Polygenic adaptation and convergent evolution across both growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers

Christina M. Bergey[1,2], Marie Lopez[3,4,5], Genelle F. Harrison[6], Etienne Patin[3,4,5], Jacob Cohen[2], Lluis Quintana-Murci[3,4,5,*], Luis Barreiro[6,*], and George H. Perry[1,2,7,*]

[1]Department of Anthropology, Pennsylvania State University, University Park, Pennsylvania, U.S.A.

[2]Department of Biology, Pennsylvania State University, University Park, Pennsylvania, U.S.A.

[3]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France.

[4]Centre National de la Recherche Scientifique UMR 2000, Paris, France.

[5]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France.

[6]Université de Montréal, Centre de Recherche CHU Sainte-Justine, Montréal, Canada.

[7]Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania, U.S.A.

[*] *Co-senior authors*

April 18, 2018

Corresponding authors: C.M.B. (cxb585@psu.edu) and G.H.P. (ghp3@psu.edu)

Abstract:

Different human populations facing similar environmental challenges have sometimes evolved convergent biological adaptations, for example hypoxia resistance at high altitudes and depigmented skin in northern latitudes on separate continents. The pygmy phenotype (small adult body size), a characteristic of hunter-gatherer populations inhabiting both African and Asian tropical rainforests, is often highlighted as another case of convergent adaptation in humans. However, the degree to which phenotypic convergence in this polygenic trait is due to convergent vs. population-specific genetic changes is unknown. To address this question, we analyzed high-coverage sequence data from the protein-coding portion of the genomes (exomes) of two pairs of populations, Batwa rainforest hunter-gatherers and neighboring Bakiga agriculturalists from Uganda, and Andamanese rainforest hunter-gatherers (Jarawa and Onge) and Brahmin agriculturalists from India. We observed signatures of convergent positive selection between the Batwa and Andamanese rainforest hunter-gatherers across the set of genes with annotated 'growth factor binding' functions ($p < 0.001$). Unexpectedly, for the rainforest groups we also observed convergent and population-specific signatures of positive selection in pathways related to cardiac development (e.g. 'cardiac muscle tissue development'; $p = 0.003$). We hypothesize that the growth hormone sub-responsiveness likely underlying the pygmy phenotype may have led to compensatory changes in cardiac pathways, in which this hormone also plays an essential role. Importantly, we did not observe similar patterns of positive selection on sets of genes associated with either growth or cardiac development in the agriculturalist populations, indicating that our results most likely reflect

a history of convergent adaptation to the similar ecology of rainforest hunter-gatherers rather than a more common or general evolutionary pattern for human populations.

Introduction

Similar ecological challenges may repeatedly result in similar evolutionary outcomes, and many instances of phenotypic convergence arising from parallel changes in the same genetic loci have been uncovered (reviewed in [11, 18, 52]). Many examples of convergent genetic evolution reported to date are for simple monogenic traits, for example depigmentation in independent populations of Mexican cave fish living in lightless habitats [21, 48] and persistence of the ability to digest lactose in adulthood in both European and African agriculturalist/pastoralist humans [53]. Most biological traits, however, are highly polygenic. Since the reliable detection of positive selection in aggregate on multiple loci of individually small effect (i.e., polygenic adaptation) is relatively difficult [12, 46, 47, 51, 56], the extent to which convergent genetic changes at the same loci and functional pathways or changes affecting distinct genetic pathways may underlie these complex traits is less clear.

Human height is a classic example of a polygenic trait with approximately 800 known loci significantly associated with stature in Europeans collectively accounting for 27.4% of the heritable portion of height variation in this population [30]. A stature phenotype also represents one of most striking examples of convergent evolution in humans. Small body size (or the "pygmy" phenotype, e.g. average adult male stature <155 cm) appears to have evolved independently in rainforest hunter-gatherer populations from Africa, Asia, and South America [43], as groups on different continents do not share common ancestry to the exclusion of nearby agriculturalists [35, 49]. Positive correlations between stature and the degree of admixture with neighboring agriculturalists have confirmed that the pygmy

phenotype is, at least in part, genetically mediated and therefore potentially subject to natural selection [5, 25, 44, 45]. In one rainforest hunter-gatherer population, the Batwa from Uganda, an admixture mapping analysis identified 16 genetic loci specifically associated with the pygmy phenotype [44]. While these genomic regions were enriched for genes involved in the growth hormone pathway and for variants associated with stature in Europeans, there was no significant overlap between the pygmy phenotype-associated regions and the strongest signals of positive selection in the Batwa genome. Rather, subtle shifts in allele frequencies were observed across these regions in aggregate, suggesting a history of polygenic adaptation for the Batwa pygmy phenotype [44].

Here, we investigate population-specific and convergent patterns of positive selection in African and Asian hunter-gatherer populations using genome-wide sequence data from two sets of populations: the Batwa rainforest hunter-gatherers of Uganda in East Africa and the nearby Bakiga agriculturalists [28], and the Jarawa and Onge rainforest hunter-gatherers of the Andaman Islands in South Asia and the Uttar Pradesh Brahmin agriculturalists from mainland India [36, 37]. We specifically test whether convergent or population-specific signatures of positive selection, as detected both with 'outlier' tests designed to identify strong signatures of positive selection and tests designed to identify signatures of polygenic adaptation, are enriched for genes with growth-related functions. After studying patterns of convergent- and population-specific evolution in the Batwa and Andamanese hunter-gatherers, we then repeat these analyses in the paired Bakiga and Brahmin agriculturalists to evaluate whether the evolutionary patterns most likely relate to adaptation to hunter-gatherer subsistence in rainforest habitats, rather than being more generalized evolutionary patterns for human populations.

Results

We sequenced the protein coding portions of the genomes (exomes) of 50 Batwa rainforest hunter-gatherers and 50 Bakiga agriculturalists (dataset originally reported in [28]), identified single nucleotide polymorphisms (SNPs), and analyzed the resultant data alongside those derived from published whole genome sequence data for 10 Andamanese rainforest hunter-gatherers and 10 Brahmin agriculturalists (dataset from [36]). We restricted our analysis to exonic SNPs, for comparable analysis of the Asian whole genome sequence data with the African exome sequence data. To polarize allele frequency differences observed between each pair of hunter-gatherer and agriculturalist populations, we merged these data with those from outgroup comparison populations from the 1000 Genomes Project [4]: exome sequences of 30 unrelated British individuals from England and Scotland (GBR) for comparison with the Batwa/Bakiga data, and exome sequences of 30 Luhya individuals from Webuye, Kenya (LWK) for comparison with the Andamanese/Brahmin data. Outgroup populations were selected for genetic equidistance from the test populations. While minor levels of introgression from a population with European have been observed for the Batwa and Bakiga [28, 42], PBS is relatively robust to low levels of admixture [24].

To identify regions of the genome that may have been affected by positive selection in each of our test populations, we computed the population branch statistic (PBS; [58]) for each exonic SNP identified among or between the Batwa and Bakiga, and Andamanese and Brahmin populations (Fig. S1, S2; Table S5). PBS is an estimate of the magnitude of allele frequency change that occurred along each population lineage following divergence of the most closely related populations, with the allele frequency information from the outgroup population used to polarize frequency changes to one or both branches. Larger PBS values for a population reflect greater allele frequency change on that branch, which in some cases could reflect a history of positive selection [58].

For each analyzed population, we computed a PBS selection index for each gene by comparing the mean PBS for all SNPs located within that gene to a distribution of values estimated by shuffling SNP-gene associations (without replacement) and re-computing the mean PBS value for that gene 10,000 times (Table S7, S8). The PBS selection index is the percentage of permuted values that is higher than the actual (observed) mean PBS value for that gene. Per-gene PBS selection index values were not significantly correlated with gene size (linear regression of log adjusted selection indices against gene length: adjusted $R^2 = -2.74 \times 10^{-5}$, F-statistic $p = 0.81$; Fig. S3), suggesting that this metric is not overtly biased by gene size.

Convergent evolution can operate at different scales, including on the same mutation or amino acid change, different genetic variants between populations but within the same genes, or across a set of genes involved in the same biomolecular pathway or functional annotation. Given that our motivating phenotype is a complex trait and signatures of polygenic adaptation are expected to be relatively subtle and especially difficult to detect at the individual mutation and gene levels, in this study we principally consider patterns of convergence versus population specificity at the functional pathway/annotation level. We do note that when we applied the same approaches described in this study to individual SNPs, we identified several individual alleles with patterns of convergent allele frequency evolution between the Batwa and Andamanese that may warrant further study (Table S6), including a nonsynonymous SNP in the gene *FIG4*, which when disrupted in mice results in a phenotype of small but proportional body size [8]. However, likely related to the above-discussed challenges of identifying signatures of polygenic adaptation at the locus-specific level, the results of our individual SNP and gene analyses were otherwise largely unremarkable, and thus our the remainder of our report and discussion focuses on pathway-level analyses.

Outlier signatures of strong convergent and population-specific selection

The set of genes with the lowest (outlier) PBS index values for each population may be enriched for genes with histories of relatively strong positive natural selection. We used a permutation-based analysis to test whether curated sets of genome-wide growth-associated genes (4 lists tested separately ranging 266-3,996 genes in length; 4,888 total; Suppl. Text) or individual Gene Ontology (GO) annotated functional categories of genes (GO categories with fewer than 50 genes were excluded) have significant convergent excesses of genes with low PBS selection index values (< 0.01) in both of two cross-continental populations, for example the Batwa and Andamanese. Specifically, we first used Fisher's exact tests to estimate the probability that the number of genes with PBS selection index values < 0.01 was greater than that expected by chance, for each functional category set of genes and population. We then reshuffled the PBS selection indices across all genes 1,000 different times for each population to generate distributions of permuted enrichment p-values for each functional category set of genes. We compared our observed Batwa and Andamanese Fisher's exact test p-values to those from the randomly generated distributions as follows. We computed the joint probability of the null hypotheses for both the Andamanese and Batwa being false as $(1 - p_{Batwa})(1 - p_{Andamanese})$, where p_{Batwa} and $p_{Andamanese}$ are the p-values of the Fisher's exact test, and we compared this joint probability estimate to the same statistic computed for the p-values from the random iterations. We then defined the p-value of our empirical test for convergent evolution as the probability that this statistic was more extreme (lower) for the observed values than for the randomly generated values. The resultant p-value summarizes the test of the null hypothesis that both results could have been jointly generated under random chance. While each individual population's outlier-based test results are not significant after multiple test correction, this joint approach provides increased power to

identify potential signatures of convergent selection by assessing the probability of obtaining two false positives in these independent samples.

Several GO biological processes were significantly overrepresented—even when accounting for the number of tests performed—among the sets of genes with outlier signatures of positive selection in both the Batwa and Andamanese hunter-gatherer populations (empirical test for convergence $p < 0.005$; Table 1; Fig. 1A). These GO categories include 'skeletal system morphogenesis' (GO:0048705; empirical test for convergence $p < 0.001$; $q < 0.001$; Batwa: genes observed = 5, expected = 2.2, Fisher's exact $p = 0.069$; Andamanese: observed = 6, expected = 2.95, Fisher's exact $p = 0.074$).

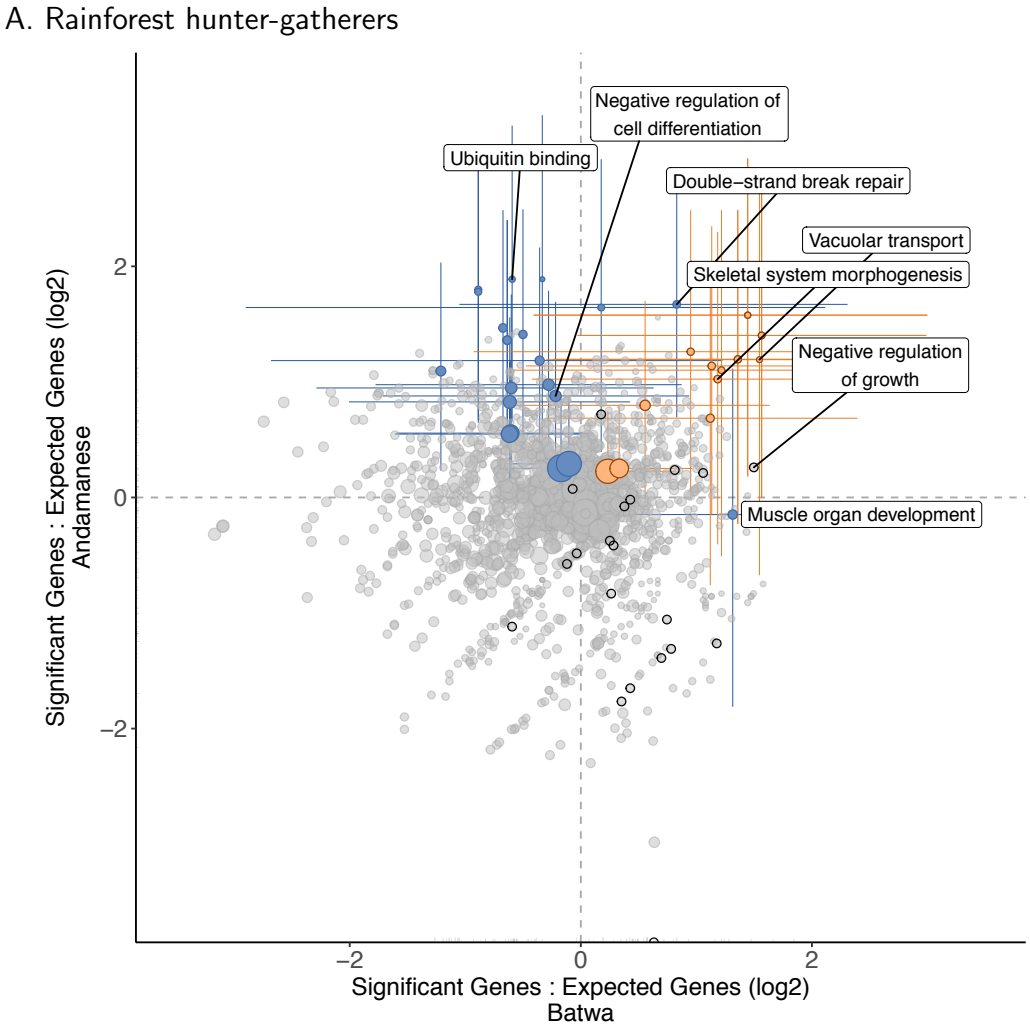

Figure 1: (Continued on the following page.)

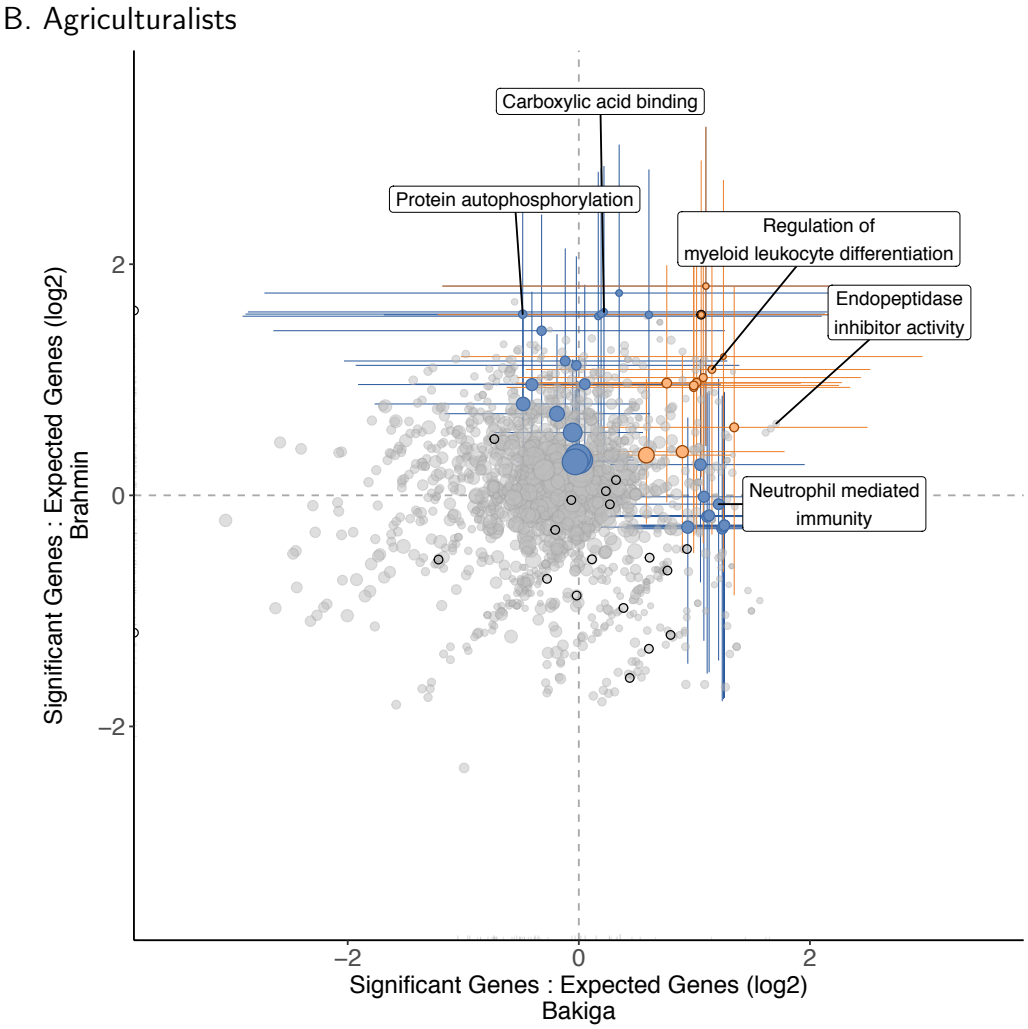

Figure 2: Gene Ontology (GO) functional categories' ratios of expected to observed counts of outlier genes (with PBS selection index < 0.01) in the Batwa and Andamanese rainforest hunter-gatherers (A) and Bakiga and Brahmin agriculturalist control (B). Results shown for GO biological processes and molecular functions. Point size is scaled to number of annotated genes in category. Terms that are significantly overrepresented for genes under positive selection (Fisher $p < 0.01$) in either population shown in blue and for both populations convergently (empirical permutation-based $p < 0.005$) shown in orange. Colored lines represent 95% CI for significant categories estimated by bootstrapping genes within pathways. Dark outlines indicate growth-associated terms: the 'growth' biological process (GO:0040007) and its descendant terms, or the molecular functions 'growth factor binding,' 'growth factor receptor binding,' 'growth hormone receptor activity,' and 'growth factor activity' and their descendants.

Other functional categories of genes were overrepresented in the sets of outlier loci for

one of these hunter-gatherer populations but not the other (Fig. 1A; Table S2, S14). The top population-specific enrichments for genes with outlier PBS selection index values for the Batwa were associated with growth and development: 'muscle organ development' (GO:0007517; observed genes: 10; expected genes: 4.02; $p = 0.007$) and 'negative regulation of growth' (GO:0045926; observed = 7; expected = 2.48; $p = 0.012$). Significantly overrepresented GO biological processes for the Andamanese included 'negative regulation of cell differentiation' (GO:0045596; observed genes: 18; expected genes: 9.79; $p = 0.009$). However, these population-specific enrichments were not significant following multiple test correction (false discovery rate $q = 0.71$ for both Batwa terms and $q = 0.22$ for the Andamanese result).

In contrast, no GO functional categories were observed to have similarly significant convergent excesses of 'outlier' genes with signatures of positive selection across the two agriculturalist populations as that observed for the rainforest hunter-gatherer populations (Fig. 1B; Table S9), and both the top ranked GO categories from the convergent evolution analysis and the population-specific analyses were absent any obvious connections to skeletal growth. The top-ranked functional categories with enrichments for genes with outlier PBS selection index values for the individual agriculturalist populations included 'neutrophil activation involved in immune response' for the Bakiga (GO:0002283; observed = 13; expected = 5.43; $p = 0.003$; $q = 0.41$) and 'protein autophosphorylation' for the Brahmin (GO:0046777; observed = 11; expected = 3.71; $p = 0.0012$; $q = 0.16$; Table S14).

Signatures of convergent and population-specific polygenic adaptation

Outlier-based approaches such as that presented above are expected to have limited power to identify signatures of polygenic adaptation [12, 46, 47, 51, 56] which is our expectation

for the pygmy phenotype [44]. Unlike the previous analyses in which we identified functional categories with an enriched number of genes with outlier PBS selection index values, for our polygenic evolution analysis we computed a "distribution shift-based" statistic to instead identify functionally-grouped sets of loci with relative shifts in their distributions of PBS selection indices. Specifically, we used the Kolmogorov-Smirnov (KS) test to quantify the distance between the distribution of PBS selection indices for the genes within a functional category to that of the genome-wide distribution. Significantly positive shifts in the PBS selection index distribution for a particular functional category may reflect individually subtle but consistent allele frequency shifts across genes within the category, which could result from either a relaxation of functional constraint or a history of polygenic adaptation. Our approach is similar to another recent method that was used to detect polygenic signatures of pathogen-mediated adaptation in humans [13]. As above, we identified functional categories with convergently high KS values between cross-continental groups by repeating these tests 1,000 times on permuted gene-PBS values and computing the joint probability of both null hypotheses being false for the two populations. We then compared this value from the random iterations to the same statistic computed with the observed KS p-values for each functional category. For example, for the Batwa and Andamanese, we tallied the number of random iterations for which the joint probability of both null hypotheses being false was more extreme (lower) than those of the random iterations. In this way we tested the null hypothesis that both of our observed p-values could have been jointly generated by random chance.

The GO molecular function with the strongest signature of a convergent polygenic shift in PBS selection indices across the Batwa and Andamanese populations was 'growth factor binding' (Table S3, S15; Fig. 3A; GO:0019838; Batwa $p = 0.021$; Andamanese $p = 0.027$; Fisher's combined $p = 0.0048$; empirical test for convergence $p < 0.001$; $q < 0.001$), and the top GO biological process was 'organ growth' (GO:0035265; Batwa $p = 0.028$; Andamanese

$p = 0.045$; Fisher's combined $p = 0.0095$; empirical test for convergence $p = 0.001$; $q = 1.0$). The other top Batwa-Andamanese convergent GO biological processes are not as obviously related to growth, but instead involve muscles, particularly heart muscles. A significant convergent shift in PBS selection indices across both hunter-gatherer populations was observed for 'cardiac muscle tissue development' (GO:0048738; Batwa $p = 0.046$; Andamanese $p = 0.003$; Fisher's combined $p = 0.001$; empirical test for convergence $p = 0.003$).

In contrast, when this analysis was repeated on the agriculturalist populations, no growth- or muscle-related functional annotations were observed with significantly convergent shifts in both populations (Fig. 3B; Table S16). The GO categories with evidence of potential convergent evolution between the agriculturalists were the biological processes 'leukocyte differentiation' (GO:0002521; Bakiga $p = 0.0086$; Brahmin $p = 0.0149$; Fisher's combined $p = 0.00128$; convergence empirical $p = 0.016$) and 'protein autophosphorylation' (GO:0046777; Bakiga $p = 0.033$; Brahmin $p = 0.0099$; Fisher's combined $p = 0.003$; convergence empirical $p = 0.036$) and the molecular function 'serine-type peptidase activity' (GO:0008236; Bakiga $p = 0.039$; Brahmin $p = 0.011$; Fisher's combined $p = 0.003$; convergence empirical $p = 0.0038$).

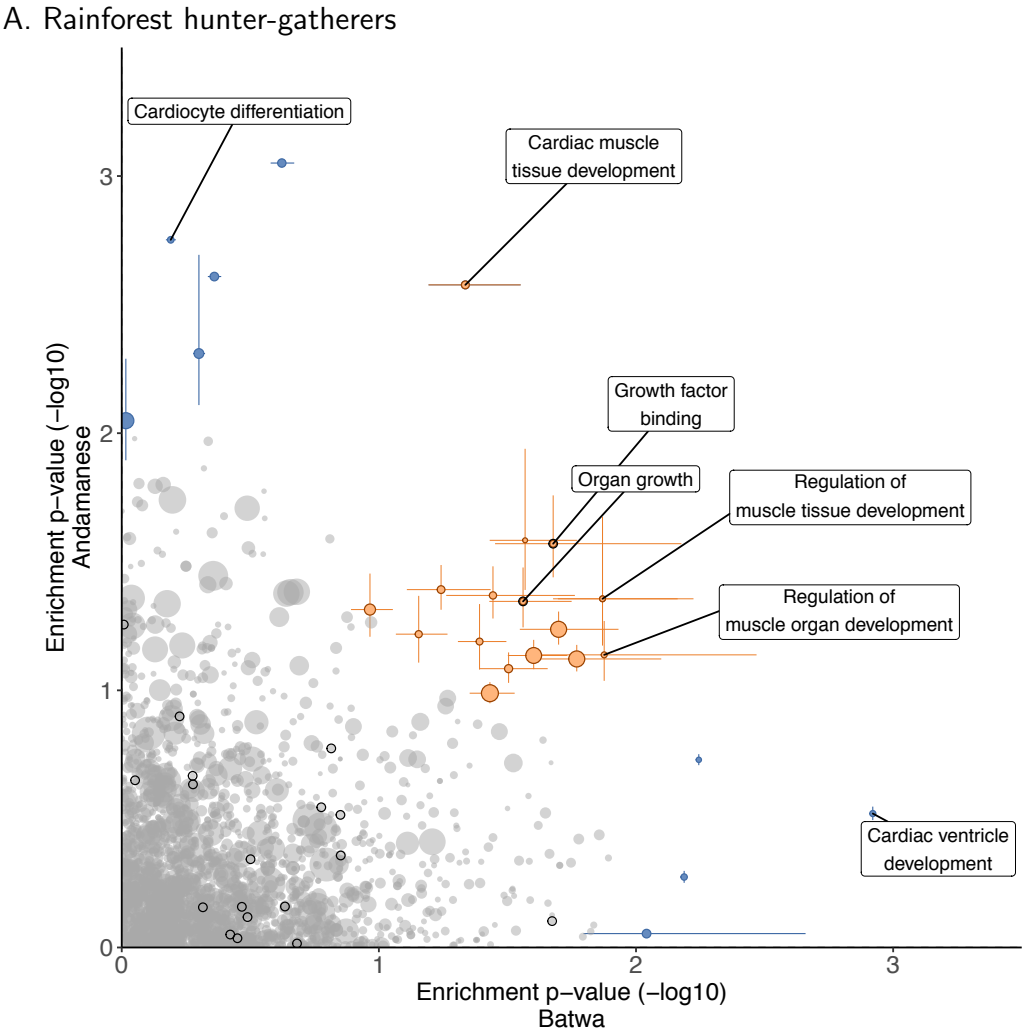

Figure 3: (Continued on the following page.)

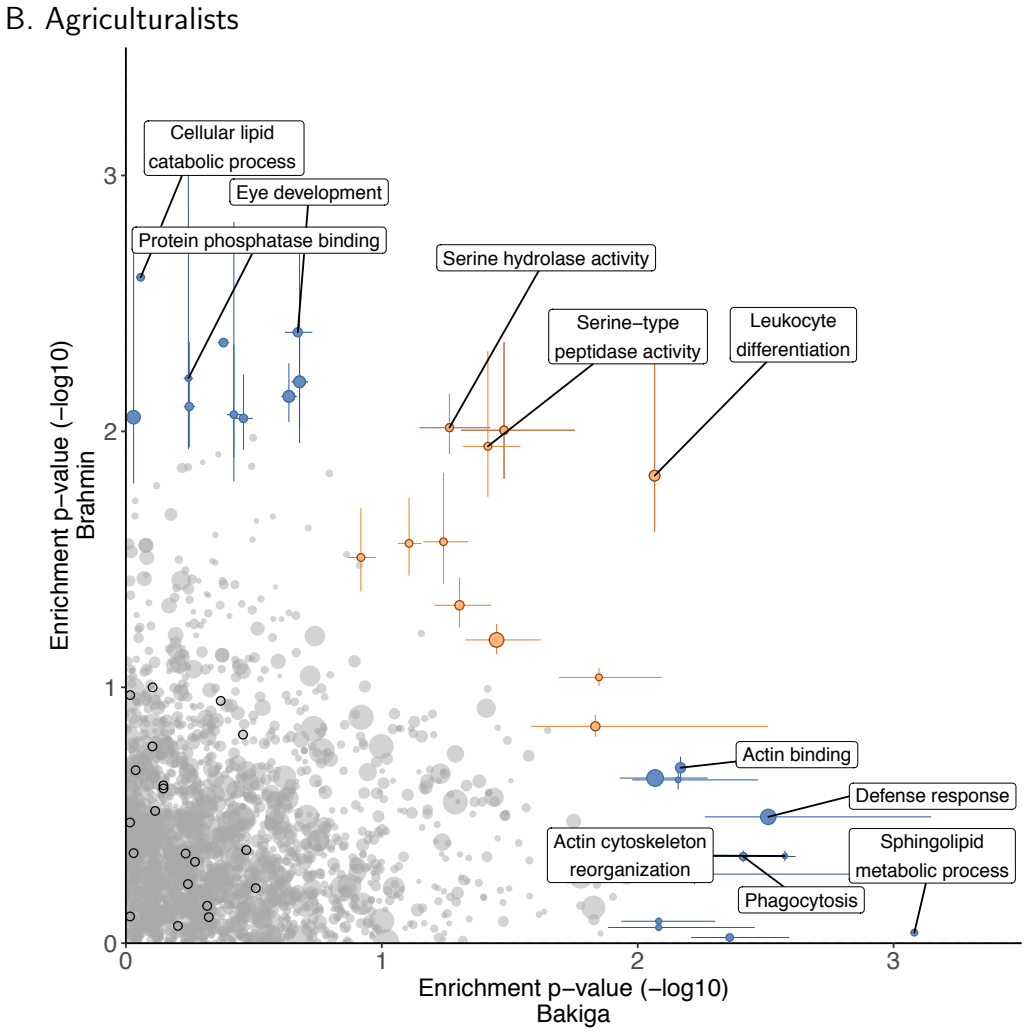

Figure 4: Gene Ontology (GO) functional categories' distribution shift test p-values, indicating a shift in the genes' PBS selection index values, in the Batwa and Andamanese rainforest hunter-gatherers (A) and Bakiga and Brahmin agriculturalist control (B). Results shown for GO biological processes and molecular functions. Point size is scaled to number of annotated genes in category. Terms that are significantly enriched for genes under positive selection (Kolmogorov-Smirnov $p < 0.01$) in either population shown in blue and for both populations convergently (empirical permutation-based $p < 0.005$) shown in orange. Colored lines represent 95% CI for significant categories estimated by bootstrapping genes within pathways. Dark outlines indicate growth-associated terms: the 'growth' biological process (GO:0040007) and its descendant terms, or the molecular functions 'growth factor binding,' 'growth factor receptor binding,' 'growth hormone receptor activity,' and 'growth factor activity' and their descendants.

We also used Bayenv, a Bayesian linear modeling method for identifying loci with allele

frequencies that covary with an ecological variable [12, 22], to assess the level of consistency with our convergent polygenic PBS shift results. Specifically, we used Bayenv to test whether the inclusion of a binary variable indicating subsistence strategy would increase the power to explain patterns of genetic diversity for a given functional category of loci over a model that only considered population history (as inferred from the covariance of genome-wide allele frequencies in the dataset.) We converted Bayes factors into per-gene index values via permutation of SNP-gene associations (Table S11) and identified GO terms with significant shifts in the Bayenv Bayes factor index distribution [12, 22] (Table S17). The top results from this analysis included 'growth factor activity' (GO:0008083; $p = 0.006$; $q = 0.11$), categories related to enzyme regulation (e.g. 'enzyme regulator activity'; GO:0030234; $p = 0.003$; $q = 0.01$), and categories related to muscle cell function (e.g. 'microtubule binding'; GO:0008017; $p = 0.003$; $q = 0.10$). There were more GO terms that were highly ranked ($p < 0.05$) in both the hunter-gatherer PBS shift-based empirical test of convergence and the Bayenv analysis than expected by chance (for biological processes GO terms: observed categories in common = 13, expected = 8.03, Fisher's exact test $p = 9.67 \times 10^{-5}$; for molecular function GO terms: observed categories in common = 4, expected = 1.45, Fisher's exact test $p = 0.045$).

While we did not observe any significant population-specific shifts in PBS selection index values for growth-associated GO functional categories in any of our studied populations (Table 4; Suppl. Text), for each individual rainforest hunter-gatherer population we did observe nominal shifts in separate biological process categories involving the heart (Fig. 3A). For the Batwa, 'cardiac ventricle development' (GO:0003231) was the top population-specific result (median PBS index = 0.272 vs. genome-wide median PBS index = 0.528; $p = 0.001$; $q = 0.302$). For the Andamanese, 'cardiocyte differentiation' (GO:0035051) was also ranked highly (median PBS index = 0.353 vs. genome-wide median PBS index = 0.552; $p = 0.002$; $q = 0.232$). We note that while these are separate population-specific signatures, 17 genes are shared between the above two cardiac-related pathways (of 61 total 'cardiocyte

differentiation' genes total, 28%; of 71 total 'cardiac ventricle development' genes, 24%; Table S18).

In contrast, cardiac development-related GO categories were not observed among those with highly-ranked population-specific polygenic shifts in selection index values for either the Bakiga or Brahmin agriculturalists (Fig. 3B; Table S19). The only GO term with a significant population-specific shift in the agriculturalists after multiple test correction was molecular function 'carboxylic acid binding' in the Brahmins (GO:0031406; $p = 7.30 \times 10^{-5}$; $q = 0.005$).

Discussion

The independent evolution of small body size in multiple different tropical rainforest environments worldwide presents a natural human model for comparative study of the genetic and evolutionary bases of growth and body size. Through an evolutionary genomic comparison of African and Asian rainforest hunter-gatherer populations to one another and with nearby agriculturalists, we have gained additional, indirect insight into the genetic structure of body size, a fundamental biological trait. Specifically, we identified a signature of potential convergent positive selection on the growth factor binding pathway that could partially underlie the independent evolution of small body size in African and Asian rainforest hunter-gatherers.

Unexpectedly, we also observed signatures of potential polygenic selection across functional categories of genes related to heart development for the rainforest hunter-gatherer populations, both convergently and on a population-specific basis. To a minor extent, the growth factor- and heart-related functional categories highlighted in our study do overlap: of the 123 total genes annotated across the three heart-related categories ('cardiac muscle tissue development' GO:0048738, 'cardiac ventricle development' GO:0003231, and 'cardiocyte

differentiation' GO:0035051), nine (7.3%) are also included among the 66 annotated genes in the 'growth factor binding' category (GO:0019838). However, even after excluding these nine genes from our dataset, we still observed similar polygenic PBS shifts in the Batwa and Andamanese for both growth factor- and heart-related functional categories (Suppl. Text), demonstrating that our observations are not driven solely by cross-annotated genes.

We hypothesize that the evolution of growth hormone sub-responsiveness, which appears to at least partly underlie short stature in some rainforest hunter-gatherer populations [19, 20, 32, 33, 50] may in turn have also resulted in strong selection pressure for compensatory adaptations in cardiac pathways. The important roles of growth hormone (GH) in the heart are evident from studies of patients deficient in the hormone. For example, patients with GH deficiency are known to be at an increased risk of atherosclerosis and mortality from cardiovascular disease [10] and have worse cardiac function [3]. More broadly, shorter people have elevated risk of coronary artery disease [41], likely due to the pleiotropic effects of variants affecting height and atherosclerosis development [39]. Such health outcomes may relate to the important roles that GH plays in the development and function in the myocardium [15, 34], which contains a high relative concentration of receptors for GH [31]. We hypothesize that the adaptive evolution of growth hormone sub-responsiveness underlying short stature in rainforest hunter-gatherers may have necessitated compensatory adaptations in the cardiac pathways reliant on growth hormone.

An alternative explanation for our finding of potential convergent selection on cardiac-related pathways relates to the nutritional stress of full-time human rainforest habitation. Especially prior to the ability to trade forest products for cultivated goods with agriculturalists, the diets of full-time rainforest hunter-gatherers may have been calorically and nutritionally restricted on at least a seasonal basis [43]. Caloric restriction has a direct functional impact on cardiac metabolism and function, with modest fasting in mice leading to the depletion of myocardial phospholipids, which potentially act as a metabolic reserve to en-

sure energy to essential heart functions [23]. In human rainforest hunter-gatherers, selection may have favored variants conferring cardiac phenotypes optimized to maintain myocardial homeostasis during the nutritional stress that these populations may have experienced in the past.

An important caveat to our study is the lack of statistical significance for our population-specific analyses after controlling for the multiplicity of tests resulting from hierarchically nested GO terms. The absence of strong signals of positive selection that are robust to the multiple testing burden likely reflects both the expected subtlety of evolutionary signals of selection on polygenic traits and the restriction of our dataset to gene coding region sequences. However, our comparative approach to identify signatures of convergent evolution is more robust. Therefore, while we cannot yet accurately estimate the extent to which signatures of positive selection that potentially underlie the evolution of the pygmy phenotype occurred in the same versus distinct genetic pathways between the Batwa and Andamanese, we do feel confident in our findings of convergent growth-related and cardiac-related pathways evolution. The concurrent signatures of convergent evolution across these two pathways in both African and Asian rainforest hunter-gatherers is an example of the insight into a biomedically-relevant phenotype that can be gained from the comparative study of human populations with non-pathological natural variation.

Materials and Methods

Dataset generation and sourcing

Sample collection, processing, and sequencing have been previously described [28, 44]. Briefly, sampling of biomaterials (blood or saliva) from Batwa rainforest hunter-gatherers and Bakiga agriculturalists of southwestern Uganda took place in 2010, and DNA was extracted [44]. Collection was approved by the Institutional Review Boards (IRBs) of both the University

of Chicago (#16986A) and Makerere University, Kampala, Uganda (#2009-137), and local community approval and individual informed consent were obtained before collection. DNA samples of 50 Batwa and 50 Bakiga adults were included in the present study for exome capture using the Nextera Rapid Capture Expanded Exome kit and sequencing [28].

Variants were identified as described in [28]. Briefly, reads were aligned to the hg19/GRCh37 genome with BWA v.0.7.7 mem with default settings [27], PCR duplicates were detected with Picard Tools v.1.94 (http://broadinstitute.github.io/picard), and re-alignment around indels and base quality recalibration was done with GATK v3.5 [14] using the known indel sites from the 1000 Genomes Project [4]. Variants were called individually with GATK HaplotypeCaller [14], and variants were pooled together with GATK GenotypeGVCF and filtered using VQSR. Only biallelic SNPs with a minimum depth of 5x and less than 85% missingness that were polymorphic in the entire dataset were retained for analyses.

Variant data for the Andamanese individuals (Jarawa and Onge) and an outgroup mainland Indian population (Uttar Pradesh Brahmins) from [36] were downloaded in VCF file format from a public website. To ensure the exome capture-derived African and whole genome shotgun sequencing-derived Asian datasets were comparable, we restricted our analysis to only exonic SNPs.

Merging with 1000 Genomes data

We chose outgroup comparison populations from the 1000 Genomes Project [4] to be equally distantly related to the ingroup populations: Reads from a random sample of 30 unrelated individuals from British in England and Scotland (GBR) and Luhya in Webuye, Kenya (LWK) were chosen for the Batwa/Bakiga and Andamanese/Brahmin datasets, respectively. We re-called variants in each 1000 Genomes comparison population at loci that were variable in the ingroup populations using GATK UnifiedGenotyper [14]. Variants were filtered to exclude those with QD < 2.0, MQ < 40.0, FS > 60.0, HaplotypeScore > 13.0, MQRankSum

$<$ -12.5, or ReadPosRankSum $<$ -8.0. We removed SNPs for which fewer than 10 of the 30 individuals from the 1000 Genomes datasets had genotypes.

Computation of the Population Branch Statistic (PBS) and the per-gene PBS index

Using these merged datasets, we computed F_{ST} between population pairs using the unbiased estimator of Weir and Cockerham [55], transformed it to a measure of population divergence $[T = -log(1 - F_{ST})]$, and then calculated the Population Branch Statistic (PBS), after [58]. PBS was computed on a per-SNP basis. We computed an empirical p-value for each SNP, simply the proportion of coding SNPs with PBS greater than this SNP's value, which we adjusted for FDR.

SNPs were annotated with gene-based information using ANNOVAR [54] with refGene (Release 76) [40] and PolyPhen [1] data. As the Andamanese/Brahmin dataset spanned the genome and the Batwa/Bakiga exome dataset included off target intronic sequences as well as untranslated regions (UTRs), and microRNAs, we restricted our analysis to only exonic SNPs. For both the Batwa/Bakiga and Andamanese/Brahmin datasets, we computed a "PBS selection index" for each gene as follows. We compared the mean PBS for all SNPs located within that gene to a distribution of values estimated by shuffling SNP-gene associations (without replacement) and re-computing the mean PBS value for that gene 10,000 times. We defined the PBS selection index of the gene as the percentage of these empirical mean values that is higher than its observed mean PBS value. When identifying outlier genes, gene-based indices were adjusted for FDR.

To identify SNPs with allele frequency that correlate to subsistence strategy (hunter-gatherer: Andamanese and Batwa; agriculturalists: Bakiga and Brahmin), we used Bayenv2.0 [22] to assess whether the addition of a binary variable denoting subsistence strategy im-

proved the Bayesian model that already took into account covariance between samples due to ancestry. As with the PBS results, we computed each gene's index by sampling new values for each SNP from the distribution of all Bayes factors and comparing the gene's actual average to those of the bootstrapped replicates.

Creation of *a priori* lists of growth-related genes

To test the hypothesis that genes with known influence on growth would show increased positive selection in rainforest hunter-gatherer populations, we curated *a priori* lists of growth-related genes as described fully in the Supplemental Text. Briefly, we obtained the following gene lists: i) 3,996 genes that affect growth or size in mice (MP:0005378) from the Mouse/Human Orthology with Phenotype Annotations database [6]; ii) 266 genes associated with abnormal skeletal growth syndromes in the Online Mendelian Inheritance in Man (OMIM) database (https://omim.org), as assembled by [57]; iii) 427 genes expressed substantially more highly in the mouse growth plate, the cartilaginous region on the end of long bones where bone elongation occurs, than in soft tissues [lung, kidney, heart; >= 2.0 fold change; [29]]; and iv) 955 genes annotated with the Gene Ontology "growth" biological process (GO:0040007). As the GH/IGF1 pathway is a major regulator of growth and disruptions to the pathway have been implicated in the pygmy phenotype, we also collected lists of genes associated with GH1 and IGF1 respectively from the OPHID database of proteinprotein interaction (PPI) networks [7]. Separately, we also used a list of genes found to be associated with the pygmy phenotype in the Batwa [44].

Statistical overrepresentation and distribution shift tests

Using the PBS and Bayenv indices, we next tested for a statistical over-representation of extreme values ($p < 0.01$) for the above *a priori* gene lists as well as all Gene Ontology

(GO) terms using the topGO package of Bioconductor [2], gene-to-GO mapping from the org.Hs.eg.db package [9], and Fisher's exact test in "classic" mode (i.e., without adjustment for GO hierarchy). We similarly performed a statistical enrichment test using the Kolmogorov-Smirnov test again in "classic" mode, which tested for a shift in the distribution of the PBS or Bayenv statistic, rather than a excess of extreme values. In all cases, we pruned the GO hierarchy to exclude GO terms with fewer than 50 annotated genes to reduce the number of tests, leaving 1,742 and 1,816 GO biological processes and 266 and 285 GO molecular functions tested for the African and Asian datasets, respectively. To further reduce the number of redundant tests, we also computed the semantic similarity between GO terms to remove very similar terms. We computed the similarity metric of [26] as implemented in the GoSemSim R package [59], a measure of the overlapping information content in each term using the annotation statistics of their common ancestor terms, and then clustered based on these pairwise distances between GO terms using Ward Hierarchical Clustering. We then pruned GO terms by cutting the tree at a height of 0.5 and retaining the term in each cluster with the lowest p-value. With this reduced set of GO overrepresentation and distribution shift results, we adjusted the p-value for FDR.

Identification of convergent evolution

We used two methods to identify convergent evolution: i.) computation of simple combined p-values for SNPs, genes, and GO overrepresentation and distribution shift tests using Fisher's and Edgington's methods, and ii.) a permutation based approach to identify GO pathways for which both the Batwa and Andamanese overrepresentation or distribution shift test results are more extreme than is to be expected by chance (the "empirical test for convergence"). These two approaches are summarized below.

We searched for convergence between Batwa and Andamanese individuals by computing the joint p-value for PBS on a per-SNP, per-gene, and per-GO term basis. We calculated all

joint p-values using Fisher's method (as the sum of the natural logarithms of the uncorrected p-values for the Batwa and Andamanese tests [38]) as well as via Edgington's method (based on the sum of all p-values [17]). Meta-analysis of p-values was done via custom script and the metap R package [16].

We also assessed the probability of getting two false positives in the Batwa and Andamanese selection results by shuffling the genes' PBS indices 1,000 times and performing GO overrepresentation and distribution shift tests on these permuted values. We compared the observed Batwa and Andamanese p-values to this generated distribution of p-values, as described above. We computed the joint probability of both null hypotheses being false for the Andamanese and Batwa as $(1-p_{Batwa})(1-p_{Andamanese})$, where p_{Batwa} and $p_{Andamanese}$ are the p-values of the Fisher's exact test or of the Kolmogorov-Smirnov test for the outlier- and shift-based tests, respectively, and we compared the joint probability to the same statistic computed for the p-values from the random iterations. The empirical test for convergence p-value was simply the number of iterations for which this statistic was more extreme (lower) for the observed values than for the randomly generated values.

Script and data availability

All scripts used in the analysis are available at https://github.com/bergeycm/rhg-convergence-analysis and released under the GNU General Public License v3. Exome data for the Batwa and Bakiga populations have previously been deposited in the European Genome-phenome Archive under accession code EGAS00001002457. Extended data tables are available at https://doi.org/10.18113/S1N63M.

References

[1] 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*. 7:248–249.

[2] Alexa A, Rahnenfuhrer J. 2016. topGO: enrichment analysis for gene ontology. *R package version*. 2.

[3] Arcopinto M, Salzano A, Giallauria F, et al. (20 co-authors). 2017. Growth hormone deficiency is associated with worse cardiac function, physical performance, and outcome in chronic heart failure: Insights from the T.O.S.CA. GHD study. *PLoS ONE*. 12:1–12.

[4] Auton A, Abecasis GR, Altshuler DM, et al. (776 co-authors). 2015. A global reference for human genetic variation. *Nature*. 526:68–74.

[5] Becker NSA, Verdu P, Froment A, Le Bomin S, Pagezy H, Bahuchet S, Heyer E. 2011. Indirect evidence for the genetic determination of short stature in African Pygmies. *American Journal of Physical Anthropology*. 145:390–401.

[6] Blake JA, Eppig JT, Kadin JA, et al. (39 co-authors). 2017. Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse. *Nucleic Acids Research*. 45:D723–D729.

[7] Brown KR, Jurisica I. 2005. Online predicted human interaction database. *Bioinformatics*. 21:2076–2082.

[8] Campeau PM, Lenk GM, Lu JT, et al. (16 co-authors). 2013. Yunis-Varón syndrome is caused by mutations in FIG4, encoding a phosphoinositide phosphatase. *American Journal of Human Genetics*. 92:781–791.

[9] Carlson M. 2017. org.Hs.eg.db: Genome wide annotation for Human.

[10] Carroll PV, Christ ER, Bengtsson BÅ, et al. (13 co-authors). 1998. Growth Hormone Deficiency in Adulthood and the Effects of Growth Hormone Replacement: A Review. *The Journal of Clinical Endocrinology & Metabolism*. 83:382–395.

[11] Christin PA, Weinreich DM, Besnard G. 2010. Causes and evolutionary significance of genetic convergence. *Trends in Genetics*. 26:400–405.

[12] Coop G, Witonsky D, Di Rienzo A, Pritchard JK. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics*. 185:1411–1423.

[13] Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L. 2013. Evidence for polygenic adaptation to pathogens in the human genome. *Molecular Biology and Evolution*. 30:1544–1558.

[14] DePristo MA, Banks E, Poplin R, et al. (18 co-authors). 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. 43:491–8.

[15] Devesa J, Almengló C, Devesa P. 2016. Multiple effects of growth hormone in the body: Is it really the hormone for growth? *Clinical Medicine Insights: Endocrinology and Diabetes*. 9:47–71.

[16] Dewey M. 2017. metap: meta-analysis of significance values.

[17] Edgington ES. 1972. An Additive Method for Combining Probability Values from Independent Experiments. *The Journal of Psychology*. 80:351–363.

[18] Elmer KR, Meyer A. 2011. Adaptation in the age of ecological genomics: Insights from parallelism and convergence. *Trends in Ecology and Evolution*. 26:298–306.

[19] Geffner ME, Bailey RC, Bersch N, Vera JC, Golde DW. 1993. Insulin-like growth factor-I unresponsiveness in an Efe Pygmy. *Biochemical and Biophysical Research Communications*. 193:1216–1223.

[20] Geffner ME, Bersch N, Bailey RC, Golde DW. 1995. Insulin-like growth factor I resistance in immortalized T cell lines from African Efe Pygmies. *Journal of Clinical Endocrinology and Metabolism*. 80:3732–3738.

[21] Gross JB, Borowsky R, Tabin CJ. 2009. A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish Astyanax mexicanus. *PLoS Genetics*. 5.

[22] Günther T, Coop G. 2013. Robust identification of local adaptation from allele frequencies. *Genetics*. 195:205–220.

[23] Han X, Cheng H, Mancuso DJ, Gross RW. 2004. Caloric restriction results in phospholipid depletion, membrane remodeling, and triacylglycerol accumulation in murine myocardium. *Biochemistry*. 43:15584–15594.

[24] Huerta-Sánchez E, DeGiorgio M, Pagani L, et al. (16 co-authors). 2013. Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. *Molecular Biology and Evolution*. 30:1877–1888.

[25] Jarvis JP, Scheinfeldt LB, Soi S, et al. (12 co-authors). 2012. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genetics*. 8.

[26] Jiang JJ, Conrath DW. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*. .

[27] Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25:1754–60.

[28] Lopez M, Kousathanas A, Quach H, et al. (12 co-authors). 2018. The demographic history and mutational load of African hunter-gatherers and farmers. *Nature Ecology & Evolution*. 2:721–730.

[29] Lui JC, Nilsson O, Chan Y, Palmer CD, Andrade AC, Hirschhorn JN, Baron J. 2012. Synthesizing genome-wide association studies and expression microarray reveals novel genes that act in the human growth plate to modulate height. *Human Molecular Genetics*. 21:5193–5201.

[30] Marouli E, Graff M, Medina-Gomez C, et al. (372 co-authors). 2017. Rare and low-frequency coding variants alter human adult height. *Nature*. 542:186–190.

[31] Mathews LS, Enberg B, Norstedt G. 1989. Regulation of rat growth hormone receptor gene expression. *The Journal of Biological Chemistry*. 264:9905–10.

[32] Merimee TJ, Rimoin DL, Cavalli-Sforza LC, Rabinowitz D, McKusick VA. 1968. Metabolic effects of human growth hormone in the African pygmy. *The Lancet*. 292:194–195.

[33] Merimee TJ, Rimoin DL, Cavalli-Sforza LL. 1972. Metabolic studies in the African pygmy. *The Journal of Clinical Investigation*. 51:395–401.

[34] Meyers DE, Cuneo RC. 2003. Controversies Regarding the Effects of Growth Hormone on the Heart. *Mayo Clinic Proceedings*. 78:1521–1526.

[35] Migliano AB, Romero IG, Metspalu M, et al. (24 co-authors). 2013. Evolution of the Pygmy Phenotype: Evidence of Positive Selection from Genome-wide Scans in African, Asian, and Melanesian Pygmies. *Human Biology*. 85:251–284.

[36] Mondal M, Casals F, Majumder PP, Bertranpetit J. 2016. Further confirmation for unknown archaic ancestry in Andaman and South Asia . *bioRxiv*. .

[37] Mondal M, Casals F, Xu T, et al. (11 co-authors). 2016. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature Genetics*. 48:1066–1070.

[38] Mosteller F, Fisher R. 1948. Questions and Answers. *The American Statistician*. 2:30–31.

[39] Nelson CP, Hamby SE, Saleheen D, et al. (56 co-authors). 2015. Genetically Determined Height and Coronary Artery Disease. *New England Journal of Medicine*. 372:1608–1618.

[40] O'Leary NA, Wright MW, Brister JR, et al. (55 co-authors). 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. 44:D733–D745.

[41] Paajanen TA, Oksala NK, Kuukasjärvi P, Karhunen PJ. 2010. Short stature is associated with coronary heart disease: A systematic review of the literature and a meta-analysis. *European Heart Journal*. 31:1802–1809.

[42] Patin E, Siddle KJ, Laval G, et al. (18 co-authors). 2014. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nature Communications*. 5.

[43] Perry GH, Dominy NJ. 2009. Evolution of the human pygmy phenotype. *Trends in Ecology and Evolution*. 24:218–225.

[44] Perry GH, Foll M, Grenier JC, et al. (14 co-authors). 2014. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences*. 111:E3596–603.

[45] Perry GH, Verdu P. 2016. Genomic perspectives on the history and evolutionary ecology of tropical rainforest occupation by humans. *Quaternary International*. 448:150–157.

[46] Pritchard JK, Di Rienzo A. 2010. Adaptation not by sweeps alone. *Nature Reviews Genetics*. 11:665–667.

[47] Pritchard JK, Pickrell JK, Coop G. 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Current Biology*. 20:R208–R215.

[48] Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ. 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature Genetics*. 38:107–111.

[49] Rasmussen M, Guo X, Wang Y. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science*. 334:94–8.

[50] Rimoin D, Merimee T, Rabinowitz D, Cavalli-Sforza L, McKusick V. 1969. Peripheral subresponsiveness to human growth hormone in the African pygmies. *The New England Journal of Medicine*. 281:1383–1388.

[51] Stephan W. 2016. Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*. 25:79–88.

[52] Stern DL. 2013. The genetic causes of convergent evolution. *Nature Reviews Genetics*. 14:751–764.

[53] Tishkoff SA, Reed FA, Ranciaro A, et al. (19 co-authors). 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*. 39:31–40.

[54] Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 38:e164.

[55] Weir B, Cockerham C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution*. 38:1358–1370.

[56] Wellenreuther M, Hansson B. 2016. Detecting Polygenic Evolution: Problems, Pitfalls, and Promises. *Trends in Genetics*. 32:155–164.

[57] Wood AR, Esko T, Yang J, et al. (445 co-authors). 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*. 46:1173–1186.

[58] Yi X, Liang Y, Huerta-Sanchez E, et al. (70 co-authors). 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science (New York, N.Y.)*. 329:75–8.

[59] Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. 2010. GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*. 26:976–978.

Acknowledgments

The authors would like to thank the Batwa and Bakiga communities and all individuals who participated in this study, and J.A. Hodgson and E.C. Reeves for helpful discussion. This work was supported by NIH R01-GM115656 (to G.H.P and L.B.B.), 1 F32 GM125228-01A1 (to C.M.B), and ANR AGRHUM ANR-14-CE02-0003-01 (to L.Q.-M.). M.L. was supported by the Fondation pour la Recherche Médicale (FDT20170436932). This research was conducted with Advanced CyberInfrastructure computational resources provided by The Institute for CyberScience at The Pennsylvania State University. The authors declare no competing financial interests.

Supplemental Material for: Polygenic adaptation and convergent evolution across both growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers

Christina M. Bergey[1,2], Marie Lopez[3,4,5], Genelle F. Harrison[6], Etienne Patin[3,4,5], Jacob Cohen[2], Lluis Quintana-Murci[3,4,5,*], Luis Barreiro[6,*], and George H. Perry[1,2,7,*]

[1]Department of Anthropology, Pennsylvania State University, Pennsylvania, U.S.A.

[2]Department of Biology, Pennsylvania State University, Pennsylvania, U.S.A.

[3]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France.

[4]Centre National de la Recherche Scientifique UMR 2000, Paris, France.

[5]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France.

[6]Université de Montréal, Centre de Recherche CHU Sainte-Justine, H3T 1C5 Montréal, Canada.

[7]Huck Institutes of the Life Sciences, Pennsylvania State University, Pennsylvania, U.S.A.

[*] *Co-senior authors*

April 18, 2018

Corresponding authors: C.M.B. (cxb585@psu.edu) and G.H.P. (ghp3@psu.edu)

Contents

List of Figures

List of Tables

1 Supplemental Text

Positive selection in growth-associated genes We examined whether gene-specific signatures of strong positive selection (using an "outlier-based"" designation of genes with PBS index values < 0.01) in the rainforest populations were enriched for known functional associations with growth using *a priori* lists of 4,888 total growth-related genes, consisting of (with some redundancy among individual categories, as expected): i) 3,996 genes that affect growth or size in mice (MP:0005378) from the Mouse/Human Orthology with Phenotype Annotations database [1]; ii) 266 genes associated with abnormal skeletal growth syndromes in the Online Mendelian Inheritance in Man (OMIM) database (https://omim.org/), as assembled by [5]; iii) 427 genes expressed substantially more highly in the mouse growth plate, the cartilaginous region on the end of long bones where bone elongation occurs, than in soft tissues (lung, kidney, heart; >= 2.0 fold change; [3]; and iv) 955 genes annotated with the Gene Ontology "growth" biological process (GO:0040007). Separately, we also considered in our analyses the set of 166 genes located within the 16 genomic regions previously associated with the pygmy phenotype in the Batwa, using an admixture mapping approach [4], as well as GH1- and IGF1-associated genes using data from OPHID database of proteinprotein interaction (PPI) networks [2].

We used each of the curated *a priori* growth-related gene lists for testing the hypothesis that such loci are enriched for genes with signatures of strong positive selection (outlier PBS selection index values) or have a shift in the distribution of PBS selection index values consistent with subtle polygenic adaptation in the Batwa and Andamanese rainforest hunter-gatherer but not the Bakiga and Brahmin agriculturalist populations. We identified 202, 188, 291, and 252 outlier strong selection candidate genes (with PBS index values < 0.01) in each of the Batwa, Bakiga, Andamanese, and Brahmin populations, respectively. Genes in the *a priori* growth-related gene lists were not significantly overrepresented among PBS

outliers in any populations, except for those associated with mouse growth phenotype in the Brahmin (68 observed, 47.7 expected; Fisher $p = 0.0179$) (Table S9). Though the lack of over-representation of growth-related gene lists among loci with outlier signatures of strong positive natural selection related to growth is perhaps unsurprising considering the polygenic phenotype, our distribution shift-based test also showed no significant shifts in the distribution of PBS indices for any population (Table S11). Genes in genomic regions previously associated with the pygmy phenotype in the Batwa [4] were enriched for genes with outlier PBS selection index values in the Batwa (outlier-based test: 5 observed, 1.39 expected; Fisher $p = 0.017$; Table S9) and the phenotype-associated genes' PBS distribution was shifted relative to the genome-wide distribution (distribution shift-based test: Kolmogorov-Smirnov test $p = 0.056$; Table S11). We found no evidence that genes associated with GH1 and IGF1 were enriched for outlier or polygenic selection.

Impact of cross-annotated genes between growth factor- and cardiac-related pathways To assess whether shared genes in GO categories relating to the heart and growth factor binding were responsible for the significant shift in PBS selection index values for genes in these annotations, we compared the distributions of PBS selection indices before and after removing 9 genes common to heart pathways and growth factor binding. The heart GO terms assessed were: 'cardiocyte differentiation' (GO:0035051), with a shift in the Andamanese hunter-gatherers; 'cardiac ventricle development' (GO:0003231), with a shift in the Batwa hunter-gatherers; and 'cardiac muscle tissue development' (GO:0048738) with a convergent shift in the Batwa and Andamanese. Of the 123 heart related genes contained in these pathways, 9 were also annotated to the GO molecular function 'growth factor binding' (GO:0019838): *ACVR1*, *EGFR*, *ENG*, *FGFR2*, *FGFRL1*, *LTBP1*, *SCN5A*, *TGFBR1*, and *TGFBR3*.

After removing the 9 shared genes, the mean PBS selection index for the Andamanese

among genes annotated to 'cardiocyte differentiation' decreased slightly from 0.444 to 0.443 and the pre- and post-filtration distributions were not significantly different (Kolmogorov-Smirnov $D = 0.023$, $p = 1$). Similarly, the mean PBS selection index for the Batwa for genes in 'cardiac ventricle development' decreased slightly from 0.654 to 0.652, and the distributions were not significantly different ($D = 0.044$, $p = 1$). Finally, for 'cardiac muscle tissue development', the mean PBS selection index for the Andamanese increased from 0.450 to 0.453, and for the Batwa increased from 0.474 to 0.486. Again the pre- and post-filtering distributions were not significantly different for the Andamanese ($D = 0.015$, $p = 1$) or Batwa ($D = 0.015$, $p = 1$).

Similarly, after removing 9 shared genes, the mean PBS selection index for genes annotated to 'growth factor binding' (GO:0019838) for the Batwa increased slightly from 0.437 to 0.440 and for the Andamanese decreased from 0.455 to 0.437. Again, the pairs of distributions were not significantly different (Batwa: $D = 0.030$, $p = 1$; Andamanese: $D = 0.036$, $p = 1$).

2 Figures

Fig. S1: Population Branch Statistic (PBS) schematic.

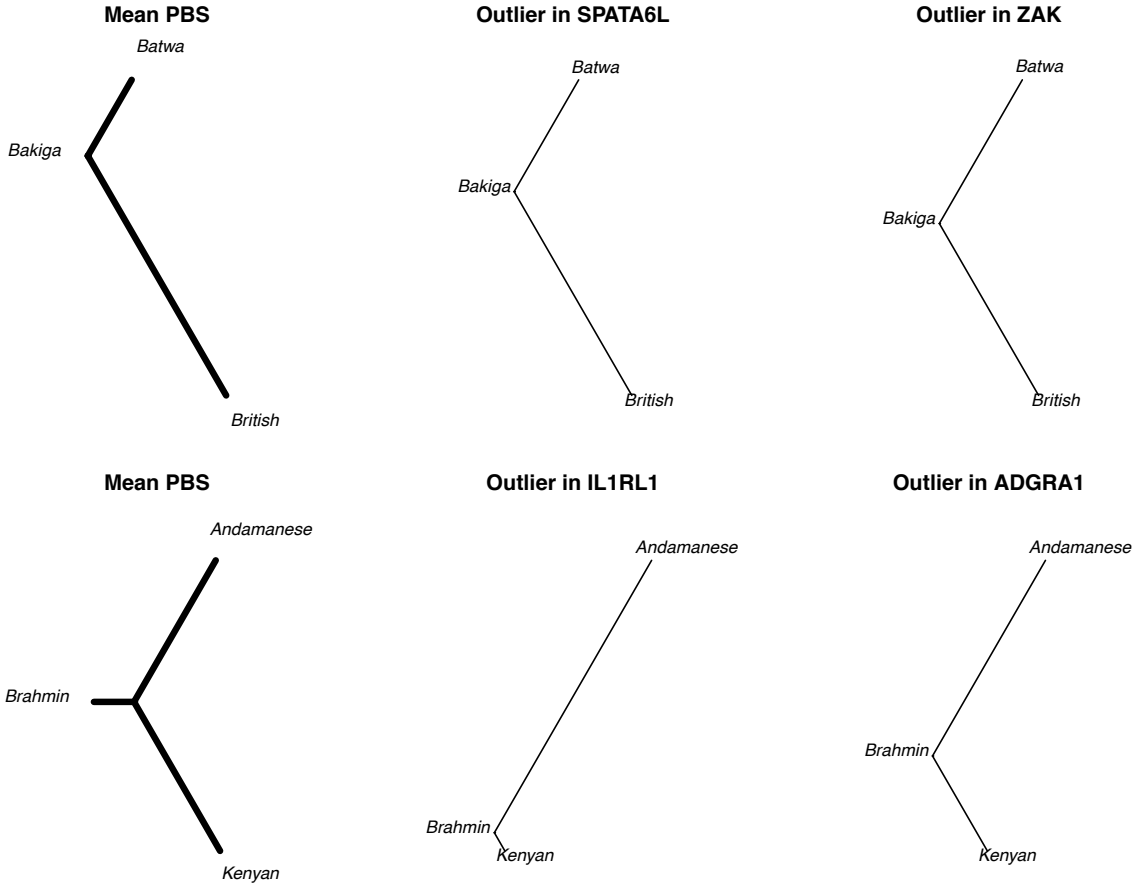

Figure S1: Mean values of the Population Branch Statistic (PBS; left) for the African dataset (Batwa, Bakiga, and outgroup British populations; upper row) and Asian dataset (Andamanese, Brahmin, and outgroup Kenyan populations; lower row). Middle and right columns contain PBS values for two outlier SNPs in each population.

Fig. S2: Population Branch Statistic (PBS) by SNP.

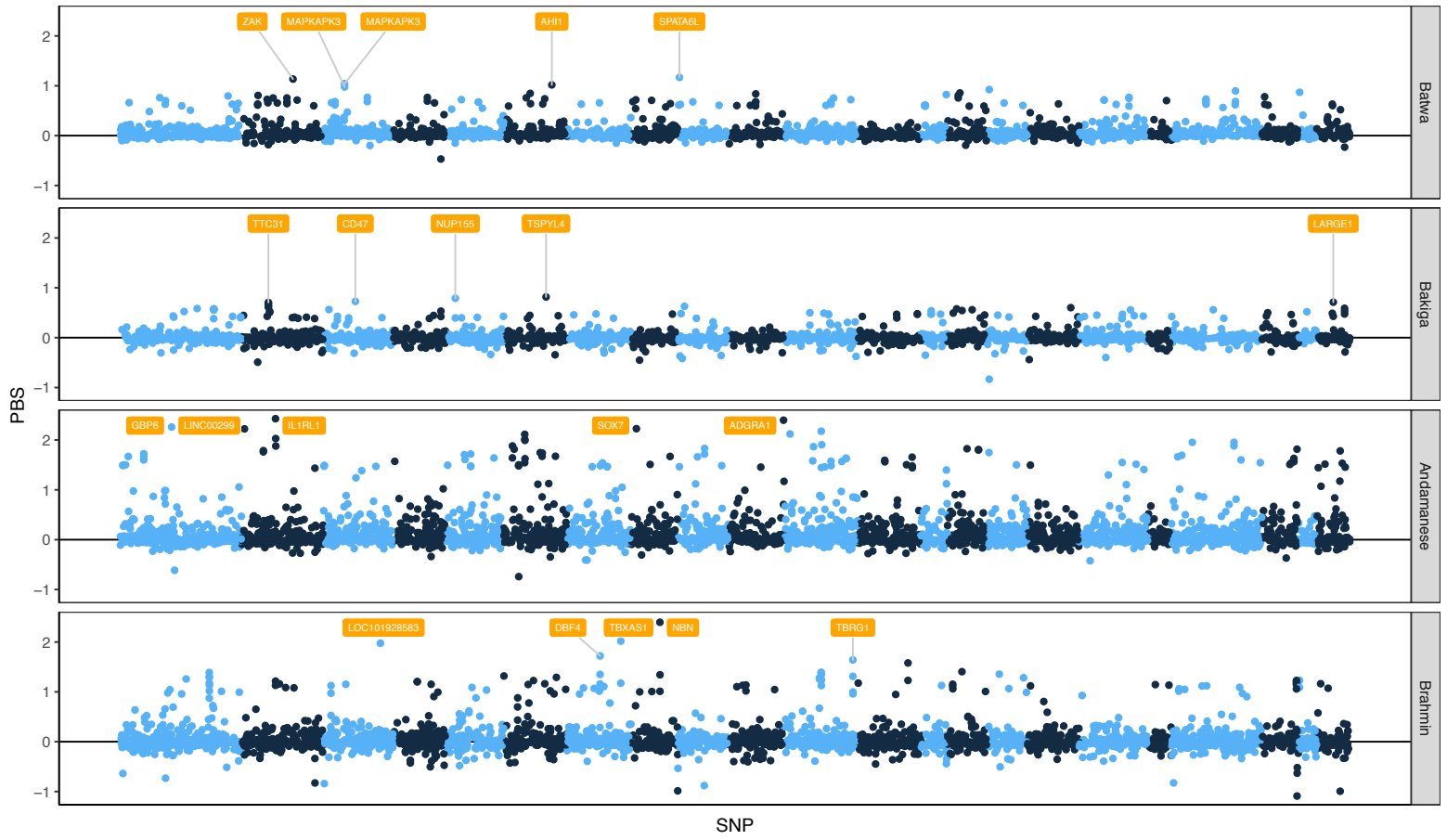

Figure S2: Population Branch Statistic (PBS) values plotted across the genome for the four focal populations. The genes containing the SNPs with the 5 highest PBS values in each population are labeled.

Fig. S3: Population Branch Statistic (PBS) by gene SNP count.

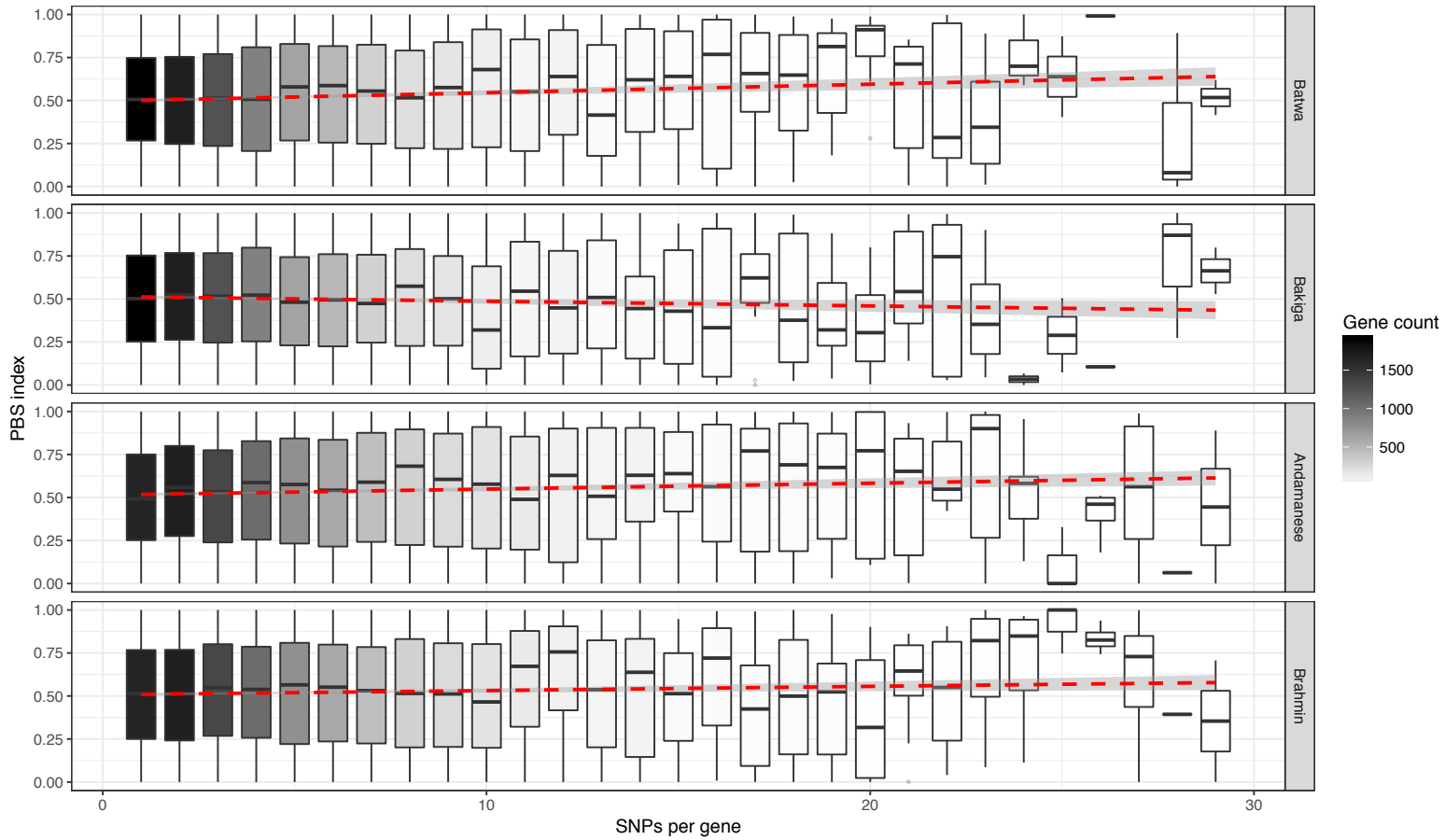

Figure S3: Population Branch Statistic (PBS) selection index values plotted by number of SNPs in gene. Color indicates number of genes with that SNP count. Only SNP counts from 1 to 30 shown.

3 Tables

Table S1: Gene Ontology (GO) biological processes with evidence of convergent enrichment for strong positive selection in the hunter-gatherer populations, as measured by outlier Population Branch Statistic (PBS) values. No molecular functions were found to be convergently enriched. Joint p-values were computed via a permutation-based method, and those with joint empirical $p < 0.005$ are shown.

			Batwa:				*Andamanese:*			
	GO	Joint p	Exp.	Obs.	p	adj. p	Exp.	Obs.	p	adj. p
GO:0048705	skeletal system morphogenesis	0.000	2.20	5	0.069	1.000	2.95	6	0.074	1.000
GO:1901617	organic hydroxy compound biosynthetic processes	0.000	2.28	5	0.077	1.000	3.18	7	0.039	1.000
GO:0007034	vacuolar transport	0.002	1.37	4	0.047	1.000	1.75	4	0.097	1.000
GO:0035107	appendage morphogenesis	0.002	1.69	5	0.027	1.000	2.27	6	0.025	0.901
GO:0035108	limb morphogenesis	0.002	1.69	5	0.027	1.000	2.27	6	0.025	0.901
GO:0048736	appendage development	0.003	1.95	5	0.045	1.000	2.62	6	0.047	1.000
GO:0060173	limb development	0.003	1.95	5	0.045	1.000	2.62	6	0.047	1.000
GO:0030326	embryonic limb morphogenesis	0.004	1.47	4	0.058	1.000	2.01	6	0.015	0.901
GO:0035113	embryonic appendage morphogenesis	0.004	1.47	4	0.058	1.000	2.01	6	0.015	0.901

Table S2: Gene Ontology (GO) biological processes with evidence of population-specific enrichment for strong positive selection in the hunter-gatherer populations, as measured by outlier Population Branch Statistic (PBS) values. Results with $p < 0.01$ shown.

GO		Exp.	Obs.	p	adj. p
Batwa RHG - Biological Processes:					
GO:0007517	muscle organ development	10	4.02	0.0069	0.708
GO:0045926	negative regulation of growth	7	2.48	0.0118	0.708
Andamanese RHG - Biological Processes:					
GO:0006302	double-strand break repair	10	3.14	0.0011	0.171
GO:0070085	glycosylation	12	4.34	0.0013	0.171
GO:0000723	telomere maintenance	8	2.3	0.002	0.175
GO:0033365	protein localization to organelle	25	14.09	0.0036	0.189
GO:1903827	regulation of cellular protein localization	19	9.66	0.0036	0.189
GO:0007569	cell aging	6	1.62	0.0052	0.208
GO:0009101	glycoprotein biosynthetic process	12	5.28	0.0065	0.208
GO:0034613	cellular protein localization	41	27.87	0.0067	0.208
GO:0051179	localization	116	97.3	0.0079	0.208
GO:0060249	anatomical structure homeostasis	13	6.09	0.0079	0.208
GO:0045596	negative regulation of cell differentiation	18	9.79	0.009	0.215
Batwa RHG - Molecular Functions:					
GO:0003723	RNA binding	26	17.66	0.028	0.732
GO:0043167	ion binding	83	70.68	0.034	0.732
Andamanese RHG - Molecular Functions:					
GO:0043130	ubiquitin binding	7	1.89	0.0026	0.177
GO:0008233	peptidase activity	17	9.58	0.0153	0.383

Table S3: Gene Ontology (GO) biological processes (BP) and molecular functions (MF) with evidence of convergent distribution shifts in PBS selection index values in the hunter-gatherer populations. Joint p-values were computed via a permutation-based method, and those with joint empirical $p < 0.1$ are shown.

				Batwa:		*Andamanese:*	
		GO	Joint p	p	adj. p	p	adj. p
BP	GO:0035265	organ growth	0.001	0.028	0.997	0.045	1.000
	GO:1903828	negative regulation of cellular protein localization	0.001	0.036	0.997	0.043	1.000
	GO:0016202	regulation of striated muscle tissue development	0.003	0.014	0.997	0.044	1.000
	GO:0048634	regulation of muscle organ development	0.003	0.013	0.997	0.073	1.000
	GO:0048738	cardiac muscle tissue development	0.003	0.046	0.997	0.003	1.000
	GO:0048863	stem cell differentiation	0.003	0.070	0.997	0.060	1.000
	GO:1901861	regulation of muscle tissue development	0.003	0.014	0.997	0.044	1.000
MF	GO:0019838	growth factor binding	0.000	0.021	0.817	0.027	0.784
	GO:0019199	transmembrane receptor protein kinase activity	0.001	0.027	0.817	0.026	0.784

Table S4: Gene Ontology (GO) biological processes (BP) and molecular functions (MF) with evidence of population-specific distribution shifts in PBS selection index values in the hunter-gatherer populations. No molecular functions were found to be significantly shifted for the Batwa. Results with $p < 0.01$ are shown.

GO		p	adj. p
Batwa RHG - Biological Processes:			
GO:0003231	cardiac ventricle development	0.001	0.302
GO:0061351	neural precursor cell proliferation	0.007	0.348
GO:0034976	response to endoplasmic reticulum stress	0.009	0.348
Andamanese RHG - Biological Processes:			
GO:0016579	protein deubiquitination	0.001	0.232
GO:0035051	cardiocyte differentiation	0.002	0.232
GO:0048738	cardiac muscle tissue development	0.003	0.232
GO:1901800	positive regulation of proteasomal protein catabolic process	0.004	0.262
GO:0006508	proteolysis	0.009	0.453
Andamanese RHG - Molecular Functions:			
GO:0005085	guanyl-nucleotide exchange factor activity	0.005	0.278

References

[1] Blake JA, Eppig JT, Kadin JA, et al. (39 co-authors). 2017. Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse. *Nucleic Acids Research*. 45:D723–D729.

[2] Brown KR, Jurisica I. 2005. Online predicted human interaction database. *Bioinformatics*. 21:2076–2082.

[3] Lui JC, Nilsson O, Chan Y, Palmer CD, Andrade AC, Hirschhorn JN, Baron J. 2012. Synthesizing genome-wide association studies and expression microarray reveals novel genes that act in the human growth plate to modulate height. *Human Molecular Genetics*. 21:5193–5201.

[4] Perry GH, Foll M, Grenier JC, et al. (14 co-authors). 2014. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences*. 111:E3596–603.

[5] Wood AR, Esko T, Yang J, et al. (445 co-authors). 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*. 46:1173–1186.