is a summary statistic based on the number of different haplotypes in the sample.

The third family of methods is based on population differentiation. F_{ST} (37); calculated following the diploid method in Weir 1996 (p. 178) and ΔDAF (18) are estimates of population differentiation based on derived allele frequencies. XP-CLR (19) is a multi-locus allele-frequency-differentiation statistic between two populations.

Additional statistics like segregating sites per 30-kb window and the nucleotide diversity and others (Table 1) are listed as descriptive statistics. A thorough description of the tests is given in the original literature (see Table 1) and in diverse excellent reviews on the topic (38,39).

COMPUTATIONAL FRAMEWORK AND DESCRIPTION OF 1000 GENOMES SOURCE DATA

A framework to calculate diverse summary statistics (Table 1) from 1000 genomes data was developed (Figure 1). A detailed description of how the statistics were implemented is given (Supplementary Material). A genome-wide overview of the results stored in the database for selected summary statistics is given (Supplementary Table S1). As described in the 1000 genomes Phase 1 paper (1), the quality of the 1000 genomes low coverage data has improved considerably over the pilot phase (2), but a number of limitations need to be kept in mind for population genomic analysis: (i) singletons and other rare variants are still underrepresented, (ii) the accessibility of the genome with the used short-read-sequencing technologies ~94% and (iii) the reported phasing switch error every 250 kb (median, Supplementary Figure S5 in (1)) likely underestimates the length of long-shared haplotypes expected to occur around recent selective sweeps. Despite of these drawbacks which are mainly due to the nature of the low coverage approach, the short-read technology and differences in read depth (40), this dataset has important advantages over genotyping data, most importantly (i) a higher SNP density, (ii) the overcoming of ascertainment bias and (iii) a larger number of individuals per population, when compared to previous datasets (HapMap II and HGDP). We used phased data from the CEU, the CHB and the YRI populations from the integrated Phase 1 variant set (April 2012), with 97, 85 and 88 individuals, respectively. From the input vcf (variant call format) file we extracted exclusively the low-coverage VSQR SNP calls in order to avoid any bias that might result from differences between low-coverage calls and high-coverage exome SNP calls. Indels were not used. Ancestral states in this data set were identified using a 4-way alignment of humans, chimp, orangutan and rhesus macaque, provided by the 1000 genomes consortium (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/).

AVAILABILITY OF DATABASE

All data is available via our entry page: http://hsb.upf.edu. A search mask gives the user easy access to the results for a specific gene or a genomic region of choice. The 'submit' button leads the user to a UCSC-style genome browser (http://pgb.ibe.upf.edu/) which is a custom installation of the UCSC Genome Browser (41,42). This UCSC Genome Browser installment allows for a visual inspection of the data, and for an integration of our data with many other available datasets. The raw scores of the tracks can be conveniently downloaded using the UCSC

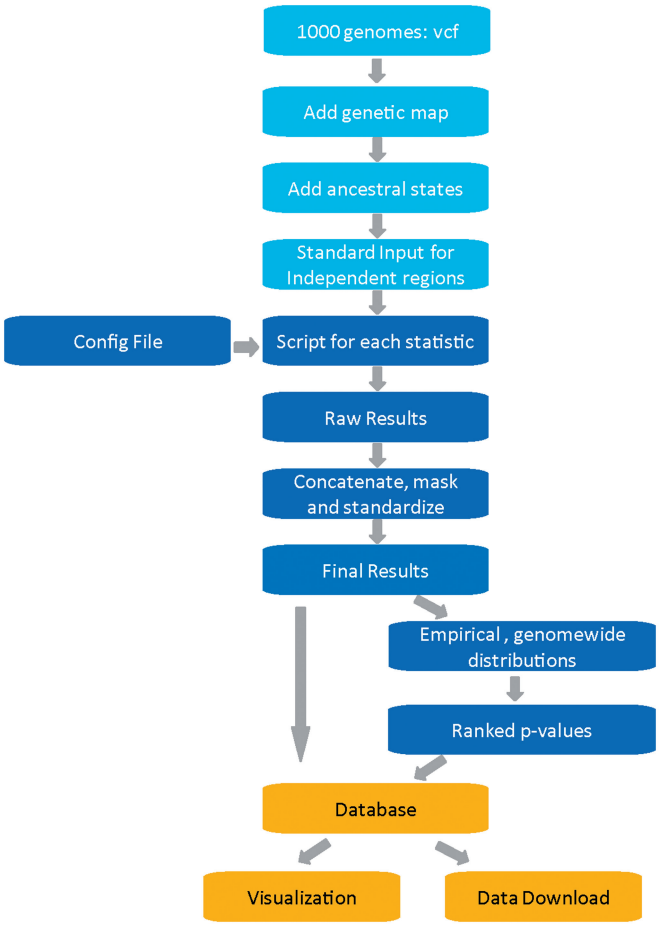

Figure 1. Schematic workflow developed in order to calculate diverse genome-wide summary statistics informative for the action of selection and to build a database in order to share and visualize the results.

Table function (43) and is integrated with the Galaxy platform (galaxyproject.org). Using the 'configure' function on the browser page, the tracks can be further customized and using 'right click' the visualized genomic regions can be downloaded as a picture in .png format. For every statistical test, we provide two tracks, one for the raw scores and one for ranked scores. The purpose of the rank score tracks is to provide a comparison to the rest of the genome. Conveniently, the rank scores are presented in such a way that they present a peak (instead of a valley) in regions under positive selection. They are calculated using an outlier approach (22,44) by sorting all the scores genome-wide and determining the −log10 of the rank divided by the number of values in the distribution, taking the upper tail for most of the tests, or the lower tail for Tajima's D, Fay and Wu's H, Fu and Li's F and D, R_2 and Fu's F (see Table 1 and a more detailed description on the entry page). The main purpose of the entry page is to provide a channel of communication with users, following the guidelines in (45). It serves as a platform for updates, questions and feedback (46). Therefore the page also provides documentation on the tracks and on the tests implemented as well as a FAQ and a feedback section.

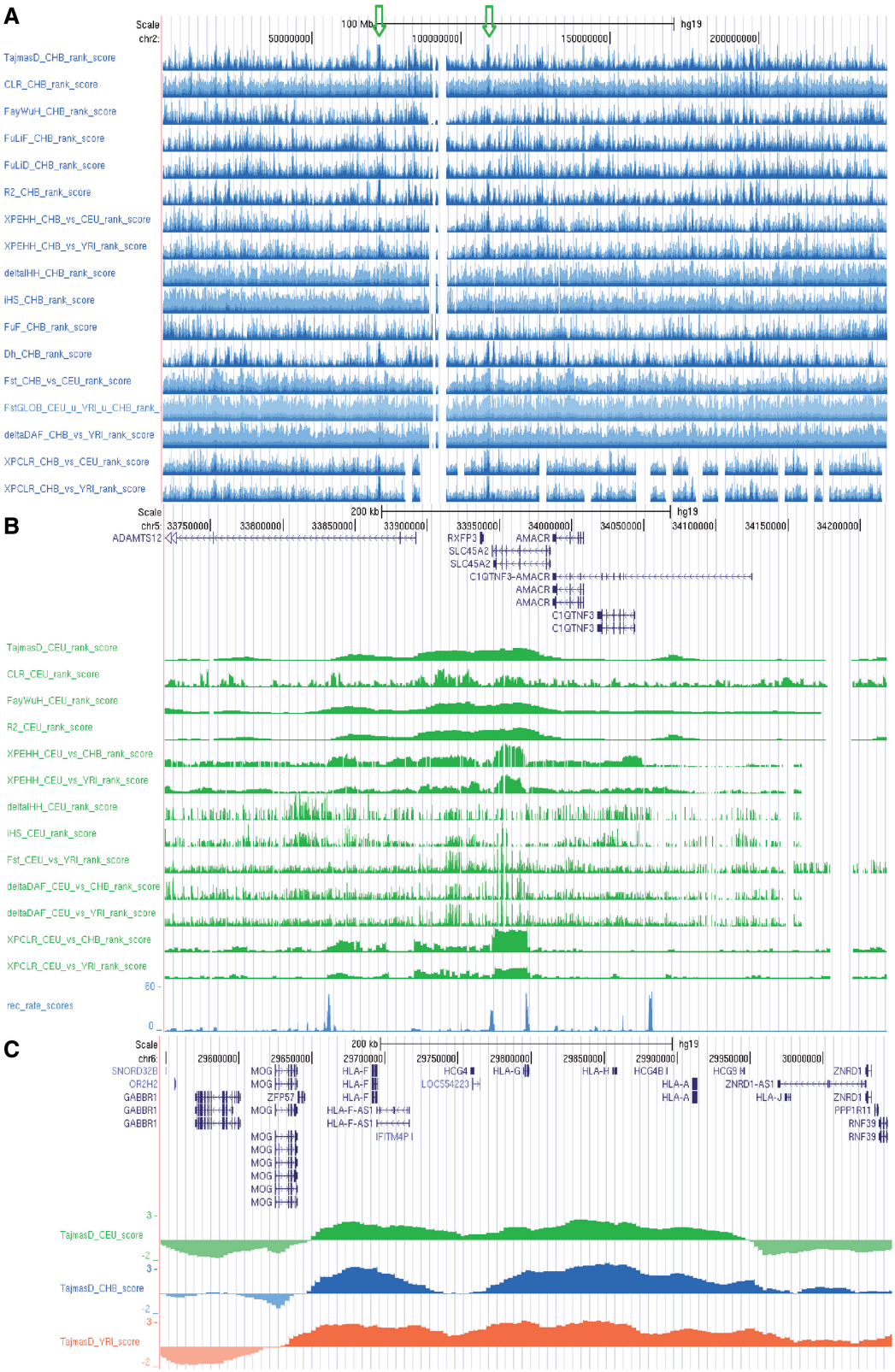

Figure 2. Examples of genomic regions under selection in the 1000 genomes selection browser. Tracks of statistics from different populations are visualized in colour (CEU in green, CHB in red and YRI in blue). Additional examples are given at http://hsb.upf.edu (A) The p- and q-arms of chromosome 2 (−log10 of empirically ranked scores). Recurrent peaks at around 72.5 Mb (left green arrow) and 109.5 Mb (right green arrow) indicate the loci *CYP26B1*/*EXOC6B* and *EDAR*, respectively. (B) Signature of positive selection around *SLC45A2*, another established skin colour gene, in the CEU population (0.5-Mb window; −log10 of empirically ranked scores). (C) Widespread balancing selection in the *HLA* region indicated by strongly positive scores for Tajima's D in all three analysed human populations (0.5-Mb window).

EXAMPLE APPLICATIONS

First, we exemplify the use of the database by extracting results for a number of established loci under selection: *EDAR* (47), *LCT* (46), *SLC45A2* (48), *CD36* (49), *HERC2* (50), *SLC24A5* (51), *CD5* (52) and *APOL1* (53). A loci-specific summary of statistical tests is given (Supplementary Table S2). Interestingly, for any given locus, only a subset of statistical tests shows an extreme outlier score. This is consistent with differences in the architecture of selective sweeps. iHS scores near to certain very pronounced selective sweeps (e.g. *LCT* and *SLC24A5*) failed to compute due to inherent properties of the statistics, because either (i) the selected haplotype was near fixation or (ii) the EHH did not drop below the defined threshold in a given window. Examples for both positive (*SLC45A2*) and balancing (*HLA* region) selection are visualized in Figure 2. As expected, Tajima's D scores around *HLA* (54) as well as the *ABO* locus (55) (data not shown) were pronouncedly elevated in all three analyzed populations, a pattern which is compatible with the action of balancing selection.

COMPARISON TO OTHER WEB RESOURCES

As for positive selection based on between-species comparisons, the Selectome database (http://bioinfo.unil.ch/selectome/; (56)) presents results based on the dN/dS method using a branch-site specific likelihood test. As for recent natural selection within modern humans, a number of web resources are available. For previous datasets, e.g. the HapMap 2 and HGDP projects, several positive selection statistics are available in form of the haplotter tool (http://haplotter.uchicago.edu/; (24)) and in form of the HGDP selection browser (http://hgdp.uchicago.edu/; (57)). For the 1000 genomes project data, the online tool ENGINES (http://spsmart.cesga.es; (58)) is useful for the analysis of allele frequencies and a recent study presented a method to calculate corrected summary statistics from low coverage sequencing data (40). dbPSHP (http://jjwanglab.org/dbpshp) offers a large number of statistical tests in a SNP-specific manner for HapMap 3 and 1000 genomes datasets. Complementary to these databases, our database gives a large number of region- and SNP-specific scores (depending on the test statistic) based on resequencing data (1000 genomes Phase 1), with a special focus on genome-wide significance (by the ranked scores) and the visualization of several statistics in parallel (Figure 2).

CONCLUSIONS

By applying a large number of summary statistics to data from the 1000 genomes project, we have built a timely and expandable resource for the population genomics research community. An associated user-friendly genome browser gives a visual impression of the genetic variation in a genomic region of interest and offers functionality for an array of down-stream analyses. While this resource will not replace a thorough, case by case analysis of selection, we expect that it will prove useful for the research community through the large number of test statistics and the fine-grained character of resequencing data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thankfully acknowledge contributions from Anna Ramírez-Soriano, Arcadi Navarro, Francesc Calafell, Elena Bosch, Chris Tyler-Smith, Gonçalo Abecasis, Roger Bartomeus Peñalver and the 1000 Genomes Project (1000genomes.org). The authors also thank Txema Heredia and the National Institute of Bioinformatics (http://www.inab.org) for computational support.

FUNDING

Ministerio de Ciencia y Tecnología (Spain); Direcció General de Recerca, Generalitat de Catalunya (Grup de Recerca Consolidat 2009 SGR 1101); Subprogram BMC [BFU2010-19443 awarded to J.B.]; Post-doctoral scholarship from the Volkswagenstiftung [Az: I/85 198 to J.E.]; Spanish government [BFU-2008-01046; SAF2011-29239]; The Spanish government FPI scholarships [BES-2009-017731 and BES-2011-04502 to G.M.D. and M.P., respectively]; PhD fellowship from 'Acción Estratégica de Salud, en el marco del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008-2011' from Instituto de Salud Carlos III (to P.L.). Funding for open access charge: Prof. Jaume Bertranpetit.

Conflict of interest statement. None declared.

REFERENCES

1. The 1000 Genomes Project Consortium, Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
2. The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
3. Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
4. Nielsen,R., Williamson,S., Kim,Y., Hubisz,M.J., Clark,A.G. and Bustamante,C. (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.*, **15**, 1566–1575.
5. Fay,J.C. and Wu,C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.
6. Fu,Y.X. and Li,W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.
7. Ramos-Onsins,S.E. and Rozas,J. (2002) Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.*, **19**, 2092–2100.
8. Sabeti,P.C., Varilly,P., Fry,B., Lohmueller,J., Hostetter,E., Cotsapas,C., Xie,X., Byrne,E.H., McCarroll,S.A., Gaudet,R. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

9. Voight,B.F., Kudaravalli,S., Wen,X. and Pritchard,J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.
10. Sabeti,P.C., Reich,D.E., Higgins,J.M., Levine,H.Z.P., Richter,D.J., Schaffner,S.F., Gabriel,S.B., Platko,J.V., Patterson,N.J., McDonald,J.G. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
11. Wall,J.D. (1999) Recombination and the power of statistical tests of neutrality. *Genet. Res.*, **74**, 65–79.
12. Wall,J.D. (2000) A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.*, **17**, 156–163.
13. Fu,Y.X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**, 915–925.
14. Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY.
15. Rozas,J., Gullaud,M., Blandin,G. and Aguadé,M. (2001) DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. *Genetics*, **158**, 1147–1155.
16. Kelly,J.K. (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.
17. Weir,B.S. and Cockerham,C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
18. Hofer,T., Ray,N., Wegmann,D. and Excoffier,L. (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann. Hum. Genet.*, **73**, 95–108.
19. Chen,H., Patterson,N. and Reich,D. (2010) Population differentiation as a test for selective sweeps. *Genome Res.*, **20**, 393–402.
20. Nei,M. and Li,W.H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, **76**, 5269–5273.
21. Barton,N.H. (1998) The effect of hitch-hiking on neutral genealogies. *Genet. Res.*, **72**, 123–133.
22. Akey,J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.*, **19**, 711–722.
23. Pavlidis,P., Jensen,J.D., Stephan,W. and Stamatakis,A. (2012) A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Mol. Biol. Evol.*, **29**, 3237–3248.
24. Sabeti,P.C., Schaffner,S.F., Fry,B., Lohmueller,J., Varilly,P., Shamovsky,O., Palma,A., Mikkelsen,T.S., Altshuler,D. and Lander,E.S. (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.
25. Barrett,R.D.H. and Hoekstra,H.E. (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.*, **12**, 767–780.
26. Dall'olio,G.M., Laayouni,H., Luisi,P., Sikora,M., Montanucci,L. and Bertranpetit,J. (2012) Distribution of events of positive selection and population differentiation in a metabolic pathway: the case of asparagine N-glycosylation. *BMC Evol. Biol.*, **12**, 98.
27. Luisi,P., Alvarez-Ponce,D., Dall'olio,G.M., Sikora,M., Bertranpetit,J. and Laayouni,H. (2012) Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. *Mol. Biol. Evol.*, **29**, 1–40.
28. Vitti,J.J., Cho,M.K., Tishkoff,S.A. and Sabeti,P.C. (2012) Human evolutionary genomics: ethical and interpretive issues. *Trends Genet.*, **28**, 137–145.
29. Hernandez,R.D., Kelley,J.L., Elyashiv,E., Melton,S.C., Auton,A., McVean,G., Sella,G. and Przeworski,M. (2011) Classic selective sweeps were rare in recent human evolution. *Science*, **331**, 920–924.
30. Grossman,S.R., Andersen,K.G., Shlyakhter,I., Tabrizi,S., Winnicki,S., Yen,A., Park,D.J., Griesemer,D., Karlsson,E.K., Wong,S.H. *et al.* (2013) Identifying recent adaptations in large-scale genomic data. *Cell*, **152**, 703–713.
31. Smith,J. and Haigh,J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.*, **23**, 23–35.
32. Braverman,J.M., Hudson,R.R., Kaplan,N.L., Langley,C.H. and Stephan,W. (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, **140**, 783–796.
33. Kim,Y. and Nielsen,R. (2004) Linkage disequilibrium as a signature of selective sweeps. *Genetics*, **167**, 1513–1524.
34. Barreiro,L.B., Laval,G., Quach,H., Patin,E. and Quintana-Murci,L. (2008) Natural selection has driven population differentiation in modern humans. *Nat. Genet.*, **40**, 340–345.
35. Williamson,S.H., Hubisz,M.J., Clark,A.G., Payseur,B.A., Bustamante,C.D. and Nielsen,R. (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genet.*, **3**, e90.
36. Ramírez-Soriano,A., Ramos-Onsins,S.E., Rozas,J., Calafell,F. and Navarro,A. (2008) Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*, **179**, 555–567.
37. Weir,B.S. and Hill,W.G. (2002) Estimating F-statistics. *Annu. Rev. Genet.*, **36**, 721–750.
38. Nielsen,R. (2004) Population genetic analysis of ascertained SNP data. *Hum. Genomics*, **1**, 218–224.
39. Bamshad,M. and Wooding,S.P. (2003) Signatures of natural selection in the human genome. *Nat. Rev. Genet.*, **4**, 99–111.
40. Korneliussen,T.S., Moltke,I., Albrechtsen,A. and Nielsen,R. (2013) Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinform.*, **14**, 289.
41. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,A.D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.
42. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
43. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
44. Kelley,J.L., Madeoy,J., Calhoun,J.C., Swanson,W. and Akey,J.M. (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.*, **16**, 980–989.
45. Dall'Olio,G.M., Marino,J., Schubert,M., Keys,K.L., Stefan,M.I., Gillespie,C.S., Poulain,P., Shameer,K., Sugar,R., Invergo,B.M. *et al.* (2011) Ten simple rules for getting help from online scientific communities. *PLoS Comput. Biol.*, **7**, e1002202.
46. Enattah,N.S., Sahi,T., Savilahti,E., Terwilliger,J.D., Peltonen,L. and Järvelä,I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.*, **30**, 233–237.
47. Bryk,J., Hardouin,E., Pugach,I., Hughes,D., Strotmann,R., Stoneking,M. and Myles,S. (2008) Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. *PLoS One*, **3**, e2209.
48. Branicki,W., Brudnik,U., Draus-Barini,J., Kupiec,T. and Wojas-Pelc,A. (2008) Association of the SLC45A2 gene with physiological human hair colour variation. *J. Hum. Genet.*, **53**, 966–971.
49. Fry,A.E., Ghansa,A., Small,K.S., Palma,A., Auburn,S., Diakite,M., Green,A., Campino,S., Teo,Y.Y., Clark,T.G. *et al.* (2009) Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. *Hum. Mol. Genet.*, **18**, 2683–2692.
50. Duffy,D.L., Montgomery,G.W., Chen,W., Zhao,Z.Z., Le,L., James,M.R., Hayward,N.K., Martin,N.G. and Sturm,R.A. (2007) A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am. J. Hum. Genet.*, **80**, 241–252.
51. Lamason,R.L., Mohideen,M.-A.P.K., Mest,J.R., Wong,A.C., Norton,H.L., Aros,M.C., Jurynec,M.J., Mao,X., Humphreville,V.R., Humbert,J.E. *et al.* (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, **310**, 1782–1786.

52. Carnero-Montoro,E., Bonet,L., Engelken,J., Bielig,T., Martínez-Florensa,M., Lozano,F. and Bosch,E. (2012) Evolutionary and functional evidence for positive selection at the human CD5 immune receptor gene. *Mol. Biol. Evol.*, **29**, 811–823.
53. Genovese,G., Friedman,D.J., Ross,M.D., Lecordier,L., Uzureau,P., Freedman,B.I., Bowden,D.W., Langefeld,C.D., Oleksyk,T.K., Uscinski Knob,A.L. *et al.* (2010) Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*, **329**, 841–845.
54. Hedrick,P.W. and Thomson,G. (1983) Evidence for balancing selection at HLA. *Genetics*, **104**, 449–456.
55. Calafell,F., Roubinet,F., Ramírez-Soriano,A., Saitou,N., Bertranpetit,J. and Blancher,A. (2008) Evolutionary dynamics of the human ABO gene. *Hum. Genet.*, **124**, 123–135.
56. Proux,E., Studer,R.A., Moretti,S. and Robinson-Rechavi,M. (2009) Selectome: a database of positive selection. *Nucleic Acids Res.*, **37**, D404–D407.
57. Pickrell,J.K., Coop,G., Novembre,J., Kudaravalli,S., Li,J.Z., Absher,D., Srinivasan,B.S., Barsh,G.S., Myers,R.M., Feldman,M.W. *et al.* (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.*, **19**, 826–837.
58. Amigo,J., Salas,A. and Phillips,C. (2011) ENGINES: exploring single nucleotide variation in entire human genomes. *BMC Bioinformatics*, **12**, 105.