D1080–D1089 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 18 October 2018*
doi: 10.1093/nar/gky959

PopHumanScan: the online catalog of human genome adaptation

Jesús Murga-Moreno †, Marta Coronado-Zamora †, Alejandra Bodelón, Antonio Barbadilla * and Sònia Casillas *

Institut de Biotecnologia i de Biomedicina and Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

Received August 15, 2018; Revised September 21, 2018; Editorial Decision October 03, 2018; Accepted October 04, 2018

ABSTRACT

Since the migrations that led humans to colonize Earth, our species has faced frequent adaptive challenges that have left signatures in the landscape of genetic variation and that we can identify in our today's genomes. Here, we (i) perform an outlier approach on eight different population genetic statistics for 22 non-admixed human populations of the Phase III of the 1000 Genomes Project to detect selective sweeps at different historical ages, as well as events of recurrent positive selection in the human lineage; and (ii) create *PopHumanScan*, an online catalog that compiles and annotates all candidate regions under selection to facilitate their validation and thoroughly analysis. Well-known examples of human genetic adaptation published elsewhere are included in the catalog, as well as hundreds of other attractive candidates that will require further investigation. Designed as a collaborative database, PopHumanScan aims to become a central repository to share information, guide future studies and help advance our understanding of how selection has modeled our genomes as a response to changes in the environment or lifestyle of human populations. PopHumanScan is open and freely available at https://pophumanscan.uab.cat.

INTRODUCTION

Since the split with chimpanzees, and especially since the migrations that led humans to colonize almost every single place on Earth, our species has faced frequent environmental and social changes that have shaped the variation patterns of our genomes through the action of natural selection (1). These environmental challenges include, for example, extreme cold temperatures in much of the Americas and Eurasia during the last ice age, limiting exposure to sunlight as we moved to higher latitudes or contact with new pathogens. Part of the incorporated genetic innovations may have been introgressed from archaic hominins that left Africa before us, including Neanderthals and Denisovans, with whom we encountered and interbred before they got extinct. Around 1–6% of any modern non-African human genome can be traced back to the genomes of these archaic populations (2). Another dramatic change occurred within the past 10 000 years coinciding with the transition from a hunting-gathering lifestyle to farming. Selection pressures for adapting to large settlements and new diets favored genetic variants associated with innate immune response, fatty acid metabolic efficiency, and lactose tolerance, among others (3).

These selection pressures left signatures in the landscape of genetic variation that can be identified in our today's genomes (4). Starting from single-locus studies to the first large-scale catalogs of genetic variation (5–8), dozens of targets of positive selection have been identified, providing important insights into recent human evolutionary history (3,9,10). Even though genome-wide HapMap genotyping data is able to disentangle the effects of demography and selection better than single-locus approaches, it still has the problem of ascertainment bias, which may alter the site frequency spectrum (SFS) of analyzed single nucleotide polymorphisms (SNPs) (11). The availability of the most comprehensive worldwide nucleotide variation dataset so far from the 1000 Genomes Project (1000GP) (12,13), based on whole-genome re-sequencing, provides the human lineage with an abundant, ascertained variation dataset on which to test molecular population genetics hypotheses and eventually pinpoint targets of positive selection in one or more human populations that escape from the background evolutionary dynamics of genetic variation (14).

To gain deeper understanding of how environmental and social challenges have shaped our genomes through the action of natural selection, here we (i) perform a genome-wide scan of selection on the latest version of the 1000GP

*To whom correspondence should be addressed. Tel: +34 93 586 8958; Fax: +34 93 581 2011; Email: sonia.casillas@uab.cat
Correspondence may also be addressed to Antonio Barbadilla. Tel: +34 93 586 8941; Fax: +34 93 581 2011; Email: antonio.barbadilla@uab.cat
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

data by surveying distinctive signatures of genomic variation left by different selective events, and (ii) create an online catalog of all candidate genomic regions under selection to facilitate their validation and thorough analysis. As far as we are concerned, dbPSHP (15) is the only previous online database that compiles putative positively selected loci in human evolution. In their case, regions were extracted from curated publications based on genotyping data—instead of whole-genome re-sequencing data—of the HapMap III (8) and the 1000GP Pilot 1 (12), and the last update is reported as far as May 2014. On the other hand, after the publication of the 1000 Genomes Selection Browser 1.0 (16), Pybus *et al.* developed a machine-learning framework—Hierarchical Boosting—that combines the results of multiple tests for detecting positive selection to classify genomic regions into different selection regimes (17). They analyzed within-species polymorphism data for three populations of the 1000GP Phase I (12), and the resulting scores were made available as UCSC Hub Tracks. Here, we perform an outlier approach on the greatest number of population genetic statistics and sampled populations available so far. This genome-wide scan of selection is able to detect sweeps at different historical ages, as well as evidence of recurrent selection in the human lineage since the split between our species and chimpanzees. Results have been made available in a collaborative, online database, *PopHumanScan*, which is aimed at compiling and annotating adaptation events along the human evolutionary history. Well-known examples of human genetic adaptation published elsewhere are included in the catalog, as well as hundreds of other attractive candidates that will require a more thoroughly analysis. PopHumanScan graphically represents each signature of selection within the empirical distributions of the corresponding DNA diversity statistic across populations. It also provides structural and functional annotations of the region, links to external databases and cross-references to 268 publications.

POPHUMANSCAN ANALYSIS PIPELINE

We have designed and implemented a custom pipeline (Figure 1) to perform a genome-wide scan of selection. Specifically, the pipeline processes eight different neutrality tests calculated either in sliding windows along the genome or for each protein-coding gene, for 22 non-admixed human populations. The genomic regions identified should show signatures that are compatible with natural selection having driven the evolution of the region at one or different timescales, from recent selective sweeps to recurrent selection since the split between our species and chimpanzees. These candidate regions under selection are further characterized with structural and functional annotations of that particular region. Furthermore, 268 articles reporting evidences of natural selection in genomic regions and genes using different statistical methods have been manually curated and cross-referenced to the candidate regions detected with our pipeline.

Pre-processing of the PopHuman data

Population genomic data were retrieved from PopHuman (18) for 22 non-admixed populations of the Phase III of the 1000GP (13) (Supplementary Table S1), mapped to GRCh37/hg19. Specifically, values for seven different neutrality tests have been obtained for each population in 186 549 10-kb non-overlapping sliding windows along the autosomes and the X chromosome (Figure 1). In addition, the McDonald and Kreitman test (MKT) (19), as well as the proportion of substitutions that are adaptive (α) (20,21), were calculated on the protein-coding genes overlapping the candidate regions under selection identified with the other seven statistics and that showed some variability in both polymorphism and divergence, according to gene annotations from GENCODE release 27 (22) and PopHuman polymorphism and divergence data (Figure 1). MKT-derived calculations were performed using the R package iMKT (https://github.com/BGD-UAB/iMKT; last accessed: February 2018). In total, eight different neutrality tests were performed. They identify different types of signatures that remain visible in the genomic sequences for a timescale ranging from few thousands of years after a single sweep selection event to several million years for the case of recurrent selection (4,23).

Linkage Disequilibrium (LD). Selection signatures in the LD, e.g. long haplotypes, were detected using two complementary measures: *iHS* (24) and *XP-EHH* (9). *iHS* has good power to detect selective sweeps with haplotypes at moderate frequency (50–80%), while *XP-EHH* is more powerful for detecting selective sweeps when the selected haplotype has a frequency >80%. In the case of *XP-EHH*, which analyzes pairs of populations, only pairs CEU-YRI, CEU-CHB and YRI-CHB were considered, and the population showing the evidence of selection was identified with the locus-specific branch length method (*LSBL*) (25). These long-range haplotypes persist for relatively short periods of type, typically up to ∼30 kya, and thus this signature allows us to identify recent selection events only (4).

Site Frequency Spectrum (SFS). Five statistics have been considered: four are based on both the allele frequency spectrum and the levels of variability—Fay and Wu's *H* (26), Fu and Li's *D* and *F* (27), and Tajima's *D* (28)—, and the other one is based on population differentiation—F_{ST} (29,30)—. Fay and Wu's *H* detects an excess of high-frequency derived SNPs, compatible with an incomplete sweep or recombination breaking swept linked SNPs. Fu and Li's *D* and *F* and Tajima's *D* assess the lack or excess of rare alleles (or singletons). Lack of rare alleles is compatible with balancing selection, while an excess is normally explained by either positive or weakly deleterious selection. F_{ST} detects population-specific selective events that changed the genetic composition of the affected population. It analyzes pairs of populations. The pairs CEU-YRI, CEU-CHB and YRI-CHB were considered, and the population showing the evidence of selection was identified with the *LSBL* (25). SFS signatures can persist in the genomes for a longer period than LD, and thus selective events identified by shifts in the SFS might have occurred up to ∼80 kya (4).

Protein changes. Recurrent selection since the split between our species and chimpanzees (<6 mya) is detected using a test based on comparisons of polymorphism and

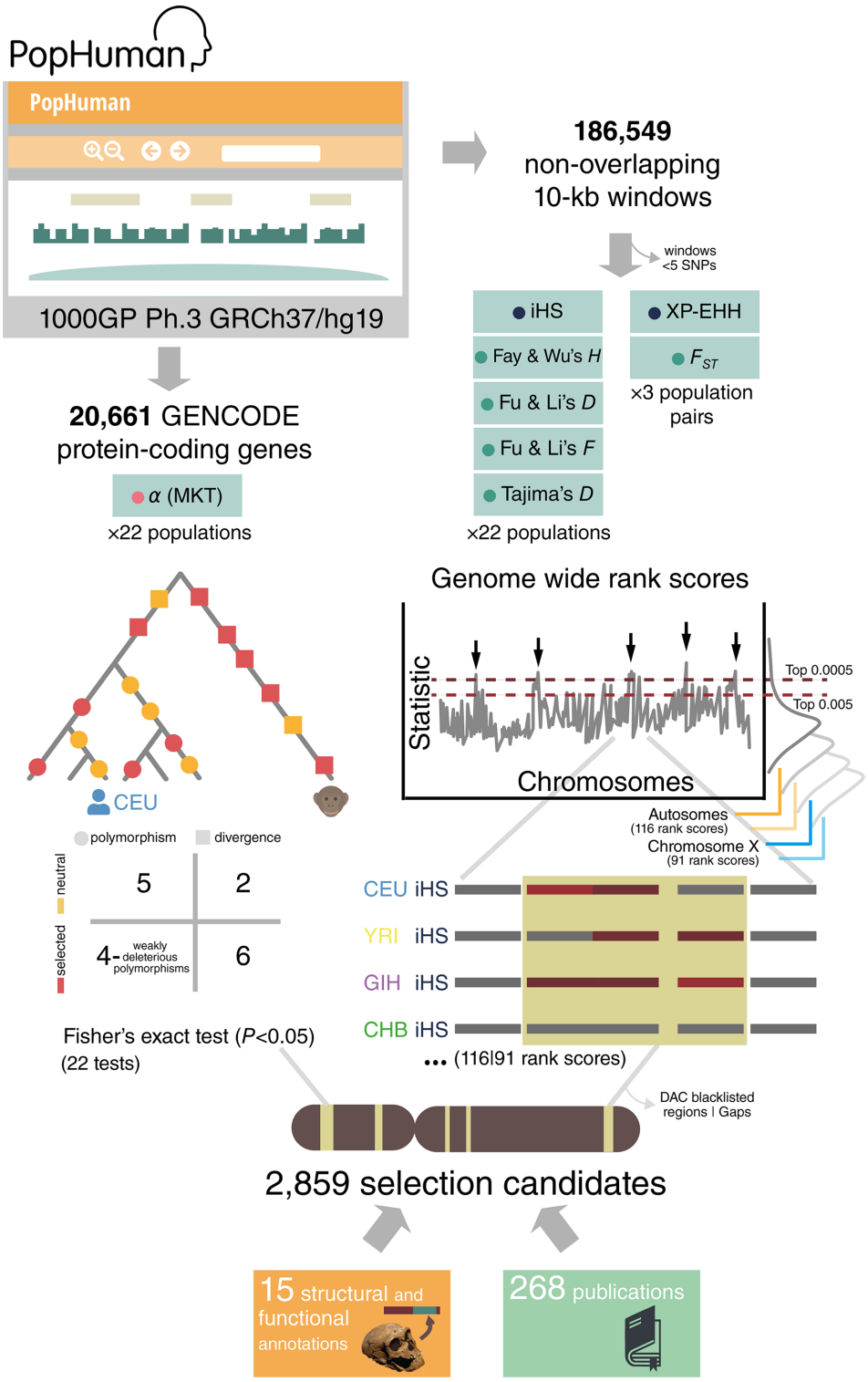

Figure 1. PopHumanScan pipeline. Starting from population genomic data retrieved from PopHuman, 8 different neutrality tests are analyzed in 22 non-admixed human populations (or 3 population pairs). Tests are color-coded depending on the type of signature they are able to detect: *Linkage Disequilibrium* (*LD*), *Site Frequency Spectrum* (*SFS*) and *Protein Changes*. The significance of each test is assessed either with a Fisher's exact test or a rank score, for each of the 22 populations (or 3 population pairs) independently, and independently for autosomes and the X chromosome. Finally, candidate regions under selection are structurally and functionally annotated, and cross-referenced with 268 publications.

divergence—*MKT* (19)—, and the result of the test is summarized with the estimator α (20,21). For this calculation, we used an MKT-based methodology that corrects for the presence of non-synonymous slightly deleterious segregating sites in order to avoid underestimating α (31).

Genome-wide scan of selection

For the parameter α of MKT, evidence of positive selection for protein-coding genes was inferred when $\alpha > 0$ and the Fisher's Exact Test for the 2×2 MKT contingency table was significant (P-value < 0.05) (Figure 1). Because the other seven selection statistics have not been associated with a simple parametric distribution, candidate windows under selection were identified as the most extreme values (within the 0.05% tail) in the corresponding empirical distribution. These empirical distributions were performed independently for each of the 22 populations (or three population pairs in the case of *XP-EHH* and F_{ST}), and independently for the autosomes and the X chromosome (to account for different demographic histories and the different effective population size of the autosomes compared to the X chromosome; chromosome Y was not analyzed). In total, 116 empirical distributions were obtained for autosomal regions, and 91 for the X chromosome (data of *iHS* and *XP-EHH* was not available for the X chromosome in PopHuman) (Figure 1).

From the initial 186 549 10-kb non-overlapping windows from PopHuman for each population and statistic, those containing <5 segregating sites were discarded (<0.2%) (Supplementary Figure S1). Then, an empirical P-value was assigned to each of the remaining windows for each of the 116 combinations of population (or population pair) and statistic, separately for the autosomes as a whole and the X chromosome. Specifically, for each window i in a population (or population pair), p is the quantile of that window for statistic j, that is, its empirical P-value. In the case of Tajima's D, Fu and Li's D and F, and Fay and Wu's H, two-tailed P-values were calculated. Once the significance for each individual 10-kb window in the genome was assessed, a candidate region under selection was defined as being a contiguous genomic region containing at least one 10-kb significant window (P-value < 0.0005) and spanning adjacent windows with P-values < 0.005 (Supplementary Figure S2). In addition, this region may span stretches <20-kb of contiguous nucleotides not analyzed in PopHuman (i.e. because they contain non-accessible bases according to the Pilot-style Accessibility Mask of the 1000GP (13,18)). This outlier approach was designed to face the unique features and limitations of our PopHuman source data and to be highly conservative defining candidate regions under selection. We expect that it likely results in an enriched set of genomic regions that have been targets of natural selection along the human evolutionary history (11), and refer to the outlier regions as candidate regions showing signatures of selection.

Once candidate selected regions (or genes) were assigned for the 22 populations (or three population pairs) and eight statistics, they were collapsed according to their coordinates into a joint set of 2879 candidate regions under selection genome-wide. Of these, 20 regions were removed because they were completely located in DAC Blacklisted regions (i.e. regions of the reference genome which are troublesome for high throughput sequencing aligners) or partially overlapped genomic gaps, as obtained from the UCSC (32). Therefore, a total of 2859 regions were finally considered.

Structural and functional annotations

The final 2859 candidate regions under selection were structurally and functionally characterized according to 15 different annotations categorized into five groups, extracted from the UCSC (32) and two publicly available databases (33,34) (Figure 1).

Sequencing. (i) *Mappability* was assessed as the percentage of bases in the region that do not present any troublesome to high-throughput sequencing aligners according to the DAC Blacklisted regions of the UCSC (35). (ii) *Distance to closest GAP* was computed as the distance (in Mb) to the closest gap (32).

Regulation. (iii) *CpG Islands* (36), (iv) *Vista Enhancers* (37), (v) *Transcription Factor Binding Sites* (TFBSs) (32) and (vi) *ORegAnno Regulatory Elements* (38) were computed as the total number of elements contained in the region, as well as the number of SNPs contained in the overlapping elements.

Comparative genomics. Evolutionary conservation of the regions was assessed by considering the results of three different algorithms—phastCons, PhyloP and GERP—on the multiple alignments of the genomes of 100 vertebrate species (39). (vii) *PhyloP Evolutionary Conservation* and (viii) *GERP Constrained Elements* were assessed as the percentage of bases in the region that have a score >2 for the given statistic (i.e. constrained sites) (32). (ix) *phastCons Evolutionary Conservation* was calculated as the percentage of bases that overlap phastCons conserved elements (32).

Structural variation. (x) *InvFEST Inversions* (34), (xi) *DGV Structural Variants* (40), (xii) *Repeat Masker* (41), (xiii) *Segmental Duplications* (42) and (xiv) *TRF Simple Tandem Repeats* (43) were assessed as the percentage of bases in the region that overlap these genomic elements.

Archaic introgression. (xv) *Archaic introgression* was assessed as the percentage of bases in the region that overlap either Neanderthal or Denisova introgressed haplotypes (33).

Published references

A total of 268 publications from 1954 to 2018 reporting either specific loci or multiple regions from a genome-wide scan of selection in the human genome were cross-referenced with our final 2859 candidate regions under selection (Figure 1 and Supplementary Table S2). Of these, 132 publications were directly extracted from the dbPSHP database (15), while the other 136 were manually curated here. Exhaustive information from the main text and/or supplementary figures and tables was extracted for each

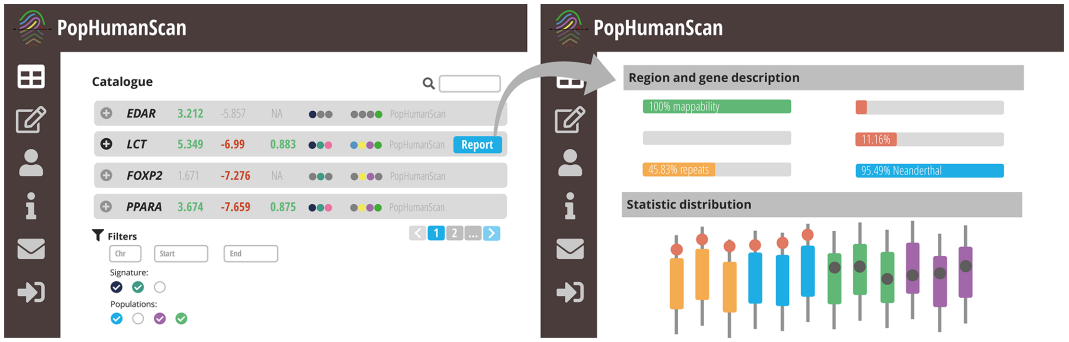

Figure 2. Simplified representation of the PopHumanScan interface. The main PopHumanScan table is displayed to the left, while the complete report for a particular candidate region under selection is displayed to the right.

reported loci, including the genomic coordinates, affected population(s), statistic(s), type of selection and PubMed ID. Genomic coordinates were lifted over to GRCh37/hg19 using the LiftOver tool of the UCSC (32), or deduced from protein-coding gene location, if necessary.

OVERVIEW OF THE POPHUMANSCAN ONLINE CATALOG

In addition to the exhaustive genome-wide selection scan that has been performed, we have also created PopHumanScan, a collaborative, online database that is aimed at compiling and annotating adaptation events along the human evolutionary history (Figure 2). PopHumanScan reports each evidence of selection with the empirical distributions of the corresponding DNA diversity statistic across the human genome and among populations, structural and functional annotations of the region, links to external databases, as well as cross-references to 268 publications.

Implementation

PopHumanScan is currently running under Apache on a CentOS 7.2 Linux x64 server with 16 Intel Xeon 2.4 GHz processors and 32 GB RAM. It is mainly built on PHP as backend framework. It also includes AJAX for specific file requests and MySQL for data storage. The client-side is build on JavaScript and uses several JavaScript libraries, including jQuery, the jQuery plugin DataTables and Plotly.js, as well as a custom Bootstrap 4 framework.

The PopHumanScan catalog

Main table. All 2859 candidate regions under selection are displayed as rows in an interactive table (Figure 2, left). The information displayed in each row includes: (i) the genomic coordinates of the candidate locus, (ii) genes contained in or partially overlapping the region (if any), (iii) the most extreme value for each of the eight statistics considered (i.e. most extreme value in any 10-kb window included in the region, for any population or population pair; green (positive) and red (negative) values are outliers ($P < 0.0005$) in the corresponding empirical distribution), (iv) color-coded dots depicting different types of selection signatures (i.e. ● *LD*, ● *SFS* and/or ● *Protein Changes*), (v) color-coded dots depicting the meta-population(s) that show signatures of selection (i.e. ● *Europe* (*EUR*), ● *Africa* (*AFR*), ● *South-Asia* (*SAS*) and/or ● *East-Asia* (*EAS*)) and (vi) the source that contributed the candidate region under selection. At the time of writing, all 2859 regions came uniquely from our genome-wide selection scan (i.e. source labeled as *PopHumanScan*), but additional data sources by contributors from the scientific community are expected once PopHumanScan is published (see next section). By clicking the ⊕ icon at the beginning of each row, detailed information of the particular candidate region under selection is displayed, including the values for all significant statistics in all target populations (or population pairs) and an overview of the main structural and functional annotations and cross-referenced publications (i.e. non-gray buttons represent overlapping annotations or cross-referenced publications), as well as access to the complete report for the corresponding candidate region under selection. At last, several filters are available at the bottom of the page to narrow the search.

Complete report. A complete report for each candidate region can be accessed from the main table (Figure 2, right). The first section of the report displays all the structural and functional annotations of the region, together with links to external databases: (i) *PopHuman* (18), which complements the population genomics information; (ii) *HaploReg* (44), which allows the exploration of evolutionary conservation, expression eQTSs, epigenomic data and regulatory annotations; and (iii) *Ensembl* (45), which allows the exploration of the LD of the region, among others. The second section lists all the genes contained in or partially overlapping the region (if any). For each encoded gene, a short description of the gene and associated *Gene Ontology* terms for the *Biological Process* classification (46) are provided, along with links to external databases: *Ensembl* (45), *NCBI* (47), *Uniprot* (48), *UCSC* (32), *Expression Atlas* (49), *OMIM* (50), *Open Targets* (51) and *HumanMine* (52). The third section contains cross-referenced publications that support the selection evidence found in the region. The fourth section contains an interactive graph showing recombination rate values in cM/Mb along the chromosome in which the region is located, calculated from the recombination map by Bhérer *et al.* (53) and extracted from PopHuman (18). The specific location of the candidate region under selection is indicated

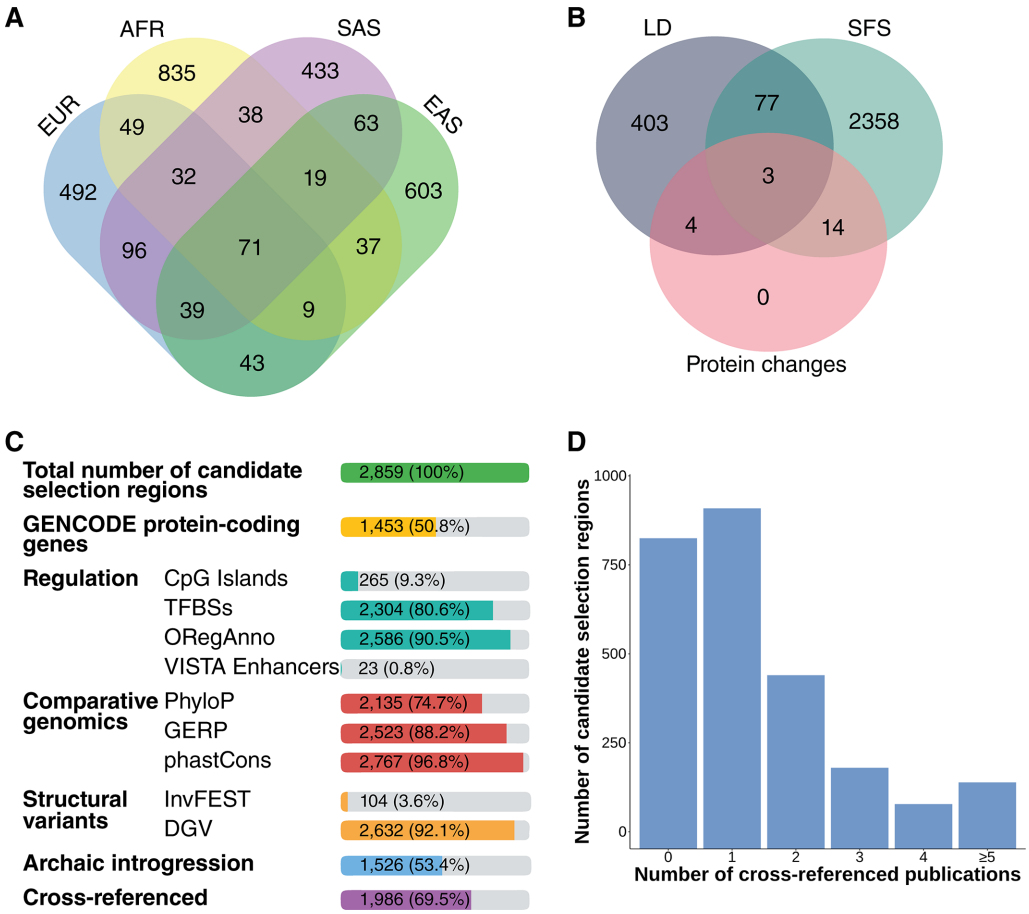

Figure 3. Summary of the contents of PopHumanScan. (**A**) Number of candidate regions under selection unique and shared among the four meta-populations: *Europe (EUR)*, *Africa (AFR)*, *Sourth-Asia (SAS)* and *East-Asia (EAS)*. (**B**) Number of candidate regions under selection unique and shared among the three different signature types: *Linkage Disequilibrium (LD)*, *Site Frequency Spectrum (SFS)* and *Protein Changes*. (**C**) Number of candidate regions under selection overlapping different structural and functional annotations. (**D**) Number of candidate regions under selection cross-referenced with 0, 1, 2, 3, 4 or ≥5 published papers.

with dashed vertical lines, and the solid horizontal line represents the average recombination rate value in the candidate region. At last, in the fifth section boxplots show the distribution of each significant statistic in all the populations (or population pairs). Highlighted values correspond to those in the candidate region, and those in red are outliers of the empirical distribution ($P < 0.0005$).

Utilities and support resources

Contributing to PopHumanScan. PopHumanScan has been devised as a collaborative database. In order to incorporate information contributed by the scientific community, two password-protected tools have been implemented. The first one allows users to add additional candidate regions under selection in the catalog. All contributed regions will be subjected to manual curation and clearly labeled with a data source tag. The second tool allows manually cross-referencing candidate regions already present in the database.

Help and tutorial. This section documents the data used and the procedures implemented in PopHumanScan, as well as instructions on how to contribute to it. Interestingly, it also contains a complete tutorial introducing to the usage of the database through a step-by-step worked example.

CONTENTS OF POPHUMANSCAN

At the time of writing, the PopHumanScan database contains 2859 candidate regions under selection derived from the genome-wide selection scan pipeline presented here. Regions are distributed homogeneously along the autosomes and the X chromosome (Table 1 and Supplementary Figure S3). Of these, 1453 regions (50.8%) overlap GENCODE protein-coding genes, and 1986 regions (69.5%) are cross-referenced with at least one publication (Table 1 and Figure 3).

Selection signatures in meta-populations

The total number of candidate regions showing signatures of selection in the four meta-populations is: 831 (29.1%) in EUR, of which 413 (49.7%) overlap protein-coding genes; 1090 (38.1%) in AFR, of which 580 (53.2%) overlap protein-coding genes; 791 (27.7%) in SAS, of which 401 (50.7%)

Table 1. Summary of the candidate regions under selection included in PopHumanScan

		Regions with selection signatures in meta-populations				Regions with different types of signatures				
Chromosome	Number of candidate regions	*European* (EUR)	*African* (AFR)	*South-Asian* (SAS)	*East-Asian* (EAS)	*Linkage disequilibrium (LD)*	*Site Frequency Spectrum (SFS)*	*Protein Changes*	Regions overlapping protein-coding genes	Regions cross-referenced with publications
1	214	57	84	48	66	46	176	2	123	152
2	253	77	95	60	81	56	203	2	131	191
3	201	61	81	49	54	34	173	0	116	153
4	241	62	110	68	59	42	207	1	111	173
5	166	46	62	41	52	32	140	1	85	119
6	171	42	83	42	49	29	146	1	81	118
7	144	53	58	49	49	25	125	0	74	115
8	164	41	57	51	48	32	135	0	68	108
9	96	28	33	22	34	12	85	1	51	59
10	141	45	47	45	48	21	126	1	70	103
11	120	35	37	34	44	22	99	2	62	86
12	142	44	52	35	42	24	123	1	87	105
13	82	17	31	27	21	9	75	0	23	51
14	74	19	29	25	22	15	61	0	39	57
15	79	34	29	17	25	12	70	1	47	57
16	88	20	20	41	30	21	68	1	49	69
17	86	30	32	34	24	18	71	1	64	69
18	56	15	23	14	18	7	49	0	24	43
19	67	21	29	21	27	5	62	4	42	38
20	59	19	22	14	23	10	52	1	29	49
21	34	13	11	10	16	7	28	0	11	26
22	39	8	16	8	15	8	35	1	27	32
X	142	44	49	36	37	ND	142	0	39	13
TOTAL	**2859**	**831**	**1090**	**791**	**884**	**487**	**2451**	**21**	**1453**	**1986**
		29.1%	***38.1%***	***27.7%***	***30.9%***	***17.0%***	***85.7%***	***0.7%***	***50.8%***	***69.4%***

ND = Not Determined

overlap protein-coding genes; and 884 (30.9%) in EAS, of which 424 (48.0%) overlap protein-coding genes (Table 1). Most of the regions (82.5%) show signatures that are unique to one single meta-population (Figure 3A): 492 (17.2%) show signatures that are unique in EUR, 835 (29.2%) are unique in AFR, 433 (15.1%) are unique in SAS and 603 (21.1%) are unique in EAS. Of the 1090 regions showing signatures in AFR, 76.6% are unique to AFR; while a lesser percentage—59.2, 54.7 and 68.2%—of the regions showing signatures in EUR, SAS and EAS, respectively, are unique to their meta-population. About one third (29.0%) of the candidate regions under selection are shared across populations within the same meta-population. This percentage is higher for candidate regions showing both LD and SFS signatures (52.7%), it is 33.6% for candidate regions showing LD signatures only and 27.1% for candidate regions showing SFS signatures only.

Types of selection signature

The total number of candidate regions showing distinct types of signatures of selection is: 487 (17.0%) for *LD*; 2451 (85.7%) for *SFS*; and 21 (0.7%) for *Protein Changes* (i.e. recurrent selection since the split between humans and chimpanzees) (Table 1). Most of the regions (96.6%) show one single signature of selection (Figure 3B): 403 (14.1%) show LD signatures only; and 2358 (82.5%) show SFS signatures only. All genes showing evidence of recurrent selection also show signatures in either LD and/or SFS, as only genes overlapping candidate regions under selection detected by LD and/or SFS were tested for α (MKT). These results would indicate that the statistics we used in our genome-wide scan of selection look at different characteristics of the genetic variability of the region, and that they are largely complementary.

Structural description of the regions

Region length. Most of the candidate regions under selection (63.6%) span one single 10-kb window, and the variable *lengths of candidate regions* follows a reversed J-shaped distribution (Supplementary Figure S4A).

Distance between consecutive regions. The average distance between consecutive candidate regions is ∼1 Mb, and the distribution of distances is also reversed J-shaped (Supplementary Figure S4B).

Recombination. The average recombination rate of the candidate regions is 0.71 cM/Mb, and the distribution of recombination rates is again reversed J-shaped (Supplementary Figure S4C). There is a strong, negative, non-linear association between recombination rate and both region length (Supplementary Figure S5A) and distance between consecutive candidate regions (Supplementary Figure S5B).

Functional description of the regions

Regulation. Most of the candidate regions (90.5%) contain at least one regulatory element annotated in the ORegAnno database, and 80.6% contain TFBSs (Figure 3C). On the contrary, VISTA enhancers are much less abundant in the genome and they are only found in 23 of the 2859 candidate regions (0.8%). CpG Islands are also in shortage and they are present in 9.3% of the regions.

Comparative genomics. Nearly all (96.8%) candidate regions overlap phastCons conserved elements. In the case of GERP and PhyloP, 88.2 and 74.7% of the regions, respectively, overlap constrained bases with score >2.

Structural variation. The Database of Genomic Variants (DGV) (40) is a very exhaustive database of structural variants annotated in the human genome. One or more elements annotated in this database are present in 92.1% of the candidate regions under selection. On the contrary, only 104 regions (3.6%) overlap validated polymorphic inversions from the manually curated InvFEST database (34).

Archaic introgression. A total of 1526 of the candidate regions (53.4%) overlap haplotypes introgressed from either neanderthals or denisovans. This percentage is expected, as introgressed haplotypes persisting in different present-day human individuals cover 46.7% of the reference genome (33).

Cross-references with publications. A percentage of 69.5% of the candidate regions are cross-referenced with at least one publication, and 36.0% are cross-referenced more than once (Figure 3D).

Gene ontology analysis

Our candidate regions overlap a total of 1447 unique GENCODE protein-coding genes. These were functionally classified into Gene Ontology (GO) terms (46) according to the PANTHER GO-Slim annotation dataset using the PANTHER Classification System (54) (Supplementary Figure S6). In addition, statistically over- and under-represented functions were analyzed using the complete GO annotation dataset (46) using the same tool (Supplementary Tables S3–5). Interestingly, among all Biological Process categories, *regulation of neuron projection development* is over-represented (fold enrichment 1.88, False Discovery Rate (FDR) 1.23E-02), in addition to *cellular component organization* (fold enrichment 1.24, FDR 1.59E-03) (Supplementary Table S4). Finally, several Cellular Component categories are statistically over-represented, including *presynaptic membrane* (fold enrichment 2.72, FDR 4.32E-02) (Supplementary Table S5). In spite of finding some statistically over-represented GO categories in our genes list, selection signatures seem to be heterogeneous and a detailed analysis of each candidate region is required to understand the real story under each selective event.

POPHUMANSCAN WITH AN EXAMPLE: SELECTION AT THE LACTASE LOCUS

The introduction of agriculture and cattle domestication in the Middle East and North Africa ∼10 000 years ago lead to strong selection pressure for the ability to digest milk as adults. This is accomplished if the enzyme lactase that metabolizes lactose, encoded by the *LCT* gene, maintains high levels into adulthood, a characteristic that is called lactase persistence. Several variants near the *LCT* locus show some of the strongest signals of selection in the human genome for those populations that have traditionally practiced dairying, including a genetic variant in an intron of the gene *MCM6*, upstream of *LCT* (3).

The *LCT* locus is found inside the longest candidate region under selection reported in PopHumanScan (∼1 Mb). The region is located in the long arm of chromosome 2 and contains 8 GENCODE protein-coding genes, including *LCT* and *MCM6* (Figure 4). Our genome-wide scan of selection has detected signatures at four different statistics that span the three types of signatures: LD (*iHS* and *XP-EHH*), SFS (Fu and Li's *D*) and Protein Changes (α). LD signatures involve basically EUR and AFR populations, while the signature of recurrent selection is more general to the four meta-populations. The region contains thousands of TFBSs and hundreds of ORegAnno regulatory elements, it overlaps evolutionary constrained elements, and >95% of the region overlaps haplotypes introgressed from Neanderthals. It has been reported in 24 published articles (of the set of 268 that we considered).

CONCLUSION

In summary, our exhaustive approach combining eight different statistics to detect candidate regions under selection in 22 non-admixed human populations has been able to locate distinct signatures in 2859 regions that stand out from the background genomic variability, including abnormally long haplotypes, shifts in the SFS or excess of non-synonymous substitutions between our species and chimpanzees. Many of these regions probably manifest the footprints of selective sweeps that occurred at different historical ages, or recurrent selection that has been taking place during the last millions of years. The PopHumanScan online database is going to facilitate the thorough analysis of candidate regions under selection in the human genome by putting together all these evidences of selection with structural and functional annotations of the regions and cross-references to previously published articles. Furthermore, the database can incorporate new data from the scientific community through specific build-in utilities. All in all, PopHumanScan aims to become a central repository to share information, guide future studies and contribute to the research on human genome adaptation.

DATA AVAILABILITY

Scripts for the PopHumanScan analysis pipeline are available as Jupyter Notebooks at https://github.com/BGD-UAB/PopHumanScan. All data, tools and support resources provided by the PopHumanScan database are freely

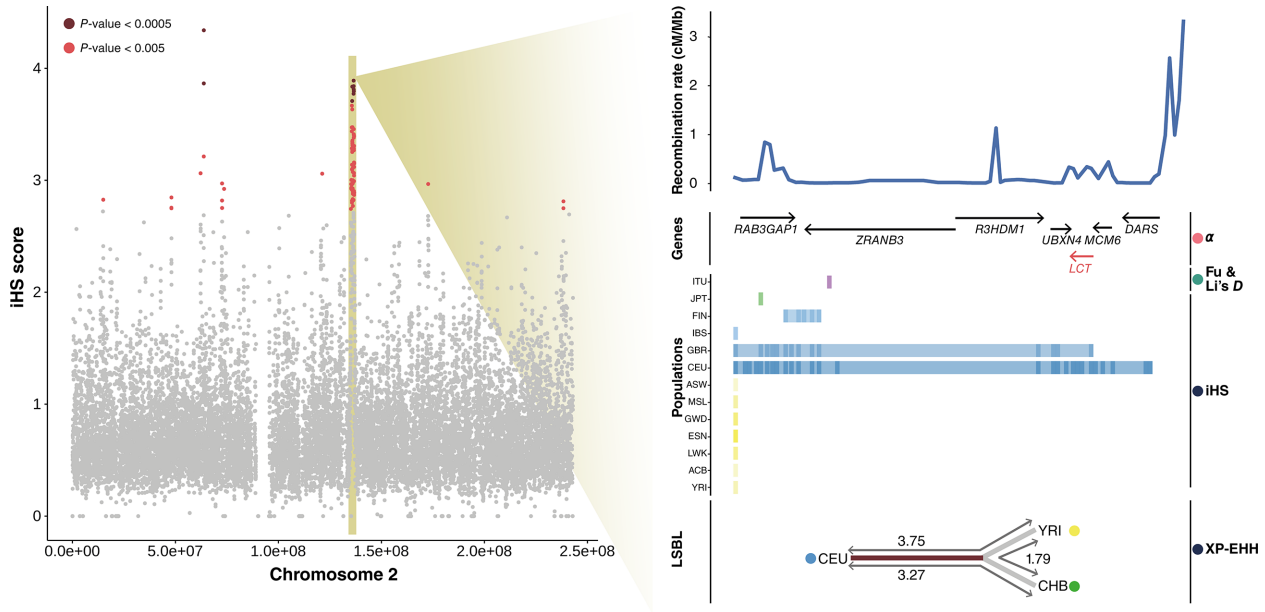

Figure 4. Signatures of selection detected at the lactase locus. The distribution of *iHS* values for the CEU population in 10-kb windows along chromosome 2 are displayed to the left; windows with a *P*-value < 0.0005 or *P*-value < 0.005 in the empirical distribution are highlighted. The candidate region under selection including the *LCT* gene is zoomed-in to the right, where all significant signatures at four different statistics spanning three different signature types are represented: *Protein Changes*, *Site Frequency Spectrum (SFS)* and *Linkage Disequilibrium (LD)*. Signatures in each population are colored according to its meta-population: *Europe (EUR)*, *Africa (AFR)*, *South-Asia (SAS)* and *East-Asia (EAS)*.

available at https://pophumanscan.uab.cat. Log-in information to contribute data to PopHumanScan is available upon request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Carla Giner for helpful comments on the PopHumanScan data and implementation, and Esteve Sanz for help with the informatics infrastructure in which PopHumanScan is implemented. We also thank two anonymous referees for very helpful comments on the PopHumanScan implementation and manuscript.

FUNDING

Ministerio de Economía y Competitividad (Spain) [CGL2017-89160P to Mauro Santos, A.B.]; AGAUR (Generalitat de Catalunya) [2017SGR-1379 to Alfredo Ruiz]; Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement (Generalitat de Catalunya) [FI-DGR2015 to M.C-.Z.]; Departament de Genètica i de Microbiologia (UAB) [PIF to J.M-.M.]; Servei de Genòmica i Bioinformàtica de la UAB. Funding for open access charge: Ministerio de Economía y Competitividad (Spain).
Conflict of interest statement. None declared.

REFERENCES

1. Nielsen,R., Akey,J.M., Jakobsson,M., Pritchard,J.K., Tishkoff,S. and Willerslev,E. (2017) Tracing the peopling of the world through genomics. *Nature*, **541**, 302–310.
2. Racimo,F., Sankararaman,S., Nielsen,R. and Huerta-Sánchez,E. (2015) Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.*, **16**, 359–371.
3. Fan,S., Hansen,M.E.B., Lo,Y. and Tishkoff,S.A. (2016) Going global by adapting local: a review of recent human adaptation. *Science*, **354**, 54–59.
4. 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.
5. Hinds,D.A., Stuve,L.L., Nilsen,G.B., Halperin,E., Eskin,E., Ballinger,D.G., Frazer,K.A. and Cox,D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
6. International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
7. International HapMap Consortium, Frazer,K, Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
8. International HapMap 3 Consortium, Altshuler,D.M., Gibbs,R.A., Peltonen,L., Altshuler,D.M., Gibbs,R.A., Peltonen,L., Dermitzakis,E., Schaffner,S.F., Yu,F. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
9. 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.
10. Akey,J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.*, **19**, 711–722.
11. 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.

12. 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.
13. The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
14. Johnson,K.E. and Voight,B.F. (2018) Patterns of shared signatures of recent positive selection across human populations. *Nat. Ecol. Evol.*, **2**, 713–720.
15. Li,M.J., Wang,L.Y., Xia,Z., Wong,M.P., Sham,P.C. and Wang,J. (2014) dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Res.*, **42**, D910–D916.
16. Pybus,M., Dall'Olio,G.M., Luisi,P., Uzkudun,M., Carreño-Torres,A., Pavlidis,P., Laayouni,H., Bertranpetit,J. and Engelken,J. (2014) 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.*, **42**, D903–D909.
17. Pybus,M., Luisi,P., Dall'Olio,G.M., Uzkudun,M., Laayouni,H., Bertranpetit,J. and Engelken,J. (2015) Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, **31**, 3946–3952.
18. Casillas,S., Mulet,R., Villegas-Mirón,P., Hervas,S., Sanz,E., Velasco,D., Bertranpetit,J., Laayouni,H. and Barbadilla,A. (2018) PopHuman: the human population genomics browser. *Nucleic Acids Res.*, **46**, D1003–D1010.
19. McDonald,J.H. and Kreitman,M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 652–654.
20. Charlesworth,B. (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.*, **63**, 213–227.
21. Smith,N.G. and Eyre-Walker,A. (2002) Adaptive protein evolution in Drosophila. *Nature*, **415**, 1022–1024.
22. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
23. Casillas,S. and Barbadilla,A. (2017) Molecular Population Genetics. *Genetics*, **205**, 1003–1035.
24. 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.
25. Shriver,M.D., Kennedy,G.C., Parra,E.J., Lawson,H.A., Sonpar,V., Huang,J., Akey,J.M. and Jones,K.W. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum. Genomics*, **1**, 274–286.
26. Fay,J.C. and Wu,C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.
27. Fu,Y.X. and Li,W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.
28. Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
29. Wright,S. (1950) Genetical structure of populations. *Nature*, **166**, 247–249.
30. Hudson,R.R., Slatkin,M. and Maddison,W.P. (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.
31. Mackay,T.F.C., Richards,S., Stone,E.A., Barbadilla,A., Ayroles,J.F., Zhu,D., Casillas,S., Han,Y., Magwire,M.M., Cridland,J.M. *et al.* (2012) The Drosophila melanogaster genetic reference panel. *Nature*, **482**, 173–178.
32. Casper,J., Zweig,A.S., Villarreal,C., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Karolchik,D. *et al.* (2018) The UCSC genome browser database: 2018 update. *Nucleic Acids Res.*, **46**, D762–D769.
33. Vernot,B., Tucci,S., Kelso,J., Schraiber,J.G., Wolf,A.B., Gittelman,R.M., Dannemann,M., Grote,S., McCoy,R.C., Norton,H. *et al.* (2016) Excavating neandertal and denisovan DNA from the genomes of melanesian individuals. *Science*, **352**, 235–239.
34. Martínez-Fundichely,A., Casillas,S., Egea,R., Ràmia,M., Barbadilla,A., Pantano,L., Puig,M. and Cáceres,M. (2014) InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res.*, **42**, D1027–D1032.
35. Derrien,T., Estellé,J., Marco Sola,S., Knowles,D.G., Raineri,E., Guigó,R. and Ribeca,P. (2012) Fast computation and applications of genome mappability. *PLoS One*, **7**, e30377.
36. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
37. Pennacchio,L.A., Ahituv,N., Moses,A.M., Prabhakar,S., Nobrega,M.A., Shoukry,M., Minovitsky,S., Dubchak,I., Holt,A., Lewis,K.D. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.
38. Lesurf,R., Cotto,K.C., Wang,G., Griffith,M., Kasaian,K., Jones,S.J.M., Montgomery,S.B., Griffith,O.L. and Open Regulatory Annotation Consortium. (2016) ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.*, **44**, D126–D132.
39. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
40. MacDonald,J.R., Ziman,R., Yuen,R.K.C., Feuk,L. and Scherer,S.W. (2014) The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.
41. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.
42. Bailey,J.A., Gu,Z., Clark,R.A., Reinert,K., Samonte,R.V., Schwartz,S., Adams,M.D., Myers,E.W., Li,P.W. and Eichler,E.E. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.
43. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
44. Ward,L.D. and Kellis,M. (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.*, **44**, D877–D881.
45. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
46. The Gene Ontology Consortium. (2017) Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.
47. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
48. UniProt Consortium,T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, D158–D169.
49. Papatheodorou,I., Fonseca,N.A., Keays,M., Tang,Y.A., Barrera,E., Bazant,W., Burke,M., Füllgrabe,A., Fuentes,A.M.-P., George,N. *et al.* (2018) Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.*, **46**, D246–D251.
50. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
51. Koscielny,G., An,P., Carvalho-Silva,D., Cham,J.A., Fumis,L., Gasparyan,R., Hasan,S., Karamanis,N., Maguire,M., Papa,E. *et al.* (2017) Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.*, **45**, D985–D994.
52. Lyne,R., Sullivan,J., Butano,D., Contrino,S., Heimbach,J., Hu,F., Kalderimis,A., Lyne,M., Smith,R.N., Štěpán,R. *et al.* (2015) Cross-organism analysis using InterMine. *Genes*, **53**, 547–560.
53. Bhérer,C., Campbell,C.L. and Auton,A. (2017) Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat. Commun.*, **8**, 14994.
54. Mi,H., Huang,X., Muruganujan,A., Tang,H., Mills,C., Kang,D. and Thomas,P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189.