PLOS ONE

Detecting Genome-wide Signals of Human Adaptation to Tropical Forests in a Convergent Evolution Framework

--Manuscript Draft--

Manuscript Number:	PONE-D-14-14972
Article Type:	Research Article
Full Title:	Detecting Genome-wide Signals of Human Adaptation to Tropical Forests in a Convergent Evolution Framework
Short Title:	Human genomic adaptations to tropical forests
Corresponding Author:	Carlos Eduardo Guerra Amorim, Ph.D Columbia University New York, NY UNITED STATES
Keywords:	HGDP; Positive selection; Pygmy phenotype; local adaptation; Amerindian; African; convergent evolution
Abstract:	Tropical forests are believed to be very harsh environments for human life. It is unclear if human beings would have ever subsisted in those environments without external resources. It is therefore possible that humans have developed recent biological adaptations in response to specific selective pressures to cope with this challenge. To understand such biological adaptations we analyzed genome-wide SNP data under a Bayesian statistics framework, looking for outlier markers with overly large extent of differentiation between populations living in a tropical forest, as compared to genetically related populations living outside the forest. The most significant positive selection signals were found in genes related to lipid metabolism, the immune system, body development, and RNA Polymerase III transcription initiation. The results are discussed in the light of putative tropical forest selective pressures, namely food scarcity, high prevalence of pathogens, difficulty to move, and inefficient thermoregulation. Agreement between our results and previous studies on pygmy phenotype, a putative prototype of forest adaptation, were found, suggesting that a few genetic regions previously described as associated with short stature may be evolving under similar positive selection in Africa and the Americas. In general, convergent evolution was less pervasive than local adaptation in one single continent, suggesting that Africans and Amerindians may have followed different routes to adapt to similar environmental selective pressures.
Order of Authors:	Carlos Eduardo Guerra Amorim, Ph.D
	Josephine T Daub
	Francisco Mauro Salzano
	Matthieu Foll
	Laurent Excoffier
Additional Information:	
Question	**Response**
Financial Disclosure Please describe all sources of funding that have supported your work. A complete funding statement should do the following: Include **grant numbers and the URLs** of any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who	Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) to CEGA and FMS <http://cnpq.br/>. Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, Programa de Apoio a Núcleos de Excelência (FAPERGS/PRONEX) to FMS <http://www.fapergs.rs.gov.br/>. CAPES Foundation grant (BEX 8279/11-0) to CEGA <http://www.capes.gov.br/>. Swiss NSF grants (31003A-143393 and PDFMP3_130309) to LE <http://www.snf.ch/>. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

received the funding. **Describe the role** of any sponsors or funders in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. If they had no role in any of the above, include this sentence at the end of your statement: "*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*" If the study was **unfunded**, provide a statement that clearly indicates this, for example: "*The author(s) received no specific funding for this work.*" * typeset	
Competing Interests You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests. Do any authors of this manuscript have competing interests (as described in the PLOS Policy on Declaration and Evaluation of Competing Interests)? **If yes**, please provide details about any and all competing interests in the box below. Your response should begin with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests:* **If no** authors have any competing interests to declare, please enter this statement in the box: "*The authors have declared that no competing interests exist.*" * typeset	The authors have declared that no competing interests exist.
Ethics Statement You must provide an ethics statement if	N/A

your study involved human participants, specimens or tissue samples, or vertebrate animals, embryos or tissues. All information entered here should **also be included in the Methods section** of your manuscript. Please write "N/A" if your study does not require an ethics statement. **Human Subject Research (involved human participants and/or tissue)** All research involving human participants must have been approved by the authors' Institutional Review Board (IRB) or an equivalent committee, and all clinical investigation must have been conducted according to the principles expressed in the Declaration of Helsinki. Informed consent, written or oral, should also have been obtained from the participants. If no consent was given, the reason must be explained (e.g. the data were analyzed anonymously) and reported. The form of consent (written/oral), or reason for lack of consent, should be indicated in the Methods section of your manuscript. Please enter the name of the IRB or Ethics Committee that approved this study in the space below. Include the approval number and/or a statement indicating approval of this research. **Animal Research (involved vertebrate animals, embryos or tissues)** All animal work must have been conducted according to relevant national and international guidelines. If your study involved non-human primates, you must provide details regarding animal welfare and steps taken to ameliorate suffering; this is in accordance with the recommendations of the Weatherall report, "The use of non-human primates in research." The relevant guidelines followed and the committee that approved the study should be identified in the ethics statement. If anesthesia, euthanasia or any kind of animal sacrifice is part of the study, please include briefly in your statement which substances and/or methods were applied.	

Please enter the name of your Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board, and indicate whether they approved this research or granted a formal waiver of ethical approval. Also include an approval number if one was obtained. **Field Permit** Please indicate the name of the institution or the relevant body that granted permission.	

Detecting Genome-wide Signals of Human Adaptation to Tropical Forests in a Convergent Evolution Framework

Carlos Eduardo G. Amorim[1,2,3,4,5]*, Josephine T. Daub[1,2], Francisco M. Salzano[3], Matthieu Foll[2,6,¶], Laurent Excoffier[1,2,¶]

1 Computational and Molecular Population Genetics Laboratory, Institute of Ecology and Evolution, 3012, Berne, Switzerland, **2** Swiss Institute of Bioinformatics, 1015, Lausanne, Switzerland, **3** Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, 91501-970, Rio Grande do Sul, Brazil, **4** CAPES Foundation, Ministry of Education of Brazil, Brasília, 70040-020, Brazil, **5** Current address: Department of Biological Sciences, Columbia University, 10027, New York, NY, USA, **6** School of Life Science, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Keywords: Cholesterol metabolism; HGDP; Immune system; Positive selection; Pygmy phenotype

Funding: CNPq (Brazil); CAPES (Brazil); FAPERGS, PRONEX (Brazil); SNSF (Switzerland)

*** E-mail:** cg2827@columbia.edu

¶ Joint senior authors

Abstract Tropical forests are believed to be very harsh environments for human life. It is unclear if human beings would have ever subsisted in those environments without external resources. It is therefore possible that humans have developed recent biological adaptations in response to specific selective pressures to cope with this challenge. To understand such biological adaptations we analyzed genome-wide SNP data under a Bayesian statistics framework, looking for outlier markers with overly large extent of differentiation between populations living in a tropical forest, as compared to genetically related populations living outside the forest. The most significant positive selection signals were found in genes related to lipid metabolism, the immune system, body development, and RNA Polymerase III transcription initiation. The results are discussed in the light of putative tropical forest selective pressures, namely food scarcity, high prevalence of pathogens, difficulty to move, and inefficient thermoregulation. Agreement between our results and previous studies on pygmy phenotype, a putative prototype of forest adaptation, were found, suggesting that a few genetic regions previously described as associated with short stature may be evolving under similar positive selection in Africa and the Americas. In general, convergent evolution was less pervasive than local adaptation in one single continent, suggesting that Africans and Amerindians may have followed different routes to adapt to similar environmental selective pressures.

INTRODUCTION

Tropical forests are characterized by a high diversity of plants, with tall trees, dense canopies and low light penetration [1]. Their climate is generally warm with minimum temperatures well above the freezing point and mean annual rainfall above 1,000 mm [2]. Despite being one of the most productive environments of the world,

tropical forests provide only few resources for humans [3]. Indeed, in these environments plants invest most of their energy in structure maintenance and not into the reproductive parts that are the most edible parts for humans and their preys [2]. In addition, the instability of food resources in response to the high seasonality of rain falls raises the costs of foraging, further reducing its capacity to support human life [2,3].

Besides food limitation, other characteristics of tropical forests may also contribute to the hostility of these environments. For instance, tropical areas harbor on average 70% higher pathogen diversity as compared to more temperate areas [4]. As a consequence, infant and child mortality rates among tropical forest dwellers should be high [5]. Moreover, the small differences between air and skin relative humidities and high temperature, coupled with little air movement, make sweat production and evaporation difficult in tropical forests, potentially compromising thermoregulation [6].

Due to the hostility of this environment, it is unclear if humans have ever subsisted in tropical forests without depending on external resources, such as agriculture or possible exchanges with neighboring populations. Evidence of societies living in such harsh conditions is scarce for contemporary modern humans [2], as well as for early *Homo* [7]. Nonetheless, it is possible that humans have developed recent biological adaptations to tropical forests. A few examples of such adaptations have indeed been documented, the most well-known being the pygmy phenotype. Short-statured individuals may have, for instance, advantages to cope with food limitation, thermoregulation, and mobility hardship in a dense forest [6]. However, it has also been suggested that this phenotype could be a by-product of selection for early onset of reproduction [8], which could enable populations to overcome problems related to their life history and increased mortality [9].

To investigate whether tropical forest dwellers have developed specific biological adaptations to this harsh environment, we searched for genome-wide signals of positive selection in populations from the Americas and Africa, specifically aiming at identifying convergent evolution signals, that is a significant signal of positive selection occurring at the same genomic region or biological pathway in populations belonging to two distinct evolutionary lineages. For doing so, we investigated populations living in tropical forests and others, genetically related, living outside these environments using publicly available genome-wide SNP data and a robust and sensitive F_{ST}-based method for inference of positive selection that explicitly includes a convergent selection model.

SUBJECTS AND METHODS

Populations and samples

Genome-wide single nucleotide polymorphism (SNP) data were downloaded for seven populations included in the Human Genetic Diversity Panel database (HGDP; [10,11]). Two African (Biaka and Mbuti pygmies) and two American (Surui and Karitiana) tropical forest populations were selected, as well as three other populations from these continents to serve as non-tropical forest comparisons (Mandenka and Yoruba in Africa; Pima in America). Considering the genetic similarity between Mandenka and Yoruba [12], these populations were grouped into a single set, hereafter called "West Africa", to increase sample size and the statistical power of the analyses. More information on the chosen populations can be found in [3, 10, 11, 13, 14]. We excluded atypical and duplicated samples, keeping only those present in the H1048 subset of [15]

The six populations were combined into four distinct population sets (PS), each including one tropical forest and one control population per continent as follows:

PS1: West Africa, Biaka; Pima, Surui.

PS2: West Africa, Mbuti; Pima, Surui.

PS3: West Africa, Biaka; Pima, Karitiana.

PS4: West Africa, Mbuti; Pima, Karitiana.

Each PS was analyzed separately to find loci and genomic regions that would putatively be under natural selection. The rationale behind the use of different population sets is to look for congruent signals of adaptation across data sets, and thus to eliminate signals potentially due to particular tropical forest populations, which have been shown to present high rates of genetic drift due to their small effective population sizes [16, 17].

Genetic data

Data on 660,918 SNPs were downloaded from the Stanford University HGDP-CEPH SNP genotyping data supplement 1 ([12]; <ftp://ftp.cephb.fr/hgdp_supp1/>). We initially discarded 250 markers that were monomorphic in all populations, those that presented only missing data, and those located on either the Y-chromosome, the pseudoautosomal region of the sex chromosomes, or the mtDNA, leaving us with 660,668 SNPs. After selecting the different PSs as described above, those markers with minor allele frequency below 5% at all populations joined into one were discarded, yielding 582,074, 581,855, 584,205, and 577,345 SNPs for population sets PS1-PS4, respectively.

Detecting outlier SNPs

A modified version of BayeScan [18] was used to identify candidate targets for natural selection. The original methodology implemented with this software is based on the multinomial-Dirichlet likelihood-based approach [19] implemented via a Markov

chain Monte Carlo (MCMC) algorithm [20] and uses an island model [21] – in which the subpopulations' allele frequencies are correlated through a common migrant gene pool from which they differ by varying degrees – to calculate a population specific F_{ST} coefficient. Logistically transformed F_{ST} coefficients are then decomposed into a population-specific component (β), shared by all loci, and a locus-specific component (α), shared by all the populations [18, 20]. Selection is inferred when α is significantly different from zero. For each locus, two alternative evolutionary models including α (selection) or neutrality can thus be explored. The posterior probability of each model (selection vs. neutrality) is estimated with a reversible-jump MCMC algorithm [18] and indicates how likely the model with selection is in comparison to the neutral one. Significantly positive values are indicators of an overly large level of differentiation of a given SNP and thus positive selection; whereas significantly negative values of α are indicative of balancing selection. Further information on this methodology can be found in the BayeScan manual or other methodological papers on the F-model [18-20].

The newest version of BayeScan [22] includes a hierarchical island model accounting for the relative closer similarity of certain populations in comparison to others, as should be the case of populations that are sampled in a given continent and therefore share part of their history. According to the hierarchical island model, each continent has a specific migrant pool. A F_{SC} coefficient measures the differentiation of each population within the continental pool of migrants and a F_{CT} coefficient measures the differentiation of each continent within the overall meta-population. In this regard, when considering two pairs of populations in two different continents one ends up with four alternative selection models for each locus: (1) neutral variability; (2) selection in one continent; (3) selection in the other continent; and (4) selection in both continents.

BayeScan estimates for each marker the posterior probability of each one of the four tested models. These posterior probabilities are then transformed into *q*-values for each marker in order to control for the False Discovery Rate (FDR; [23]) considering the probability of a SNP being under selection regardless of which model (selection in one or both continents). FDR is defined as the expected proportion of false positives among outlier markers. See further detailing of the methods and new implementations of Bayescan at [22]

In this analysis, we considered all SNPs with *q*-values lower than 0.1 as potentially significant. This procedure yielded a list of outlier SNPs for each PS, which were then considered as candidate loci for natural selection targets. We finally assume hereafter that convergent selection occurred when selection presents a higher probability than neutrality and model 4 has a higher posterior probability than models 2 and 3.

Gene and regulatory elements annotation

All SNPs were assigned to genes using PLINK v1.07 ([24]; available at http://pngu.mgh.harvard.edu/purcell/plink). SNPs not present in coding regions were assigned to a given gene if located less than 50 kb away from it. When more than one gene was within this range, the closest gene was chosen for the subsequent analyses. The amount of outlier SNPs falling in genic regions (or <50kb away from them) was compared to the amount of SNPs in non-genic regions first taking into account only protein coding genes and then all regulatory elements as described ahead. This proportion was then compared to the distribution of the 660,668 HGDP SNPs in genic or non-genic regions with a chi-square test using R [25] to check if these outlier SNPs were enriched for genic SNPs in comparison to all available markers.

The hg19 assembly coordinates (NCBI Build 37.3) of 19,683 protein-coding genes located on the human autosomal and X chromosomes were obtained from the NCBI Entrez Gene website ([26]; <http://www.ncbi.nlm.nih.gov/gene>, accessed on February 7, 2013). Seventeen genes presented multiple locations; in these cases we took the outermost start and end positions.

The original positions of the SNPs on the hg18 reference genome (NCBI Built 36.3) obtained from the original dataset ([26]; <ftp://ftp.cephb.fr/hgdp_supp1/>) were remapped on hg19 with the NCBI Genome Remapping Service (<http://www.ncbi.nlm.nih.gov/genome/tools/remap>). In doing so, we were not able to remap 74 SNPs, which were then excluded from the following analyses.

Information on the DNaseI Hypersensivity sites were obtained online (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/>, volume 2, accessed on April 4, 2013) for the hg19 assembly and added to the annotation file. These clusters show DNaseI hypersensitive areas assayed in 125 human cell types by the ENCODE Project [27] and may be indicative of regulatory regions. This information was used to calculate the distance of an outlier SNP to a putative functional region.

Detecting clusters of outlier SNPs

A sliding-window approach was implemented with R to identify significant clusters of candidate SNPs and remove isolated loci. We considered consecutive windows of size 500 kb-wide with 25 kb overlap. The *q*-value associated to each window was assumed as the 95% quantile of all included SNPs. Low density windows, i.e. those with less than one fifth of the average SNP density per chromosome, were set as non-significant. A graphical representation of this procedure was plotted with R taking into

account the distribution of the SNPs, their particular *q*-value, the window *q*-value, and the best supported model of selection for each outlier SNPs. For the outlier SNPs, we also included information on their nearest gene if it was at most 50 kb apart.

The sliding-window approach yielded a second set of outlier markers. This SNP set is more refined than the one with the first candidates, since it ignores the low SNP density regions of the genome and those candidates that are isolated, highlighting genomic regions with a higher density of outlier SNPs.

Gene set analysis

We applied a gene set enrichment analysis as described by Daub et al. [28] to find pathway level signals of selection. In short this method tests if genes in a gene set show a shift in the distribution of a certain selection score. In our case we took as selection score the probability for selection in Africa (p_{af}), the Americas (p_{am}), and both continents (p_{co}, convergent selection). We also added the score p_{sl}, which is the probability of any selection ($p_{sl} = p_{af} + p_{am} + p_{co}$). As we use one value per gene in the enrichment test, we transformed the SNP-based scores to gene-based scores. First, SNPs were assigned to genes as describe above: a SNP was assigned to a gene if the SNP's location was within a gene transcript; otherwise it was assigned to the closest gene within 50kb distance. For each gene and each selection model, we took the highest selection score of all SNPs assigned to this gene. After removing 1,922 genes with no SNPs assigned, a list of 17,761 genes remained for PS1 and PS4, 17,766 for PS2, and 17,769 for PS3.

We collected 2,336 gene sets from the NCBI Biosystems database ([29], http://www.ncbi.nlm.nih.gov/biosystems, downloaded 3 Sep 2013). After removing genes that were not part of the abovementioned gene list, excluding gene sets containing less than 10 genes and pooling (nearly) identical gene sets into union sets, 1,216 (PS1

and PS2) and 1,217 (PS3 and PS4) sets remained and they were used as input in the enrichment tests.

For all selection models, we computed the SUMSTAT [30] score for each gene set, which is the sum of the p_{af}, p_{am}, p_{co}, or p_{sl} values of all genes in a set. Potential candidates for selection are gene sets that score a high SUMSTAT value. To assess significance we compared the SUMSTAT score of each tested gene set with a null distribution created from random gene sets of equal size. To improve computation time, the creation of the null distribution was done with a sequential random sampling method. We first tested all sets against 10,000 randomizations. Next, for each tested gene set, we counted the number of random sets with the same or higher SUMSTAT score. Only for those sets with a count smaller than 5,000, we expanded the null distribution with another 10,000 randomizations. This process was continued until we reached a maximum of 500,000 randomizations.

Selecting the highest score among the SNPs in or near a gene can induce a bias, as it is more likely that genes with high SNP density have an extreme value assigned. We corrected for this potential bias by placing genes in bins of similar SNP densities and constructing the null distribution from random gene sets with the same bin distribution as the gene set being tested.

We removed the overlap among candidate gene sets by applying a 'pruning' method according to which we iteratively assigned shared genes between sets to the highest scoring gene set. As these tests were not independent anymore, we estimated the q-value of these pruned sets empirically. All sets scoring a q-value below 20% (before and after pruning) were reported.

We used *R* [25] and Cytoscape [31] to visualize the significant pathways and their overlap using a layout that was inspired by the Cytoscape plugin EnrichmentMap [32].

RESULTS

According to the F_{ST}-based Bayesian approach implemented via BayeScan [18, 22] 1,482, 1,222, 1,579, and 1,365 outlier SNPs were identified in populations sets PS1 to PS4, respectively. From these, 75 SNPs are significant regardless of which PS is analyzed. Using a threshold of 50 kb, 568, 474, 620, and 517 genes were associated with those significant SNPs from PS1 to PS4 respectively, of which 57 genes were found to be co-occurring in all four PSs (namely *ABLIM3, ACSS2, AKAP6, ANKRD26, ARHGEF10, ATIC, BCAT1,* C20orf111*,* C2orf73*, CBLN1, CNTN4, CNTNAP5, COL22A1, CPA5, CRTC3, CWH43, DCUN1D4, DHCR7, EPHB4, FAM188B, FKBP6, GALNT16, GLIS3, GLRB, GPC6, GRIK2, HLA-DPA1, HMG20B, HSF2, IQGAP1, KIAA1598, KLHL29, LRRC66, MASTL, MPST, NAALADL2, NRG1, NRP2, NSUN5, PARK2, PKIB, PPP2R2C, RABGGTB, RAD51B, RBFOX1, RBM9, RBMS3, RFX3, ROBO2, SCP2, SGCB, SHISA6, SPAG16, SPATA13, SPATA18, ST6GAL1,* and *TRG@).* We confirmed with a random permutation test (N=5,000) that the number of outlier SNPs (75) and genes (57) shared by all population sets is higher than expected by chance (p-value < 2e-4).

The sliding-window procedure yielded 440, 399, 505, and 471 SNPs with a significant signal of positive selection for each PS. There were no cases in which balancing selection could be inferred, i.e. α never reached a significant negative value, which may be due to the characteristics of the genetic system employed (e.g. a possible ascertainment bias), a lack of power to detect markers under balancing selection [20]

and not necessarily to the absence of this phenomena in the evolutionary history of these populations.

Outlier SNPs are significantly enriched (p-value < 0.00017) for genic SNPs in comparison to the whole HGDP SNP set. On average 72.0% of the outlier SNPs after the sliding-window procedure are located in or less than 50 kb away from protein-coding regions considering all four PSs, while for all SNPs in the HGDP database this proportion is only 63.6%. Moreover, all outliers were located at maximum 45.8 kb apart from a gene or from a putative regulatory element indicated by a hypersensibility to DNaseI.

Figure 1 shows the Manhattan plots of the distribution of the SNPs in the genome for the different PSs and their estimated *q*-value, as well as the best model for selection (color-coded) into the sliding-window approach. It was possible to identify seven clusters of outlier SNPs co-occurring in all four PSs (indicated by gray-shaded boxes at Figure 1 and at Supplementary Figures 1-6).

For the first two clusters, the most likely model of selection for the majority of outlier SNPs is positive selection in the Americas; however, for a few SNPs convergent evolution cannot be ruled out (Figure 1). While the first cluster includes the gene *SCP2* (Supplementary Figure 1), the second cluster lies in an inter-genic region (Supplementary Figure 2). The next two clusters of significant SNPs occur on the same chromosome and selection in Africa is the best-supported model in most cases, although convergent selection is likely for a few SNPs (Supplementary Figure 3). Cluster 3 is associated with gene *CWH43*, while cluster 4 comprises SNPs that fall into genes such as *DCUN1D4, LRRC66,* and *SGCB*. Convergent evolution is the best-supported model of selection for the majority of SNPs of the remaining three clusters (Supplementary Figure S4-S6). Cluster 5 is associated with the following three genes: *C5orf34*, *NNT*, and *CCL28;*

which may or may not present significant signal of selection depending on the analyzed PS (Supplementary Figure 4). Cluster 6 includes genes *HSF2* and *PKIB* for all PSs, and the analysis of PS1 and PS3 revealed an additional significant SNP located at *SERINC1.* The last cluster (no. 7) includes two SNPs, one located in a *FKBP6* intron and the other located 0.12 kb apart from *NSUN5* and 3.55 kb from *TRIM50*. The abovementioned 14 genes that include or are close (<50 kb) to SNPs that present signals of positive selection in the sliding-windows approach for all four PSs are described in Table 1.

Fifteen genomic regions have been found to be associated with the pygmy phenotype by means of covariation between allele frequencies and body height in Africa in another study [33]. We found signals of positive selection in the African continent in four of these regions. By comparing Mbuti to West Africa, Mendizabal et al. [33] identified two genomic regions candidate for selection. One of them, located in the long arm of chromosome 10, is among those with the highest *q*-values in our study (Figure 1). It includes 14 outlier SNPs in the following 11 genes: *P4HA1*, *NUDT13*, *ECD*, *FAM149B1*, *DNAJC9*, *MRPS16*, *ANXA7*, *ZMYND17*, *USP54*, *PPP3CB*, and *TTC18*. Two outlier SNPs in this region – rs2271904 and rs4294502 – are both non-synonymous mutations in *ECD* and *TTC18* respectively. The second region is defined by two SNPs (rs7174731 and rs7181518) in *TRIP4*. In a third region, we found two genes with positive selection signals considering West Africans in comparison to Biaka instead of Mbuti: *USP46* and *MLL3*. An additional region suggested by Mendizabal et al. [33] to be associated with the pygmy phenotype presents signals of convergent adaptation using all four different combinations of populations in our study. They found significant SNPs in *NNT*, while we observed a more diffuse significant signal also including *CCL28* in some cases (Supplementary Figure S4).

As selection is acting on phenotypes instead of genotypes, we applied a gene set analysis [28] to detect pathways involved in adaption to living in tropical forests. We found 43 pathways scoring a q-value below 20% for at least one of the population sets and selection models (Supplementary Figure7), most of them scoring significant for selection in Africa. Among those we find a large cluster of 22 gene sets involved in immune response, nine of which show signals of convergent selection, and a cluster of 11 gene sets involved in RNA Polymerase III transcription initiation. After removing the overlapping genes between the pathways ('pruning'), 14 significant pathways remain significant, including “Apoptosis” and “Cholesterol Biosynthesis” among other immune and nervous systems-related pathways (Figure 2). Four of these pathways present a consistent signal of positive selection across different PSs or selection models, namely “PD1-signaling”, “IL 12-mediated signaling events”, “RNA Polymerase III Transcription Initiation From Type 3 Promoter”, and “Chemokine receptors bind chemokines”.

DISCUSSION

Tropical forests are believed to be very harsh habitats for human beings [3] In addition to being almost deprived from energy-rich food and edible plants [2], these environments are very propitious for the development of diseases [4] and might also compromise thermoregulation [6]. In this work we used a Bayesian method [18, 22] to identify SNPs in a genome-wide dataset showing overly large or low extent of differentiation between tropical and non-tropical forest populations. We sought to infer signals of convergent evolution by comparing native populations from tropical forests (Biaka, Mbuti, Karitiana, and Surui) with genetically related populations living elsewhere (Mandenka, Yoruba, and Pima) combined in four different population sets

(PSs). The analysis suggested some SNPs, genes, and a few biological pathways in which convergence could be inferred. These outlier SNPs are found to be enriched for genic loci. Those few cases that have fallen within an intergenic region, such as Cluster 2 (Supplementary Figure 1), could always be associated (distance < 50 kb) to a putative regulatory element described by the ENCODE project. These results suggest that regulatory regions could also be involved in some recent human adaptations [34]. Genomic regions where the same signal of selection could be identified by employing different combinations of populations (clusters 1-7 at Figure 1) suggest that the inferred signal is due to environmental selective pressures and adaptation rather than to the demographic history of the populations and those regions are the main focus of our discussion.

In a previous study involving Native Americans, Hünemeier et al. [35] suggested that *ABCA1*, a gene encoding a cholesterol efflux regulatory protein, was evolving under positive selection due to limited food resources that Native American encountered during their history. We also found significant signals of positive selection in genes that are related to lipid circulation and metabolism, such as *SCP2* and *CWH43* (Table 1) and in a genetic pathway related to cholesterol biosynthesis (Figure 2 and Supplementary Figure 7). Besides playing a role in nutrition, these genes could also be involved in immunological response, since cholesterol plays an important role in various infectious processes such as virus invasion and replication [36] as well as in resistance against malaria [37].

In this regard, Sabeti et al. [38] already noticed a preponderance of genes related to the immune system in available genome-wide scans for positive selection. This prevalence was further confirmed by Williamson et al. [39], López Herráez et al. [40],

and Daub et al. [28]. Besides the two abovementioned genes – *SCP2* and *CWH43*, which have a possible role in immunity – the protein encoded by *CCL28* modulates immunity to HIV infection and skin-related inflammatory diseases. Additionally, the selective pressure of this class of genes may also be important at the multi-genic level, since a network of 22 pathways involved in immune response presents signals of positive selection (Supplementary Figure 7). Two of these pathways remain significant after pruning for more than one PS, namely "PD-1 signaling" and "IL 12-mediated signaling events", indicating that they both have independently a strong selective signal, which would deserve further investigations for their role in adaptations to tropical environments.

Another category of genes that frequently presents signals of positive selection is fertility, more specifically, spermatozoid development [38]. In this regard, *FKBP6,* a male fertility factor, is found here to be potentially evolving under positive selection in our current analyses (Supplementary Figure 6).

Heat-shock transcription factors, such as that encoded by *HSF2* (Cluster 6; Supplementary Figure 5), are activated by stress and respond to elevated temperatures. One of the consequences of inefficient thermoregulation is the increase of body temperature. The observed positive selection signals at SNPs found in this gene could be due to an adaptation to the tropics, initiating gene(s) transcription in response to high body temperatures. Another study with African-, European-American, and Chinese populations also found a significant signal of positive selection in heat shock genes [39], suggesting that this category might have some importance in human adaptation to different environments worldwide.

It is generally accepted that the pygmy phenotype might have evolved as an adaptation to life in dense tropical forests, to thermoregulation, and to food scarcity [6] or as a by-product of selection for early onset of reproduction [8]. Our research design enables the comparison of two African pygmy populations with two other non-pygmy populations from the same continent, from which inferences about the differences in selective pressures that African pygmy and non-pygmy populations are subjected to can be drawn. Clusters 3 and 4 (Supplementary Figure 3) are the two main regions where we found signals for positive selection in Africa. The first cluster presents a gene involved in lipid metabolism (*CWH43*, discussed above) and the second includes the genes *DCUN1D4*, *LRRC66*, and *SGCB*, which are found in a region known to be associated to severe limb-girdle muscular Duchenne-like dystrophy [41].

Other regions of interest for the study of the pygmy phenotype are those four in which we found positive selection signals and are associated with the pygmy phenotype according to another study [33]. From the genes found in this region, three are notable for being also associated with height, bone development or the pygmy phenotype in previous studies. Those genes are *PPP3CB,* which encodes a subunit of calcineurin, a protein that regulates bone formation by osteoblast differentiation [42]; *TRIP4*, a gene with positive selection in pygmy populations [43] that relates to a category (thyroid hormone receptor) that was suggested to be associated with this phenotype [40]; and *MLL3*, involved in histone modification, a category associated with height in a genome-wide association study [44].

In general, non-convergent positive selection signals were more pervasive than signals of convergent evolution at the SNP, gene and pathway levels. This suggests that the similar selective pressures imposed by the tropical forest environment in Africa and

the Americas have targeted different genes and pathways, what could be a result of different genetic background of the populations from these two continents and also a result of the differences in the time-scales that adaptation took place, since the Americas were peopled more recently than Africa [16, 17]. As it is particularly true for the pygmy phenotype [43], putative adaptations to the tropical forests – even though they might be similar at the phenotypic level – could be the result of selection acting on different genetic targets. Nonetheless, due to the power limitation associated with detecting convergent evolution [22], this model cannot be ruled out for those regions where at least one SNP was assigned to it among others SNPs assigned to a model of positive selection in one continent, which is the case of Clusters 1 to 4. Therefore, it is likely that additional genes evolving under convergent evolution in these populations could be found.

CONCLUSIONS

The F_{ST}-based hierarchical Bayesian method used in our study enabled us to detect a number of regions with positive selection signals suggesting that the following biological functions and pathways may play a role in human adaptations to tropical forest: lipid metabolism, immunology, body development, and heat stress response. The same signal found in different population sets suggests that they are due to environmental adaptation rather than to the demographic history of the sampled populations. Further refinement of these analyses with e.g. full genome or exome sequence information could reveal which particular mutations are responsible for these adaptations. Moreover, the few cases in which convergent evolution could be inferred contrast to the bigger amount of genes with non-convergent positive selection signals,

suggesting that Africans and Amerindians may have followed different routes to adapt to similar environmental selective pressures.

ACKNOWLEDGEMENT

The authors would like to thank Drs. Maria Luiza Petzl-Erler, Eduardo Tarazona Santos and Nelson J. R. Fagundes for their valuable comments on an early version of the manuscript and Prof. Sandro L. Bonatto for his contribution to the project. Detailed results regarding the list with all outlier SNPs before and after the sliding-window approach, nearest gene (if located in a genic region), *q*-value and best-supported model of selection as well as the table with gene set enrichment are available upon request.

AUTHOR CONTRIBUTIONS

Conceived and designed the study: CEGA MF LE. Analyzed the data: CEGA JTD. Contributed material and analytical tools: JTD MF LE. Wrote the paper: CEGA JTD FMS LE.

REFERENCES

1. Ratnam J, Bond WJ, Fensham RJ, Hoffmann WA, Archibald S et al. (2011) When is a 'forest' a savanna, and why does it matter? Glob Ecol Biogeogr 20: 653-60.
2. Bailey RC, Head G, Jenike M, Owen B, Rechtman R, Zechnter E (1989) Hunting and gathering in tropical rain forest: is it possible? Am Anthropologist 91: 59-82.

3. Hart TB, Hart JA (1986) The ecological basis of hunter-gatherer subsistence in African rain forests: the Mbuti of Eastern Zaire. Hum Ecol 14: 29-55.

4. Guernier V, Hochberg ME, Guégan JF (2004) Ecology drives the worldwide distribution of human diseases. PLoS Biol 2: e141.

5. Ohenjo N, Willis R, Jackson D, Nettleton C, Good K, Mugarura B (2006) Health of indigenous people in Africa. Lancet 367: 1937-46.

6. Perry GH, Dominy NJ (2009) Evolution of the human pygmy phenotype. Trends Ecol Evol 24: 218-25.

7. Mercader J (2002) Forest people: the role of African rainforests. Evol Anthropol 11: 117-24.

8. Migliano AB, Vinicius L, Lahr MM (2007) Life history trade-offs explain the evolution of human pygmies. Proc Natl Acad Sci U S A 104: 20216-9.

9. Walker RS, Hamilton MJ (2008) Life-history consequences of density dependence and the evolution of human body size. Curr Anthr 49: 115-22.

10. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V et al. (2002) A human genome diversity cell line panel. Science 296: 261-2.

11. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK et al. (2002) Genetic structure of human populations. Science 298: 2381-5.

12. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-4.

13. Lee RB, Daly R (2004) The Cambridge Encyclopedia of Hunters and Gatherers. Cambridge University Press. 534p.

14. ISA – Instituto Socioambiental (2013) Povos Indígenas do Brasil <http://pib.socioambiental.org> Website accessed on April 26, 2013.

15. Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70: 841-7.
16. Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N et al. (2007) Genetic variation and population structure in Native Americans. PLoS Genet 3: e185.
17. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A et al. (2009) The genetic structure and history of Africans and African Americans. Science 324: 1035-44.
18. Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180: 977-93.
19. Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol 63: 221-30.
20. Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13: 969-80.
21. Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103: 285-98.
22. Foll M, Gaggiotti O, Daub JT, Excoffier L (2014) Hierarchical Bayesian model of population structure reveals convergent adaptation to high altitude in human populations. arXiv:1402.4348 [q-bio.PE]
23. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57: 289–300.
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-75.

25. R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R-project.org/.

26. Maglott D, Ostell J, Pruitt KD, Tatusova T (2011) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 39 (Database issue): D52-7.

27. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.

28. Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L et al. (2013) Evidence for Polygenic Adaptation to Pathogens in the Human Genome. Mol Biol Evol 30: 1544–1558.

29. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J et al. (2010). The NCBI BioSystems database. Nucleic Acids Res 38:D492-496.

30. Tintle NL, Borchers B, Brown M, Bekmetjev A. 2009. Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16. BMC Proc 3 Suppl 7: S96.

31. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431-432.

32. Merico D, Isserlin R, Stueker O, Emili A, Bader GD (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One 5: e13984.

33. Mendizabal I, Marigorta UM, Lao O, Comas D (2012) Adaptive evolution of loci covarying with the human African Pygmy phenotype. Hum Genet 131: 1305-17.

34. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S et al. (2013) Identifying recent adaptations in large-scale genomic data. Cell 152: 703-13.

35. Hünemeier T, Amorim CEG, Azevedo S, Contini V, Acuña-Alonzo V (2012) Evolutionary responses to a constructed niche: ancient Mesoamericans as a model of gene-culture coevolution. PLoS One 7: e38862.

36. Lee CJ, Lin HR, Liao CL, Lin YL (2008) Cholesterol effectively blocks entry of flavivirus. J Virol 82: 6470-80

37. Combes V, Coltel N, Alibert M, van Eck M, Raymond C et al. (2005) ABCA1 gene deletion protects against cerebral malaria: potential pathogenic role of microparticles in neuropathology. Am J Pathol 166: 295-302.

38. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P et al. (2006) Positive natural selection in the human lineage. Science 312: 1614-20.

39. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3: e90.

40. López Herráez D, Bauchet M, Tang K, Theunert C, Pugach I et al. (2009) Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PLoS One 4: e7888.

41. Kaindl AM, Jakubiczka S, Lücke T, Bartsch O, Weis J et al. (2005) Homozygous microdeletion of chromosome 4q11-q12 causes severe limb-girdle muscular dystrophy type 2E with joint hyperlaxity and contractures. Hum Mutat 26: 279-80.

42. Sun L, Blair HC, Peng Y, Zaidi N, Adebanjo OA et al. (2005) Calcineurin regulates bone formation by the osteoblast. Proc Natl Acad Sci USA 102: 17130-5.

43. Migliano AB, Romero IG, Metspalu M, Leavesley M et al. (2013) Evolution of the pygmy phenotype: evidence of positive selection fro genome-wide scans in African, Asian, and Melanesian pygmies. Hum Biol 85: 251-84.

44. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-8.

45. Kriska T, Pilat A, Schmitt JC, Girotti AW (2010) Sterol carrier protein-2 (SCP-2) involvement in cholesterol hydroperoxide cytotoxicity as revealed by SCP-2 inhibitor effects. J Lipid Res 51: 3174-84.

46. Umemura M, Fujita M, Yoko-O T, Fukamizu A, Jigami Y (2007) *Saccharomyces cerevisiae* CWH43 is involved in the remodeling of the lipid moiety of GPI anchors to ceramides. Mol Biol Cell 18: 4304-16.

47. Castelletti E, Lo Caputo S, Kuhn L, Borelli M, Gajardo J et al. (2007) The mucosae-associated epithelial chemokine (MEC/CCL28) modulates immunity in HIV infection. PLoS One 2: e969.

48. Ezzat MH, Sallam MA, Shaheen KY (2009) Serum mucosa-associated epithelial chemokine (MEC/CCL28) in atopic dermatitis: a specific marker for severity. Int J Dermatol 48: 822-9.

49. Meimaridou E, Kowalczyk J, Guasti L, Hughes CR, Wagner F et al. (2012) Mutations in NNT encoding nicotinamide nucleotide transhydrogenase cause familial glucocorticoid deficiency. Nat Genet 44: 740-2.

50. Sandqvist A, Björk JK, Akerfelt M, Chitikova Z, Grichine A et al. (2009) Heterotrimerization of heat-shock factors 1 and 2 provides a transcriptional switch in response to distinct stimuli. Mol Biol Cell 20: 1340-7.

51. Chung S, Tamura K, Furihata M, Uemura M, Daigo Y et al. (2009) Overexpression of the potential kinase serine/ threonine/tyrosine kinase 1 (STYK 1) in castration-resistant prostate cancer. Cancer Sci 100: 2109-14.

52. Zhang W, Zhang S, Xiao C, Yang Y, Zhoucun A (2007) Mutation screening of the FKBP6 gene and its association study with spermatogenic impairment in idiopathic infertile men. Reproduction 133: 511-6.

53. Doll A, Grzeschik KH (2001) Characterization of two novel genes, WBSCR20 and WBSCR22, deleted in Williams-Beuren syndrome. Cytogenet Cell Genet 95: 20-7.

54. Micale L, Fusco C, Augello B, Napolitano LM, Dermitzakis ET et al. (2008) Williams-Beuren syndrome TRIM50 encodes an E3 ubiquitin ligase. Eur J Hum Genet 16: 1038-49.

FIGURE LEGENDS

Figure 1: Manhattan plots of the distribution of SNPs (*x*-axis) and corresponding *q*-values (*y*-axis) for inferring selection. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black or grey), in Africa (blue), in the Americas (green), and in both continents (convergent evolution, red). Different sets of populations were used in the analysis yielding four different population sets (PS1-4). Congruent clusters of outlier SNPs considering all four PSs are highlighted with a grey box.

Figure 2: Gene sets significantly enriched for selection signals. Nodes are gene sets that score a q-value ≤ 20% after pruning in the gene set enrichment test (See Supplementary Figure 7 for the equivalent plot regarding the results before pruning). Gene sets are connected when at least one of them shares >33% of its genes with the other set before pruning. The size of a node scales with the size of the gene set. Each quadrant of a node represents results on one of the four selection scores, while the four rings in a node correspond to the four population sets (different combination of populations; see Methods). The color of each quarter of a ring corresponds to the significance of the test result (dark colored: q≤10%, light colored: 10%<q≤20%, white: q>20%).

TABLES

Table 1. Genes with signals of positive selection suggesting human adaptations to tropical forests in Africa and the Americas.

Gene	Cluster	Biological function in mammalians or associated human diseases
SCP2	1	Involved in cholesterol trafficking and metabolism [45].
CWH43	3	Enhance lipid remodeling to ceramides [46].
DCUN1D4	4	Unknown.
LRRC66	4	Unknown.
SGCB	4	Is located in a genomic region where a microdeletion causes limb-girdle muscular dystrophy type 2E with joint hyperlaxity and contractures [41].
C5orf34	5	Unknown.
CCL28	5	Modulate immunity to viral infection [47] and skin-related inflammatory diseases [48].
NNT	5	Produces high concentrations of NADPH at mitochondria and the resulting energy is used for biosynthesis and in free-radical detoxification [49].
HSF2	6	Involved in the activation of heat-shock response genes under conditions of heat [50].
PKIB	6	Associated to the aggressive phenotype of prostate cancer [51].
SERINC1	6	Unknown.
FKBP6	7	May play a role in modifying the susceptibility to idiopathic spermatogenic impairment [52].
NSUN5	7	Deleted in Williams-Beuren syndrome (vascular system and calcium metabolism problems) [53].
TRIM50	7	May be involved in the Williams-Beuren syndrome [54].

SUPPORTING INFORMATION

Supplementary Figure 1A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 1. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 1A; PS2: 1B; PS3: 1C; PS4: 1D)**.**

Supplementary Figure 2A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 2. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 2A; PS2: 2B; PS3: 2C; PS4: 2D)**.**

Supplementary Figure 3A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 4. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 3A; PS2: 3B; PS3: 3C; PS4: 3D).

Supplementary Figure 4A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 5. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 4A; PS2: 4B; PS3:45C; PS4: 4D).

Supplementary Figure 5A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 6. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 5A; PS2: 5B; PS3: 5C; PS4: 5D).

Supplementary Figure 6A-D: Manhattan plots of the distribution of SNPs (*x*-axis) and their correspondent *q*-values (log-transformed at *y*-axis) for inferring positive selection in Chromosome 7. The sliding-window *q*-value is indicated by a yellow continuous line. With a False Discovery Rate of 0.1 (dashed line), SNPs are color-coded according to the best supported model of selection, namely neutrality (black), selection in Africa (blue), selection in the Americas (green), and in both continents (convergent evolution, red). When an outlier SNP was located less than 50 kb apart from a gene, the closest gene name was written next to it. Different sets of populations were used in the analysis yielding four different population sets (PS1: 6A; PS2: 6B; PS3: 6C; PS4: 6D).

Supplementary Figure 7: Gene sets significantly enriched for selection signals. Nodes are gene sets that score a q-value ≤ 20% (before pruning). Gene sets are connected when at least one of them shares >33% of its genes with the other set. The width of the connecting lines represents the amount of similarity between sets; the size of a node scales with the size of the gene set. Each quadrant of a node represents results on one of the four selection scores, while the four rings in a node correspond to the four population sets (different combination of populations; see Methods). The color of each quarter of a ring corresponds to the significance of the test result (dark colored: q≤10%, light colored: 10%<q≤20%, white: q>20%).

Figure
Click here to download high resolution image

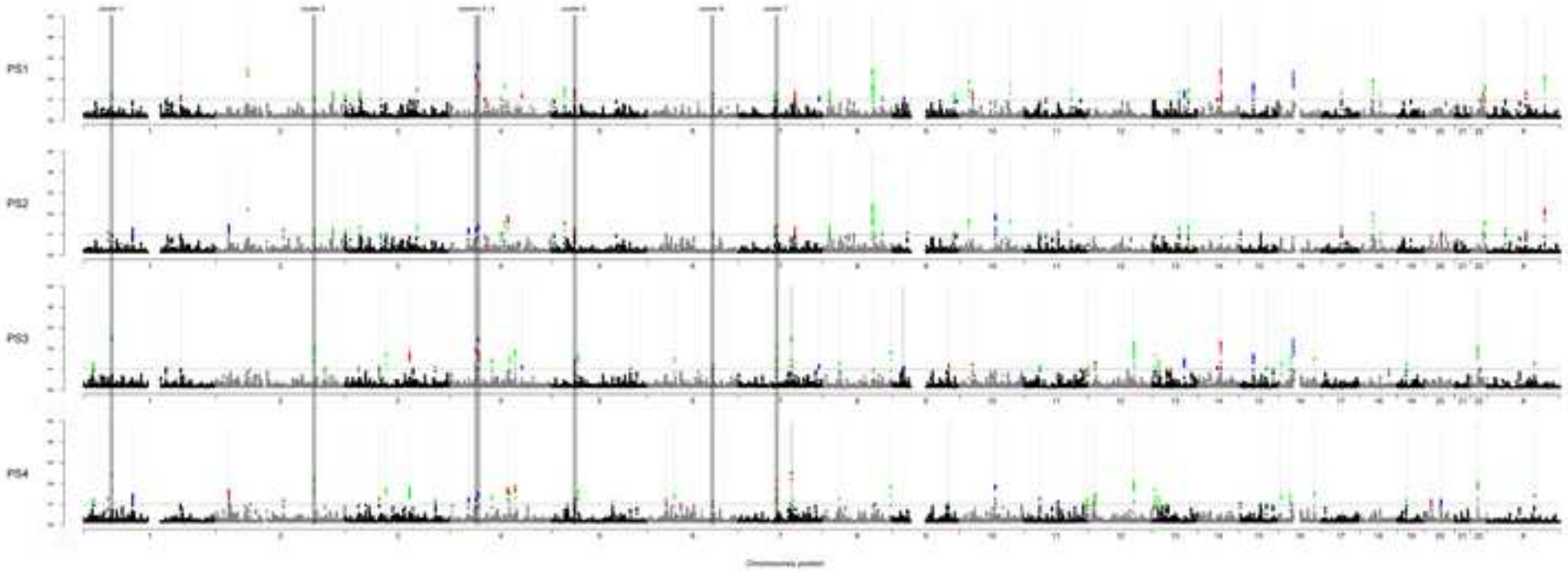
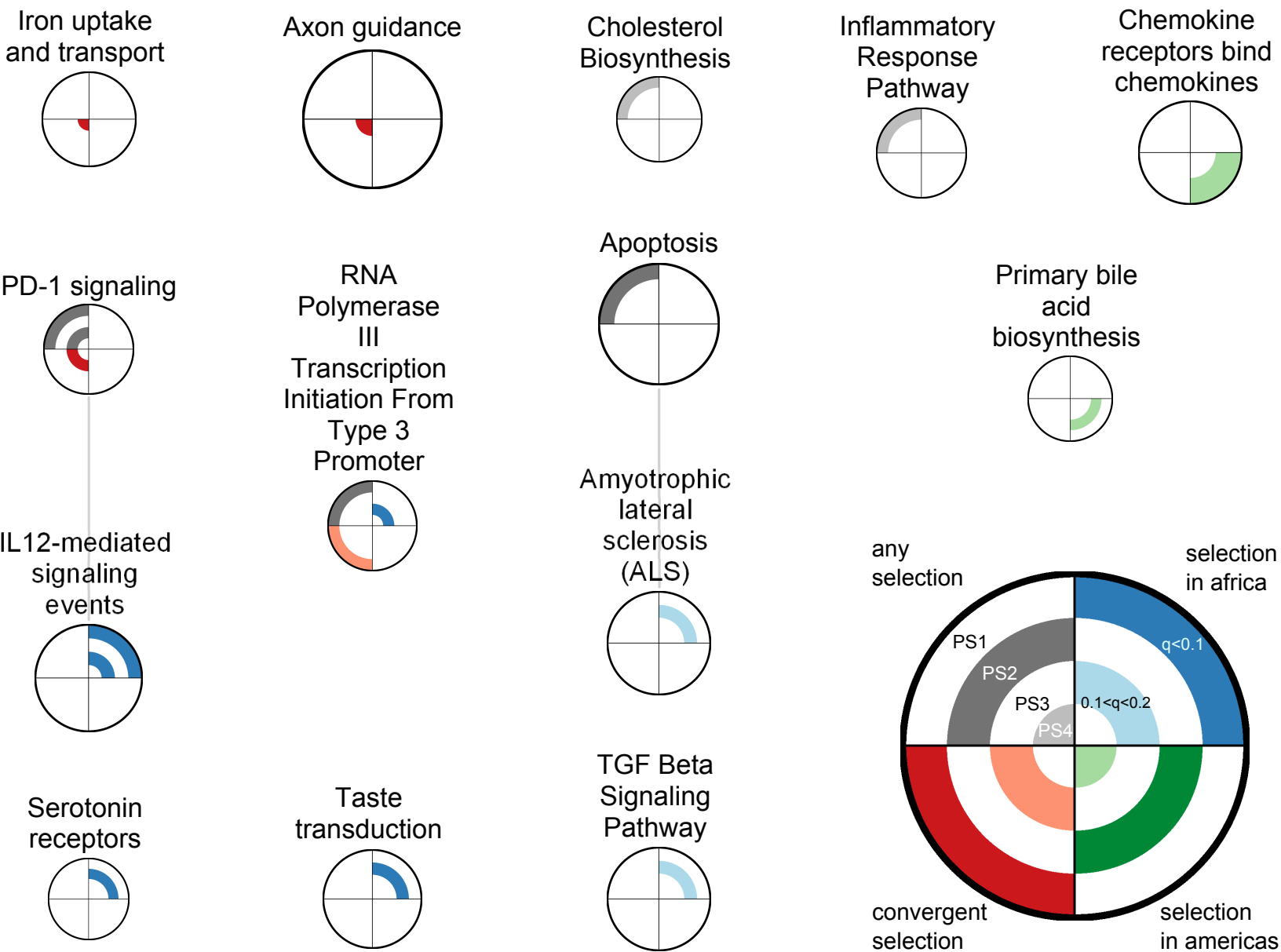


Figure
Click here to download Figure: Fig2.eps



Supporting Information

Click here to download Supporting Information: FigS1.tiff

Supporting Information

Click here to download Supporting Information: FigS2.tiff

Supporting Information

Click here to download Supporting Information: FigS3.tiff

Supporting Information

Click here to download Supporting Information: FigS4.tiff

Supporting Information

Click here to download Supporting Information: FigS5.tiff

Supporting Information

Click here to download Supporting Information: FigS6.tiff

Supporting Information

Click here to download Supporting Information: FigS7eps.eps