# Supporting Information

## Perry et al. 10.1073/pnas.1402875111

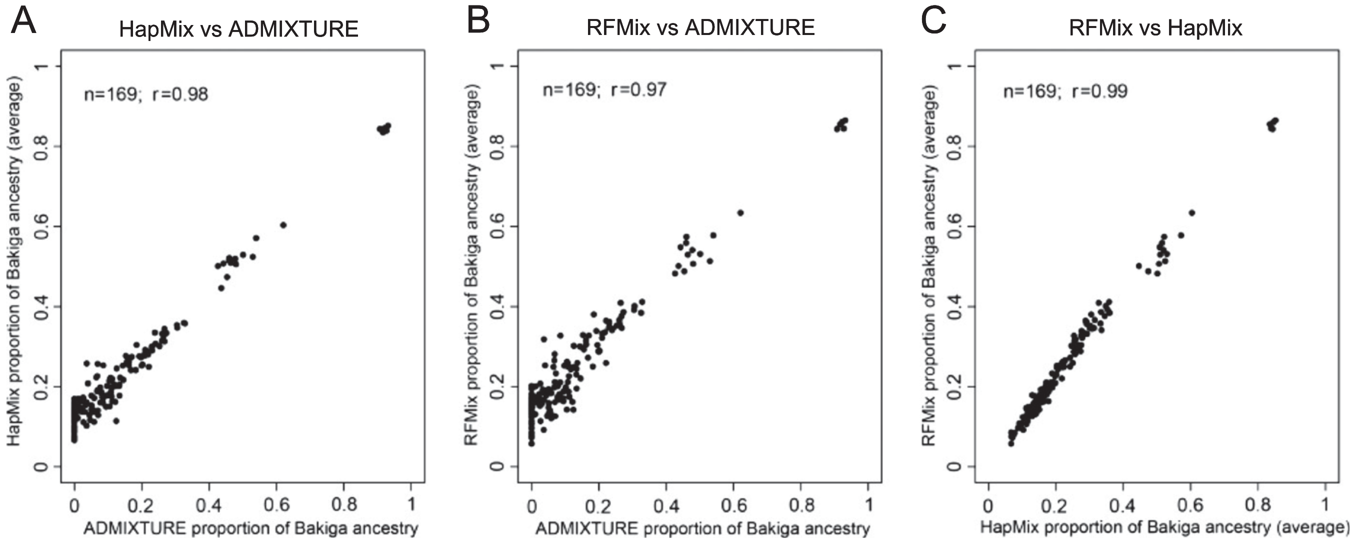
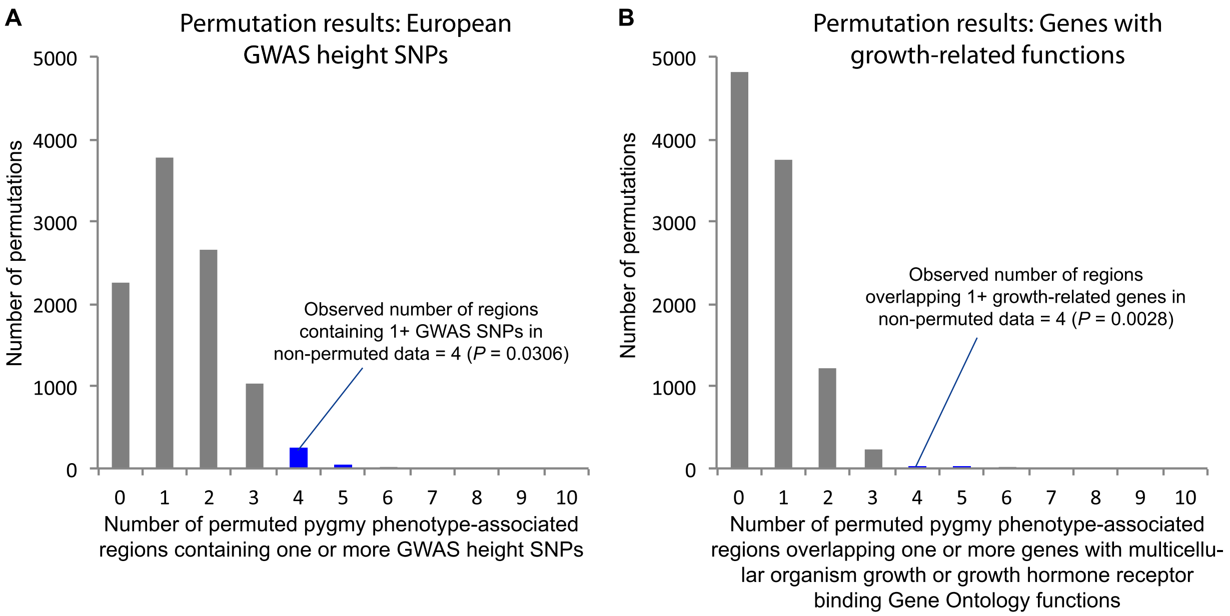


**Fig. S1.** Correlation between ADMIXTURE and average HapMix and RFMix ancestry estimates. (*A*) Correlation between the levels of Bakiga ancestry among all Batwa individuals estimated using ADMIXTURE (*x* axis) and those obtained when averaging local ancestry proportions across all SNPs using HAPMIX (*y* axis). (*B*) Correlation between the levels of Bakiga ancestry among all Batwa individuals estimated using ADMIXTURE (*x* axis) and those obtained when averaging local ancestry proportions across all SNPs using RFMix (*y* axis). (*C*) Correlation between the levels of Bakiga ancestry among all Batwa individuals estimated by ADMIXTURE (*x* axis) and those estimated by HAPMIX (*y* axis).



**C** Summary of all permutation analyses performed with GWAS SNPs and GO genes

| # pygmy phenotype-associated regions | Comparison dataset | # of SNPs/ regions in comparison dataset | # of phenotype-associated regions overlapping 1+ test features | Number of permutations (of 10,000) in which N out of the 16 pygmy phenotype associated regions overlaped one or more test features | | | | | | | | | | | Emperical p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| 16 | European height GWAS SNPs | 433 | 4 | 2262 | 3763 | 2643 | 1026 | 253 | 46 | 7 | 0 | 0 | 0 | 0 | 0.0306 |
| 16 | IgG glycosylation GWAS SNPs | 699 | 1 | 1118 | 2882 | 3189 | 1870 | 734 | 174 | 29 | 4 | 0 | 0 | 0 | 0.8882 |
| 16 | Crohns disease GWAS SNPs | 210 | 0 | 4216 | 3935 | 1508 | 296 | 41 | 4 | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | Type 2 diabetes GWAS SNPs | 207 | 1 | 5607 | 3407 | 868 | 106 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0.4393 |
| 16 | Multiple Sclerosis GWAS SNPs | 175 | 0 | 4490 | 3807 | 1352 | 314 | 32 | 5 | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | BMI GWAS SNPs | 110 | 0 | 6130 | 3131 | 654 | 74 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | Genes with multicellular organism growth Gene Ontology functions (GO0040014) | 78 | 3 | 6023 | 3205 | 691 | 77 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0081 |
| 16 | Genes with growth hormone receptor binding Gene Ontology function (GO0060396) | 30 | 2 | 7710 | 2057 | 217 | 14 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0233 |
| 16 | Combined genes in GO0040014 and GO0060396 | 101 | 4 | 4807 | 3738 | 1209 | 218 | 25 | 2 | 1 | 0 | 0 | 0 | 0 | 0.0028 |
| 16 | Batwa extreme iHS regions | 105 | 1 | 5003 | 3567 | 1189 | 208 | 27 | 6 | 0 | 0 | 0 | 0 | 0 | 0.4997 |

**D** Summary of all permutation analyses performed with Fst and iHS data

| | | | | | Actual data | | | | Permutations | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pygmy phenotype associated regions | # regions | Statistic | Population | # genome-wide SNPs | # SNPs in regions | Median Fst or \|iHS\| in regions | Median Fst or \|iHS\| outside of regions | Fst or \|iHS\| inside regions - Fst or \|iHS\| outside regions | # permutations | avg (Fst or \|iHS\| inside regions - Fst or \|iHS\| outside regions) | # permutations with larger inside-outside difference than actual data | Emperical p-value |
| Batwa - FDR 0.25 | 16 | Fst | Batwa-Bakiga | 826589 | 5190 | 0.01596433 | 0.0138925 | 0.002071824 | 1000 | -0.000193804 | 31 | 0.031 |
| Batwa - FDR 0.25 | 16 | Fst | Batwa-Baka | 804350 | 5027 | 0.01572692 | 0.01358969 | 0.002137231 | 1000 | 0.000140935 | 40 | 0.04 |
| Batwa - FDR 0.25 | 16 | Fst | Baka-Nzebi/Nzime | 809564 | 4069 | 0.007728587 | 0.008541248 | -0.000812661 | 1000 | -3.49E-06 | 809 | 0.809 |
| Lachance et al. and Jarvis et al. | 66 | Fst | Batwa-Bakiga | 826589 | 640 | 0.01501706 | 0.01389836 | 0.001118701 | 1000 | 0.00043829 | 412 | 0.412 |
| Lachance et al. and Jarvis et al. | 66 | Fst | Batwa-Baka | 804350 | 608 | 0.01421534 | 0.01359129 | 0.000624048 | 1000 | 0.000495335 | 433 | 0.433 |
| Lachance et al. and Jarvis et al. | 66 | Fst | Baka-Nzebi/Nzime | 809564 | 621 | 0.01753567 | 0.008538085 | 0.008997585 | 1000 | 0.000240872 | 0 | 0 |
| Batwa - FDR 0.25 | 16 | iHS | Batwa | 670543 | 3985 | 0.782858 | 0.67489 | 0.107968 | 1000 | -0.002853879 | 62 | 0.062 |
| Batwa - FDR 0.25 | 16 | iHS | Bakiga | 745829 | 4604 | 0.6595975 | 0.671036 | -0.0114385 | 1000 | -0.000433612 | 629 | 0.629 |
| Batwa - FDR 0.25 | 16 | iHS | Baka | 693983 | 4303 | 0.648094 | 0.684587 | -0.036493 | 1000 | 0.002887768 | 820 | 0.82 |
| Batwa - FDR 0.25 | 16 | iHS | Nzebi/Nzime | 747898 | 4613 | 0.686978 | 0.67253 | 0.014448 | 1000 | -0.004692552 | 218 | 0.218 |
| Lachance et al. and Jarvis et al. | 66 | iHS | Batwa | 670543 | 486 | 0.706432 | 0.675528 | 0.030904 | 1000 | 0.0226089 | 414 | 0.414 |
| Lachance et al. and Jarvis et al. | 66 | iHS | Bakiga | 693983 | 472 | 0.851274 | 0.684242 | 0.167032 | 1000 | 0.009689946 | 10 | 0.01 |
| Lachance et al. and Jarvis et al. | 66 | iHS | Baka | 745829 | 541 | 0.842882 | 0.670839 | 0.172043 | 1000 | 0.01433248 | 10 | 0.01 |
| Lachance et al. and Jarvis et al. | 66 | iHS | Nzebi/Nzime | 747898 | 542 | 0.8324985 | 0.6725295 | 0.159969 | 1000 | 0.005571687 | 6 | 0.006 |

**Fig. S2.** Observed and permuted distributions of the number of pygmy phenotype-associated regions that overlap European genome-wide association study (GWAS) height SNPs or genes with growth and development functions, and permutation analyses of $F_{ST}$ and absolute iHS values within and outside of pygmy phenotype-associated regions. (*A–C*) We used permutation analyses to test the significance of observed overlaps between the 16 Batwa pygmy phenotype-associated regions and various GWAS SNP and functional gene datasets. Briefly, we created 10,000 sets of size-matched regions (*Methods*) and determined the number of regions in each permutation that overlapped at least one feature from a tested dataset. (*A*) Distribution of the number of permuted pygmy phenotype associated regions that contain one or more European stature GWAS SNPs. Permutations in which an equal or greater number of regions contained ≥1 GWAS height SNP as that observed in the actual 16 pygmy phenotype associated regions (1) are highlighted in blue. The empirical *P* value ($P = 0.0306$) was determined based on the total number of permutations with an equal or greater number of regions containing ≥1 GWAS SNP. (*B*) Distribution of the number of permuted regions that overlap one or more genes with multicellular organism growth or growth hormone receptor binding Gene Ontology functions compared with the number of actual regions observed to overlap 1+ such genes (4; $P = 0.0028$). (*C*) Summary of all permutation analyses performed, with the counts of permutation observations equal to or greater than the observed result from the actual data highlighted in blue, and significant *P* values ($<0.05$) highlighted in red. (*D*) We also used permutation analyses to test the significance of the difference between median $F_{ST}$ and |iHS| values within pygmy phenotype-associated regions—both the 16 Batwa pygmy phenotype-associated regions identified by our admixture mapping approach and the 66 stature-associated regions for west central African rainforest hunter-gatherers reported by Jarvis et al. (2) and Lachance et al. (3)—compared with the rest of the genome. As above, we created 1,000 sets of size-matched regions (*Methods*) and determined $F_{ST}$ and |iHS| values within and outside these regions in each permuted dataset compared with the observed actual values.

1. Patin E, et al. (2014) The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nat Commun* 5:3163.
2. Jarvis JP, et al. (2012) Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet* 8(4):e1002641.
3. Lachance J, et al. (2012) Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 150(3):457–469.

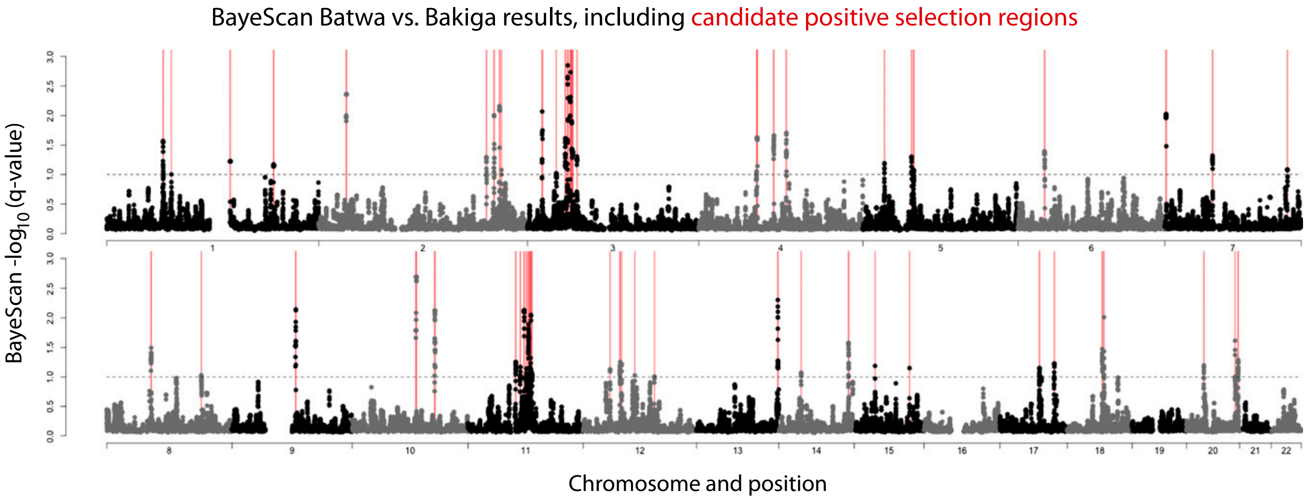
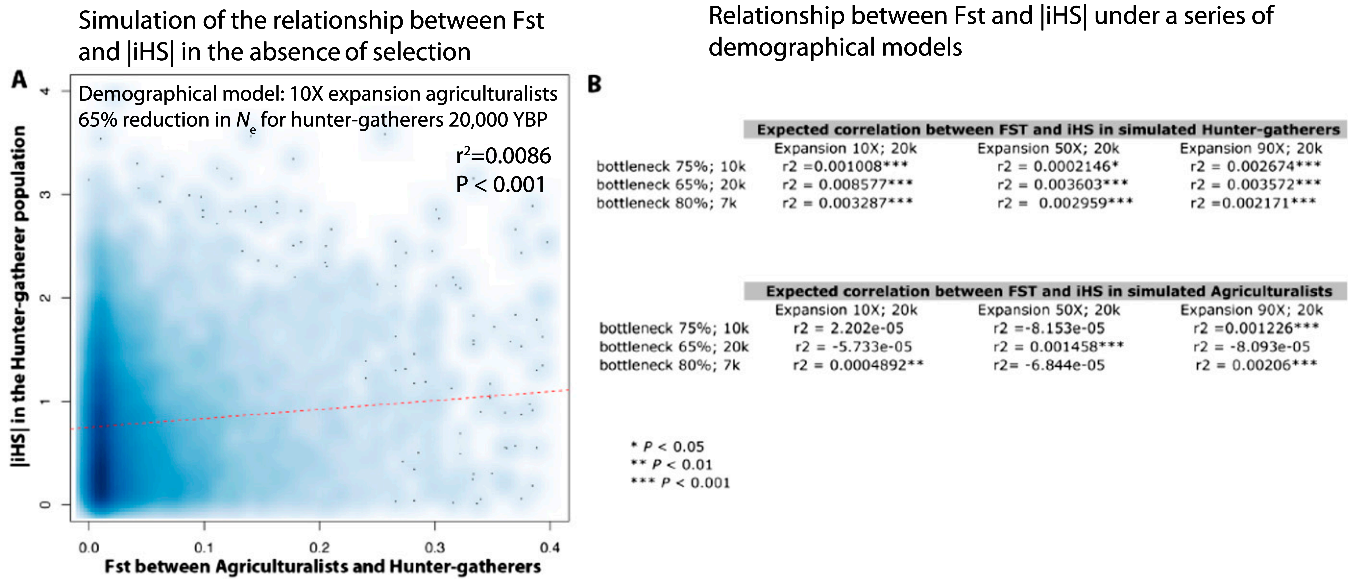


**Fig. S3.** Candidate regions targeted by positive selection identified by BayeScan. Manhattan plot showing the negative $\log_{10}$ *q* values (*y* axis), indicating the level of support for a history of positive selection at the locus, for all genome-wide SNPs with a minor allele frequency (MAF) >5%. The red lines indicate the regions identified by our sliding-window approach as harboring an excess of highly differentiated SNPs (i.e., SNPs with *q* values <0.1).



Simulation of the relationship between Fst and |iHS| in the absence of selection

Relationship between Fst and |iHS| under a series of demographical models

| Expected correlation between FST and iHS in simulated Hunter-gatherers | Expansion 10X; 20k | Expansion 50X; 20k | Expansion 90X; 20k |
|---|---|---|---|
| bottleneck 75%; 10k | r2 =0.001008*** | r2 = 0.0002146* | r2 = 0.002674*** |
| bottleneck 65%; 20k | r2 = 0.008577*** | r2 = 0.003603*** | r2 = 0.003572*** |
| bottleneck 80%; 7k | r2 = 0.003287*** | r2 = 0.002959*** | r2 =0.002171*** |

| Expected correlation between FST and iHS in simulated Agriculturalists | Expansion 10X; 20k | Expansion 50X; 20k | Expansion 90X; 20k |
|---|---|---|---|
| bottleneck 75%; 10k | r2 = 2.202e-05 | r2 =-8.153e-05 | r2 =0.001226*** |
| bottleneck 65%; 20k | r2 = -5.733e-05 | r2 = 0.001458*** | r2 = -8.093e-05 |
| bottleneck 80%; 7k | r2 = 0.0004892** | r2= -6.844e-05 | r2 = 0.00206*** |

* *P* < 0.05
** *P* < 0.01
*** *P* < 0.001

**Fig. S4.** Simulations show that the observed increases in $F_{ST}$ values and absolute integrated haplotype score (iHS) values within the pygmy phenotype-associated regions are not expected under neutral evolution. We used the program ms (1) to perform simulations under a series of neutral demographical models that are in accordance with the known demographical history of Bantu-speaking agriculturalist and African rainforest hunter-gatherer populations. We wanted to test whether the observed increase in Batwa |iHS| values within the Batwa pygmy phenotype-associated regions was simply the result of increased $F_{ST}$ values within these same regions (which in itself is already supportive of a history of positive selection on these regions). To do so, we performed simulations under a series of demographical models that are in accordance with the known demographical history of Bantu-speaking agriculturalist and African rainforest hunter-gatherer populations, based on the range of values that lead to the 2% of models that best fit the genome-wide genetic diversity data for these populations, as defined by Patin et al. (2). Specifically, we simulated two populations that split 60,000 y ago (3), followed by an expansion in the agriculturalist ancestral population 16,000–22,000 y ago ($N_e$ increase of 10–90 times) and a bottleneck in the hunter-gatherer ancestral population 7,000–22,000 y ago ($N_e$ reduction of 65–80%) (2). Under each demographical model, we simulated 10 independent regions of 1 Mb each, with recombination and mutation rates set at $1 \times 10^{-8}$ per base pair. We sampled 50 chromosomes from each of the populations. (*A*) Example results, with observed correlations between $F_{ST}$ values between the two simulated populations for all SNPs (*x* axis) and absolute iHS values (*y* axis) for the same SNPs in the hunter-gatherer population for a simulation with a 10 times expansion in the agriculturalist ancestors beginning 20,000 y ago and a 65% reduction in $N_e$ for the hunter-gatherer group also beginning 20,000 y ago. (*B*) Observed correlations between $F_{ST}$ values and absolute iHS in both population groups under a number of other demographical parameters. Overall, although significant correlations between $F_{ST}$ and absolute iHS values are observed for some combinations of demographical models, the correlation estimates are remarkably small, never exceeding an $r^2$ of 0.0086. Such low correlation estimates are not sufficient to explain the observed increase in absolute iHS values in our real data, given the observed changes in $F_{ST}$, even when considering the combination of demographical parameters leading to the strongest $r^2$ between simulated $F_{ST}$ and |iHS| values. Specifically, given the observed mean $F_{ST}$ difference of 0.0897 between SNPs within vs. outside the 16 pygmy phenotype-associated regions in our real data, the most extreme simulated results suggest an expected mean difference in Batwa |iHS| values of only 0.0028, which is 31 times lower than the observed result for SNPs within vs. outside the pygmy phentoype-associated regions. Thus, these simulations confidently show that under a neutral model of evolution one would not expect to observe concomitant increases in $F_{ST}$ and absolute iHS within the Batwa pygmy phenotype-associated regions to the extent observed in our data.

1. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.
2. Patin E, et al. (2014) The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nat Commun* 5:3163.
3. Patin E, et al. (2009) Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet* 5(4):e1000448.

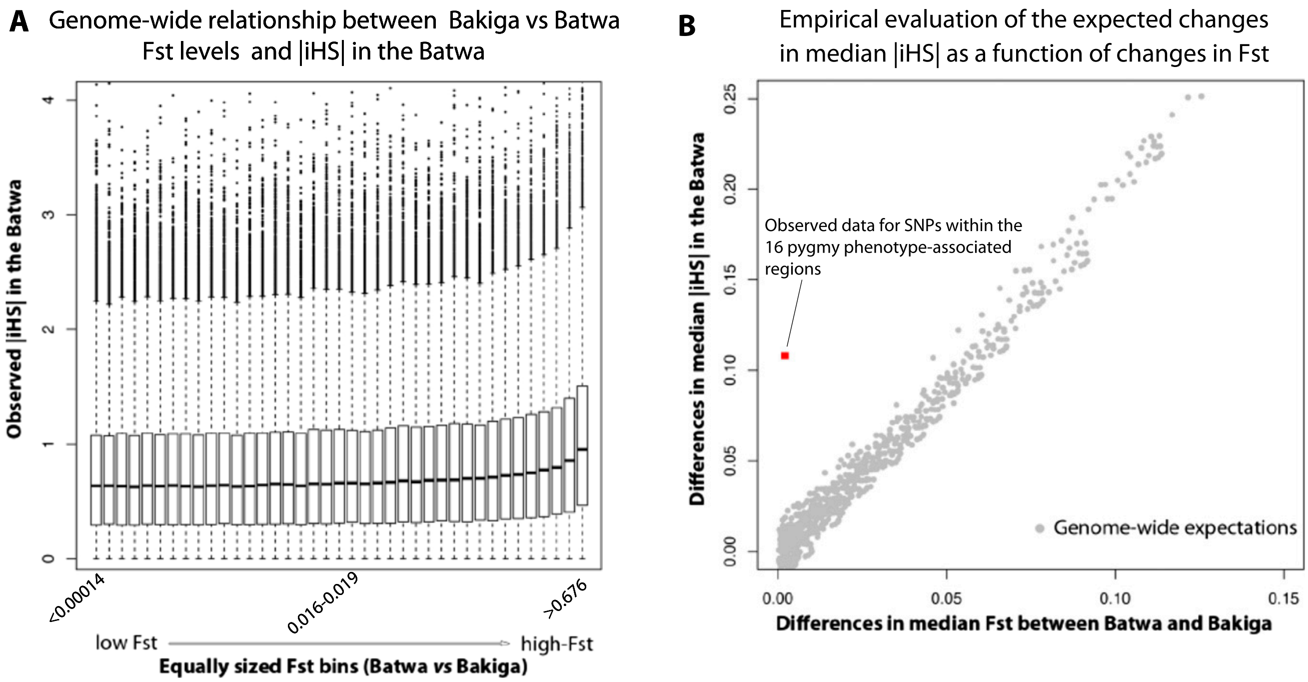


**Fig. S5.** Empirical evaluation of the expected changes in median |iHS| as a function of changes in $F_{ST}$. (*A*) Genome-wide correlation between Bakiga vs. Batwa $F_{ST}$ (*x* axis) and absolute iHS values in the Batwa population (*y* axis). Genome-wide patterns are expected to primarily reflect demographic events, and, to a lesser degree, the effects of natural selection. We observed a significant genome-wide correlation between Batwa vs. Bakiga $F_{ST}$ and |iHS| values in the Batwa, albeit only explaining a very small proportion of the variance ($r^2 = 0.011$; $P < 0.01$). The association is driven principally by high-$F_{ST}$ SNPs, which suggests that some of the signal is likely due to natural selection rather than demography alone. (*B*) Expected changes in absolute iHS values in the Batwa population (*y* axis) given changes in Bakiga vs. Batwa $F_{ST}$ values. Based on the genome-wide correlation shown in *A*, we calculated the expected changes in median iHS values in the Batwa as a function of increased $F_{ST}$ values between the Batwa and the Bakiga (gray dots). The red square shows the pattern observed when contrasting SNPs within the Batwa pygmy phenotype-associated regions than the remainder of the genome. This result suggests that the genome-wide correlation between |iHS| and $F_{ST}$ values is too small to account for the observed increase in |iHS| within the Batwa pygmy phenotype-associated regions.

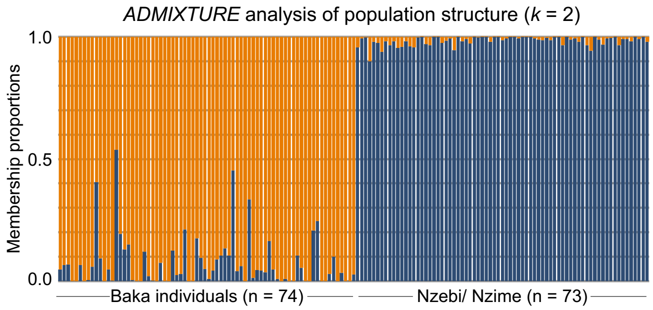


**Fig. S6.** ADMIXTURE results for the Baka and Nzebi/Nzime populations. Population structure analysis based on autosomal SNPs. Each individual is represented as a vertical line, with population origins indicated below the lines. Cluster membership proportions are depicted in orange (inferred proportion of Baka ancestry) and blue (inferred proportion of Nzebi/Nzime ancestry).

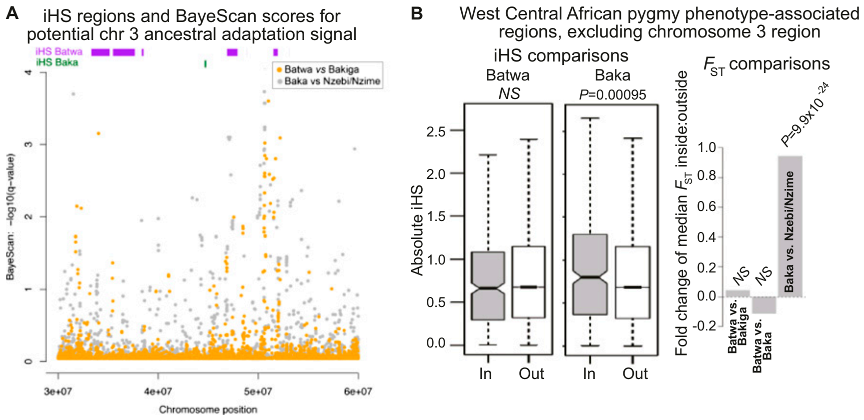


**Fig. S7.** Potential region of ancestral adaptation on chromosome 3. (*A*) BayeScan *q* values (*y* axis) indicating the level of support for a past history of positive selection for all SNPs on the Chr3:40–60 Mb region when evaluating the level of population differentiation between the Batwa and the Bakiga (orange) and between the Baka and the Nzebi/Nzime (gray). The bars on top of the figure delineate genomic regions for which the |iHS| values are among the top 99th percentile of the genome-wide distribution in the Batwa (purple) and the Baka populations (green). (*B*) Boxplot and fold-change comparisons of |iHS| and median $F_{ST}$ values for SNPs located inside (In) vs. outside (Out) of the west central African pygmy phenotype-associated regions following removal of the chromosome 3 locus harboring the potential signal of ancestral adaptation.

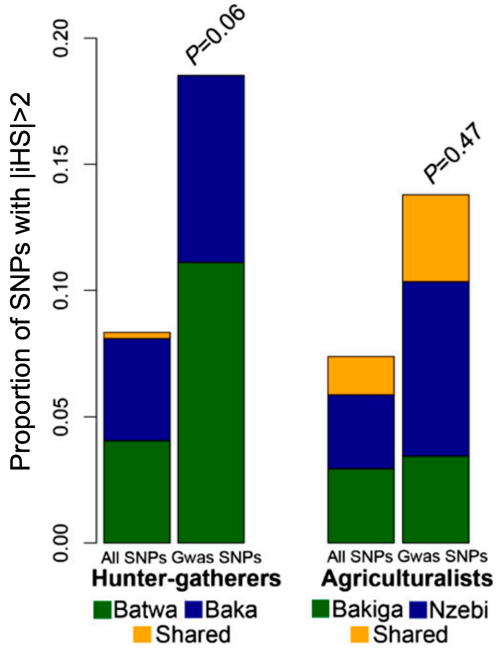


**Fig. S8.** Cross-population comparison of the proportion of candidate African stature SNPs with absolute iHS > 2 relative to the remainder of the genome. We considered a set of 30 candidate height-associated SNPs for individuals of African descent that were also included on our 1 million (1M) SNP array. We tested whether these loci were enriched for high iHS values (absolute iHS > 2) relative to the genomic background. The barplots on the left show the proportion of SNPs with absolute iHS > 2 in either the Batwa or Baka for the 30 height-associated SNPs (i.e., GWAS SNPs) and the proportion observed for the remainder of the SNPs in the genome (i.e., All SNPs). The green fraction shows the proportion that is due to SNPs with |iHS| > 2 only in the Batwa, blue shows the proportion that is due to SNPs with |iHS| > 2 only in the Baka, and orange the proportion that is due to SNPs with |iHS| > 2 in both the Batwa and the Baka. The barplots on the right show similar data but for the two Bantu-speaking agricultural populations: the Bakiga and the Nzebi/Nzime.

## Dataset S1. Sample information

Dataset S1

**Dataset S2. Genomic regions associated with the pygmy phenotype in Batwa east central African rainforest hunter-gathers**

Dataset S2

**Dataset S3. Summary of genomic regions harboring potential signatures of strong, recent positive selection based on iHS and BayeScan results, and Gene Ontology enrichment results**

Dataset S3

**Dataset S4. Genomic regions previously reported to be associated with the pygmy phenotype in west central African rainforest hunter-gathers**

Dataset S4