Published online 3 September 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 3 **1**
doi: 10.1093/nargab/lqaa061

A fully integrated machine learning scan of selection in the chimpanzee genome

Jessica Nye[1], Mayukh Mondal[1], Jaume Bertranpetit[1,*] and Hafid Laayouni [1,2,*]

[1]Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain and [2]Bioinformatics Studies , ESCI-UPF, Pg. Pujades 1, 08003 Barcelona, Spain

Received November 03, 2019; Revised June 11, 2020; Editorial Decision July 18, 2020; Accepted July 31, 2020

ABSTRACT

After diverging, each chimpanzee subspecies has been the target of unique selective pressures. Here, we employ a machine learning approach to classify regions as under positive selection or neutrality genome-wide. The regions determined to be under selection reflect the unique demographic and adaptive history of each subspecies. The results indicate that effective population size is important for determining the proportion of the genome under positive selection. The chimpanzee subspecies share signals of selection in genes associated with immunity and gene regulation. With these results, we have created a selection map for each population that can be displayed in a genome browser (www.hsb.upf.edu/chimp_browser). This study is the first to use a detailed demographic history and machine learning to map selection genome-wide in chimpanzee. The chimpanzee selection map will improve our understanding of the impact of selection on closely related subspecies and will empower future studies of chimpanzee.

INTRODUCTION

Although chimpanzees are the closest genetic relative to humans, sharing much of our genetic information, we still understand little about their evolutionary history. In recent years, a more comprehensive picture of their demographic history has been elucidated (1–3). Chimpanzees and bonobos diverged around a million years ago. The two major branches of the chimpanzee lineage began to split from each other some 500 thousand years ago (kya). Today, we identify four subspecies of chimpanzee, and understand that *Pan troglodytes verus* and *P.t. ellioti* are more closely related and diverged from each other first (∼250 kya), followed by *P.t. troglodytes* and *P.t. schweinfurthii* (diverging <150 kya). Although the four subspecies are genetically and geographically distinct, it is clear that there has been an extensive gene flow among chimpanzee subspecies and introgression with bonobo (4).

The demographic history of a species is an important key in understanding their evolution. The size of a population's breeding pool can indicate how many positive selective events are likely to have taken place (5) and how strong the effects of genetic drift may in fact be (6). Furthermore, introgression between species and gene flow from a close subspecies may be a source for beneficial genetic material, as reviewed by Arnold and Martin (7).

Beyond demography, unique selective events are likely to have impacted the genomes of chimpanzee based on their habitats. The four subspecies live in two distinct regions of Africa. *Pan troglodytes verus* (western chimpanzee) lives in Ivory Coast and Guinea, while the other three subspecies live in central Africa. Specifically, *P.t. ellioti* (Nigeria–Cameroon) lives in its namesake countries, *P.t. schweinfurthii* (eastern) inhabits seven countries but primarily in the Democratic Republic of Congo and *P.t. troglodytes* (central) inhabits five countries but primarily Gabon. The selective pressures that these populations face are likely due to their unique habitats. In different regions, these populations will experience exposure to divergent pathogens, differing quantity and quality of resources, and, most importantly, for a social animal, separate cultures. All these factors together are likely to be responsible for some of the genetic differences we observe between chimpanzees.

Here, we present the first comprehensive scan of the chimpanzee genome that integrates varied selective simulations that encompass complete and ongoing selection occurring between present time and some 60 kya. These simulations are interrogated by 15 statistical tests, and with a random forest machine learning approach we map positive selection to better understand the unique evolutionary history of our genetic cousins, unveiling their adaptive history through the unique or shared signals of positive selection.

*To whom correspondence should be addressed. Tel: +1 34 93 316 0845; Email: hafid.laayouni@upf.edu
Correspondence may also be addressed to Jaume Bertranpetit. Email: jaume.bertranpetit@upf.edu
Present address: Mayukh Mondal, Institute of Genomics, University of Tartu, Riia 23b Tartu, 51010 Tartu, Estonia.

© The Author(s) 2020. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

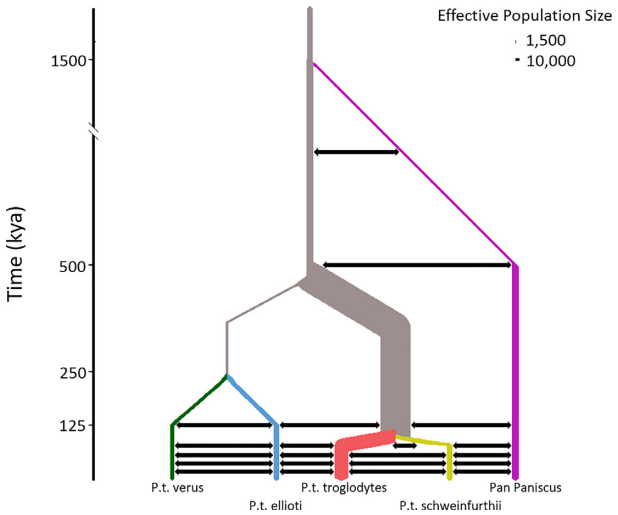

Figure 1. Demographic model indicating population divergence times, effective population sizes and migration within the *Pan* clade. Migration (black arrows) indicates bidirectional migrational pulses over time. Population size is indicated in the upper right-hand corner.

MATERIALS AND METHODS

Genome sequences

Full genome sequences of chimpanzee and bonobo were obtained from the Great Ape Genome Project (3,4) as vcf files aligned to chimpanzee genome release Pantro4. The sample sizes are 18 *P.t. troglodytes*, 19 *P.t. schweinfurthii*, 10 *P.t. ellioti*, 12 *P.t. verus* and 10 *P. paniscus*. The ancestral states for each single-nucleotide variant (SNV) were extracted from the 1000 Genomes data (8) using the human–gorilla–chimpanzee alignment from Ensemble release 54. The data were pruned to exclude missing sites, missing ancestral information, introgressed regions (4), and insertions or deletions for a total of 1 022 493 SNVs. The data were phased using Shapeit (9) (for additional information, see Section 1.1 in the Supplementary Data).

Demographic model

We used a demographic model adapted from (4) (the formulation of the demographic model is explained in detail in Section 1.2 in the Supplementary Data, Supplementary Figure S1 and Supplementary Table S1). Briefly, the demographic model (Figure 1) was adapted to include all four subspecies of chimpanzee and bonobo with all meaningful admixture events, including migration and introgression. The model was additionally adapted to account for the long-term effective population sizes for each subspecies before the bottleneck at current time, as in (10), resulting in the current estimates for effective population sizes as 39 925 for *P.t. troglodytes*, 12 829 for *P.t. schweinfurthii*, 12 364 for *P.t. ellioti* and 10 742 for *P.t. verus*. *Pan troglodytes troglodytes* and *P.t. schweinfurthii* diverged from each other 136 350 ya, *P.t. ellioti* and *P.t. verus* diverged 498 462 ya and the ancestors of the two major branches of chimpanzee split 512 050 ya.

The selection of the demographic model is important for a study of this type. Luckily, our demographic model has been estimated from the full genome sequences included in this study. This results in the fact that the used demography is the most accurate depiction of the complex relationships between analyzed subspecies as estimated from the analyzed samples.

Simulations

We used the coalescent simulator msms (11). For the neutral simulations, we simulated 2000 replicates of 600 000 bp sections. We matched the sample sizes of each subspecies resulting in sample sizes of $n = 18$ for *P.t. troglodytes*, $n = 19$ for *P.t. schweinfurthii*, $n = 10$ for *P.t. ellioti* and $n = 12$ for *P.t. verus*. Each subspecies was modeled independently because each subspecies is an independent lineage where selection and divergence have been impacting each group uniquely for >5000 generations.

For the selective scenarios, we used the demographic model introduced above. We chose to simulate selection at the following generation time points: 600, 900, 1200, 1500, 1800, 2100 and 2400 (generation time is 25 years), in order to take advantage of the predicted power of positive selection statistics, discussed below. To do this, we used the tag -e. We allowed for hard selection (where selection begins on a singleton in the populations) using the tag -SAA.

We allowed for the selection coefficient to vary between 0.05 and 0.55. These coefficients were chosen to combat the power of drift. As genetic drift is a more powerful evolutionary force than selection in small populations, the short timescale of the most recent selection events required the use of high coefficients. We independently selected 1000 simulations where the final allele frequency (FAF) was fixed (complete sweep) and 1000 simulations where the FAF was between 0.6 and <1 (incomplete) for each subspecies using the tag -oTrace. We set the selected allele in the center of the 600 000 bp region, resulting in a total of 7000 hard incomplete and 7000 hard complete, or 14 000 simulations per subspecies of length 600 000 bp. The full msms code is presented in Section 1.2 in the Supplementary Data.

Statistical tests

In order to differentiate regions under selection, we calculated a suite of statistical tests based on site frequency spectrum, linkage disequilibrium, descriptive statistics and population differentiation (Table 1). We chose 15 statistics based on the results from (25), which employed a machine learning approach to search and classify selection in the human genome using a combination of signals from various statistics. All statistics were calculated genome-wide for both real and simulated datasets. The window-based statistics were calculated in windows of 30 kb with a 3 kb sliding window. Windows were dropped if there were <5 SNVs in the window, to avoid the possibility of poor statistical power in that area. The window- and SNV-based statistics were combined by selecting the median value of each SNV-based statistic per window. All statistics were calculated using scripts provided by Pybus *et al.* (25). Only windows in the center of the simulated 600 000 bp regions that contained the selected allele were used for the final model. The genome-wide real data and the neutral simulations for each

Table 1. Tests calculated in the chimpanzee genomes and used to train the random forest algorithm in order to calculate a composite score of selection

Principle	Method
Site frequency spectrum	Tajima's D (12)
	Fu and Li's D (13)
	Fu and Li's F (13)
	R_2 (14)
Linkage disequilibrium	iHS (15)
	$EHH_{Average}$ (16)
	Wall's B (17)
	Wall's Q (18)
	Fu's F (19)
	Z_a (20)
	Z_{nS} (21)
	ZZ (20)
Population differentiation	ΔDAF (22)
	XP-EHH (23
Descriptive statistics	π (nucleotide diversity) (24)

statistic are presented in Supplementary Figures S2–S22, and for simulated neutral and selective scenarios in Supplementary Figures S23–S43.

Random forest algorithm

We employed a machine learning approach in order to differentiate between regions of the genome, split into 30 kb windows with a 3 kb sliding window, that are neutral and regions that have evidence of a selective sweep. Random forest uses decision trees as a base classifier. This learning method is used for classification and regression of data. We chose a random forest method because it is an improvement on previous machine learning approaches such as decision trees (which run the risk of overparameterization when using correlated input statistics) (26,27) or bagging (i.e. bootstrap aggregating, which runs the risk of basing the regression model on too many similar trees) (28). Our input data consist of many correlated statistics (Figure 2), all of which have benefits and disadvantages. The random forest algorithm is an extension of bagging, in that it constructs an entire forest with trees of random structure at each node by randomly selecting training instances for each tree (i.e. each tree is trained with a different set of simulations and a unique random subset of the calculated statistics). This ensures that the ultimate regression model is based on a sufficient mixture of the underlying data that avoids both overparameterization and bias being built into the underlying model (29). New statistical advances are using new methods to extract information from complex dataset; for example, approximate Bayesian computation is used to infer past demographic events from genome-wide sequencing data (30) and deep learning methods are applied to identify the best demographical models (31) or to jointly infer natural selection and demography (32). The random forest model has been observed to be particularly apt for use in computationally difficult problems like selection in genomes (32,33) or models where single predictors have little power (34). For more information, see Section 1.3 in the Supplementary Data.

We used the R package randomForest version 4.6-14 (35) in R version 3.5.2 (36). The model was trained with the 15 calculated statistics for both the neutral simulations and the selective simulations. We modeled extensive selection scenarios occurring between present time and 60 kya, when the test statistics are robust (8–9,11–21). We grew forests of 5000 random trees; all other parameters were set to default. The measures of accuracy for random forest algorithms are out-of-bag (OOB) error rates. These values are calculated by randomly selecting a subset of the statistics for random simulated regions to train and test the accuracy of the model based on the non-sampled instances. OOB error rates are similar to the more common cross-validation procedures, as they are both bootstrapping methods. An OOB rate indicates the probability that the instance belongs to any class. The random forest was modeled independently for each subspecies.

According to our OOB error estimates, ∼500 trees are adequate (Supplementary Figures S45–S48). The output of a random forest is the majority voting score. Each randomly generated tree categorizes the input region; that is, an output score of 0.60 selection and 0.40 neutral indicates that 60% of the 5000 trees categorized the input region as under selection while 40% of the trees categorized the input region as neutral. In order to be as conservative as possible, we accepted only signals of putative positive selection where at least 95% of trees categorized the region as under selection. This cutoff is not the same as other cutoffs like a false discovery rate (FDR) in which the study accepts an error rate of 5%. In the case of random forests, each tree is trained with a unique combination of the underlying parameters, which means that not all trees have been trained with the correct scenarios to detect selection in every place in the genome; in other words, this 95% cutoff allows for 'badly' designed trees to be ignored. Error rates are calculated separately using OOB (Supplementary Table S6). Our final model is a function of the 15 selection tests that give, for each window, a value that tells what portion of the trees in the forest predicts that positive selection has acted in that given window.

Negative selection

To remove false positives due to background selection, we perform the McDonald–Kreitman (MK) test (37) using the program PopGenome version 2.7.1 (38) in R version 3.5.2 (36). We split the genome into gene regions using Ensembl release 90 annotations. The outgroup was set as bonobo and all sample sizes were matched ($N = 10$).

Gene ontology

Gene ontology (GO) analysis was performed using PANTHER classification system version 14.1 (39) with Ensemble 95 annotations. In all cases, gene lists were uploaded and analyzed using the *P. troglodytes* organism annotations and statistical overrepresentation test using a Fischer's exact test and correction by FDR.

Genes under selection

To extract the genes within windows under selection, we used intersectBed from BEDTools version 2.28.0 (40) using

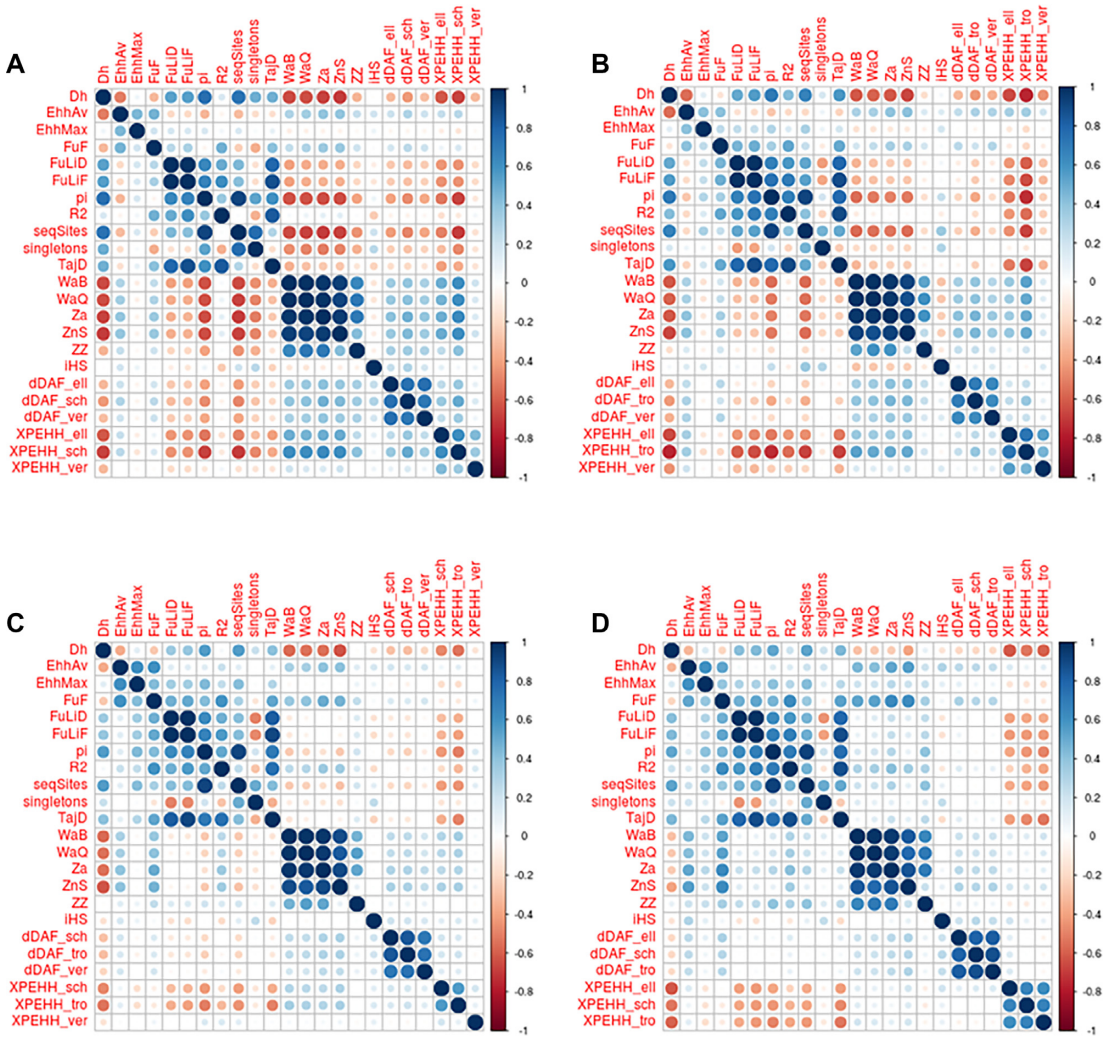

Figure 2. Correlation between calculated statistics for each subspecies: (**A**) *P.t. troglodytes*, (**B**) *P.t. schweinfurthii*, (**C**) *P.t. ellioti* and (**D**) *P.t. verus*.

the .gtf file from Ensembl release 90. The gene overlap was plotted using jvenn (41). We then used the desktop version of the Ensembl variant effect predictor [release 93 (42)] to extract possible functional variants under selection. We extract all sites with predicted impact high, moderate and low. We extract allele frequencies for each variant site for all four subspecies in order to assess differences among them.

RESULTS

Regions identified using the random forest machine learning approach

We trained a machine learning algorithm to discern between regions of the genome that have been evolving neutrally or experienced a positive selective sweep at some point in the recent past (up to 60 kya). We calculated 15 statistics (Table 1 and Supplementary Figures S2–S43) based on site frequency spectrum, linkage disequilibrium and population differentiation. These statistics were calculated in 30 kb regions genome-wide and within neutral and selection simulations based on the demographic history of chimpanzees (see Section 1.2 in the Supplementary Data and Supplementary Figure S1 for further information). We used the simulations to train a random forest algorithm; see the 'Materials and Methods' section for additional information. We observe a low OOB error rate for each subspecies (*P.t. verus* = 2.94%, *P.t. ellioti* = 2.62%, *P.t. schweinfurthii* = 1.26% and *P.t. troglodytes* = 1.09%), indicating that we are able to discern between selective and neutral regions, based on selective simulations (see Supplementary Figures S44–S51 and Supplementary Tables S2–S5 for subspecies-specific and test-specific performances). For all four subspecies, the model has higher false negative (i.e. incorrectly categorizing as neutral) rates than false positive (i.e. incorrectly categorizing as selection) in all cases; the false positive rate is <1% (Supplementary Table S6). The false negatives indicate the power to detect selection (calculated as 1 − false negative error rate) is dependent on our study sample sizes (*P.t. verus* = 86.75%, *P.t. ellioti* = 88.00%, *P.t. schweinfurthii* = 93.17% and *P.t. troglodytes* = 93.95%).

After applying the regression model based on the training set, we compute a prediction for each window in the genome for each subspecies of chimpanzee. In order to re-

Table 2. Regions of the genome for each subspecies of chimpanzee identified as under selection (P.t.v. is *P.t. verus*, P.t.e. is *P.t. ellioti*, P.t.s. is *P.t. schweinfurthii* and P.t.t. is *P.t. troglodytes*)

Subspecies	Regions under selection	Relative genome proportion	Total number of genes	Number of unique genes	Proportion of regions without genes
P.t.v.	373	0.48%	302	257	54.4%
P.t.e.	322	0.42%	381	322	28.6%
P.t.s.	328	0.58%	356	269	33.2%
P.t.t.	743	1.11%	823	694	29.1%

main confident that we are extracting true signatures of positive selection, we test whether these regions are the target of background selection. Negative selection is a demography-free process (43) that can cause similar reductions in genetic diversity, which may be falsely picked up in our positive selection scan. While previous research indicates that positive selection is needed to explain the reduced genetic diversity around gene regions in the great apes (5), we performed the MK test (6) in all gene regions as compared with bonobo in order to confirm that our identified regions have not been confounded by negative selection. We found that none of the regions determined as under positive selection in our random forest model reached genome-wide significance for the MK test. However, a small fraction of regions (<5%; Supplementary Table S7) for each subspecies reported a neutrality index >1, indicating a decrease in non-silent divergence. We removed these regions from the remainder of our analysis. These results indicate that our input statistics effectively removed regions under strong negative selection but were confounded by regions under weak negative selection. This is likely since none of our 15 statistics is using divergence to look for a deviation from neutrality, and tests for background selection rely on divergence data.

We identify the greatest number of regions under selection in the subspecies *P.t. troglodytes*, and similar numbers for the other three subspecies (Table 2 and Supplementary Tables S8–S11). This is due in part to the underlying demography of these lineages, where *P.t. troglodytes* has a current effective population size around three times larger than the others (see Section 1.2 in the Supplementary Data for further demography details). Our study confirms previous research finding that the number of selective sweeps scales with effective population sizes (5).

Signals of selection in chimpanzee

As expected, most signals of selection intersect with coding gene regions. However, 54.4% of observed signals of selection in *P.t. verus* fall outside coding regions. For the remaining three of the four subspecies, ∼30% of regions predicted as under putative positive selection contain no genes (Table 2). Overall, the proportion of regions that contain genes is much higher than has been found in studies of selection in the human genome (25,44). Similar results were previously found in apes (45). With current knowledge and tools, it is difficult to interpret signals that fall outside coding regions. The annotation of the genome does not allow for a precise interpretation of selection signals in non-coding or regulatory regions. We expect that some of these signals are responsible for affecting protein coding genes through regulation.

Table 3. The number of regions and genes identified as under selection in more than one subspecies (P.t.v. is *P.t. verus*, P.t.e. is *P.t. ellioti*, P.t.s. is *P.t. schweinfurthii* and P.t.t. is *P.t. troglodytes*)

Subspecies	Overlapping regions	Overlapping genes
P.t.e.–P.t.s.	10	5
P.t.e.–P.t.t.	29	39
P.t.e.–P.t.v.	10	11
P.t.s.–P.t.t.	63	64
P.t.s.–P.t.v.	11	10
P.t.t.–P.t.v.	22	16
P.t.e.–P.t.s.–P.t.t.	4	2
P.t.e.–P.t.t.–P.t.v.	1	2
P.t.s.–P.t.t.–P.t.v.	5	6

After extracting the genes located within each region (Table 2 and Supplementary Table S12), we find 823 genes as being targets of selection for *P.t. troglodytes*, while the remaining three subspecies have around 300 genes each. For all four subspecies, the signals of selection are enriched for genes, based on randomly sampling the same number of regions 100 000 times for each subspecies *P*-value <0.01 (*P.t. ellioti* 305 versus 15, *P.t. schweinfurthii* 406 versus 20, *P.t. troglodytes* 824 versus 40, *P.t. verus* 228 versus 17 regions with genes). We first compared our results with a previous scan of selection in chimpanzee (43) and both the studies identify signatures of selection in 46 subspecies-specific genes (6 *P.t. ellioti*, 11 *P.t. schweinfurthii*, 22 *P.t. troglodytes* and 7 *P.t. verus*, Supplementary Table S12) of our total genes.

When comparing significant regions across subspecies (Table 3 and Figure 3), we observe that the target of selection in the subspecies is not common. In fact, we find no overlap for all four subspecies. We detect 10 genes that overlap with three subspecies and 145 genes appear in scans for two subspecies (Table 3). This leaves the majority of genes as unique to the individual subspecies (Supplementary Table S12). Indeed, these results may be an artifact that our simulations did not include selection on multiple subspecies concurrently; however, we were still able to detect common signals of selection based on deviations of the site frequency spectrum and linkage map. Altogether, this observation indicates that the selective pressures exerted on each of the four subspecies have been unique to each group in the last 60 kya. We found similar results for selection of the introgressed regions between chimpanzee and bonobos (46).

Interestingly, we find that 6 of the 10 shared genes overlapping with three subspecies are shared between *P.t. troglodytes*, *P.t. schweinfurthii* and *P.t. verus*. This is a surprising result as these populations do not live in a contiguous region in Africa. It should be noted here that *P.t. ellioti* has the smallest sample size, and our inability to de-

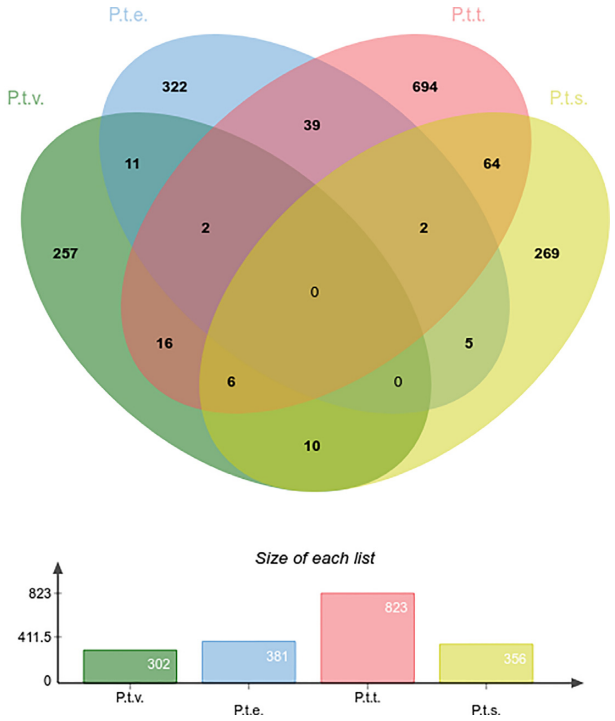

Figure 3. Venn diagram depicting the number of genes, predicted to be under putative positive selection, which overlap between subspecies. P.t.e. represents *P.t. ellioti*, P.t.s. represents *P.t. schweinfurthii*, P.t.t. represents *P.t. troglodytes* and P.t.v. represents *P.t. verus*.

tect selection in common targets of selection may be due to power more than biology; specifically, our power for this subspecies is 88%. In general, the dearth of common signals may be due to power in general, as our false negative rates are relatively high (up to 16%). Furthermore, within these regions, half of the common genes overlap; therefore, it is unlikely that all genes are driving the signal of selection. For more information about the specific genes and their functions, see Section 2.1 in the Supplementary Data and Supplementary Table S12.

For regions of selection common between two subspecies, we observe a pattern that largely matches geography of chimpanzee. The largest proportion of sharing can be observed between *P.t. troglodytes* and *P.t. schweinfurthii* (64 genes; Figure 3 and Supplementary Table S12). This large proportion of sharing would be expected because their habitats are neighboring, split by the Congo River. The second highest amount of sharing is between *P.t. troglodytes* and *P.t. ellioti.* These subspecies live in central Africa west of the Congo River, and were only recognized as separate subspecies through genetics (47,48). Their close proximity would suggest that they likely encounter similar selective pressures, leading to the observed high proportion of shared selective sweeps despite their long divergence time.

We observe seven significant (overall FDR < 0.05) GO terms for genes that are shared between two subspecies, specifically three terms for *P.t. ellioti* and *P.t. schweinfurthii*, one for *P.t. schweinfurthii* and *P.t. troglodytes*, and three for *P.t. troglodytes* and *P.t. verus* (Supplementary Table S14). The one GO term shared between *P.t. schweinfurthii* and *P.t. troglodytes* is 'U12-type spliceosomal complex' because these two subspecies share a signal of selection in Splicing Factor 3B subunits 2–4 (*SF3B2*, *SF3B3* and *SF3B4*). These genes encode for proteins that are part of a four-protein complex that is essential for the splicing of pre-mRNA (49). For a detailed description of genes and the unique divergent variants shared between two subspecies of chimpanzee, see Section 2.2 in the Supplementary Data.

By far, the majority of selective sweeps are observed to be unique to each subspecies (Figure 3). These sweeps have largely acted on gene regions (Supplementary Tables S8–S12). We observe 4 significant (FDR < 0.05) GO terms for *P.t. verus*, 7 for *P.t. ellioti*, 4 for *P.t. schweinfurthii* and 16 for *P.t. troglodytes* (Supplementary Table S14). The same GO term from above, 'U12-type spliceosomal complex', is again significant, and in addition to the three subunits of SF3B, *P.t. troglodytes* additionally has a signal in the fourth subunit *SF3B1* as well as RNA-Binding Region-Containing Protein 3 (*RNPC3*), which acts as bridge between the U11 and U12 spliceosomes (50). These signals in concert with the large signals of selection outside coding regions indicate the importance of regulation to the evolutionary history of chimpanzee.

Despite indications that regulation is an important target of selection, the majority of signals are in coding regions, and we have identified many possible functional targets of selection. Altogether, we identify 329 genes with missense substitutions, 6 genes with splice acceptors/donors, 95 genes with splice variants, 1 gene with a start loss, 1 gene with a stop loss, 9 genes with a stop gain and 371 genes with a synonymous substitution (Supplementary Table S15–S19) that are unique to the subspecies with the signal of selection. In addition, we detected 24 candidate derived alleles that are shared between subspecies having overlapping signals of selection, where the subspecies lacking the signal do not have that candidate derived allele segregating in its population (Supplementary Table S15). See Section 2.3 in the Supplementary Data for specifics.

Chimpanzee selection map

We provide a comprehensive genome-wide map of selection signals. These data can be viewed and used in the form of a USCS genome browser, available at http://hsb.upf.edu/chimp_browser/index.html, following the criteria and configuration of a published human dataset (25,44). The UCSC-style format facilitates the integration with the rich UCSC browser tracks. A search allows easy access to results for specific genes or genomic regions, and all raw data for each test and the composite random forest score can be conveniently downloaded using the UCSC Table function. We expect this to be a valuable resource for a wide range of future analyses. As such, it provides a broad picture of the action of positive selection in each genomic region in all four chimpanzee subspecies.

DISCUSSION

We have produced a descriptive genomic map for chimpanzee positive selection. With these results, we created a community resource that allows researchers to further

investigate selection at the level of subspecies in chimpanzee. These tracks can be viewed using the UCSC genome browser website; each of the calculated statistics has its own track that can be selectively viewed or downloaded using the Table function. It will be of use to researchers investigating chimpanzee phenotypes, clinicians investigating disease differences among human and great apes, and evolutionary biologists interested in speciation, among others.

These results indicate that positive selection is an important driver for differentiation between populations and eventual speciation. We find that out of all the genes with signals of positive selection, the vast majority are unique to a certain subspecies. While these signals may be decreased due to power, this indicates that many interesting differences between the subspecies appear to be mainly driven by environmental differences. Although three of the four subspecies live in seemingly ecologically similar habitats, their genomes indicate that exposures to subtle differences have resulted in differential adaptations. Our results confirm a previous study, which found that allopatric speciation was not sufficient to explaining the divergence between *P.t. ellioti* and *P.t. troglodytes* (10,36). Coupled with our results, we find evidence that selection has been a major driver for differentiation.

We observe a clear impact by the demographic history on the signals of selection. The population that has more than twice the number of signals, *P.t. troglodytes*, as compared with the other three subspecies has the highest effective size. This is an expected result because genetic drift is weaker in populations with large effective populations; as a result, selection can more easily drive beneficial alleles to high frequencies (5). We also observe a higher proportion of signals outside coding regions for *P.t. verus*, although the biological reason for this remains unclear. Moreover, GO results point to genic controls on regulation (e.g. mRNA preprocessing with significant enrichment of signals of selection indicating that regulatory regions are significant targets in driving adaptation). Demographic models used here are based on (10) and adapted to take into account meaningful admixture events and long-term effective population size differences between subspecies; however, it is worth mentioning that the robustness of our scan is sensitive to any misspecified demography [see (51) for a discussion]. Our statistical framework used neutral and positive selection to train the random forest algorithm using different selection times and coefficients. These two parameters were coupled to increase the efficiency of simulation by increasing the selection coefficient for short and recent selection times. However, this design precludes selection coefficient and timescale to be separated in the analysis and to evaluate the random forest power to detect different strengths of selection.

Our comprehensive scan of the chimpanzee genome coupled with the extensive simulations of selection scenarios occurring in the last 60 kya allows for robust categorization of regions of the genome as under selection. This work confirms and expands previous scans of selection in chimpanzee (5,32,45,52). Our data are available for use by the community in a convenient genome browser. A detailed map of interesting genetic differences between these subspecies is an important tool to use when building a better understanding the genotype–phenotype map of chimpanzees.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

Agencia Estatal de Investigación (Spain) [PID2019-110933GB-I00, CEX2018-000792-M, in part]; Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya [GRC 2017 SGR 702]; Agency for Management of University and Research Grants [FI-DGR 2015 to J.N.]; Part of the "Unidad de Excelencia María de Maeztu" [MDM-2014-0370], funded by the Ministerio de Economía, Industria y Competividad (MINECO, Spain).
Conflict of interest statement. None declared.

REFERENCES

1. Kaessmann,H., Wiebe,V. and Paabo,S. (1999) Extensive nuclear DNA sequence diversity among chimpanzees. *Science*, **286**, 1159–1162.
2. Fischer,A., Wiebe,V., Paabo,S. and Przeworski,M. (2004) Evidence for a complex demographic history of chimpanzees. *Mol. Biol. Evol.*, **21**, 799–808.
3. Prado-Martinez,J., Sudmant,P.H., Kidd,J.M., Li,H., Kelley,J.L., Lorente-Galdos,B., Veeramah,K.R., Woerner,A.E., O'Connor,T.D., Santpere,G. *et al.* (2013) Great ape genetic diversity and population history. *Nature*, **499**, 471–475.
4. de Manuel,M., Kuhlwilm,M., Frandsen,P., Sousa,V.C., Desai,T., Prado-Martinez,J., Hernandez-Rodriguez,J., Dupanloup,I., Lao,O., Hallast,P. *et al.* (2016) Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, **354**, 477–481.
5. Nam,K., Munch,K., Mailund,T., Nater,A., Greminger,M.P., Krutzen,M., Marquès-Bonet,T. and Schierup,M.H. (2017) Evidence that the rate of strong selective sweeps increases with population size in the great apes. *Proc. Natl Acad. Sci. U.S.A.*, **114**, 1613–1618.
6. Ohta,T. and Kimura,M. (1969) Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*, **63**, 229–238.
7. Arnold,M.L. and Martin,N.H. (2009) Adaptation by introgression. *J. Biol.*, **8**, 82.
8. 1000 Genomes Project Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
9. Howie,B.N., Donnelly,P. and Marchini,J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.
10. Schmidt,J.M., de Manuel,M., Marques-Bonet,T., Castellano,S. and Andrés,A.M. (2019) The impact of genetic adaptation on chimpanzee subspecies differentiation. *PLoS Genet.*, **15**, e1008485.
11. Ewing,G. and Hermisson,J. (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, **26**, 2064–2065.
12. Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
13. Fu,Y.X. and Li,W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.
14. Ramos-Onsins,S.E. and Rozas,J. (2002) Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.*, **19**, 2092–2100.
15. Voight,B.F., Kudaravalli,S., Wen,X.Q. and Pritchard,J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, 446–558.
16. 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,G.J. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

17. Wall,J.D. (1999) Recombination and the power of statistical tests of neutrality. *Genet. Res.*, **74**, 65–79.
18. Wall,J.D. (2000) A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.*, **17**, 156–163.
19. Fu,Y.X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**, 915–925.
20. Rozas,J., Gullaud,M., Blandin,G. and Aguade,M. (2001) DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. *Genetics*, **158**, 1147–1155.
21. Kelly,J.K. (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.
22. 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.
23. 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.
24. Nei,M. and Li,W.H. (1979) Mathematical model for studying genetic variation in variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. U.S.A.*, **76**, 5269–5273.
25. Pybus,M., Dall'Olio,G.M., Luisi,P., Uzkudun,M., Carreno-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.
26. Kass,G.V. (1980) An exploratory technique for investigating large quantities of categorical data. *J. Appl. Stat.*, **20**, 119–127.
27. Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) In: *Classification and Regression Trees*, Wadsworth, Inc., Belmont.
28. Krzywinski,M. and Altman,N. (2017) Classification and regression trees. *Nat. Methods*, **14**, 755–756.
29. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
30. Jay,F., Boitard,S. and Austerlitz,F. (2019) An ABC method for whole-genome sequence data: inferring Paleolithic and Neolithic human expansions. *Mol. Biol. Evol.*, **36**, 1565–1579.
31. Mondal,M., Bertranpetit,J. and Lao,O. (2019) Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nat. Commun.*, **10**, 246.
32. Sheehan,S. and Song,Y.S. (2016) Deep learning for population genetic inference. *PLoS Comput. Biol.*, **12**, e1004845.
33. Sugden,L.A., Atkinson,E.G., Fischer,A.P., Rong,S., Henn,B.M. and Ramachandran,S. (2018) Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nat. Commun.*, **9**, 703.
34. Rahman,R., Dhruba,S.R., Ghosh,S. and Pal,R. (2019) Functional random forest with applications in dose-response prediction. *Sci. Rep.*, **9**, 1628.
35. Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.
36. R Core Team. (2017) In: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
37. McDonald,J.H. and Kreitman,M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, **351**, 652–654.
38. Pfeifer,B., Wittelsbuerger,U., Ramos-Onsins,S.E. and Lercher,M.J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.*, **31**, 1929–1936.
39. Mi,H., Muruganujan,A., Ebert,D., Huang,X. and Thomas,P.D. (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.*, **47**, D419–D426.
40. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
41. Bardou,P., Mariette,J., Escudié,F., Djemiel,C. and Klopp,C. (2014) jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, **15**, 293.
42. McLaren,W., Gil,L., Hunt,S.E., Singh Riat,H., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
43. Charlesworth,B., Morgan,M.T. and Charlesworth,D. (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.
44. 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.
45. Cagan,A., Theunert,C., Laayouni,H., Santpere,G., Pybus,M., Casals,F., Prüfer,K., Navarro,A., Marquès-Bonet,T., Bertranpetit,J. *et al.* (2016) Natural selection in the great apes. *Mol. Biol. Evol.*, **33**, 3268–3283.
46. Nye,J., Laayouni,H., Kuhlwilm,M., Mondal,M., Marques-Bonet,T. and Bertranpetit,J. (2018) Selection in the introgressed regions of the chimpanzee genome. *Genome Biol. Evol.*, **10**, 1132–1138.
47. Gonder,M.K., Oates,J.F., Disotell,T.R., Forstner,M.R.J., Morales,C.J. and Melnick,D.J. (1997) A new west African chimpanzee subspecies? *Nature*, **388**, 337.
48. Pilbrow,V. (2006) Population systematics of chimpanzees using molar morphometrics. *J. Hum. Evol.*, **51**, 646–662.
49. Gozani,O., Feld,R. and Reed,R. (1996) Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes Dev.*, **19**, 233–243.
50. Benecke,H., Luhrmann,R. and Will,C.L. (2005) The U11/U12 snRNP 65K protien acts as a molecular bridge, binding the U12 snRNA and U11-59K protein. *EMBO J.*, **24**, 3057–3069.
51. Smith,M.L., Ruffley,M., Espíndola,A., Tank,D.C., Sullivan,J. and Carstens,B.C. (2017) Demographic model selection using random forests and the site frequency spectrum. *Mol. Ecol.*, **26**, 4562–4573.
52. Mitchell,M.W., Locatelli,S., Ghobrial,L., Pokempner,A.A., Sesink Clee,P.R., Abwee,E.E., Nicholas,A., Nkembi,L., Anthony,N.M., Morgan,B.J. *et al.* (2015) The population genetics of wild chimpanzee in Cameroon and Nigeria suggests a positive role for selection in the evolution of chimpanzee subspecies. *BMC Evol. Biol.*, **15**, 3.