Supporting Materials

Results

Demographic model selection: using PSMC/MSMC

We applied the pairwise sequentially Markovian coalescent (PSMC, Li and Durbin 2011) and multiple sequentially Markovian coalescent (MSMC, Schiffels and Durbin 2014) as independent means to explore the demographic history of our populations (Materials and Methods). PSMC infers effective population size over time from a single diploid genome, while MSMC measures genetic separation of populations using relative cross coalescence rates between pairs of haplotypes from two populations. We applied both PSMC and MSMC to our quality-controlled intergenic data. The PSMC curves of the farmers begin to separate from those of the Pygmies roughly 100–200 kya (**Figure S3A**), suggesting that the ancestors of the farmers and Pygmies began differentiating from each other as early as 100–200 kya, consistent with the inferred divergence time in Model-1. The MSMC curves indicate declining genetic exchange between Pygmies and farmers ~40-60 kya, suggesting that these two populations may have diverged from each other at this time (**Figure S3D-E**). To test if Model-1 and Model-2 recapitulate the divergence times between farmers and Pygmies indicated by PSMC/MSMC, we applied both methods to simulated genomes under both models (**Figure S3B-E**). Under Model-1, the PSMC curves of the simulated Pygmy and farmer genomes split at about the same time as in the PSMC analysis of the real data (**Figure S3B**), while the two simulated populations of Model-2 do not show clear separation until ~70 kya (**Figure S3C**). The MSMC curves of Model-1 and those of real data agree well, but Model-2 seems to fit the MSMC curve from the real data poorly (**Figure S3D-E**). Interestingly, the divergence times indicated using MSMC differ from those we simulated, highlighting the complexity of interpreting MSMC results. Together, however, these results suggest that Model-1 qualitatively fits the data better, and the inferred ancient divergence time in Model-1 is plausible.

Importance of controlling variation in mutation and recombination rates across the genome

Methods for detecting natural selection often rely on summaries of local genetic variation, and they may be biased by variation in mutation rate across the genome (Reich et al. 2002; Drake et al. 2005; Schaffner et al. 2005; Sainudiin et al. 2007). For example, G2D values (Nielsen et al. 2009) are correlated with local genetic diversity (Pearson correlation 0.298, $p<2.2x10^{-16}$, **Figure S4**). We addressed this by estimating and incorporating local mutation rate variation in our simulations (Materials and Methods), and our simulations can reproduce local genetic diversity in the real data (Pearson correlation=0.902, **Figure S5**). To assess whether mutation-rate heterogeneity could bias downstream inferences of selection, we compared results using two different sets of simulations under Model-1 to assign *P*-values. In the first set, the local mutation rate for each window was assigned to be the mean rate of the recombination decile to which that window belonged (**Figure S6**). In the second set, we estimated a local mutation rate for each window individually (**Figure S7**). The *P*-value distributions of G2D based on these two sets of simulations were calculated, and for both analyses we chose the top 0.5% windows in the *P*-value distributions as the top-hits. There is a clear shift to larger heterozygosity (estimated using θ/base) for the top hits in the first simulation set (**Figure S6A**), compared with the second set (**Figure S7A**). As expected, the top hits in the first simulation set tended to be windows with larger numbers of variants, while the top hits from the second set were distributed across the whole range of observed heterozygosity across the genome (**Figure S6B** vs. **Figure S7B**). This suggests that incorrectly incorporating mutation rate variation in whole-genome simulations might lead to biases toward regions with unusually high mutation rate as candidates of natural selection.

The distribution of *P*-values was sensitive to the genetic recombination map used in the simulations (**Figure S8**). In particular, the distribution of G2D p-values using the African American map (Hinch et al. 2011) is shifted more toward p=1 than using the Yoruba HapMap map, suggesting that inference using the African American map would be more conservative (**Figure S8**). To avoid potential biases due to the choice of map and/or null model, we restricted our candidates to those that are top hits using all four combinations of the two recombination maps and the two best-fit demographic models.

Because the *P*-value distributions based on the two null demographic models are highly correlated (Pearson correlation=0.984, p<2.2x10^{-16}, **Figure S9**), and the analysis based on the African American map is more conservative, unless mentioned otherwise we report *P*-values and false discovery rates obtained using Model-1 and the African American map.

Selection scan using iHS: bone synthesis and muscle-related candidates

Among the candidates of our iHS scans for signals of selection, five loci contain genes associated with bone synthesis. Except *EPHB1*, which is discussed in details in the main text, the other four are *SLCO2A1* (locus: chr3:133506737-133863702), *ZBTB38* (locus: chr3:141105569-141333249), *TSPAN5* (locus: chr4:99496207-99673561), and *GAREM* (locus: chr18:29766032-29896024). *SLCO2A1* encodes a prostaglandin transporter protein, and mutations in this gene have been shown causing Primary Hypertrophic Osteoarthropathy, a rare genetic disease that affects both skin and bones (Zhang et al. 2012). *ZBTB38* encodes a zinc finger transcriptional activator expressed in the brain, and has been associated with adult height in multiple populations (Lettre et al. 2008; Weedon et al. 2008; Wang et al. 2013). *TSPAN5* is a member of the tetraspanin protein family and is up-regulated during osteoclast differentiation (Iwai et al. 2007); knockdown of its expression dramatically inhibits osteoclastogenesis in vitro (Iwai et al. 2007; Zhou et al. 2014), suggesting its regulatory role in bone development. *GAREM* is an adapter protein in intracellular signaling cascades and has recently been associated with human height in a whole-exome sequencing association study (Kim et al. 2012). A few large F_{ST} (≥ 0.2) non-synonymous amino acid substitutions were observed within these candidate regions, but they are not suggested as functionally important by SIFT (Kumar et al. 2009) or PolyPhen-2 (Adzhubei et al. 2010). Regions near four out of these five genes, however, show high levels of differentiated SNVs in enhancer/Polycomb-repressed sequences, implying that Pygmy short stature might arise partly through *cis*-regulatory evolution (**Figure S11**).

In addition to OBSCN, two candidate loci also encompass muscle-related genes, *COX10* (locus: chr17:13911228-14241158) and *LARGE* (locus: chr22:34224706-34359718). *COX10* is a cytochrome c oxidase, and Diaz et al. (2005) reported that *COX10* knockout mice develop a slowly progressive myopathy. *LARGE* is a member of the N-acetylglucosaminyltransferase gene family, and mutations in this gene cause a form of congenital muscular dystrophy (Longman et al. 2003). Interestingly, Andersen et al. (2012) recently found evidence that variants in *LARGE* might have been positively selected for the resistance of Lassa fever in Western African populations.

Selection scan using G2D: reproduction and gene regulation-related candidates

One of our G2D candidate regions (locus: chr1:183076845-183184161) includes the gene *LAMC1*, which plays a role in reproductive development. *LAMC1* expression increases in bovine, pig, and rabbit basal lamina during follicular development (Irving-Rodgers and Rodgers 2005), and is also expressed in the human ovary (Berkholtz et al. 2006). A recent genome-wide association study reported that polymorphisms in *LAMC1* are associated with an increased risk of premature ovarian failure, which is characterized as the cessation of ovarian function before the age of 40 and could result in amenorrhea and infertility (Pyun et al. 2012). Another interesting G2D candidate region (chr19:12386669-12523799) contains cell signal transmission genes, the *ZNF* genes, which encode proteins with KRAB and zinc-finger domains. Genes in this protein family have been previously shown to be under positive selection in African Americans (Nielsen et al. 2005; Nielsen et al. 2009). It is unclear what phenotype these variants are associated with, but the role of *ZNF442* in transcriptional binding activity suggests *trans*-regulatory evolution might play a role in the adaptation of Pygmies.

Methods

Using GRCh37/hg19 for read alignment

Our genomes were assembled using the default Complete Genomics (CGI) analysis pipeline (v.1.10). Because CGI is currently not supporting GRCh38 (personal communication with the senior scientist Dr. Birgit Crain at CGI, E-mail date: 12.07.2015), our data was aligned according to GRCh37/hg19. According to the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/info/index.shtml), the major improvements in GRCh38, compared to GRCh37/hg19, are in 1) providing alternative scaffolds to better represent complex regions (e.g. centromeres) in the genome; 2) closing/reducing the known gaps in the previous releases (sizes of gaps: ~234 Mb or 7.6% in GRCh37/hg19 and ~151Mb or 4.9% in GRCh38); 3) correcting assembly errors, particularly for complex regions, in the previous releases. While realigning our data to GRCh38 might yield more data in complex regions, we do not expect this will affect our conclusions, because we excluded most complex regions and known gaps from all of our analyses (Materials and Methods).

Data quality control for genotype calls

Before any quality control filters, 13,276,198 autosomal single nucleotide variants (SNVs) were called in our samples. Unless mentioned otherwise, we analyzed only variants that were 1) fully called across all samples, 2) not in any known or called indels, 3) not in any known or CGI called copy number variants, 4) not in any known segmental duplication regions, and 5) aligned against chimpanzee (PanTro3, Hg19). Databases used for steps 3, 4, and 5 were downloaded from UCSC Genome Browser in May 2013. We used Hg19 coordinates, using the UCSC Genome Browser liftOver program if necessary. After filtering, our data consist of 10,865,288 SNVs.

Demographic inference using *∂a∂i*

∂a∂i is a forward time simulator of allele frequency spectrum (AFS) based on a diffusion approximation (Kimura 1964). To ensure genotype quality for demographic inference using our sample, SNVs were removed if they overlapped with any known repetitive genomic regions based on the UCSC Genome Browser databases, Self Chain (if sequence identity > 0.9) and RepeatMasker. We also excluded sites that are within known copy number variants (CNVs; Database of Genomic Variants, as of May 2013) as well as the CGI called CNVs. Sites within genes or 1,000 flanking base pairs were excluded to minimize possible effects of natural selection. Coordinates of genes were from the RefSeq genes database, downloaded from the UCSC Genome Browser in May 2013. We used the remaining 1,575,394 SNVs from a total of 325,957,426 non-genic base pairs to build an unfolded AFS. Ancestral states were inferred using chimpanzee as the outgroup, using human-chimpanzee alignment (PanTro3, Hg19). The estimated sequence divergence between human and chimpanzee based on these non-genic sequences is 1.14%. We used the ∂a∂i implementation of a context-dependent substitution model to statistically correct the unfolded AFS to mitigate possible biases due to ancestral state misidentification (Hernandez et al. 2007). To estimate demographic parameters, the derivative-based BFGS algorithm was used to optimize the composite log-likelihood.

To test if including sites within putatively functional non-genic regions could bias the AFS, SNPs within the top 12 strongest signals (i.e. not including the three types: 13_Heterochrom/low signal, 14_Repetitive/CNV, 15_Repetitive/CNV) of ENCODE elements (Gerstein et al. 2012) were removed from our original non-genic data, resulting in a 20% reduction of the data (from ~1.5 millions to ~1.2 millions). To quantitatively assess for deviations between the AFS of the two data sets, from the original data set we computationally generated 1,000 bootstraps (random sampling with replacement), in which each bootstrap has the same number of SNPs as in the ENCODE-filtered data set. For each entry in the AFS, we then assessed where the ENCODE-filtered data set was within the distribution of values obtained from the bootstraps of the original data (the third row of Figure S1). We found that the two AFS from the 1.5 millions and 1.2 millions SNPs are neither qualitatively nor quantitatively different

(Supporting Materials, Figure S1). Thus, we expect that using the ENCODE-filtered AFS will not change any of our conclusions of demographic inference.

Haplotype phasing

Haplotype phasing was done using BEAGLE v3.1.1 (Browning and Browning 2007). To enhance phasing accuracy, we included two additional public pygmy genomes, a Bakola and a Bedzen genome (CGI Assembly Pipeline 1.10, CGA Tools 1.4) from Lachance et al. (2012), into our genome sample. In order to obtain population-specific phased haplotypes for the Pygmies, we first constructed a scaffold for each chromosome using the 36 SNP-chip samples, including 16 Biaka Pygmies genotyped by the Human Genome Diversity Project (HGDP, Li et al. 2008, Illumina 650 K), and 10 Baka and 10 Bakola Pygmies genotyped by the Hammer lab of the University of Arizona (Affymatrix Axiom 500K). The 9 Pygmy genomes were then phased using BEAGLE, with the pre-phased scaffold as the reference. The 9 Yoruba genomes were phased separately using the same framework, together with an additional 4 Luhya genomes from the CGI public data repository. The scaffold for the Yoruba genomes consisted of genotype data for 81 Yoruba and 86 Luhya samples from the 1000 Genomes Project and 21 Yoruba and 10 Luhya samples from the HGDP. All of these samples were determined to be unrelated using the identical-by-descent operation in PLINK (Purcell et al. 2007). All positions were converted into Hg19 coordinates using the UCSC LiftOver utility if necessary.

Haplotype and diplotype analyses

Hierarchical clustering for both haplotype and diplotype data was performed using the R function “hclust” in the stats package (R Development Core Team, 2012). We used the R package pegas (v.0.6, Paradis 2010) to plot haplotype network, using pairwise nucleotide differences as the distance matrix.

ENCODE regulatory elements

We downloaded the ENCODE (Gerstein et al. 2012) database (wgEncodeBroadHmmHsmmHMM) using the UCSC Genome Browser in February 2014. We used the five most reliable functional categories: Active Promoter (state 1), Strong Enhancer (states 4 and 5), Insulator (state 8), and Polycomb-repressed (state 12). This yielded 134,769 regulatory elements.

Supplementary Figures

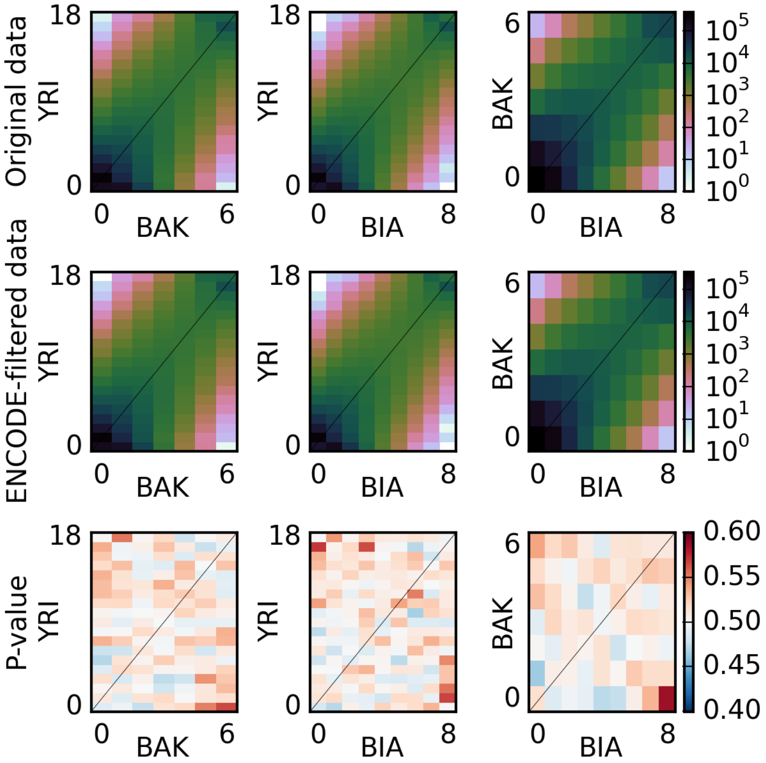


Figure S1. Evaluation of the possible effect of including ENCODE functional elements on the allele frequency spectrum (AFS). Top row: the 2-population marginal AFS of original 1.58 million intergenic SNPs, scaled to match the number of SNPS in the ENCODE-filtered subset. The middle row: the 2-population marginal AFS of the ENCODE-filtered 1.2 million intergenic SNPs, after excluding the three lowest signals among the ENCODE elements (13_Heterochrom/low signal, 14_Repetitive/CNV, 15_Repetitive/CNV). Bottom row: location of each entry of the ENCODE-filter AFS in the distribution of values from 1,000 bootstraps, each of which had 1.2 million SNPs sampled from the original intergenic data set.

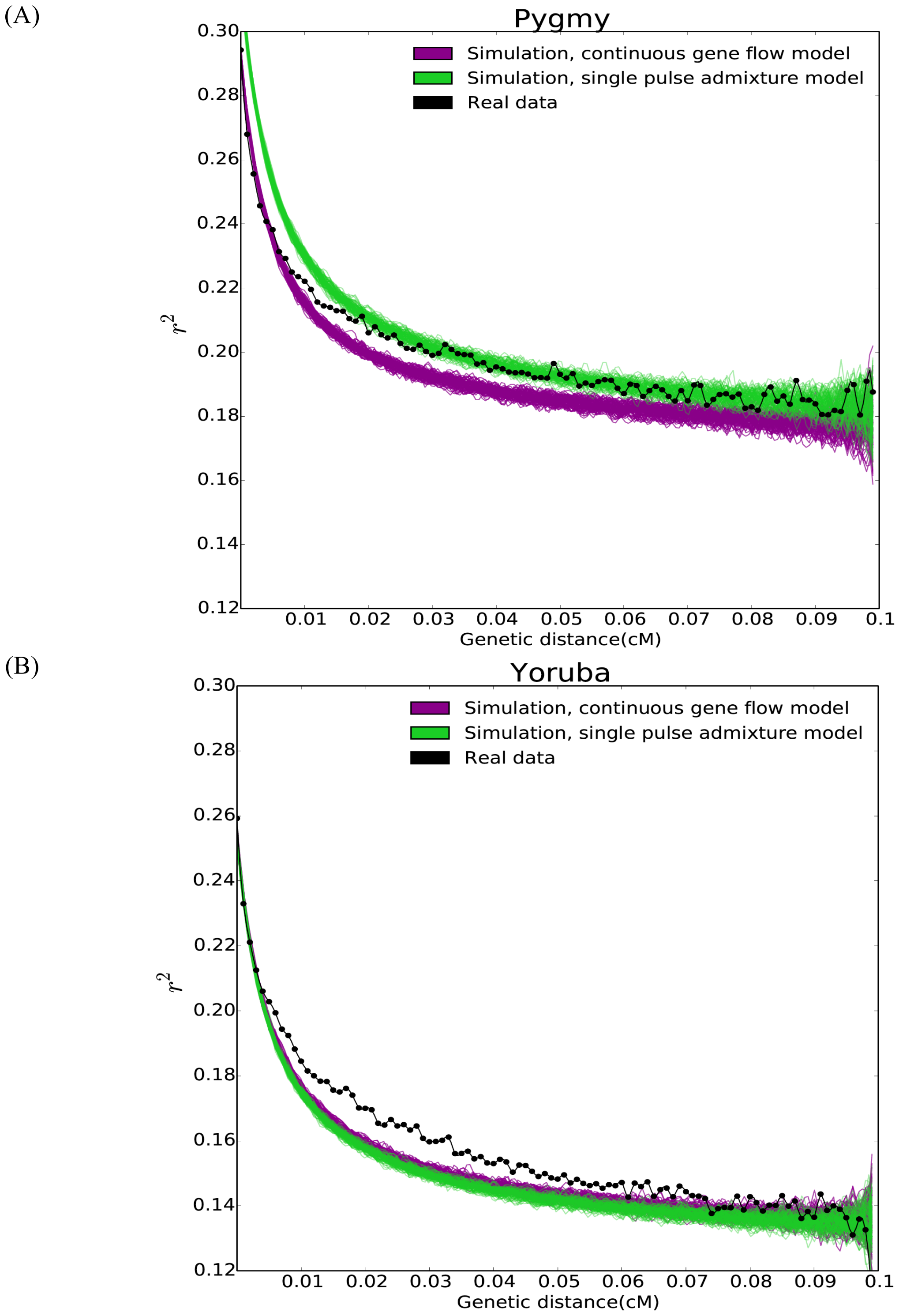
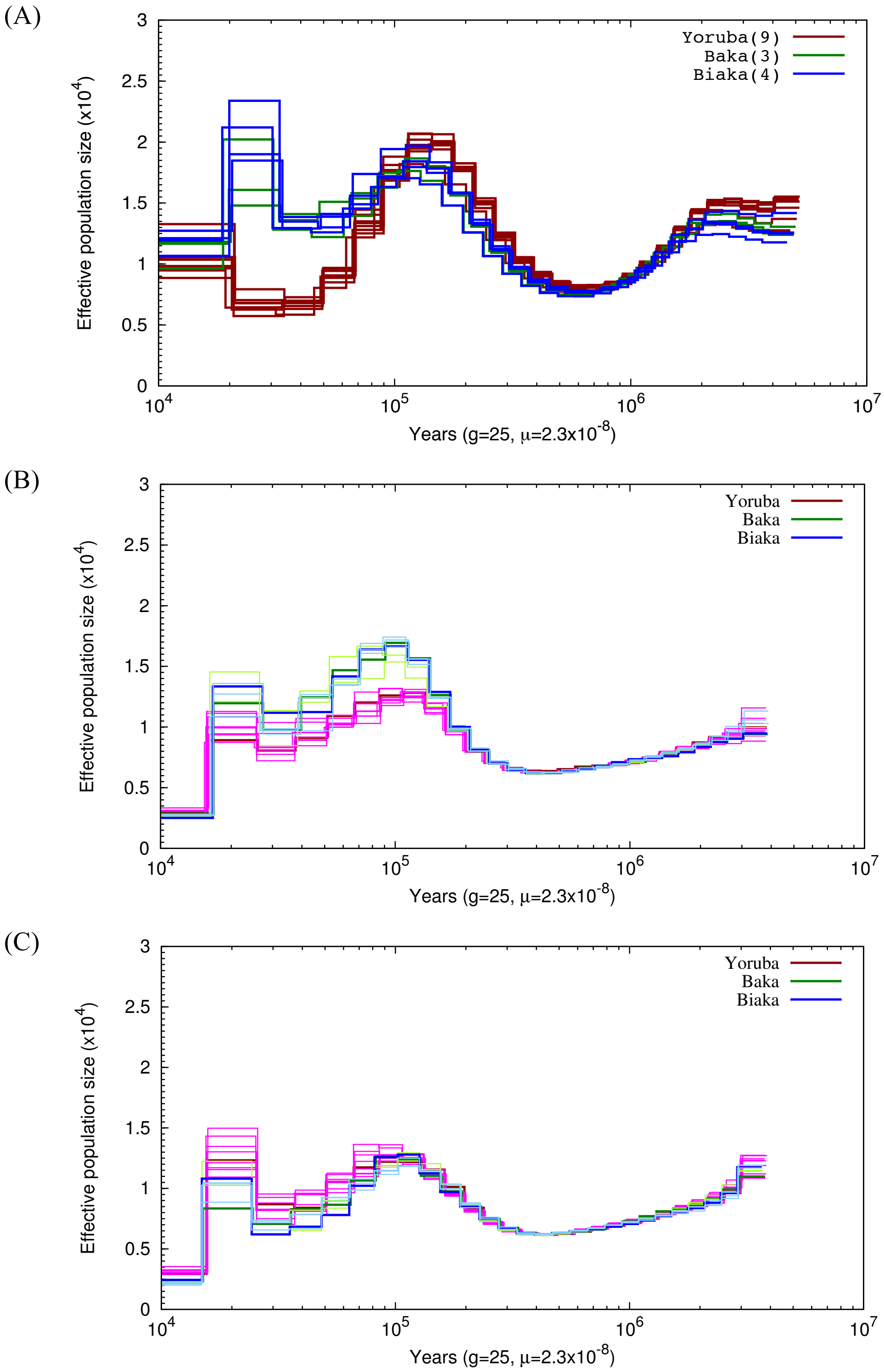


Figure S2. Decay of linkage disequilibrium in Pygmies and farmers in real and simulated data. Simulations are based on 100 models drawn from the confidence intervals of the parameter estimates for each of the two best-fit models. LD is estimated using correlation coefficient (r^2) between pairs of variants in 0.1 cM windows across the whole genome. (A) The Pygmies; (B) the Yoruba farmers.


(A)
Yoruba(9)
Baka(3)
Biaka(4)
Effective population size ($x10^4$)
0
0.5
1
1.5
2
2.5
3
10^4
10^5
10^6
10^7
Years (g=25, μ=$2.3x10^{-8}$)
(B)
Yoruba
Baka
Biaka
Effective population size ($x10^4$)
0
0.5
1
1.5
2
2.5
3
10^4
10^5
10^6
10^7
Years (g=25, μ=$2.3x10^{-8}$)
(C)
Yoruba
Baka
Biaka
Effective population size ($x10^4$)
0
0.5
1
1.5
2
2.5
3
10^4
10^5
10^6
10^7
Years (g=25, μ=$2.3x10^{-8}$)

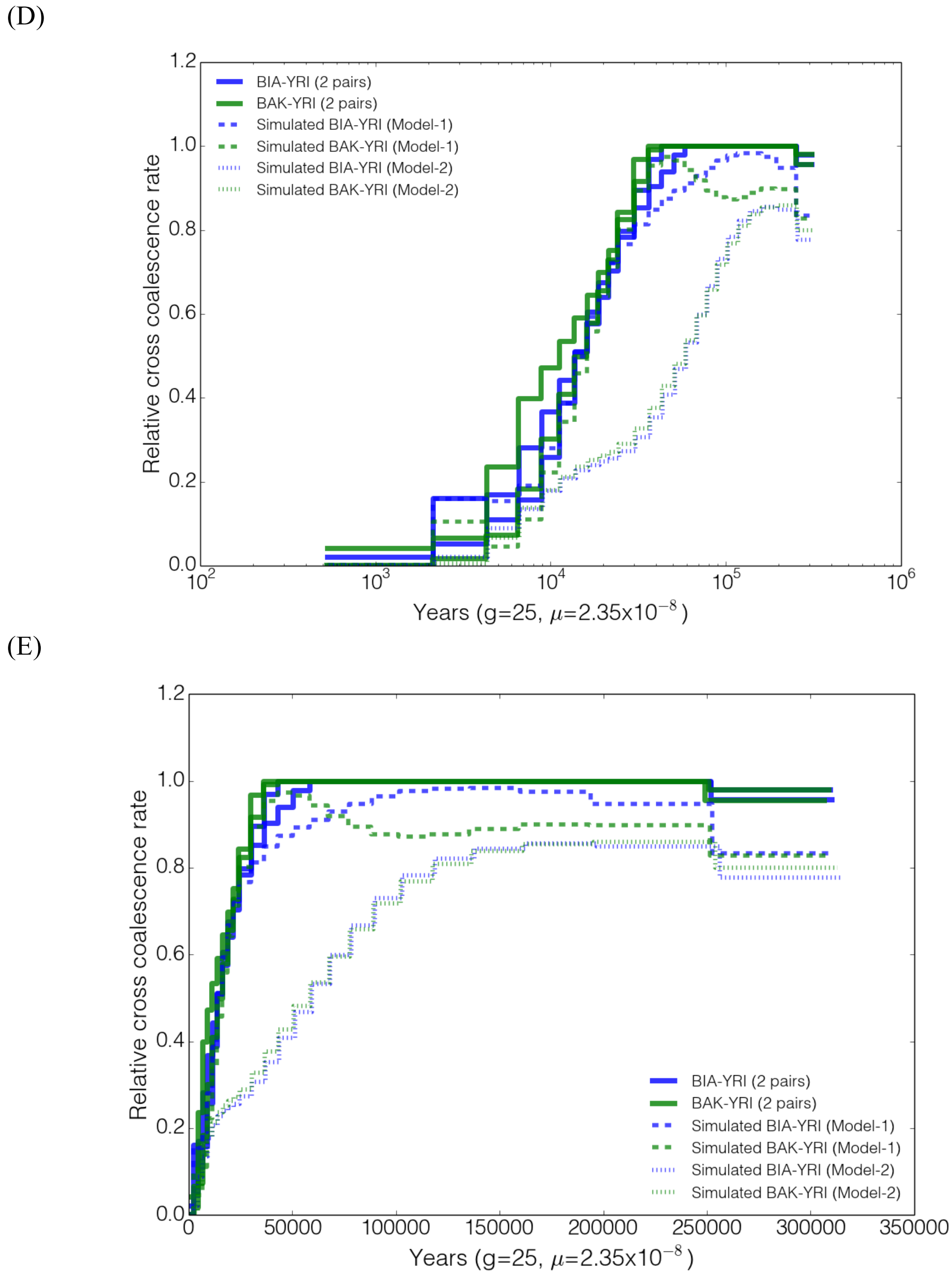


Figure S3. PSMC and MSMC analyses. (A) We performed PSMC (Li and Durbin 2011) analysis on the whole genome samples of the farmer and Pygmy populations. Each line represents a genome, plotted as the evolution of effective population size against time. The number inside parentheses in the legend indicates the sample size in each population. (B-C) PSMC analysis using simulated genomes: Model-1, the continuous asymmetric gene flow (B), Model-2, the single pulse admixture model (C). Red lines are the farmer population (Yoruba), and the green and blue lines are the two Pygmy groups (Baka and Biaka). (D) The MSMC (Schiffels and Durbin 2014) results for two random pairs of Biaka-Yoruba (blue solid lines) and of Baka-Yoruba (green solid lines), one random pair for simulated Biaka-Yoruba and Baka-Yoruba genomes from Model-1 (dash lines) and Model-2 (dot lines). Curves are plotted on a logarithm scale on x-axis. (E) The same results from (D), but plotted on an regular scale.

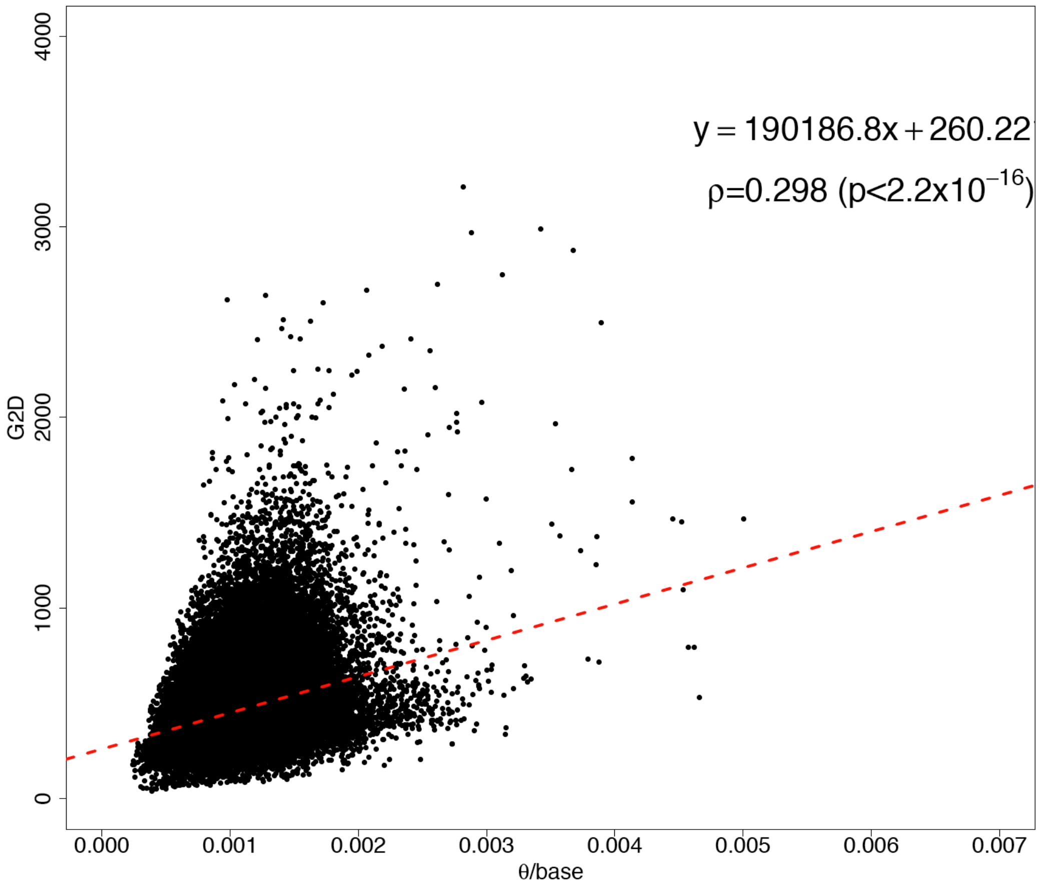


Figure S4. Dependency of G2D statistic on local heterozygosity. Each point represents a window with 500 SNVs. Red dashed line shows the result of a linear regression. Correlation is Pearson's correlation.

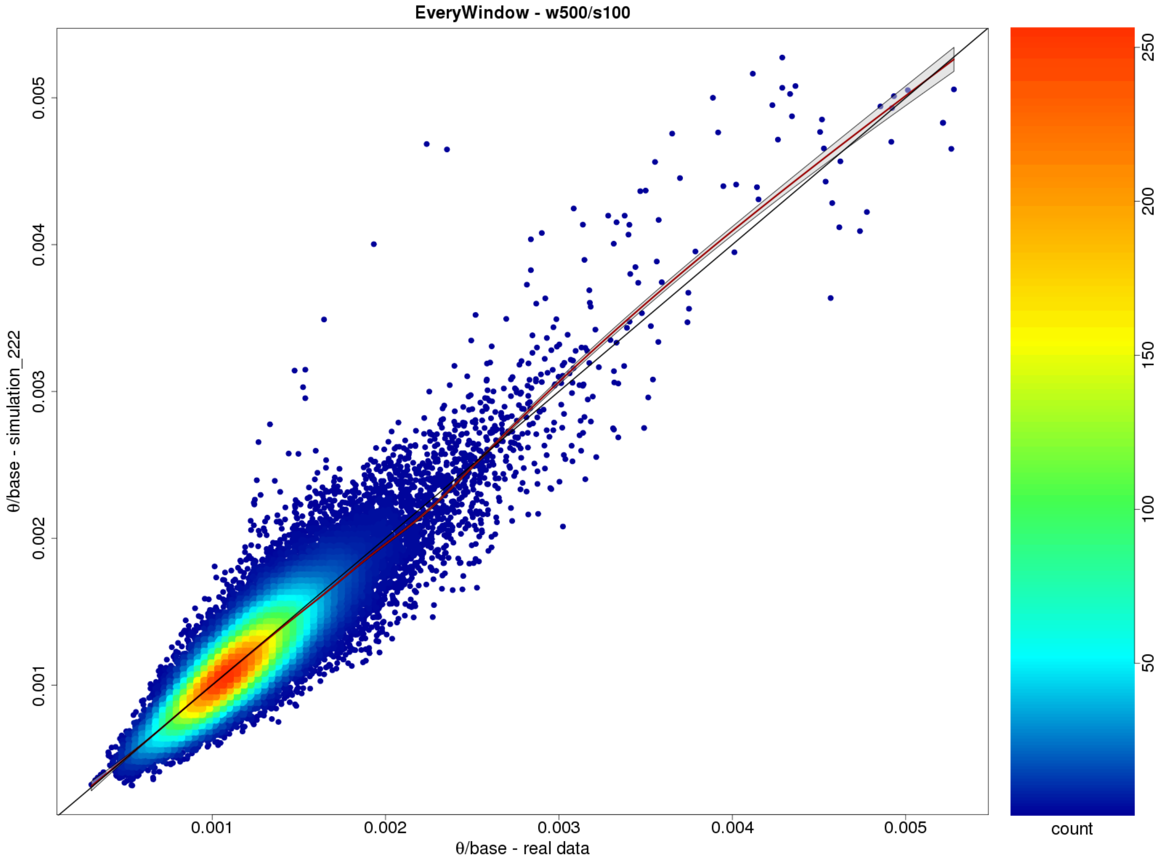


Figure S5. Correlation of per-base θ (Watterson's estimator) between windows in real and simulated whole-genome data. Window are defined as in our selection scans. Pearson correlation is 0.902.

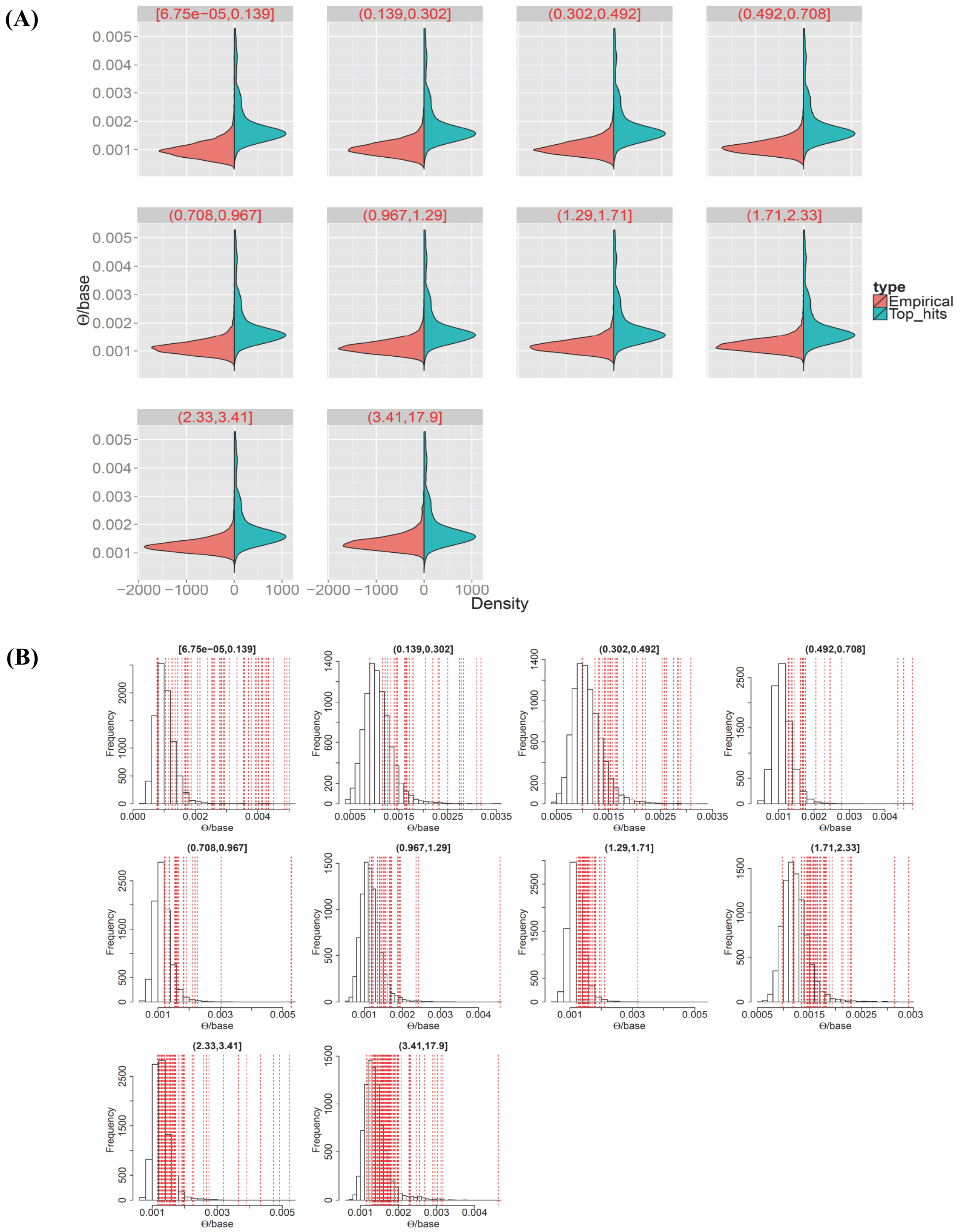


Figure S6. Biased distribution of candidates for selection (top 0.5% in *P*-value distribution) with respect to local mutational heterogeneity. The mutation parameter of each window in the simulation is assigned to be the mean rate of the recombination decile to which that window belonged. Each subpanel shows one of the 10 recombination rate deciles. (A) A clear shift to larger heterozygosity for the top hits under this simulation design. (B) The top hits (red vertical lines) tended to be windows with larger numbers of variants.

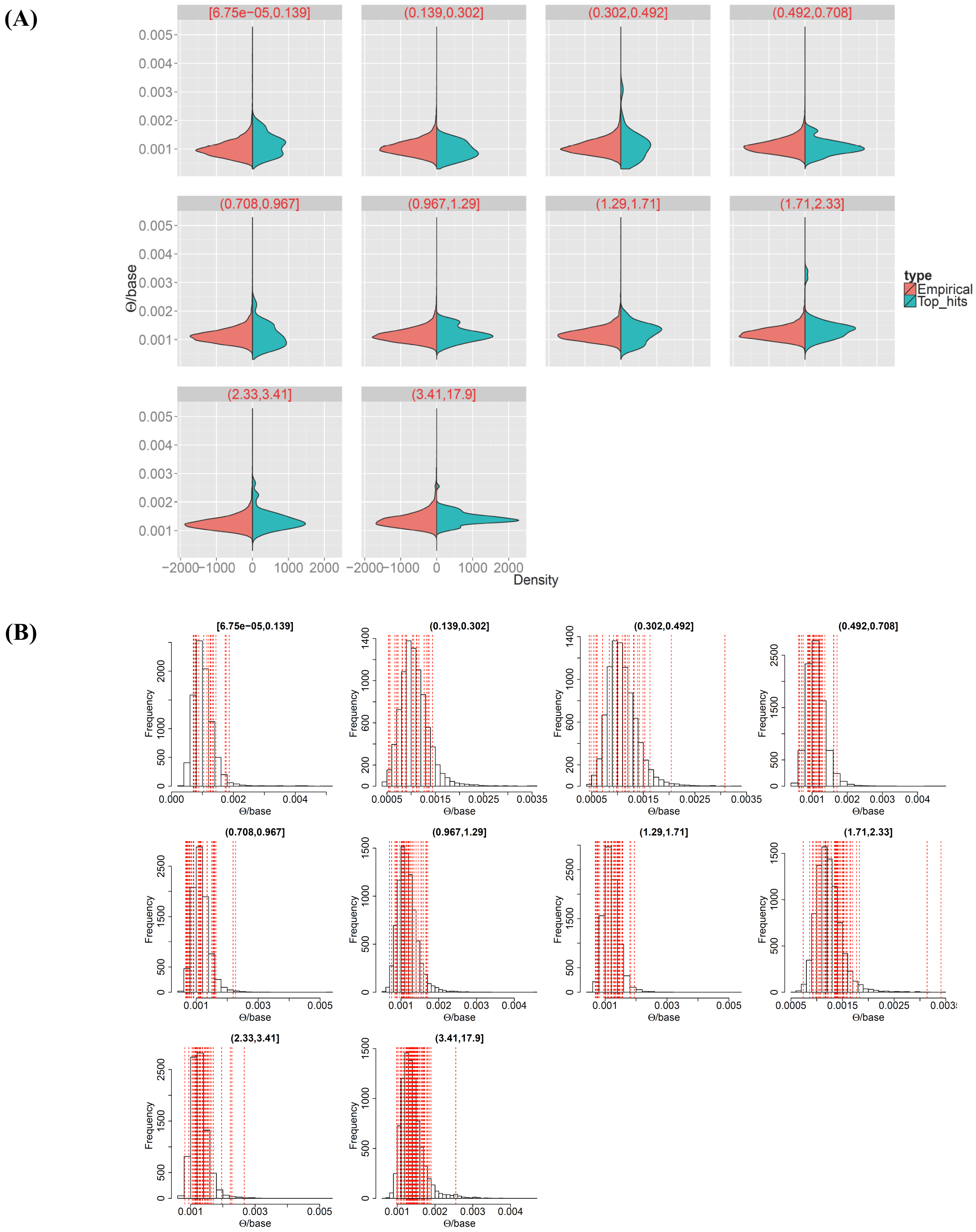


Figure S7. Unbiased distribution of candidates for selection (top 0.5% in *P*-value distribution) with respect to local mutational heterogeneity. The mutation parameter of each window in the simulation matches its local mutation rate. (A) No clear shift in heterozygosity for the top hits under this simulation design. (B) The top hits were distributed across the whole range of observed heterozygosity across the genome.

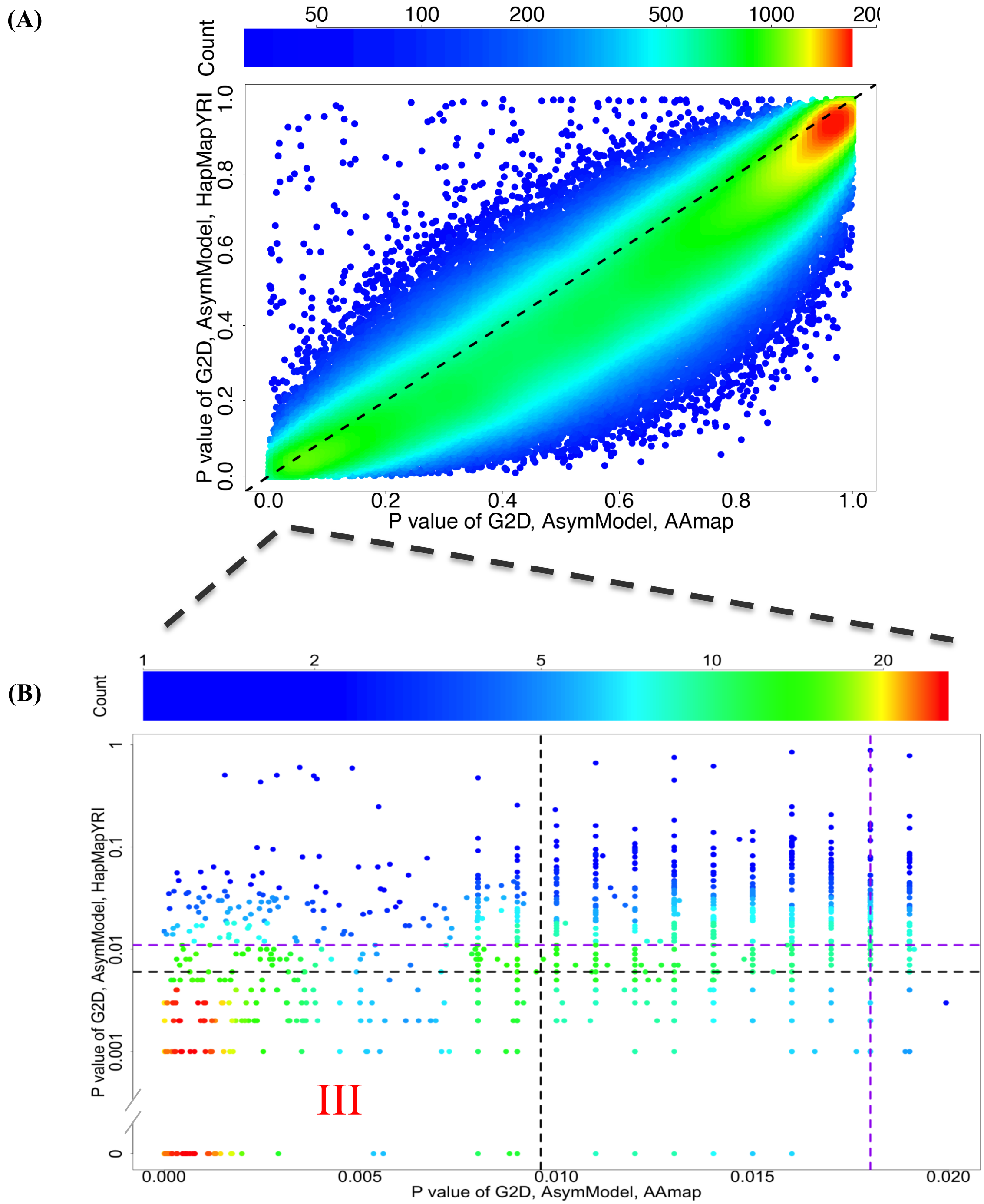


Figure S8. Dependence of top-hits in P-value approach on genetic recombination map used in simulations. (A) Shown are the *P*-values for windows simulated using the per-window mutation-rate estimation approach under Model-1 (AsymModel), but with two different published genetic recombination maps: AAmap: African American genetic recombination map (Hinch et al. 2011), HapMapYRI: Yoruba HapMap genetic recombination map (The International HapMap Consortium 2007). (B) Zoom in to the bottom-left corner of (A). Black and purple lines indicate the top 0.5% and 1% cutoff in *P*-value distributions. Only windows in Quadrant-III are robust to the choice of genetic recombination map. The color scheme represents the density of the windows on the plot. Similar results hold for simulations under Model-2 and for iHS as well

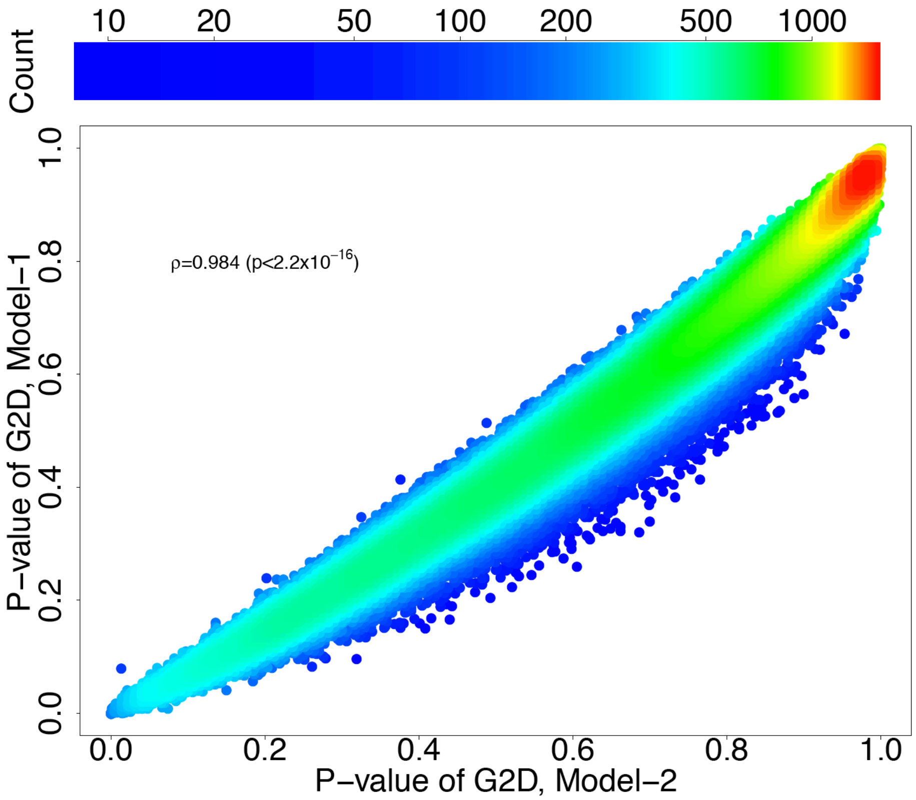


Figure S9. Correlation between P-values under our two best-fit models. Each point is a window of 500 SNVs and color represents the density of the points.

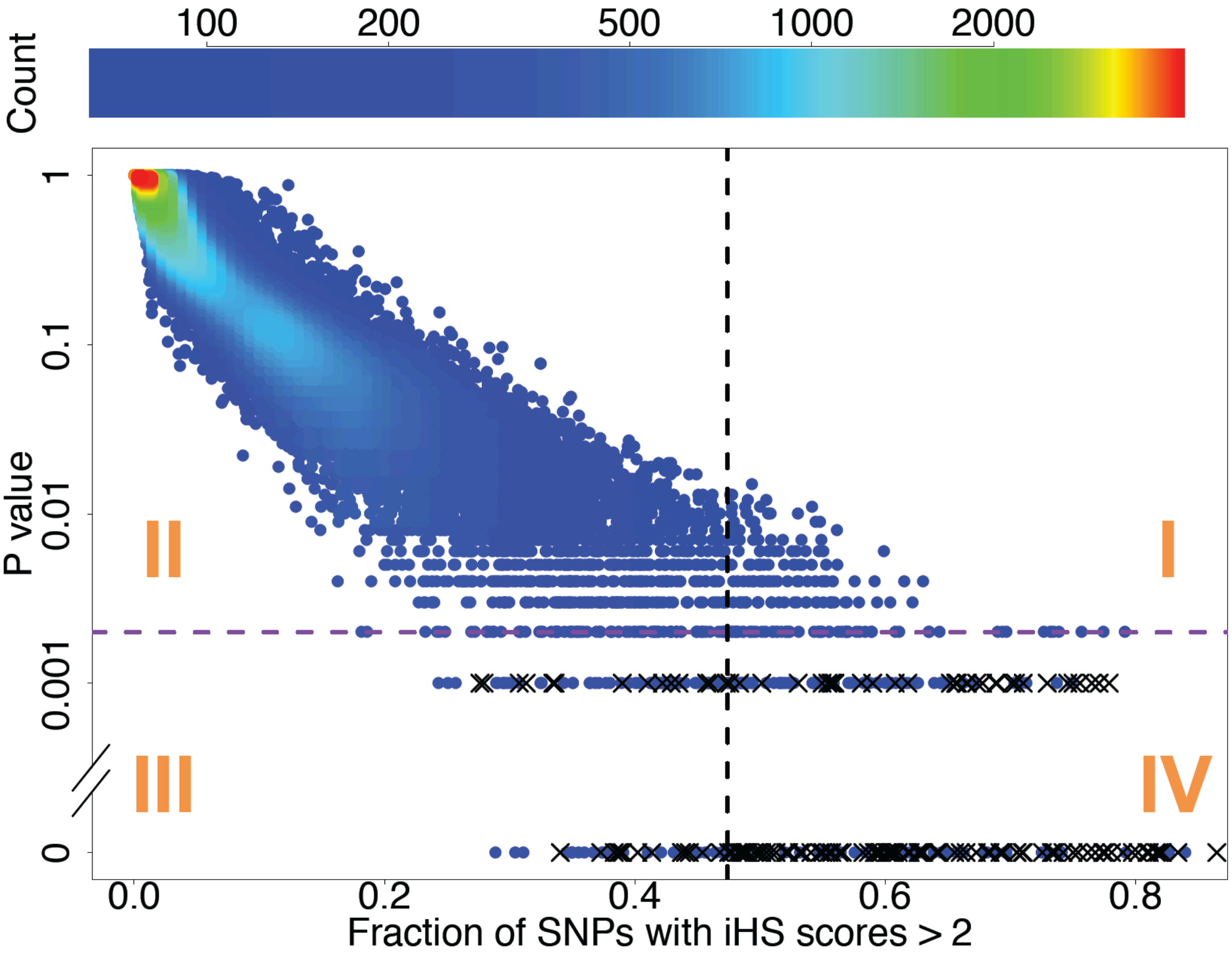


Figure S10. Importance of using *P*-values to define candidates in the iHS analysis. Each point is a window of 500 single nucleotide variants, and color represents the density of points. The vertical black line and the horizontal purple line are the top 0.5% significance cutoffs for the G2D and *P*-value distributions, respectively. Windows in Quadrant I are outliers in the iHS distribution but are not statistically significant when the effects of demography and genome architecture are controlled for. In Quadrant III are the many windows that are statistically significant even though their iHS values are modest. Cross marks (x) represent those Pygmy specific top-hits as discussed in the main text.

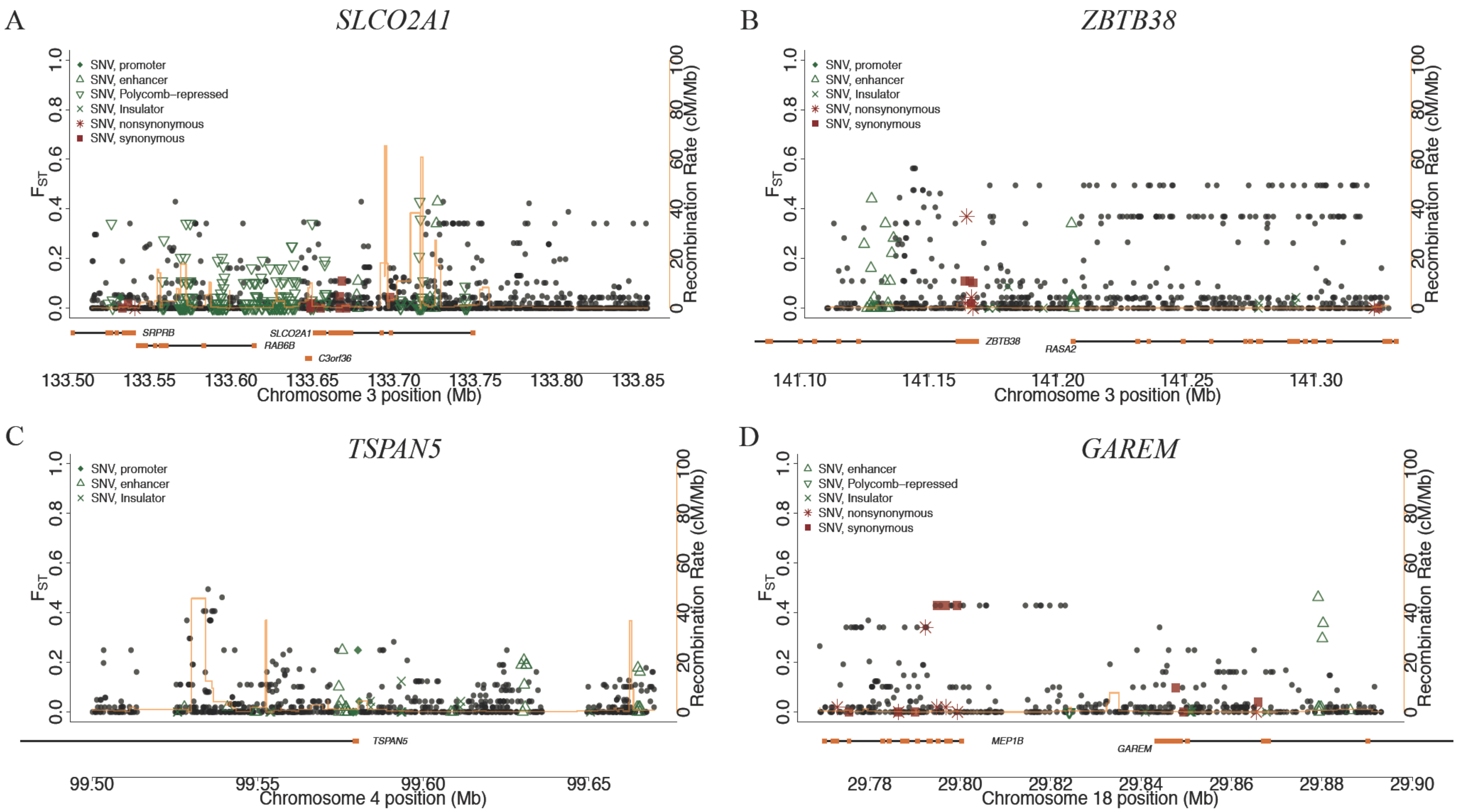


Figure S11. Elevated population differentiation (F_{ST}) in ENCODE regulatory element sequences in four candidate loci containing several bone-synthesis related genes. Each panel is titled by the bone-synthesis related gene. Green symbols indicate SNPs residing in ENCODE elements, while red symbols are variants within protein coding sequences.

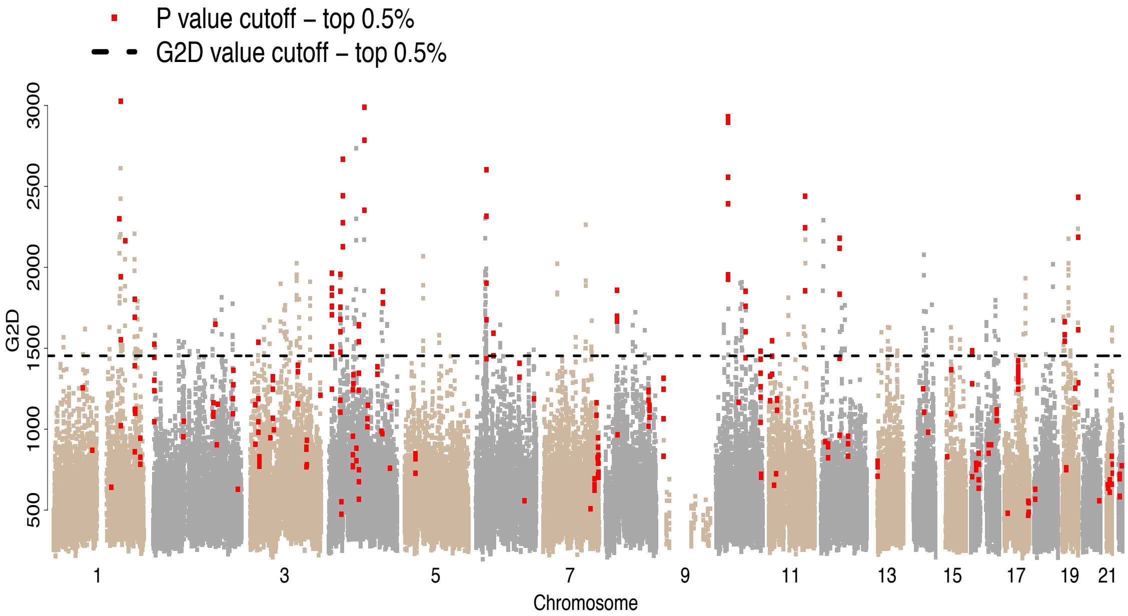


Figure S12. Genome-wide Manhattan plot of the G2D statistic. Each dot is a window of 500 single nucleotide variants. The dashed line represents the top 0.5% of the outlier threshold of the G2D statistic. Red dots are the candidates selected using the *P*-value approach (consensus windows based on the two best-fit demographic models). Most of chromosome 9 was masked by our quality control filters.

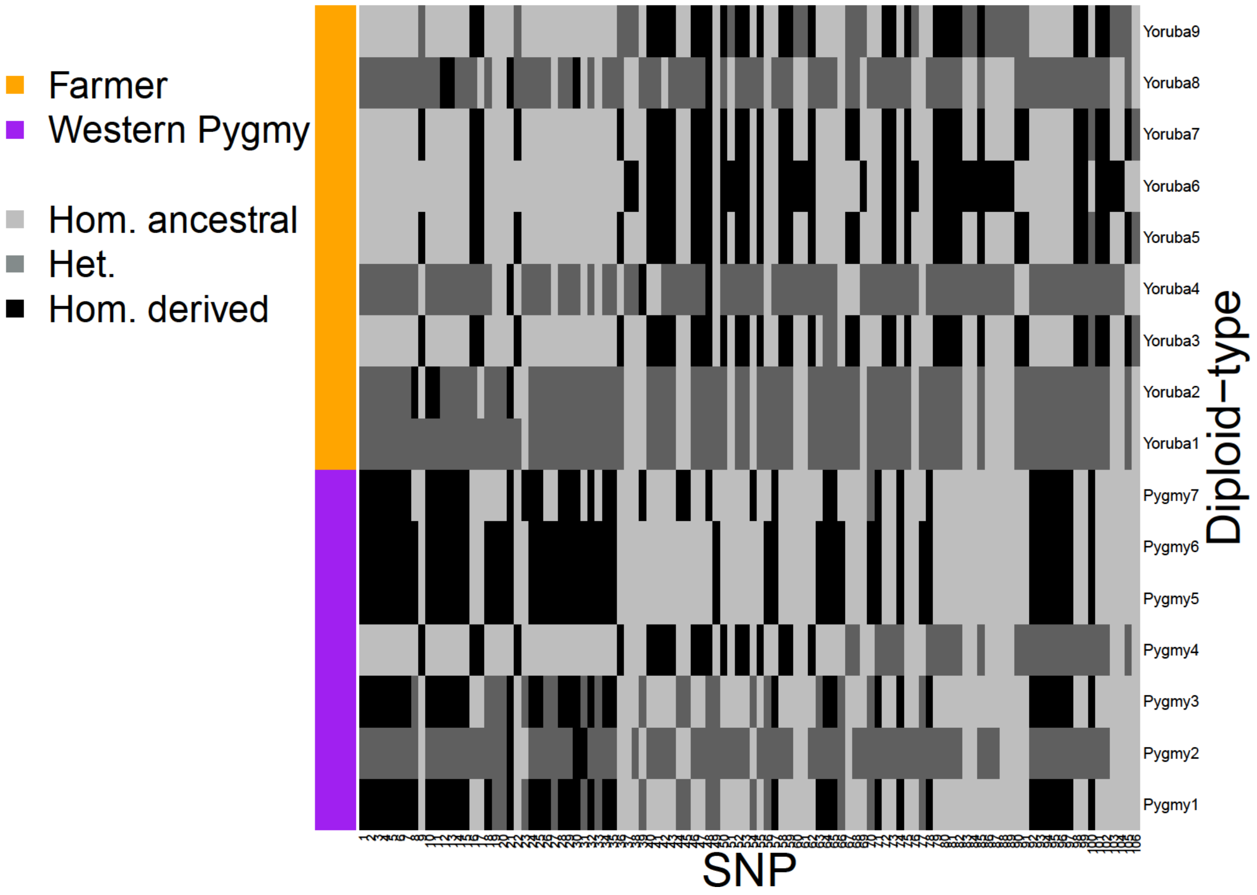


Figure S13. Candidate region of *HLA-DPA1* (chr6: 33.03-33.05 Mb) in the farmer and Pygmy samples. Columns are SNPs and rows are individual diploid-types. Light grey, dark grey, and black represent homozygous ancestral (Hom. ancestral), heterozygous (Het.), and homozygous derived (Hom. derived) genotypes.

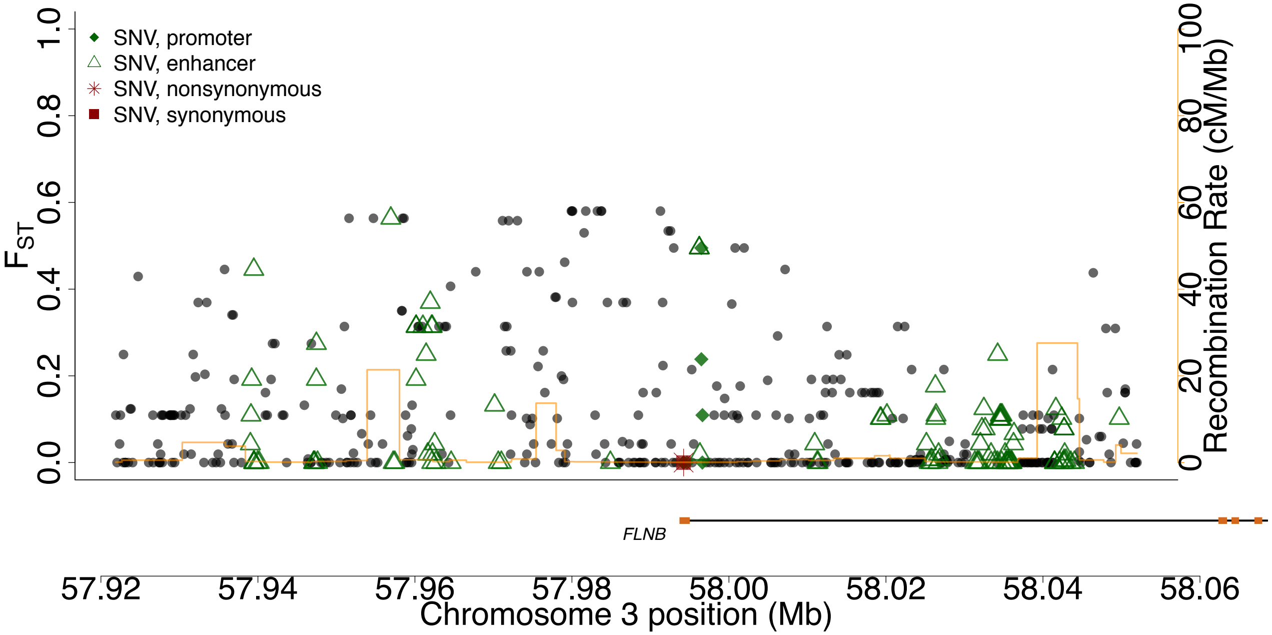
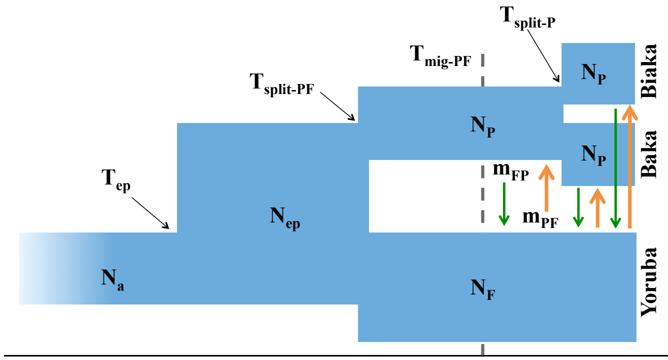
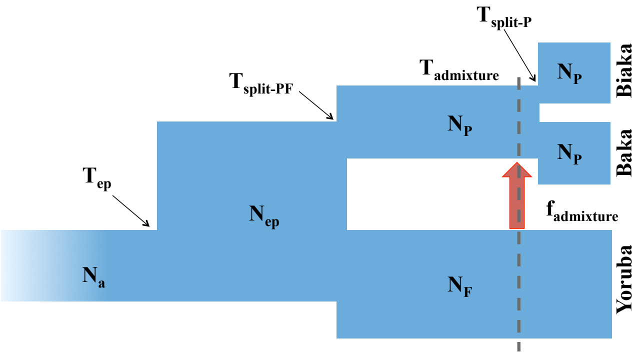
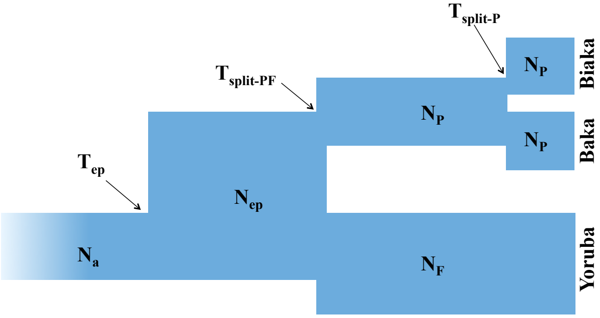
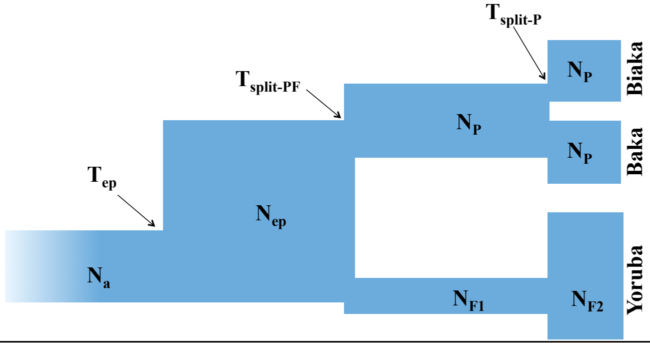
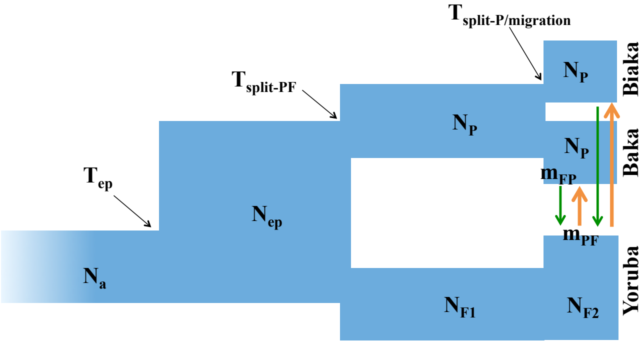
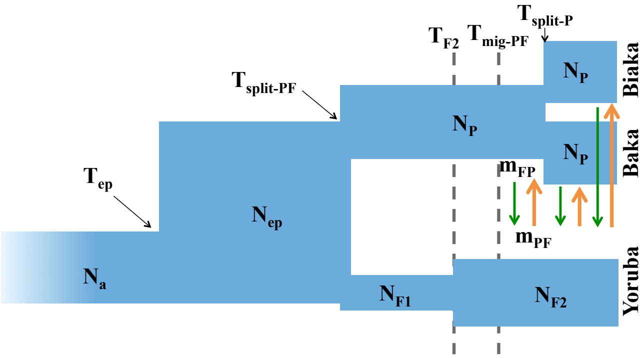


Figure S14. Elevated population differentiation (*F_{ST}*) in ENCODE regulatory element sequences the candidate locus around the gene bone-synthesis related gene *FLNB*. Green symbols indicate SNPs residing in ENCODE elements, while red symbols are variants within protein coding sequences.

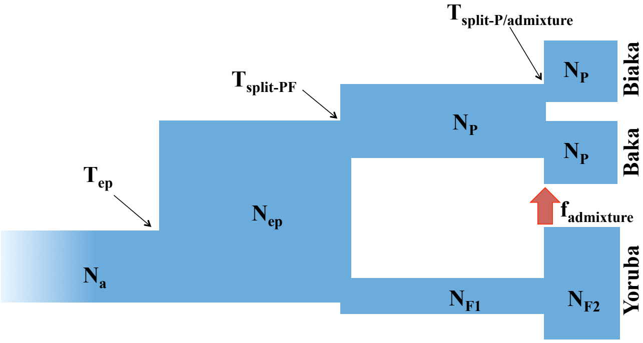
Table S1. Primary models evaluated during demographic inference using ∂a∂i (Gutenkunst *et al.* 2009). Parameter θ_a is $4N_a\mu$, where μ=2.35x10^{-8} per site per generation (Gutenkunst et al. 2009), N is effective population size, T is the time of a demographic event (in units of $2N_a$ generations, where N_a is the ancestral effective population size), m is migration rate (in units of $2N_a$), and f is admixture proportion. Events that occur at the same time are noted using "/"; otherwise, in our optimization procedure, the order of time parameters for the divergence and gene flow events is not fixed in a model. Note that we did not choose Model-5 over Model-2, even though Model-5 has a better log-likelihood than Model-2, because Model-5 is a special case of Model-1.

Model	Parameter: estimate		Log-likelihood	Optimization
Model-1	(10 parameters)		-6712	All converged.
 $T_{split-P}$ T_{mig-PF} $T_{split-PF}$ N_P Biaka N_P N_P Baka T_{ep} m_{FP} m_{PF} N_{ep} N_a N_F Yoruba	θ_a: 206127 N_{ep}: 3.02 N_F: 1.76 N_P: 0.86 T_{ep}: 0.655	$T_{split-PF}$: 0.46 T_{mig-PF}: 0.115 $T_{split-P}$: 0.015 m_{FP}: 1.22 m_{PF}: 11.9		
Model-2	(9 parameters)		-7737	All converged.
 $T_{split-P}$ $T_{admixture}$ N_P Biaka $T_{split-PF}$ N_P N_P Baka T_{ep} $f_{admixture}$ N_{ep} N_a N_F Yoruba	θ_a: 206375 N_{ep}: 2.26 N_F: 2.05 N_P: 0.79 T_{ep}: 0.686	$T_{split-PF}$: 0.266 $T_{admixture}$: 0.021 $T_{split-P}$: 0.012 $f_{admixture}$: 0.6799		

Model	Parameter: estimate		Log-likelihood	Optimization
Model-3  $T_{split-P}$ $T_{split-PF}$ T_{ep} N_P N_P N_P N_{ep} N_a N_F Biaka Baka Yoruba	(7 parameters) θ_a: 207003 N_{ep}: 2.62 N_F: 1.27 N_P: 1.35	T_{ep}: 0.668 $T_{split-PF}$: 0.048 $T_{split-P}$: 0.019	-11877	All converged.
Model-4  $T_{split-P}$ $T_{split-PF}$ T_{ep} N_P N_P N_P N_{ep} N_a N_{F1} N_{F2} Biaka Baka Yoruba	(8 parameters) θ_a: 206264 N_{ep}: 2.63 N_{F1}: 0.64 N_{F2}: 44.56	N_P: 1.31 T_{ep}: 0.677 $T_{split-PF}$: 0.047 $T_{split-P}$: 0.021	-10978	All converged.

Model	Parameter: estimate	Log-likelihood	Optimization
Model-5  $T_{split-P/migration}$ $T_{split-PF}$ T_{ep} N_P N_P N_P N_{ep} N_a m_{FP} m_{PF} N_{F1} N_{F2} Biaka Baka Yoruba	(10 parameters) θ_a: 208385 N_{ep}: 3.08 N_{F1}: 1.10 N_{F2}: 2.90 N_P: 0.97 T_{ep}: 0.617 $T_{split-PF}$: 0.147 $T_{split-P/migration}$: 0.024 M_{FP}: 1.73 M_{PF}: 22.36	-7437	All converged.
Model-6  $T_{split-P}$ T_{F2} T_{mig-PF} $T_{split-PF}$ T_{ep} N_P N_P N_P N_{ep} N_a m_{FP} m_{PF} N_{F1} N_{F2} Biaka Baka Yoruba	(12 parameters) θ_a: 206308 N_{ep}: *9.8 – 21.9 N_{F1}: 1.47 N_{F2}: *5.0 – 7.1 N_P: 0.95 T_{ep}: *0.58 – 0.63 $T_{split-PF}$: *0.3765 – 0.3785 T_{F2}: *0.1265 – 0.1285 T_{mig-PF}: *0.0165 – 0.0185 $T_{split-P}$: 0.016 M_{FP}: 1.36 M_{PF}: 10.64	-6532	*: Parameter that did not converge.

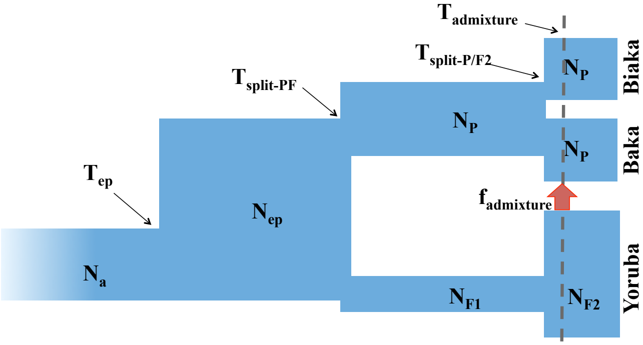
<table>
<tr><th>Model</th><th>Parameter: estimate</th><th>Log-likelihood</th><th>Optimization</th></tr>
<tr><td>Model-7



</td><td>(9 parameters)

θ_a: 206054
N_{ep}: 2.63
N_{F1}: 1.42
N_{F2}: *41 - 45
N_P: 1.06

T_{ep}: 0.677
$T_{split-PF}$: 0.127
$T_{split-P/admixture/F2}$: 0.016
$f_{admixture}$: 0.399</td><td>-8085</td><td>*: Parameter that did not converge.</td></tr>
<tr><td>Model-8



</td><td>(12 parameters)

θ_a: 206311
N_{ep}: 2.66
N_{F1}: 0.58
N_{F2}: *40 - 42
N_P: 1.25

T_{ep}: *0.768 – 0.868
$T_{split-PF}$: *0.153 – 0.253
$T_{split-P/F2}$: *0.126 – 0.246
$T_{admixture}$: *0.126 – 0.246
$f_{admixture}$: 0.116</td><td>-10837</td><td>*: Parameter that did not converge.

Also note that the estimates for $T_{split-P/F2}$ and $T_{admixture}$ are the same, suggesting these events occurred at the same time.</td></tr>
</table>

Table S2. Refitting the two candidate models with thinned data set. In the thinned data set, the polymorphisms are at least 0.01 cM apart from each other. Information criteria, AIC and BIC, were calculated for each of the two best-fit models. Model-1 is preferable over Model-2 using both AIC and BIC methods.

Model (# parameters)	**Log likelihood (full data set)**	**Log likelihood (thinned data set)**	**AIC**	**BIC**
Model-1, continuous gene flow (10)	-6712	-2803	5624	5669
Model-2, single pulse admixture (9)	-7737	-2851	5718	5758

Table S3. Overlap of selection candidates (top 0.5% in *P*-value distribution) with functional loci (i.e. protein-coding and ENCODE sequences). Enrichment of top-hit windows was found in genes, but not in ENCODE elements. *P*-values are calculated based simulations of Model-1 with the African American genetic recombination map. Similar results hold for the other three simulation conditions.

	Non-exon	**Exon**	**Non-ENCODE**	**ENCODE**
Significant	117	275	62	330
Non-significant	29392	55766	12865	72293
P-value (one-sided Fisher Exact Test)	**0.029***		0.682	

Table S4. The 35 distinct genomic regions with the strongest evidence of Pygmy-specific signals of adaptation identified by iHS. For each locus, the third column, max(|iHS|), indicates the maximum iHS score among the variants inside the locus.

Locus	Gene Name(s)	max(|iHS|)	P value	FDR	Notes
chr1:97708629-97876524	*DPYD,DPYD-AS1*	4.64	$1.00\text{x}10^{-3}$	0.16	
chr1:151752870-152059462	*TDRKH,LINGO4,RORC,C2CD4D, LOC100132111,THEM5,THEM4, S100A10,NBPF18P,S100A11,TCHHL1*	4.62	0.00	0	***S100A10*** is an regulatory element of innate immunity (Han et al. 2012).
chr1:154382442-154782105	*IL6R,SHE,TDRD10,UBE2Q1,CHRNB2, ADAR,KCNN3*	5.36	0.00	0	***IL6R*** encodes an interleukin receptor, associated with inflammatory diseases, such as rheumatoid arthritis and asthma; (Briso et al. 2008). ***ADAR*** involves A-to-I RNA editing and acts as an antiviral gene (Haralambieva et al. 2011). ***TDRD10*** is a member of methylarginine-binding proteins that have enriched expression in the germ line and are strongly associated with gametogenesis (Chen et al. 2011).
chr1:226265661-226580188	*ACBD3,MIXL1,LIN9,PARP1*	4.99	0.00	0	***ACBD3*** is a Golgi-resident protein involved in hormone-induced steroid biosynthesis in testicular Leydig cells (Fan and Papadopoulos 2013).
chr1:226865560-227209786	*ITPKB,PSEN2,ADCK3,CDC42BPA*	5.16	0.00	0	***ITPKB***, Inositol-trisphosphate 3-kinase B, which plays an active role in innate immune system (Sauer and Cooke 2010).
chr1:228103665-228842760	*WNT9A,MIR5008,WNT3A,ARF1, MIR3620,C1orf35,MRPL55,GUK1,GJC2, IBA57AS1,IBA57,C1orf145,OBSCN, TRIM11,MIR6742,TRIM17,HIST3H3, HIST3H2A,HIST3H2BB,MIR4666A, RNF187,BTNL10,RHOU,DUSP5P1*	5.11	0.00	0	***OBSCN*** is an obscurin gene and has an important role in the organization of myofibrils.
chr2:60242061-60398137	*NA*	5.10	0.00	0	
chr2:72210353-72344610	*NA*	4.93	0.00	0	
chr2:213722832-214136943	*MIR4776-2,MIR4776-1,IKZF2*	4.56	0.00	0	***IKZF2*** is a hematopoietic-specific transcription factor involved in the regulation of lymphocyte development (Stanic et al. 2014).
chr2:236917302-237037341	*AGAP1*	4.96	$1.00\text{x}10^{-3}$	0.16	
chr3:10080721-10242712	*FANCD2,FANCD2OS,BRK1,VHL,IRAK2*	5.00	$1.00\text{x}10^{-3}$	0.16	
chr3:10285605-10695444	*TATDN2,GHRLOS,LINC00852,GHRL, SEC13,ATP2B2,MIR885*	5.00	0.00	0	
chr3:111396224-111555201	*PLCXD2,PHLDB2*	5.53	0.00	0	
chr3:133506737-133863702	*SRPRB,RAB6B,C3orf36,SLCO2A1*	4.59	0.00	0	***SLCO2A1*** encodes a prostaglandin transporter protein, and mutations in this gene have been shown causing a rare genetic disease that affects both skin and bones (Zhang et al. 2012)
chr3:134572433-134716365	*EPHB1*	4.44	0.00	0	***EPHB1*** is an *Ephrin* receptor at sites of osteogenesis.
chr3:141105569-141333249	*ZBTB38,RASA2*	4.86	0.00	0	***ZBTB38*** encodes a zinc finger transcriptional activator has been associated with adult height in multiple populations (Lettre et al. 2008; Weedon et al. 2008; Wang et al. 2013).

Table S4. Continued.

chr4:97467886-97840259	*NA*	5.65	0.00	0	
chr4:99496207-99673561	*TSPAN5*	3.50	0.00	0	*TSPAN5* is a member of the tetraspanin protein family and is up-regulated during osteoclast differentiation (Iwai et al. 2007).
chr5:156183805-156461833	*SGCD,PPP1R2P3,TIMD4,HAVCR1*	4.85	0.00	0	***TIMD4*** is a T-cell immunoglobulin gene known to be associated with immune-related disorders, such as allergy and asthma (Li et al. 2014). ***HAVCR1*** has been implicated in susceptibility to allergic asthma in both mice and humans (McIntire et al. 2003) and has been hypothesized to have experienced both positive and balancing natural selection in the course of primate evolution (Nakajima et al. 2005).
chr5:156595169-156722896	*ITK,CYFIP2*	5.04	$1.00\text{x}10^{-3}$	0.16	
chr5:157044944-157310105	*SOX30,C5orf52,THG1L,LSM11,CLINT1*	5.49	$1.00\text{x}10^{-3}$	0.16	***SOX30*** is a transcription factor involved in the regulation of embryonic development, and its expression pattern is associated with testis development in mice (Han et al. 2014)
chr7:10943346-11263539	*NDUFA4,PHF14*	3.72	0.00	0	
chr7:67145543-67233803	*NA*	4.57	0.00	0	
chr7:151529659-151654789	*PRKAG2,PRKAG2-AS1,GALNTL5*	4.21	0.00	0	***GALNTL5*** is an essential functional molecule for sperm development, and the GALNTL5 mutation may cause human asthenozoospermia (Takasaki et al. 2014)
chr7:152220265-152424951	*XRCC2*	3.77	0.00	0	***XRCC2*** is an immune-related gene (Sale et al. 2001)
chr8:35242514-35469684	*UNC5D*	5.40	0.00	0	
chr8:35880031-36176084	*NA*	4.40	0.00	0	
chr8:37377237-37712337	*ZNF703,ERLIN2,LOC728024,PROSC, GPR124,BRF2*	5.08	0.00	0	***GPR124*** an immune-related gene (Fredriksson et al. 2003)
chr11:17398742-17655840	*NCR3LG1,KCNJ11,ABCC8,USH1C, OTOG*	3.68	0.00	0	***NCR3LG1*** is a ligand for natural killer cell receptors and appears to play an immunomodulatory role in response to pro-inflammatory cytokine signaling (Matta et al. 2013).
chr14:76694616-76974036	*ESRRB*	5.04	0.00	0	The gene product of ***ESRRB*** is similar to the estrogen receptor. Its function is unknown; however, a similar protein in mouse plays an essential role in placental development; (Luo et al. 1997)
chr17:13911228-14241158	*CDRT15P1,COX10-AS1,COX10,CDRT15, HS3ST3B1,MGC12916*	3.36	0.00	0	***HS3ST3B1*** encodes a heparan sulphate enzyme that alters the binding of sporozoite to hepatocytes and its subsequent development in mice infected by the malaria parasite *P. Berghei* (Brisebarre et al. 2014). Diaz et al. (2005) reported that ***COX10*** knockout mice develop a slowly progressive myopathy.
chr18:29766032-29896024	*MEP1B,GAREM*	4.58	0.00	0	***MEP1B*** encodes the subunit of Meprins, a multidomain zinc metalloprotease that has been implicated in intestinal inflammation (Claudia and Christoph 2013). ***GAREM*** is an adapter protein in intracellular signaling cascades and has recently been associated with human height in a whole-exome sequencing association study (Kim et al. 2012).

Table S4. Continued.

chr21:23493685-23662959	*NA*	4.23	0.00	0	
chr22:34224706-34359718	*LARGE*	5.39	0.00	0	Mutations in ***LARGE*** cause a form of congenital muscular dystrophy (Longman et al. 2003).
chr22:39290370-39477973	*APOBEC3A B,APOBEC3A,APOBEC3B, APOBEC3BAS1,APOBEC3C,APOBEC3, APOBEC3F,APOBEC3G*	3.99	1.00x10^{-3}	0.16	The ***APOBEC*** cluster includes seven related genes that play a role in immunity, especially against retroviruses such as HIV-1 and SIV (Bogerd et al. 2006).

Table S5. **The 7 distinct genomic regions with the strongest evidence of Pygmy-specific signals of adaptation identified by G2D.**

Position	Gene Name	Possible Associated Function	G2D	*P*-value	FDR
chr1:179361049-179468857	*AXDND1*	Bone Synthesis	2.30x10^{3}	1.82x10-^{3}	0.23
chr1:183076845-183184161	*LAMC1,LAMC2*	Ovary/Reproduction Development, skin	3.02x10^{3}	8.12x10^{-4}	0.20
chr1:219895606-220069861	*RNU5F-1*		1.80x10^{3}	3.76x10^{-4}	0.15
chr3:57918877-58055004	*FLNB*	Bone Synthesis	9.47x10^{2}	3.12x10^{-3}	0.23
chr6:32968692-33049012	*HLA-DOA, HLA-DPA1,HLA-DPB1*	Immunity	2.60x10^{3}	9.90x10^{-6}	0.03
chr7:152201610-152337563	*N/A*		8.38x10^{2}	5.54x10^{-4}	0.17
chr19:12386669-12523799	*ZNF44,ZNF563, ZNF442,ZNF799*	Cellular Signaling Transmission	1.66x10^{3}	7.92x10^{-5}	0.09

Table S6. Three Gene Ontology gene sets that show the strongest evidence of polygenic adaptation in the Pygmy groups. *P*-values are from a one-sided Mann-Whitney *U* test and Bonferroni corrected.

Name of gene set	***P*-value**	**Gene members (number of genes)**
ANTIGEN BINDING	2.31x10^{-25}	*ALB, CD1D, CHRNA7 DHCR24 FCN1, FCN2, IGHA2, IGHG1, IGHG2, IGHG3, IGHG4, IGHM, IL7R, KIR2DL3 LAG3, LILRA1 LILRA2 LILRA3 LILRB4 MSLN, PPP2R1A PPP2R1B SLAMF1 SLC7A5 SLC7A8 SLC7A9 TAPBP, TOPORS* (28)
G1 PHASE OF MITOTIC CELL CYCLE	1.75x10^{-19}	*CDC23, CDC25C, CDC6, CDK10, CDK2,CDK6, CDKN1C, E2F1, FOXO4, GFI1B, MAP3K11, TAF1, TBRG4* (13)
PATTERN RECOGNITION RECEPTOR ACTIVITY	5.04x10^{-14}	*CD14, CLEC7A, COLEC12 DMBT1, MARCO, PGLYRP1, PGLYRP2, PGLYRP3, PGLYRP4, TLR2* (10)

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature methods* **7**(4): 248-249.

Andersen KG, Shylakhter I, Tabrizi S, Grossman SR, Happi CT, Sabeti PC. 2012. Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **367**(1590): 868-877.

Berkholtz CB, Shea LD, Woodruff TK. 2006. Extracellular matrix functions in follicle maturation. In *Seminars in reproductive medicine*, Vol 24, p. 262. NIH Public Access.

Bogerd HP, Wiegand HL, Hulme AE, Garcia-Perez JL, O'Shea KS, Moran JV, Cullen BR. 2006. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. *Proceedings of the National Academy of Sciences* **103**(23): 8780-8785.

Brisebarre A, Kumulungui B, Sawadogo S, Atkinson A, Garnier S, Fumoux F, Rihet P. 2014. A genome scan for Plasmodium falciparum malaria identifies quantitative trait loci on chromosomes 5q31, 6p21. 3, 17p12, and 19p13. *Malaria Journal* **13**(1): 198.

Briso EM, Dienz O, Rincon M. 2008. Cutting edge: soluble IL-6R is produced by IL-6R ectodomain shedding in activated CD4 T cells. *The Journal of Immunology* **180**(11): 7102-7106.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**(5): 1084-1097.

Chen C, Nott TJ, Jin J, Pawson T. 2011. Deciphering arginine methylation: Tudor tells the tale. *Nature Reviews Molecular Cell Biology* **12**(10): 629-642.

Claudia B, Christoph B-P. 2013. The metalloproteases meprin alpha and meprin beta: unique enzymes in inflammation, neurodegeneration, cancer and fibrosis. *Biochemical Journal* **450**(2): 253-264.

Diaz F, Thomas CK, Garcia S, Hernandez D, Moraes CT. 2005. Mice lacking COX10 in skeletal muscle recapitulate the phenotype of progressive mitochondrial myopathies associated with cytochrome c oxidase deficiency. *Human molecular genetics* **14**(18): 2737-2748.

Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET. 2005. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature genetics* **38**(2): 223-227.

Fan J, Papadopoulos V. 2013. Evolutionary origin of the mitochondrial cholesterol transport machinery reveals a universal mechanism of steroid hormone biosynthesis in animals. *PloS one* **8**(10): e76701.

Fredriksson R, Gloriam DE, Höglund PJ, Lagerström MC, Schiöth HB. 2003. There exist at least 30 human G-protein-coupled receptors with long Ser/Thr-rich N-termini. *Biochemical and biophysical research communications* **301**(3): 725-734.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**(7414): 91-100.

Han D, Cai X, Wen J, Matheson D, Skyler J, Kenyon N, Chen Z. 2012. Innate and adaptive immune gene expression profiles as biomarkers in human type 1 diabetes. *Clinical & Experimental Immunology* **170**(2): 131-138.

Han F, Dong Y, Liu W, Ma X, Shi R, Chen H, Cui Z, Ao L, Zhang H, Cao J. 2014. Epigenetic Regulation of Sox30 Is Associated with Testis Development in Mice. *PloS one* **9**(5): e97203.

Haralambieva IH, Ovsyannikova IG, Umlauf BJ, Vierkant RA, Shane Pankratz V, Jacobson RM, Poland GA. 2011. Genetic polymorphisms in host antiviral genes: associations with humoral and cellular immunity to measles vaccine. *Vaccine* **29**(48): 8988-8997.

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL. 2011. The landscape of recombination in African Americans. *Nature* **476**(7359): 170-175.

Irving-Rodgers HF, Rodgers RJ. 2005. Extracellular matrix in ovarian follicular development and disease. *Cell and tissue research* **322**(1): 89-98.

Iwai K, Ishii M, Ohshima S, Miyatake K, Saeki Y. 2007. Expression and function of transmembrane-4 superfamily (tetraspanin) proteins in osteoclasts: reciprocal roles of Tspan-5 and NET-6 during osteoclastogenesis. *Allergol Int* **56**(4): 457-463.

Kim J-J, Park Y-M, Baik K-H, Choi H-Y, Yang G-S, Koh I, Hwang J-A, Lee J, Lee Y-S, Rhee H. 2012. Exome sequencing and subsequent association studies identify five amino acid-altering variants influencing human height. *Human genetics* **131**(3): 471-478.

Kimura M. 1964. Diffusion models in population genetics. *Journal of Applied Probability* **1**(2): 177-232.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**(7): 1073-1081.

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C. 2008. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature genetics* **40**(5): 584-591.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**(7357): 493-496.

Li N, van der Sijde MR, Bakker SJ, Dullaart RP, van der Harst P, Gansevoort RT, Elbers CC, Wijmenga C, Snieder H, Hofker MH. 2014. Pleiotropic effects of lipid genes on plasma glucose, HbA1c and HOMA-IR levels. *Diabetes*: DB_131800.

Longman C, Brockington M, Torelli S, Jimenez-Mallebrera C, Kennedy C, Khalil N, Feng L, Saran RK, Voit T, Merlini L. 2003. Mutations in the human LARGE gene cause MDC1D, a novel form of congenital muscular dystrophy with severe mental retardation and abnormal glycosylation of α-dystroglycan. *Human molecular genetics* **12**(21): 2853-2861.

Luo J, Sladek R, Bader J-A, Matthyssen A, Rossant J, Giguère V. 1997. Placental abnormalities in mouse embryos lacking the orphan nuclear receptor ERR-β. *Nature* **388**(6644): 778-782.

Matta J, Baratin M, Chiche L, Forel J-M, Cognet C, Thomas G, Farnarier C, Piperoglou C, Papazian L, Chaussabel D. 2013. Induction of B7-H6, a ligand for the natural killer cell–activating receptor NKp30, in inflammatory conditions. *Blood* **122**(3): 394-404.

McIntire JJ, Umetsu SE, Macaubas C, Hoyte EG, Cinnioglu C, Cavalli-Sforza LL, Barsh GS, Hallmayer JF, Underhill PA, Risch NJ. 2003. Immunology: hepatitis A virus link to atopic disease. *Nature* **425**(6958): 576-576.

Nakajima T, Wooding S, Satta Y, Jinnai N, Goto S, Hayasaka I, Saitou N, Guan-Jun J, Tokunaga K, Jorde L. 2005. Evidence for natural selection in the HAVCR1 gene: high degree of amino-acid variability in the mucin domain of human HAVCR1 protein. *Genes and immunity* **6**(5): 398-406.

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A. 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome research* **19**(5): 838-849.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome research* **15**(11): 1566-1575.

Paradis E. 2010. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**(3): 419-420.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**(3): 559-575.

Pyun J-A, Cha DH, Kwack K. 2012. LAMC1 gene is associated with premature ovarian failure. *Maturitas* **71**(4): 402-406.

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature genetics* **32**(1): 135-142.

Sainudiin R, Clark AG, Durrett RT. 2007. Simple models of genomic variation in human SNP density. *BMC genomics* **8**(1): 146.

Sale JE, Calandrini DM, Takata M, Takeda S, Neuberger MS. 2001. Ablation of XRCC2/3 transforms immunoglobulin V gene conversion into somatic hypermutation. *Nature* **412**(6850): 921-926.

Sauer K, Cooke MP. 2010. Regulation of immune cell development through soluble inositol-1, 3, 4, 5-tetrakisphosphate. *Nature Reviews Immunology* **10**(4): 257-271.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome research* **15**(11): 1576-1583.

Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nature genetics* **46**(8): 919-925.

Stanic B, van de Veen W, Wirz OF, Rückert B, Morita H, Söllner S, Akdis CA, Akdis M. 2014. IL-10–overexpressing B cells regulate innate and adaptive immune responses. *J Allergy Clin Immun*.

Takasaki N, Tachibana K, Ogasawara S, Matsuzaki H, Hagiuda J, Ishikawa H, Mochida K, Inoue K, Ogonuki N, Ogura A. 2014. A heterozygous mutation of GALNTL5 affects male infertility with impairment of sperm motility. *Proceedings of the National Academy of Sciences* **111**(3): 1120-1125.

Team RDC. 2012. R: A language and environment for statistical computing.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851-861.

Wang Y, Wang Zm, Teng Yc, Shi Jx, Wang Hf, Yuan Wt, Chu X, Wang Df, Wang W, Huang W. 2013. An SNP of the ZBTB38 gene is associated with idiopathic short stature in the Chinese Han population. *Clinical endocrinology* **79**(3): 402-408.

Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS. 2008. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics* **40**(5): 575-583.

Zhang Z, Xia W, He J, Zhang Z, Ke Y, Yue H, Wang C, Zhang H, Gu J, Hu W. 2012. Exome Sequencing Identifies< i> SLCO2A1</i> Mutations as a Cause of Primary Hypertrophic Osteoarthropathy. *The American Journal of Human Genetics* **90**(1): 125-132.

Zhou J, Fujiwara T, Ye S, Li X, Zhao H. 2014. Downregulation of Notch Modulators, Tetraspanin 5 and 10, Inhibits Osteoclastogenesis in Vitro. *Calcified tissue international* **95**(3): 209-217.