# Adaptive Evolution of UGT2B17 Copy-Number Variation

Yali Xue, Donglin Sun, Allan Daly, Fengtang Yang, Xue Zhou, Mengyao Zhao, Ni Huang, Tatiana Zerjal, Charles Lee, Nigel P Carter, et al.

# Adaptive Evolution of *UGT2B17* Copy-Number Variation

Yali Xue,[1] Donglin Sun,[1,4] Allan Daly,[1] Fengtang Yang,[1] Xue Zhou,[1,5] Mengyao Zhao,[1] Ni Huang,[1] Tatiana Zerjal,[1,6] Charles Lee,[2,3] Nigel P. Carter,[1] Matthew E. Hurles,[1] and Chris Tyler-Smith[1,*]

The human *UGT2B17* gene varies in copy number from zero to two per individual and also differs in mean number between populations from Africa, Europe, and East Asia. We show that such a high degree of geographical variation is unusual and investigate its evolutionary history. This required first reinterpreting the reference sequence in this region of the genome, which is misassembled from the two different alleles separated by an artifactual gap. A corrected assembly identifies the polymorphism as a 117 kb deletion arising by nonallelic homologous recombination between ~4.9 kb segmental duplications and allows the deletion breakpoint to be identified. We resequenced ~12 kb of DNA spanning the breakpoint in 91 humans from three HapMap and one extended HapMap populations and one chimpanzee. Diversity was unusually high and the time to the most recent common ancestor was estimated at ~2.4 or ~3.0 million years by two different methods, with evidence of balancing selection in Europe. In contrast, diversity was low in East Asia where a single haplotype predominated, suggesting positive selection for the deletion in this part of the world.

## Introduction

Copy-number variation (CNV)—the existence of segments of DNA longer than 1 kb with >90% sequence identity that differ in the numbers of copies between the genomes of different individuals—is now recognized as a major source of variation in the human genome.[1] It affects more nucleotides per genome than SNP variation[2] and contributes significantly to variation among normal individuals, both in levels of gene expression[3] and in phenotypes of medical relevance.[4,5] There are also substantial copy-number differences between humans and great apes[6] and in at least some cases, these appear to result from positive selection acting on genes within the differential segments;[7] CNV has therefore contributed to evolutionary changes over the last few millions of years and seems likely to have played a part in more recent evolution as well, but its impact remains poorly understood. Some CNVs show unusually high levels of population differentiation,[8,9] often but not always[10] a sign of positive selection within a subset of the populations, but definitive evidence of the influence of natural selection has been lacking. It is therefore necessary to evaluate further the contribution of CNV to the evolutionarily important variation found in humans.

The number of documented CNVs has increased substantially over the last few years and a recent worldwide survey reported 1447 CNVs in the 270 individuals from the four HapMap populations.[2] In order to determine whether a particular CNV has spread because of an evolutionary advantage or neutral genetic drift, we can investigate whether it carries the signature of positive selection, using the approaches developed for detecting selection at SNPs.[11] In addition to high levels of population differentiation, these include unusually long haplotypes, skewed allele frequency spectra, and an excess of functional changes and are best assessed by resequencing the relevant segment of DNA in multiple individuals.[11] However, in order to apply these methods to CNVs, we need to take into account some additional factors. CNVs tend to be associated with both segmental duplications (SDs) and gaps in the current genome assembly;[2] furthermore, a CNV detected by comparative genomic hybridization may have arisen by more than one mutational event.

We set out to evaluate the feasibility of applying standard methods for detecting selection to CNVs, choosing as an initial candidate from the worldwide survey[2] the CNV that involved the UDP-glucuronosyltransferase 2B17 (*UGT2B17*) gene (MIM *601903). Three considerations led to this choice. First, it has strong biomedical interest, making it a plausible target for natural selection. The UGT2B17 enzyme metabolizes steroid hormones (including testosterone) and a large number of xenobiotics, so the CNV is associated with variation in urine testosterone level,[12] male insulin sensitivity, fat mass,[13] and according to some[14,15] but not all[16,17] studies, prostate-cancer risk. These phenotypic effects are likely to be mediated by variation in expression level, which, in lymphoblastoid cells at least, is proportional to gene dosage.[18] Second, it lies in a complex genomic region showing both a >200 kb segmental duplication and an assembly gap (e.g., see UCSC Genome Browser chr4: 69,000,000–69,700,000), so would provide insights into some of the complexities of analyzing CNV structures. Third, it nevertheless promised to be both evolutionarily interesting and tractable. It showed large frequency[2,19] and expression[20] differences between

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, United Kingdom; [2]Department of Pathology, Brigham and Women's Hospital, Boston, MA, 02115, USA; [3]Harvard Medical School, Boston, MA 02115, USA
[4]Present address: Laboratory of Medical Genetics, Harbin Medical University, Harbin 150081, China
[5]Present address: Department of Children's and Adolescent Health, Public Health College of Harbin Medical University, Harbin 150086, China
[6]Present address: Station de Génétique Végétale, Ferme du Moulon, Gif-sur-Yvette, France
*Correspondence: cts@sanger.ac.uk
DOI 10.1016/j.ajhg.2008.08.004. 

populations, could be genotyped by CGH[2] or a simple PCR assay[19] and was tagged by the SNP rs3100645,[20] suggesting a single predominant pair of haplotypes. We have therefore undertaken an evolutionary analysis of this CNV.

## Material and Methods

### DNA Samples

DNAs were from the HapMap and extended HapMap populations[21] and HGDP-CEPH diversity panel,[22] in which the H952 subset[23] was used. The samples sequenced were 23 Yoruba from Ibadan, Nigeria (YRI), 23 Chinese Han from Beijing (CHB), 22 CEPH Utah residents with ancestry from northern and western Europe (CEU), and 23 Luhya from Webuye, Kenya (LWK). DNAs were purchased from the Coriell Institute for Medical Research (Camden, NJ, USA). In addition, one chimpanzee (*Pan troglodytes*) sample from the ECACC (Salisbury, Wiltshire, UK) was included as an outgroup.

### Data Generation

*Genotyping of the* UGT2B17 *Deletion*

The C and J primer pairs[19] were combined into a duplex reaction generating a 316 bp fragment from within the *UGT2B17* gene and 884 bp fragment spanning the deletion breakpoint, respectively. Individuals homozygous for the presence of *UGT2B17* showed a single 316 bp band, individuals homozygous for the deletion showed a single 884 bp band, and heterozygotes showed both bands. PCRs were carried out in a 20 μl reaction volume containing 1× Invitrogen PCR buffer with 4 mM $MgCl_2$, 200 μM of each dNTP, 0.6 U Platinum *Taq* DNA polymerase, 0.5 μM C primers, 1.5 μM J primers, and 60 ng genomic or whole-genome-amplified DNA. The cycle conditions were 95°C for 15 min, 15 cycles of 95°C for 30 s, 62°C for 30 s decreasing by 0.5°C each cycle, 72°C for 30 s, and then 20 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s, finishing with extension at 72°C for 5 min. Products were analyzed by electrophoresis on a 1.5% agarose gel containing ethidium bromide.

*Resequencing*

Proximal and distal fragments were amplified by long PCR with one primer within the SD and one in the flanking unique sequence (Table S1 available online). Reactions (15 μl) contained 1 × Invitrogen PCR buffer, 2 mM $MgSO_4$, 200 μM each dNTP, 0.6 U Platinum *Taq* DNA polymerase High Fidelity (Invitrogen, Paisley, UK) 0.4 μM each primer, and 125 ng genomic DNA. A touchdown protocol was used beginning with 2 min denaturation at 94°C, followed by 15 cycles of 94°C for 30 s, 68°C for 30 s (decreasing by 0.5°C each cycle), and 68°C for 6 min, followed by 20 cycles of 94°C for 30 s, 58°C for 30 s, and 68°C for 6 min, and finishing with extension at 68°C for 7 min. Nested PCR products of ~500–700 bp length overlapping by ~200–400 bp were then amplified with the primers in Table S1; each 15 μl PCR contained 0.5 μl of 400 × diluted long PCR products, 0.5 U Platinum Taq (Invitrogen), 1 × buffer (Invitrogen), 1.6 mM $MgCl_2$, and 10 pmol of each primer, 200 μM of each dNTP, and the cycle conditions were 94°C for 6 min, 35 cycles of 94°C for 45 s, 57°C–60°C for 45 s, 72°C for 1 min 30 s, then 72°C for 3 min. Products were sequenced on both strands by the Sanger Faculty Small Sequencing Projects Group with AB3730 capillary sequencing technology. Potential variable positions were flagged by the Mutation Surveyor v.2.0 software (SoftGenetics, PA, USA) and checked manually. Four blind duplicates were included for quality control and showed complete concordance. Only one HapMap SNP (rs3100645) was present in the sequenced region, and it also showed complete concordance with the sequencing results from the YRI, CHB, and CEU samples.

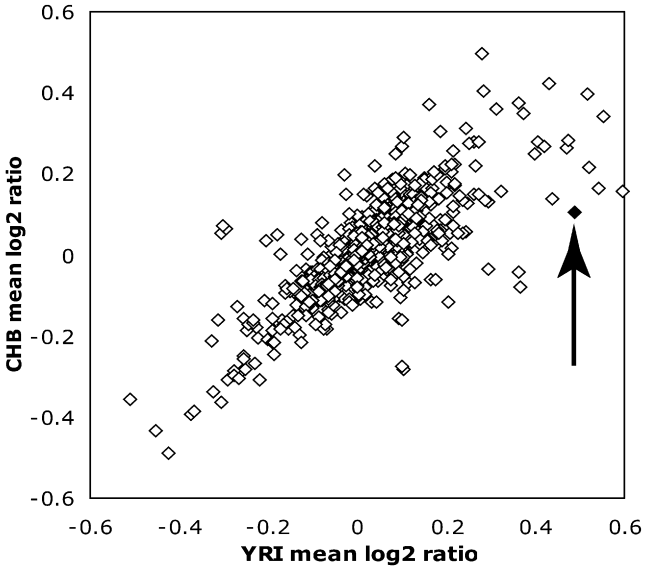
Fiber-FISH was carried out as described previously[24] with the BAC clones Chr4tp-17G4 and Chr4tp-1B5 to hybridize to GM19203 preparations, homozygous for the presence of the *UGT2B17* gene.

### Statistical Analyses

Regions chosen from the reference sequence were compared with Dotter.[25] $F_{ST}$ was calculated with the R package HIERFSTAT.[26] LD was evaluated with Haploview[27] and haplotypes were inferred with PHASE 2.1.[28,29] Median-joining networks[30] were constructed from the inferred haplotypes with Network 4.15. Coalescent simulations were performed with the program "ms"[31] incorporating best-fit demographic parameters for each population[32] without recombination because our analyses concentrated on regions of high LD and were converted to demographic settings as previously described (Supplementary Table 4 of Helgason et al.[33]). Custom C codes were used to evaluate the summary statistics Tajima's *D*,[34] Fu and Li's *D*, *D**, *F* and *F**,[35] Fay and Wu's *H*,[36] and Fu's *Fs*.[37] Time to the most recent common ancestor (TMRCA) was estimated with GENETREE.[38] This program requires infinite-sites compatible data, and so we first ran PRUNE (kindly provided by R.C. Griffiths, Oxford) on the ~3.0 kb unique region (described below) and needed to remove only one haplotype in order to produce a gene tree. We then estimated theta with a model of three populations (African [YRI+LKW], European [CEU], and Asian [CHB]) using 10,000 simulations to be 2.055 (Ne = 12,400 with chimpanzee-human split 6.5 million years ago). Finally, with this value of theta and the populations connected by the migration rates suggested by the best-fit demographic model,[32] we estimated the TMRCA by using ten runs each of 10 million simulations and chose the run with the lowest standard deviation, as recommended.[38]

## Results

Because the main criterion used thus far for identifying evolutionarily interesting CNVs has been an unusually high degree of population differentiation,[8,9] we first evaluated the *UGT2B17* CNV in the same way. Reanalysis of the WGTP CGH data[2] showed that *UGT2B17* was indeed an extreme outlier (Figure 1A). Assessment with the residuals from a least-squares regression analysis, the basis for recognizing positive selection at the *AMY1* locus,[9] ranked *UGT2B17* as the fifth most extreme outlier out of 2404 copy-number-variable clones ($p < 0.002$; Figure 1B). The *UGT2B17* CNV is a binary polymorphism, so it can also be compared with SNP data. Because it was chosen from a CNV-discovery experiment, we compared it with SNPs from SNP discovery (i.e., resequencing) experiments such as analyses of the neutral ENCODE regions[39] in the same populations. Mutational mechanisms differ between CNVs and SNPs, but evidence presented below suggests that the *UGT2B17* CNV has a single origin like most



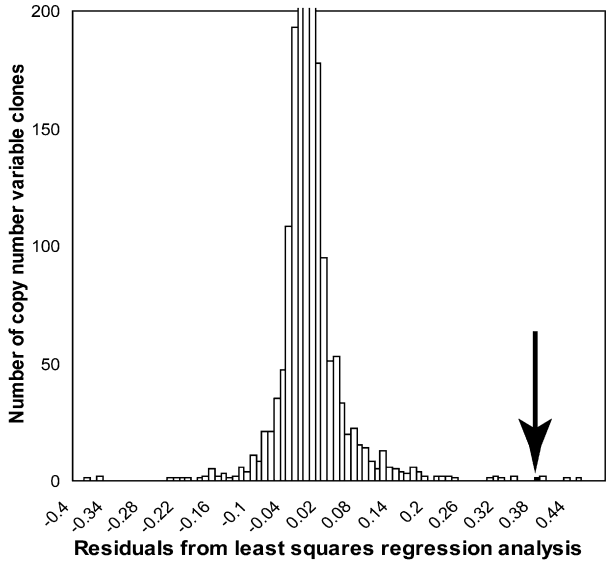




**Figure 1. The BAC Clone Containing *UGT2B17* Shows Unusually High Population Differentiation in CGH Experiments**
The mean log2 ratios of the copy-number-variable BAC clones in the YRI and CHB were compared and the *UGT2B17*-containing BAC stood out (filled symbol, arrow, top) and was an outlier when residuals from a regression analysis were examined (arrow, bottom).

SNPs, and recurrent CNV mutation would be likely to reduce $F_{ST}$, making our comparison conservative. The observed *UGT2B17* YRI-CHB $F_{ST}$ value was 0.583, and the frequency of the deleted allele in the combined YRI+CHB sample was 0.468. In the ENCODE neutral regions, there were 1036 SNPs with derived allele frequencies in this range (0.368–0.568), and 95% of them showed $F_{ST}$ values below 0.577, indicating that the observed *UGT2B17* value is significantly higher than expected if a single-tailed test and 5% significance value are used. A similar result was found with the Perlegen data[40] in which 95% of 4864 SNPs with derived allele frequencies 0.466–0.470 in the African American + Chinese American population pair showed $F_{ST}$ values $< 0.561$. However, if a two-tailed test were applied, the observed $F_{ST}$ value would not be significant in either comparison. Nevertheless, these results together seemed to indicate an unusual level of population differentiation, and we wished to investigate further the possible underlying selective pressures. The most powerful information would be obtained by resequencing a segment of DNA containing the CNV. This required understanding its structure in sufficient detail to choose the appropriate region.

### Structure and Distribution of the *UGT2B17* CNV

In the current assembly of the human reference sequence (build 36), *UGT2B17* lies in a large duplicated region consisting of ~367 kb and ~250 kb segments: The two SDs are separated by a gap and the proximal one carries a copy of *UGT2B17* with its flanking sequences, whereas the distal one does not, thereby accounting for the difference in size (Figure 2). Dotter analysis showed that the segmental duplication boundaries correspond precisely to the location of the gap in the assembly (results not shown), so we used fiber-FISH to investigate the possibility that this large duplication might represent an artifact created by misassembly. The results (Figure 2) show that a region that is predicted to be duplicated by the reference sequence is actually present only once. We therefore conclude that the reference sequence is misassembled from the two distinct alleles and, consequently, that neither the large duplication nor the gap are true features of the region. There is, however, a ~4.9 kb SD flanking the 117,339 bp region containing the *UGT2B17* gene. Because a *UGT2B17* gene is present in the chimpanzee,[41] the ancestral state is likely to be the presence of the gene, and the derived state is deduced to be the deletion. Comparison of the two copies of the SD flanking the *UGT2B17* gene with the single copy carried by the deleted allele suggests that the deletion resulted from nonallelic homologous recombination between the two ~4.9 kb repeats at a position ~3.1–3.2 kb into the SD.

A PCR assay has already been described for the presence or absence of the *UGT2B17* gene,[19] and we tested the HapMap samples using a modified version of this assay in order to compare the results with the array-CGH genotypes.[2] We observed perfect concordance, demonstrating both the reliability of the two assays and the sharing of a single structure by all the deletion alleles in this set, most likely reflecting a single origin. This is consistent with the small size of the SDs and their limited sequence identity (~95%), which would not promote nonallelic homologous recombination very efficiently.

We then used the PCR assay to test a larger number of populations representing worldwide diversity provided in the HGDP-CEPH diversity panel.[22] In agreement with previous, more limited surveys,[12,19] we observed large differences in frequency between populations (Figure 3). The

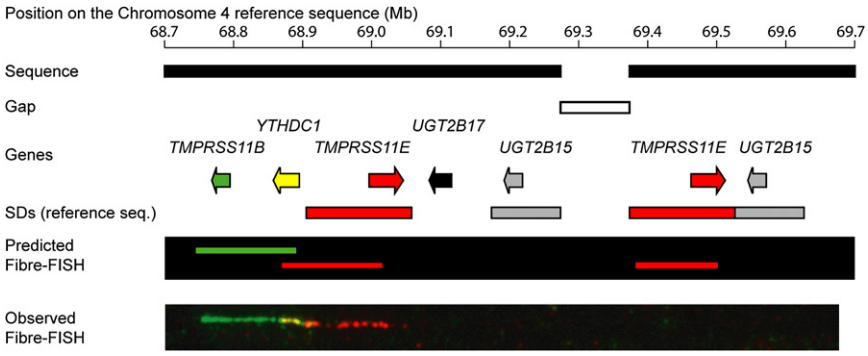



**Figure 2. Reinterpretation of the Reference Sequence surrounding *UGT2B17***
The reference sequence in this region contains a large gap (white bar) flanked by large segmental duplications (red and gray bars) carrying copies of the *TMPRSS11E* and *UGT2B15* genes. The *UGT2B17* gene lines in a 117 kb insertion into the left segmental duplication. This structure predicts that a BAC clone within the red segmental duplication (Chr4tp-1B5) should show two signals in a fiber-FISH experiment, but only one was ever observed. Consequently, we suggest that the large segmental duplication does not exist and the reference structure represents a misassembly of an allele containing (left) and lacking (right) *UGT2B17*.

deletion was in general rare in African and European populations and common in East Asian populations; for population samples more than ten, it ranged from 14% in the Yoruba from Nigeria to 92% in the Japanese; both of these samples are independent of the HapMap samples from the same populations but show similar frequencies. Interestingly, however, the deletion frequency was high in the San sample, hunter-gatherers from southern Africa. If confirmed in a larger sample, this finding would point to an unusual distribution in this population, meriting further investigation. These differences led to a worldwide $F_{ST}$ value between 0.171 and 0.198, depending on the grouping scheme (Table 1). This is considerably lower than the value obtained above in which two of the most extreme populations were used, as expected. To evaluate its significance, we again used an empirical test. We extracted the 1351 SNPs with average minor allele frequencies per population between 0.4 and 0.5 (matching 0.42–0.47 for *UGT2B17*, depending on the grouping) from the HGDP-CEPH data[42] and found that for any grouping scheme, 97.5% of the $F_{ST}$ values lay below 0.054 (Table 1). Thus *UGT2B17* shows exceptionally high population differentiation.

## Neutrality Tests

With an understanding of the structure of the region and the deletion breakpoints, we could amplify segments of DNA immediately flanking the deletion (6.4 kb proximal and 5.8 kb distal) and resequence them in a set of 91 individuals from four populations. The distal region showed low LD, both across its length and with respect to the *UGT2B17* CNV (Figure 4), and so the analyses presented below concentrated on the proximal region, where LD was high across the entire 6.4 kb. A total of 102 polymorphic sites were found, and the nucleotide diversity was $41 \times 10^{-4}$, more than five times the average for chromosome 4,[43] and far above the highest value encountered among the 308 genes resequenced by the SeattleSNPs project, $30 \times 10^{-4}$ for *ABO* (Table 2). Unusually, the highest value was in the CEU ($45 \times 10^{-4}$); the lowest was in the CHB ($13 \times 10^{-4}$) with the two African populations in between. Because the genes included in the SeattleSNPs project might be subject to purifying selection and thus be unrepresentative of a noncoding region such as the one under study, we also compared our result with the diversity of 5 kb segments from the largely noncoding ENCODE region ENr321. It was also far higher than the most diverse of

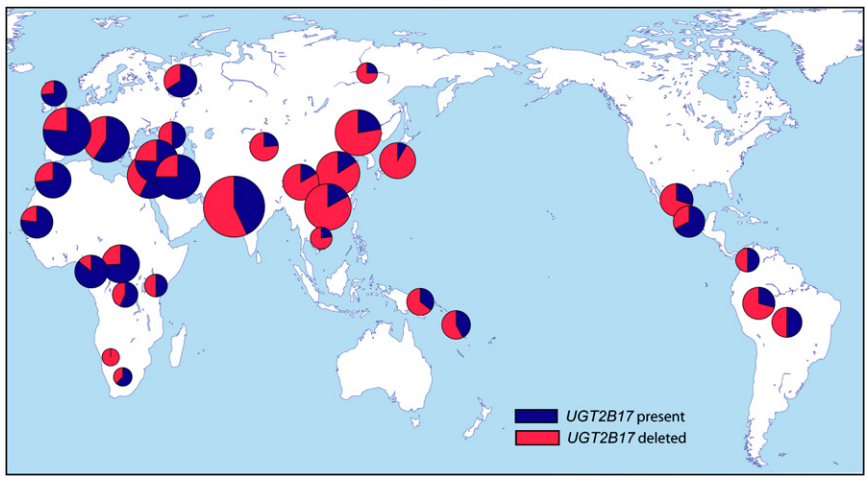



**Figure 3. Distribution of the *UGT2B17* CNV Frequency in the HGDP-CEPH Population Samples**
Note the high frequency of the gene in most African populations, intermediate frequency in Europe/West Asia, and low frequency in East Asia.

**Table 1. $F_{ST}$ Values of *UGT2B17* and Common SNPs in the HGDP-CEPH Panel**

| | $F_{ST}$ | | |
|---|---|---|---|
| | | Common SNPs[42] | |
| Grouping Scheme | *UGT2B17* | 95$^{th}$ Percentile | 97.5$^{th}$ Percentile |
| 52 populations (as HGDP) | 0.196 | 0.047 | 0.051 |
| 32 populations (as Figure 3) | 0.198 | 0.048 | 0.054 |
| 7 geographic regions[a] | 0.171 | 0.042 | 0.049 |
| 5 genetic regions (as Rosenberg et al.[56]) | 0.179 | 0.045 | 0.053 |

[a] Sub-Saharan Africa, North Africa, Middle East, Asia, Oceania, Europe, and America.

these ($16 \times 10^{-4}$). Tajima's *D* for the worldwide set was +1.42, higher than all but two of the worldwide values for the SeattleSNPs genes, whereas individual populations showed very different values: from +3.13 in the CEU (higher than any CEPH SeattleSNPs gene) and +1.60 in the YRI (higher than all but one African American SeattleSNPs gene) to −1.48 in the CHB (which has no equivalent population in SeattleSNPs, but was significantly lower than expected under neutrality). All values of Fu and Li's *D*, *D**, *F*, and *F** were positive and significantly so in many of the individual populations, particularly in the CEU; Fay and Wu's *H* was significantly negative in the CHB but did not differ from neutral expectation in the other populations. Note that these assessments of significance incorporate the appropriate Schaffner et al.'s[32] best-fit model of the demography of each population. These models include a moderate African bottleneck, a more severe out-of-Africa bottleneck shared by Asians and Europeans, and intermediate later Asian-specific and European-specific bottlenecks, together with a large postindustrial expansion. Taken at face value, these summary statistics suggested a highly unusual evolutionary history for the region, and one that differed between populations. The negative Tajima's *D* and Fay and Wu's *H* in the CHB, for example, could indicate positive selection, whereas the significantly positive Tajima's *D* in the CEU could indicate balancing selection, although before reaching such conclusions other explanations still have to be evaluated.

In particular, we needed to consider another potentially confounding factor: the segmentally duplicated nature of part of the region resequenced and thus the possibility that gene conversion might have contributed to its complex history. Two approaches could be considered: We might model a neutral evolutionary history incorporating gene-conversion events and compare our results with its predictions; however, such models are not well developed. Alternatively, we might focus on the variants exempt from gene conversion and compare their analysis with that of the whole region. To explore this second option, we first visualized the relationships of the inferred haplotypes by using a median-joining network. The network (Figures 5A and 5B) showed low levels of reticulation, reflecting the high LD throughout the 6.4 kb region and several striking and unusual features: (1) Haplotypes were spread over a large area, reflecting the unusually high diversity of the region. (2) Strong clustering was evident, with the

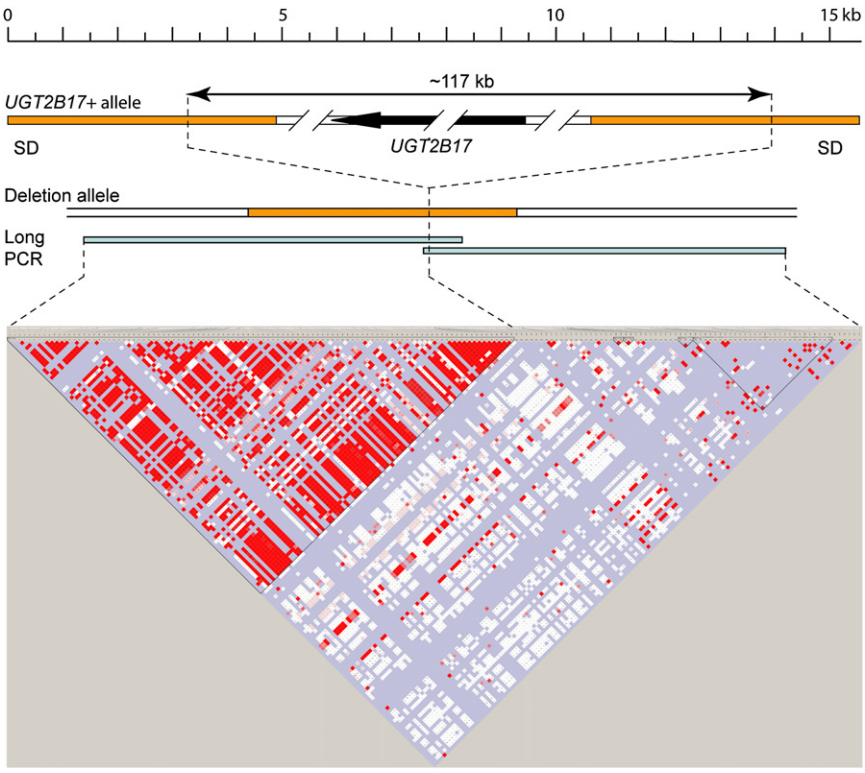



**Figure 4. Sequence Analysis of the *UGT2B17* Deletion Breakpoint**
The *UGT2B17* gene is flanked by two ~4.9 kb segmental duplications (SDs, orange) ~117 kb apart, and nonallelic homologous recombination between them led to deletion of the gene. Long PCR products were designed for amplification of the adjacent sequences, with one primer in the SD and one in the flanking unique sequence. The pattern of LD found after resequencing the amplified region in 91 individuals is shown at the bottom. Note the high proximal LD (left, predominant red color) and low distal LD (right).

**Table 2. Summary Statistics**

| | Sample Characteristics | | | | Allele Frequency Distribution Tests | | | | | | Haplotype Test |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sample Size | Polymorphic Sites | Nucleotide Diversity $\times 10^4$ | ThetaW $\times 10^3$ | Tajima's *D* | Fu and Li's *D* | Fu and Li's *D** | Fu and Li's *F* | Fu and Li's *F** | Fay and Wu's *H* (p value) | Fu's *Fs* |
| **Whole Region, All SNPs** | | | | | | | | | | | |
| All populations | 182 | 102 | *40.9* | 2.81 | *1.42* | 1.40 | 1.32 | 1.69 | 1.63 | −8.39 (0.12) | −0.42 |
| YRI | 46 | 78 | *41.0* | 2.83 | *1.60*** | 1.28** | 1.14* | 1.72* | 1.56** | −5.22 (0.15) | 3.72* |
| LWK | 46 | 86 | *35.5* | 3.12 | 0.49 | 1.64** | 1.44** | 1.44* | 1.31* | −16.32 (0.07) | 1.47 |
| CEU | 44 | 65 | *44.7* | 2.38 | *3.13*** | 1.97** | 1.73** | 2.95* | 2.66** | 1.96 (0.41) | 7.92* |
| CHB | 46 | 62 | 13.1 | 2.25 | −1.48* | 1.79** | 1.58* | 0.62 | 0.56 | −48.47 (0.00)** | 5.66 |
| Deleted, all | 90 | 34 | 5.7 | 1.07 | | | | | | | |
| Deleted, African | 32 | 27 | 8.0 | 1.07 | | | | | | | |
| Deleted, non-African | 56 | 16 | 3.5 | 0.55 | | | | | | | |
| Deleted, CHB | 43 | 9 | 2.5 | 0.33 | | | | | | | |
| **Unique Region, 3 kb** | | | | | | | | | | | |
| All populations | 182 | 27 | *16.6* | 1.56 | 0.19 | −0.57 | −0.54 | −0.30 | −0.29 | −2.73 (0.06) | 0.44 |
| YRI | 46 | 17 | *17.9* | 1.29 | 1.22* | 0.39 | 0.39 | 0.82 | 0.79 | −3.06 (0.05)* | 2.76* |
| LWK | 46 | 19 | 14.0 | 1.44 | −0.09 | 1.79** | 1.66** | 1.34* | 1.26** | −6.45 (0.02)* | 1.56 |
| CEU | 44 | 12 | *18.2* | 0.92 | *2.97*** | 1.56** | 1.48 | 2.44* | 2.32** | 0.42 (0.41) | 9.71** |
| CHB | 46 | 15 | 5.7 | 1.14 | −1.56* | 0.20 | 0.22 | −0.50 | −0.45 | −9.09 (0.01)** | 0.48 |
| Deleted, all | 90 | 6 | 1.9 | 0.39 | | | | | | | |
| Deleted, African | 32 | 4 | 2.9 | 0.33 | | | | | | | |
| Deleted, non-African | 56 | 4 | 1.0 | 0.22 | | | | | | | |
| Deleted, CHB | 43 | 4 | 1.0 | 0.23 | | | | | | | |

Italics are used to denote $p < 0.05$ by comparison to SeattleSNPs empirical data (only for All, YRI, LWK, and CEU populations; nucleotide diversity and Tajima's *D*).
*$p < 0.05$; **$p < 0.01$ by comparison to best-fit demographic model.

undeleted chromosomes falling into four distinct clusters separated by long branches; all but one of the deleted chromosomes lay in two additional clusters, close to one another. (3) Within some of the clusters, single haplotypes were present at high frequency, most notably within the deleted chromosomes in which a single haplotype made up 42/64 (66%) of one of the clusters. These common haplotypes were each present mainly in a subset of the populations.

Gene conversion typically involves segments of 200 bp to 1 kb in humans,[44] and we next identified candidate historical conversion events as those leading to sets of SNPs that were both physically clustered in the genome and phylogenetically clustered in the network. A total of 45 SNPs met these criteria, so we repeated the summary statistic calculations (results not shown) and network analyses omitting them. The unusual characteristics of high diversity in most populations, strongly positive Tajima's *D* (and other statistics) in the CEU and significantly negative Tajima's *D* and Fay and Wu's *H* in the CHB, and network clusters remained, except that most deleted chromosomes fell into a single cluster (Figures 5C and 5D). We finally restricted the analyses to the proximal ~3 kb of unique sequence lying outside the SD and thus having no potential for gene conversion with a diverged paralog. Again, the unusual characteristics remained: Diversity was higher than in 95% of the SeattleSNPs genes in the CEU and YRI, Tajima's *D* was strongly positive (+2.97) in the CEU and most other statistics were also significantly positive in this population, whereas Tajima's *D* and Fay and Wu's *H* were significantly negative in the CHB (and the other statistics were not significantly different from neutral expectation; see Table 2). Networks showed two main clusters separated by a long branch with all deleted and some nondeleted chromosomes in one cluster and the rest of the nondeleted chromosomes in the other cluster (Figures 5E and 5F).

We therefore conclude that the unusual characteristics identified by neutrality tests and network analyses are a property of the entire 6.4 kb of DNA examined and may reflect unusual selective events, differing between populations, rather than gene-conversion events.

Finally, we wished to understand the time depth of the variation in this region of the genome. Concentrating on the 3.0 kb unique region, we calculated the mean number of mutations from the root to the 182 human chromosomes by using the network in Figure 5E as 6.29 ($\rho$ statistic). There were 27 fixed differences between

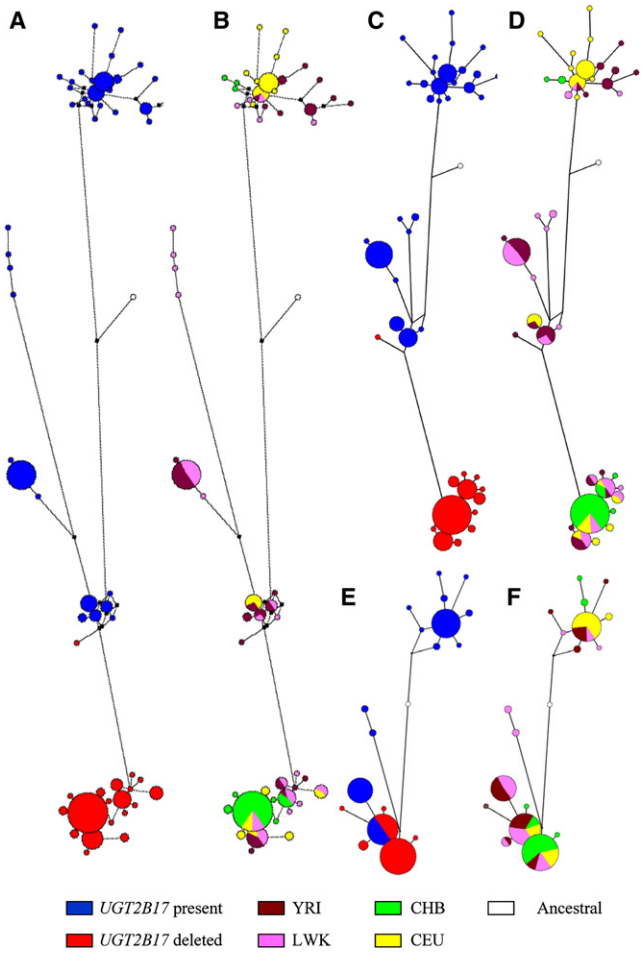



**Figure 5. Network Analysis of *UGT2B17* Inferred Haplotypes**
(A and B) A 6.4 kb region, all SNPs.
(C and D) A 6.4 kb region omitting gene conversion SNPs.
(E and F) A 3.0 kb unique region, all SNPs. Circles represent haplotypes with an area proportional to frequency. In each pair of networks, the shortest line represents on mutational step. "Ancestral" (white circle) is a reconstructed haplotype carrying the ancestral (chimpanzee) allele at each position.

chimpanzee and human in this DNA segment, so assuming that 13.5 of these had occurred on the human lineage during 6.5 million years since the chimpanzee-human common ancestor, we estimated a TMRCA of 3.0 ± 0.9 million years for the human chromosomes, an unusually ancient time. In contrast, the TMRCA for the deleted chromosomes calculated in the same way was much more recent, 180,000 ± 140,000 years. Using an alternative approach, we estimated the TMRCA of the same 3.0 kb unique region from 181 chromosomes with GENETREE at 2.4 ± 0.4 million years.

## Discussion

This work raises a number of issues concerning the technical analysis of CNVs, the biological significance of the *UGT2B17* deletion in particular, and the likely evolutionary significance of CNV in general.

Resequencing in population samples provides a "gold standard" for investigating the evolutionary history of a region of DNA because all of the variation is identified in an unbiased way. It has not previously been applied to CNVs, perhaps because their importance has only recently been appreciated and their complex structures have hindered analysis. Our study illustrates both the care needed to identify the appropriate region to resequence and how resequencing can be used fruitfully. Indeed, the reinterpretation of the reference assembly and closure of one of the remaining gaps in the human genome suggest that all remaining gaps should be reinvestigated in this light. In the discussion below, we concentrate mainly on the conclusions derived from the ~3.0 kb unique region because these are the most conservative; similar but more extreme conclusions would be derived from the full ~6.4 kb region, although they would potentially be complicated by gene conversion within the SD.

Our results point strongly to an unusual evolutionary history not accounted for by genetic drift affecting *UGT2B17* copy-number variation in both Europe and East Asia, with indications of some atypical properties in African populations as well. Diversity in the worldwide sample was very high (Table 2), and this was not due to a high mutation rate, but instead corresponded to a TMRCA of 2.4 million years or more, well above the expectation of approximately one million years for an autosomal segment[45] or 95$^{th}$ percentile of 2.1 million years in the most likely model of human evolution found by a recent study,[46] and comparable to the 2.65 million years estimated for the *ABO* gene tree.[47] The high diversity reflected the presence of two very distinct clusters of haplotypes, with deleted chromosomes confined to one cluster and forming its majority. Both clusters were found in all four populations, but at different frequencies. In the CEU, the two clusters were similar in frequency, leading to some of the highest diversity values reported from European populations and correspondingly high positive values of allele frequency distribution test statistics (Table 2). In the CHB, in contrast, the deleted chromosomes predominated and a single haplotype made up 37/46 (80%) of the chromosomes in the sample. Diversity was therefore low in this population, and allele frequency-based test statistics was negative. The two African populations showed characteristics that were intermediate in several respects, although the nondeleted chromosomes were in the majority. The maintenance of two divergent haplotype clusters for >2.4 million years is most simply interpreted as resulting from the action of long-term balancing selection.[48] Ancient balancing selection is rare in humans;[49] classical examples are found in the *HLA* and *ABO* genes, although resequencing studies have more recently identified the *KIR3DL1/S1*,[50] *SDHA*,[51] and a number of innate immunity[52] genes as likely candidates. Suggested mechanisms are heterozygous or rare-allele advantage. In order to

account for the different summary statistic results in the CHB, which share the high frequency of the deletion with a large number of East Asian populations (Figure 3), such long-term balancing selection would have to be complemented by positive selection for the deletion in this part of the world. Indeed, the distinction between positive selection and balancing selection is not absolute; balancing selection is likely to begin with positive selection for the new allele, and positive selection that has increased the frequency of the selected allele to intermediate levels will lead to summary statistics characteristic of balancing selection. Thus an alternative interpretation of the worldwide results would be that the predominant influence has been positive selection but that the selective coefficient has been lower in Europe than in Asia, leading to the currently observed intermediate frequency of the deletion in Europe. Although constant low-intensity positive selection in Europe seems unlikely to have been maintained for a long period because the climatic, biological and social environments have changed enormously in the last 20,000 years, from populations of Ice Age hunter-gathers to the farmers in the currently favorable climate,[45] more complex scenarios of selection are possible.

The target of selection is uncertain, but it includes *UGT2B17* and any other sequence in high LD with the resequenced region. The reference sequence misassembly makes LD in this region of the genome difficult to evaluate, but the documented effects of the deletion on *UGT2B17* expression[18] and metabolism[12–14] make it an excellent candidate for the target. Balancing selection might generally favor *UGT2B17* heterozygotes, or the presence or absence of the *UGT2B17* gene in different situations or in different subsets of the population. A careful examination of the network in Figure 5E also suggests the possibility of selection on UGT2B17 that is more complex than simply presence or absence. The two arms of the network consist of (1) entirely undeleted and (2) a mixture of deleted and undeleted chromosomes. Could selection have initially favored a UGT2B17 variant carried by the (2) arm undeleted chromosomes, with subsequent selection for the full deletion? The involvement of UGT2B17 in multiple biological processes including xenobiotic and steroid hormone metabolism hinders identification of a single biological candidate for the selective agent. Populations might be exposed to different foods or poisons, or different hormone levels might be favored; an intriguing but untested hypothesis is that selection might favor different levels of UGT2B17 and testosterone in males compared with females, and the balance might differ between populations. Whatever the evolutionary pressures in the distant past, the phenotypic impact of this deletion for sports medicine remains high because it strongly influences the results of androgen doping tests.[53]

There is increasing evidence that gene content varies significantly between individuals: In addition to copy-number variation that can change the number of gene segments, whole genes, or clusters of genes,[2] truncating variants arising from small indels or premature stop codons can affect individual genes, and mechanisms that oblate the expression of a gene rendering it effectively null despite its presence remain predominantly unexplored but may be common. Furthermore, this variation in gene content is of evolutionary significance. For example, an increase in copy number of the salivary amylase gene cluster has been reported in populations that depend on high-starch diets that may improve their nutrition,[9] whereas loss of the caspase-12 gene appears to have increased resistance to severe sepsis[54] and of the actinin-3 gene to have improved sprinting ability.[55] It seems likely that changes in gene content have been of great functional and evolutionary significance in the recent human past, but our understanding of them is just beginning.

## Supplemental Data

Supplemental Data include three tables and can be found with this article online at http://www.ajhg.org/.

## Acknowledgments

We thank all the sample donors for making this work possible, Nancy Holroyd and the Sanger Faculty Small Sequencing Projects Group for generating the sequence data, all CNV Consortium members for contributions, Howard M. Cann for providing the HGDP-CEPH DNA panel, Jianxiang Chi for help with FISH, and Richard Redon for advice on $F_{ST}$ calculation using R. This work was supported by The Wellcome Trust.

Received: July 22, 2008
Revised: August 6, 2008
Accepted: August 7, 2008
Published online: August 28, 2008

## Web Resources

The URLs for data presented herein are as follows:

Database of Genomic Variants, http://projects.tcag.ca/variation/
Network, http://www.fluxus-engineering.com/sharenet.htm
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim
Prof. Bob Griffiths, http://www.stats.ox.ac.uk/people/academic_staff/bob_griffiths
SeattleSNPs project, http://pga.mbt.washington.edu/
UCSC Genome Browser, http://genome.ucsc.edu/cgi-bin/hgGateway

## References

1. Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. (2006). Copy number variation: New insights in genome diversity. Genome Res. *16*, 949–961.
2. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen,

W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.
3. Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science *315*, 848–853.
4. Aitman, T.J., Dong, R., Vyse, T.J., Norsworthy, P.J., Johnson, M.D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A.J., Petretto, E., et al. (2006). Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. Nature *439*, 851–855.
5. Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J., et al. (2005). The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science *307*, 1434–1440.
6. Kehrer-Sawatzki, H., and Cooper, D.N. (2007). Structural divergence between the human and chimpanzee genomes. Hum. Genet. *120*, 759–778.
7. Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., and Eichler, E.E. (2001). Positive selection of a gene family during the emergence of humans and African apes. Nature *413*, 514–519.
8. Kidd, J.M., Newman, T.L., Tuzun, E., Kaul, R., and Eichler, E.E. (2007). Population stratification of a common *APOBEC* gene deletion polymorphism. PLoS Genet *3*, e63.
9. Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.
10. Gardner, M., Williamson, S., Casals, F., Bosch, E., Navarro, A., Calafell, F., Bertranpetit, J., and Comas, D. (2007). Extreme individual marker $F_{ST}$ values do not imply population-specific selection in humans: The *NRG1* example. Hum. Genet. *121*, 759–762.
11. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.
12. Jakobsson, J., Ekstrom, L., Inotsume, N., Garle, M., Lorentzon, M., Ohlsson, C., Roh, H.K., Carlstrŏm, K., and Rane, A. (2006). Large differences in testosterone excretion in Korean and Swedish men are strongly associated with a UDP-glucuronosyl transferase 2B17 polymorphism. J. Clin. Endocrinol. Metab. *91*, 687–693.
13. Swanson, C., Mellström, D., Lorentzon, M., Vandenput, L., Jakobsson, J., Rane, A., Karlsson, M., Ljunggren, Ö., Smith, U., Eriksson, A.L., et al. (2007). The UDP Glucuronosyltransferase 2B15 $D^{85}Y$ and 2B17 deletion polymorphisms predict the glucuronidation pattern of androgens and fat mass in men. J. Clin. Endocrinol. Metab. *92*, 4878–4882.
14. Park, J., Chen, L., Ratnashinge, L., Sellers, T.A., Tanner, J.P., Lee, J.-H., Dossett, N., Lang, N., Kadlubar, F.F., Ambrosone, C.B., et al. (2006). Deletion polymorphism of UDP-glucuronosyltransferase 2B17 and risk of prostate cancer in African American and Caucasian men. Cancer Epidemiol. Biomarkers Prev. *15*, 1473–1478.
15. Karypidis, A.H., Olsson, M., Andersson, S.O., Rane, A., and Ekström, L. (2008). Deletion polymorphism of the UGT2B17 gene is associated with increased risk for prostate cancer and correlated to gene expression in the prostate. Pharmacogenomics J. *8*, 147–151.
16. Gallagher, C.J., Kadlubar, F.F., Muscat, J.E., Ambrosone, C.B., Lang, N.P., and Lazarus, P. (2007). The UGT2B17 gene deletion polymorphism and risk of prostate cancer. A case-control study in Caucasians. Cancer Detect. Prev. *31*, 310–315.
17. Olsson, M., Lindström, S., Häggkvist, B., Adami, H.O., Bälter, K., Stattin, P., Ask, B., Rane, A., Ekström, L., and Grönberg, H. (2008). The UGT2B17 gene deletion is not associated with prostate cancer risk. Prostate *68*, 571–575.
18. McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., et al. (2006). Common deletion polymorphisms in the human genome. Nat. Genet. *38*, 86–92.
19. Wilson, W., III, Pardo-Manuel de Villena, F., Lyn-Cook, B.D., Chatterjee, P.K., Bell, T.A., Detwiler, D.A., Gilmore, R.C., Valladeras, I.C., Wright, C.C., Threadgill, D.W., et al. (2004). Characterization of a common deletion polymorphism of the *UGT2B17* gene linked to *UGT2B15*. Genomics *84*, 707–714.
20. Spielman, R.S., Bastone, L.A., Burdick, J.T., Morley, M., Ewens, W.J., and Cheung, V.G. (2007). Common genetic variants account for differences in gene expression among ethnic groups. Nat. Genet. *39*, 226–231.
21. The International HapMap Consortium (2003). The international HapMap project. Nature *426*, 789–796.
22. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.
23. Rosenberg, N.A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann. Hum. Genet. *70*, 841–847.
24. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420–426.
25. Sonnhammer, E.L., and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene *167*, GC1–GC10.
26. Goudet, J. (2005). HIERFSTAT, a package for R to compute and test variance components and *F*-statistics. Mol. Ecol. Notes *5*, 184–186.
27. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263–265.
28. Stephens, M., and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. *73*, 1162–1169.
29. Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. *68*, 978–989.
30. Bandelt, H.J., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. *16*, 37–48.
31. Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338.
32. Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Res. *15*, 1576–1583.
33. Helgason, A., Pálsson, S., Thorleifsson, G., Grant, S.F., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G.,

Reynisdottir, I., et al. (2007). Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat. Genet. *39*, 218–225.
34. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.
35. Fu, Y.-X., and Li, W.-H. (1993). Statistical tests of neutrality of mutations. Genetics *133*, 693–709.
36. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.
37. Fu, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics *147*, 915–925.
38. Bahlo, M., and Griffiths, R.C. (2000). Inference from gene trees in a subdivided population. Theor. Popul. Biol. *57*, 79–95.
39. Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.
40. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science *307*, 1072–1079.
41. The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87.
42. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.
43. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928–933.
44. Chen, J.M., Cooper, D.N., Chuzhanova, N., Férec, C., and Patrinos, G.P. (2007). Gene conversion: Mechanisms, evolution and human disease. Nat. Rev. Genet. *8*, 762–775.
45. Jobling, M.A., Hurles, M.E., and Tyler-Smith, C. (2004). Human Evolutionary Genetics (New York and Abingdon: Garland Science).
46. Fagundes, N.J., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L., and Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. USA *104*, 17614–17619.
47. Calafell, F., Roubinet, F., Ramírez-Soriano, A., Saitou, N., Bertranpetit, J., and Blancher, A. (2008). Evolutionary dynamics of the human *ABO* gene. Hum. Genet., in press. Published online July 16, 2008. 10.1007/s00439-008-0530-8.
48. Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet *2*, e64.
49. Bubb, K.L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., et al. (2006). Scan of human genome reveals no new loci under ancient balancing selection. Genetics *173*, 2165–2177.
50. Norman, P.J., Abi-Rached, L., Gendzekhadze, K., Korbel, D., Gleimer, M., Rowley, D., Bruno, D., Carrington, C.V., Chandanayingyong, D., Chang, Y.H., et al. (2007). Unusual selection on the KIR3DL1/S1 natural killer cell receptor in Africans. Nat. Genet. *39*, 1092–1099.
51. Baysal, B.E., Lawrence, E.C., and Ferrell, R.E. (2007). Sequence variation in human succinate dehydrogenase genes: Evidence for long-term balancing selection on *SDHA*. BMC Biol. *5*, 12.
52. Ferrer-Admetlla, A., Bosch, E., Sikora, M., Marquès-Bonet, T., Ramírez-Soriano, A., Muntasell, A., Navarro, A., Lazarus, R., Calafell, F., Bertranpetit, J., et al. (2008). Balancing selection is the main force shaping the evolution of innate immunity genes. J. Immunol. *181*, 1315–1322.
53. Schulze, J.J., Lundmark, J., Garle, M., Skilving, I., Ekström, L., and Rane, A. (2008). Doping test results dependent on genotype of uridine diphospho-glucuronosyl transferase 2B17, the major enzyme for testosterone glucuronidation. J. Clin. Endocrinol. Metab. *93*, 2500–2506.
54. Xue, Y., Daly, A., Yngvadottir, B., Liu, M., Coop, G., Kim, Y., Sabeti, P., Chen, Y., Stalker, J., Huckle, E., et al. (2006). Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am. J. Hum. Genet. *78*, 659–670.
55. MacArthur, D.G., Seto, J.T., Raftery, J.M., Quinlan, K.G., Huttley, G.A., Hook, J.W., Lemckert, F.A., Kee, A.J., Edwards, M.R., Berman, Y., et al. (2007). Loss of *ACTN3* gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. Nat. Genet. *39*, 1261–1265.
56. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science *298*, 2381–2385.