The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism

Maximilian P.A. Salm,[1,8,9] Stuart D. Horswell,[2] Claire E. Hutchison,[1] Helen E. Speedy,[1] Xia Yang,[3] Liming Liang,[4] Eric E. Schadt,[3] William O. Cookson,[5] Anthony S. Wierzbicki,[6] Rossi P. Naoumova,[7] and Carol C. Shoulders[1,9]

[1]Centre for Endocrinology, Barts & the London School of Medicine & Dentistry, Queen Mary University of London, London EC1M 6BQ, United Kingdom; [2]Bioinformatics & Biostatistics Group, Cancer Research UK London Research Institute, London WC2A 3LY, United Kingdom; [3]Department of Systems Biology, Sage Bionetworks, Seattle, Washington 98109, USA; [4]Department of Epidemiology, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA; [5]National Heart & Lung Institute, Imperial College London, London SW3 6LY, United Kingdom; [6]Guys & St. Thomas Hospital, King's College London, London SE1 2PR, United Kingdom; [7]Institute of Clinical Science, Imperial College London, London W12 0NN, United Kingdom

Genomic inversions are an increasingly recognized source of genetic variation. However, a lack of reliable high-throughput genotyping assays for these structures has precluded a full understanding of an inversion's phylogenetic, phenotypic, and population genetic properties. We characterize these properties for one of the largest polymorphic inversions in man (the ~4.5-Mb 8p23.1 inversion), a structure that encompasses numerous signals of natural selection and disease association. We developed and validated a flexible bioinformatics tool that utilizes SNP data to enable accurate, high-throughput genotyping of the 8p23.1 inversion. This tool was applied retrospectively to diverse genome-wide data sets, revealing significant population stratification that largely follows a clinal "serial founder effect" distribution model. Phylogenetic analyses establish the inversion's ancestral origin within the *Homo* lineage, indicating that 8p23.1 inversion has occurred independently in the *Pan* lineage. The human inversion breakpoint was localized to an inverted pair of human endogenous retrovirus elements within the large, flanking low-copy repeats; experimental validation of this breakpoint confirmed these elements as the likely intermediary substrates that sponsored inversion formation. In five data sets, mRNA levels of disease-associated genes were robustly associated with inversion genotype. Moreover, a haplotype associated with systemic lupus erythematosus was restricted to the derived inversion state. We conclude that the 8p23.1 inversion is an evolutionarily dynamic structure that can now be accommodated into the understanding of human genetic and phenotypic diversity.

[Supplemental material is available for this article.]

Genomic inversions are a ubiquitous class of structural variation, implicated in speciation (Lowry and Willis 2010), adaptation (Hoffmann and Rieseberg 2008), and human disease (Girirajan and Eichler 2010). In addition to dislocating coding sequences (Feuk 2010), inversions can influence gene expression (Weiler and Wakimoto 1995) and facilitate aberrant recombination events that cause genomic disorders (Feuk 2010). Moreover, inversions suppress local recombination, which can preserve beneficial or deleterious haplotypic configurations (Hoffmann and Rieseberg 2008; Joron et al. 2011). These characteristics are largely exemplified by the 17q21.31 inversion, which is associated with decreased microtubule-associated protein tau (*MAPT*) expression, predisposes carriers to the 17q21.31 micro-deletion syndrome, and exhibits perfect linkage disequilibrium (LD) across its ~970-kb interval (Donnelly et al. 2010). However, aside from this inversion, human inversions remain largely uncharacterized, which is partly attributable to the difficulty in assaying these structures at the population level (Feuk 2010; Alkan et al. 2011a).

The 8p23.1 inversion (*8p23-inv*) is one of the largest polymorphic inversions found in man, encompassing ~4.5 Mb (Giglio et al. 2001; Sayers et al. 2011). Conventionally *8p23-inv* is genotyped using fluorescent in situ hybridization (FISH) (Giglio et al. 2001), which is not amenable to high-throughput analyses and requires viable cells. Moreover, the size of the inversion's single copy region (~3.5 Mb) approaches the practical resolution of metaphase FISH (Raap 1998). In the small samples studied to date, the inversion has estimated frequencies of ~59% in the Yoruba, ~20%–50% in Europeans, and ~12%–27% in Asians (Broman et al. 2003; Sugawara et al. 2003; Antonacci et al. 2009), while its frequency in other ethnicities remains unspecified. At the species level, *8p23-inv* has only been observed in humans, chimpanzees, and bonobos (Antonacci et al. 2009), but whether this shared feature reflects homology or homoplasy is unresolved.

Multiple loci within the inversion are putative targets of natural selection (Barreiro et al. 2008; Deng et al. 2008; Lopez Herraez et al. 2009; Pickrell et al. 2009; Browning and Weir 2010). Furthermore, *8p23-inv* not only increases the risk of producing offspring with unbalanced chromosomal rearrangements (Hollox et al. 2008) but also encompasses numerous loci associated with autoimmune (e.g., Gregersen et al. 2009; Deng and Tsao 2010; Nordmark et al. 2011) and cardiovascular (e.g., Levy et al. 2009; Teslovich et al. 2010;

[8]Present address: Bioinformatics & Biostatistics group, Cancer Research UK London Research Institute, London WC2A 3LY, United Kingdom.
[9]Corresponding authors.
E-mail max.salm05@imperial.ac.uk.
E-mail c.shoulders@qmul.ac.uk.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.126037.111.

22:1144–1153 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

Dehghan et al. 2011) disease phenotypes. The precise molecular mechanisms underlying these associations are poorly defined. Given the profound effect of inversions on local genetic structure (Hoffmann and Rieseberg 2008), future studies should benefit from detailed characterization of *8p23-inv* (Harley et al. 2009).

Inversions are often associated with elevated LD within their boundaries due to recombination inhibition between noncollinear regions in inversion heterozygotes (Hoffmann and Rieseberg 2008). This makes high-throughput inversion-typing via genetic markers that perfectly correlate with the structure tenable (Donnelly et al. 2010). This approach was applied to *8p23-inv*, identifying candidate inversion proxy single-nucleotide polymorphisms (SNPs) in small ($n \leq 13$), ancestrally diverse HapMap samples (Antonacci et al. 2009). Furthermore, two potential inversion-associated SNP homozygosity tracts were identified in six Europeans (Bosch et al. 2009). However, these two studies generally do not concur, in either the predictive SNPs identified or the inversion-type predictions made, which may be attributable to inflated LD estimates stemming from the small sample sizes analyzed (Iles 2008). In a related approach, principal components analysis of phased SNP genotypes was used to predict inversion-type (Deng et al. 2008), but as discussed by Antonacci et al. (2009), these predictions differed considerably from FISH genotype data.

Herein we present a novel, robust high-throughput method to genotype *8p23-inv*. By use of this tool, we define population, evolutionary, and functional attributes of the inversion.

Results

Development of a novel 8p23 inversion detection method

To develop a high-throughput assay for *8p23-inv*, we typed the inversion by FISH in 68 CEU samples representing 13 CEU trios and 42 HapMap phase II founders (Supplemental Table S1; Supplemental Fig. S1). For clarity, *N* (noninverted) refers to the orientation represented in the human genome reference (hg19) and *I* to the inverted state. The inversion was transmitted according to Mendelian inheritance patterns as expected, and published inversion genotypes were confirmed (Supplemental Table S1).

No HapMap SNPs or 1000 Genome Project variants were in complete LD with *8p23-inv* (data not shown). Thus no known SNP acts as a perfect proxy for the inversion. However, 20 SNPs with a strong correlation to the inversion (i.e., $r^2 > 0.8$) (Supplemental Table S2) were identified. To evaluate their combined efficacy in predicting inversion-type, a leave-n-out cross-validation imputation analysis was performed. While the concordance between known and imputed inversion-type using this SNP set was 86.5 ± 9.2% (mean ± SD), the full range was 45%–100%. Such uncertainty precludes reliable inversion genotyping, and we therefore developed an alternative inversion-typing method based on local genetic substructure.

To explore the relationship between genetic substructure and inversions in the 8p23 region in Europeans, SNP genotypes restricted to the inversion interval in CEU founders were examined using multidimensional scaling (MDS). Individuals stratified into three groups along the first axis, in perfect correlation with their inversion-type (Fig. 1). Moreover, a tri-modal Gaussian mixture model fit to the positions along the first axis allowed effectively unsupervised clustering of this data set into their concomitant inversion-types with associated confidence values. The clustering solution is independently corroborated by three measures: connectivity, the Dunn index, and silhouette width. These concatenated processes constitute the "PFIDO" algorithm (Supplemental Fig. S2), which perfectly classified all empirically determined inversion-types ($P = 8.25 \times 10^{-17}$).

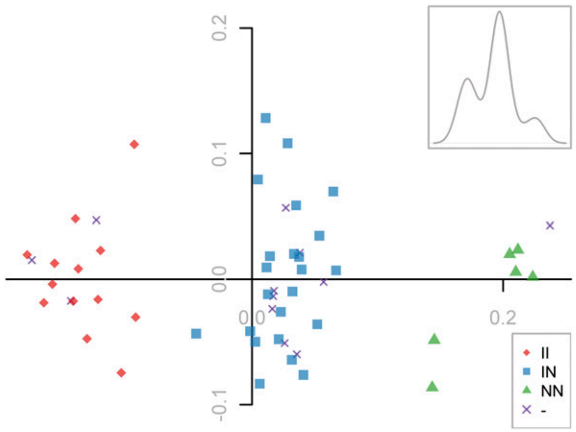

Figure 1. Genetic substructure in the *8p23* inversion region correlates perfectly with inversion status in CEU HapMap II samples. A multidimensional scaling (MDS) analysis based on 5494 SNPs typed in 54 CEU founder members (that satisfied the inclusion criteria; see Methods). The *x*- and *y*-axis correspond to the first and second "axes" derived by MDS, and each point represents an individual, labeled by inversion-type (as defined by FISH). (*Inset*) Tri-modal point distribution found in the first dimension.

To assess the algorithm's accuracy, inversion-type was predicted in a cohort of 1402 related British Caucasians, using 49 SNPs from a SNP set optimized for PFIDO (Supplemental Note; Supplemental Tables S2, S3; Supplemental Fig. S3). At a clustering threshold of $P < 0.05$, inversion-types were called for 1338 individuals (call rate > 95%). No violations of Mendelian inheritance patterns were detected, nor was Hardy-Weinberg equilibrium (HWE) breached in unrelated individuals ($P = 0.19$, $n = 233$), suggesting that the inversion-type predictions are robust. The predictions were verified by FISH in seven unrelated samples: At the designated clustering cut-off of $P < 0.05$, all were correct ($P = 9.52 \times 10^{-3}$).

The PFIDO algorithm does not rely on any specific SNP enabling retrospective analyses of existing GWAS data sets. For example, inversion-type classification of CEU reference samples was without error when restricting the input SNPs to those represented on three popular genotyping arrays, even though their SNP content overlap is ~7% (Supplemental Fig. S4). Four data sets representing ancestrally European samples (Wellcome Trust Case Control Consortium 2007) further illustrate retrospective inversion-typing (Fig. 2): the frequency of the *N* allele in the CEU sample is similar in the combined WTCCC control (42.9%), WTCCC coronary artery disease (44.4%), and WTCCC type 2 diabetes cohorts (45.6%).

Population stratification of the 8p23 inversion

To explore the relationship between the 8p23 inversion and genetic substructure in non-European populations, *8p23-inv* was typed by FISH in 17 JPT and 15 YRI HapMap II samples (Supplemental Table S1). The frequencies of *I* alleles were 21% (JPT) and 52% (YRI), concurring with previous observations (Sugawara et al. 2003; Antonacci et al. 2009), while the correlation between experimentally determined inversion-type and genetic substructure was less distinct than in Europeans, with less pronounced clustering of each inversion-type along the first MDS-derived axis (Supplemental Note). Nevertheless, PFIDO correctly categorized 2

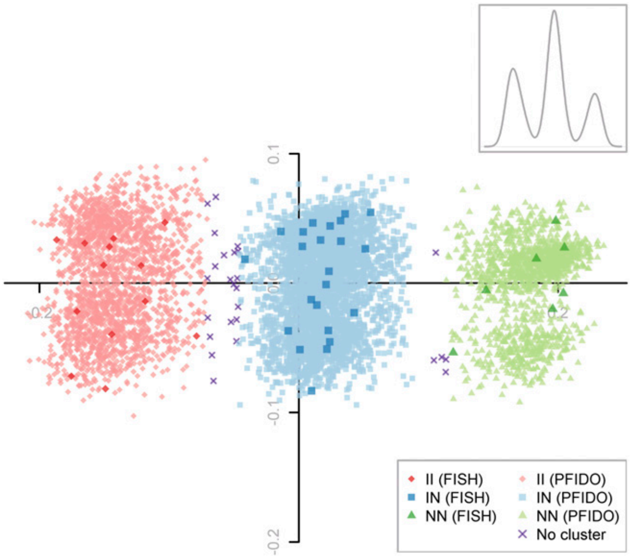

Figure 2. *8p23* inversion prediction in WTCCC cohorts. PFIDO output (based on 1043 SNPs) is shown overlaying the corresponding MDS plot, for four data sets combined. Predicted inversion-type frequencies are in HWE in all cohorts and CEU FISH-derived inversion-types served as an internal control. For further details, see Figure 1 legend.

and 3 clusters in the JPT and YRI samples, respectively (Supplemental Fig. S5). Moreover, at a PFIDO clustering threshold of $P < 0.05$, experimentally determined JPT and YRI inversion-types were all predicted correctly ($P = 1.40 \times 10^{-3}$ and $P = 3.61 \times 10^{-5}$, respectively).

To establish the worldwide distribution of *8p23-inv*, PFIDO was applied to a combined data set representing 1894 individuals sampled from 56 globally distributed populations (Supplemental Table S4). The 8p23 inversion exhibits a striking clinal distribution, strongly correlating with geographic distance from Addis Ababa (Fig. 3A), a conceivable origin of modern humans (Handley et al. 2007). The *I* allele is frequent in sub-Saharan Africa (mean = 69.7 ± 5.6%) and progressively diminishes in frequency across the Eurasian continent, effectively disappearing in the Americas (mean = 1.3 ± 2.5%). This population stratification is reflected in a global F_{st} of 0.213, suggesting greater partitioning of *8p23-inv* frequency variance between populations than within populations (Holsinger and Weir 2009). Pairwise F_{st} values (estimated in the Human Genome Diversity Project [HGDP]) also strongly correlate with the geographic distance between sampled populations (Mantel test, r = 0.7, $P = 1 \times 10^{-4}$), and there is empirically significant pairwise genetic differentiation between African and Oceanic/American samples ($P < 0.01$) (Fig. 3B).

To examine whether the inversion's distribution is explicable by nondemographic factors (e.g., positive selection), the *8p23-inv* distribution in the HGDP data set was compared with that of 19,969 autosomal SNPs, selected to represent loci of similar allele frequency under putatively neutral selection. Relative to these SNPs, the inversion's geographical correlation differs marginally ($P < 0.049$) (Supplemental Fig. S6A). Similarly, the *8p23-inv* global F_{st} (0.218) is marginally significant compared with the null distribution ($P < 0.022$) (Supplemental Fig. S6B). Thus the inversion's distribution is

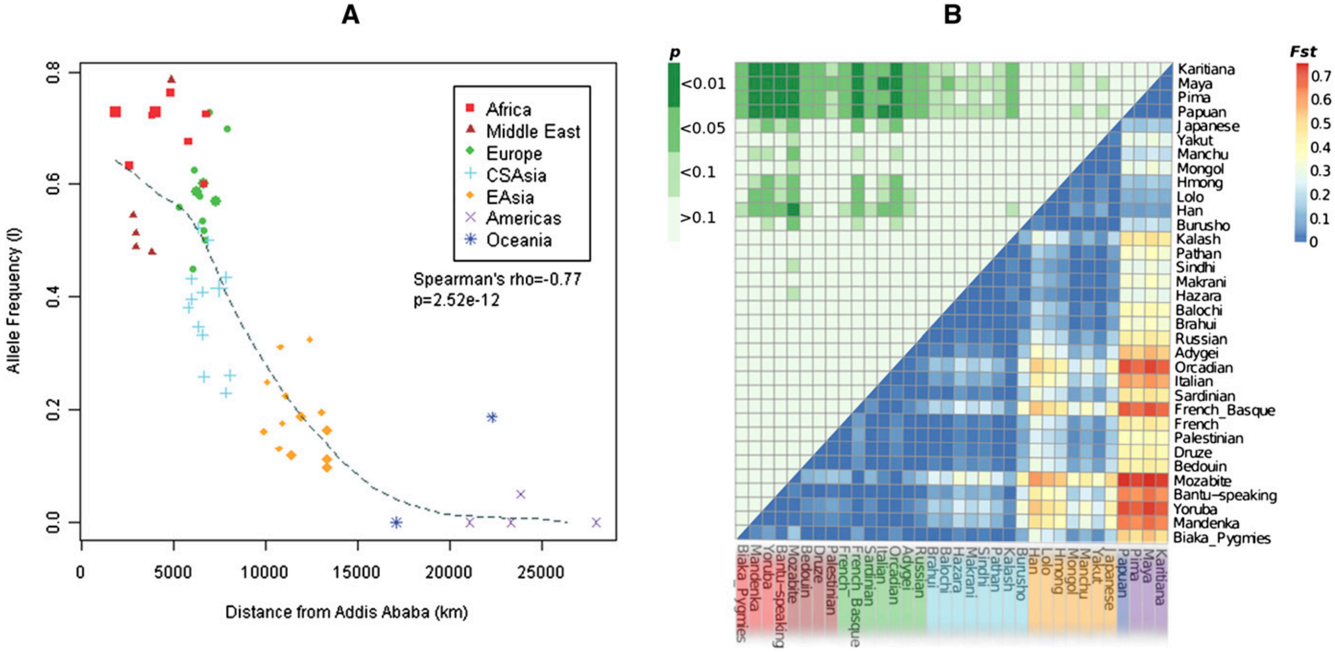

Figure 3. Population stratification of *8p23-inv*. (*A*) The worldwide frequency distribution of *8p23-inv*. The frequency of the inverted allele (*I*) in each population (Supplemental Table S4) is plotted against geographical distance from Addis Ababa (Ethiopia). Populations are color-coded according to continental origin, and larger symbols correspond to larger samples ($n > 60$). The dashed LOESS fitted line highlights the significant negative correlation between the two variables, as assessed by the Spearman's rank correlation test. Although one Oceanic sample (Tongan/Samoan) appears as an outlier, its allele frequency (19%) is consistent with this sample's previously reported genetic relationship to East Asians (Xing et al. 2010). (*B*) Pairwise F_{st} between HGDP populations, based on *8p23-inv*. The *bottom right* triangle shows pairwise F_{st} values: Each shaded box represents a pairwise population comparison, with higher values in red. The *top left* triangle (green) shows the corresponding significance of the F_{st} values relative to an empirical distribution of 1000 SNPs (Supplemental Fig. S6C). (Dark green) F_{st} values in the top 1% of the distribution (i.e., $P < 0.01$); (lighter green) F_{st} values in the top 5% (i.e., $P < 0.05$). Populations are grouped by continent of origin (as in *A*).

largely consistent with demographic models of the human expansion out of Africa, with a potential weak contribution from positive/negative selection. Nevertheless, given this distribution, population stratification must be accounted for in *8p23-inv* case-control association analyses

Inversion-specific haplotypes and their phenotypic influence

We examined the influence of *8p23-inv* on local gene expression in five data sets (Supplemental Note). *8p23-inv* was robustly associated with *BLK*, *PPP1R3B*, *XKR6*, *FAM167A*, and *CTSB* mRNA levels (Supplemental Table S5A). The association between *8p23-inv* and *PPP1R3B* expression is particularly noteworthy as the trend is consistent across populations (Supplemental Fig. S7) and *PPP1R3B* mRNA levels have been associated with serum lipid levels (Teslovich et al. 2010).

Although inversions can directly influence gene expression (Weiler and Wakimoto 1995), they also exert indirect effects by maintaining allelic configurations (Myers et al. 2007). To explore the latter, the contribution of SNPs in *cis* with inversion alleles to the expression quantitative trait loci (eQTLs) was explored using an allele-specific expression data set (Ge et al. 2009). *8p23-inv* was less significantly associated with the 14 reported "expression windows" than the originally reported SNP associations (Ge et al. 2009: supplemental table S1). Moreover, including these SNPs as covariates in the *8p23-inv* analysis left only one significant association: that between *8p23-inv* and primary transcripts from the *BLK-FAM167A* intergenic region (Supplemental Table S5B). However, another SNP identified by re-sequencing (rs1382567) (Ge et al. 2009) is more significantly associated with the *BLK-FAM167A* intergenic region expression than *8p23-inv* (Supplemental Table S5B). These data suggest that the inversion-eQTLs are primarily mediated via SNP alleles common to a specific inversion background.

Allelic expression associations with *BLK-FAM167A* have been refined to four 16-kb haplotypes (A–D); the A-haplotype accounts for differential *BLK* and *FAM167A* expression, while the A- and C-haplotypes account for differential expression from the *BLK-FAM167A* intergenic region (Ge et al. 2009). In Europeans these haplotypes are completely restricted to specific inversion backgrounds; the B- and D-haplotypes reside on an *I* background, while the A- and C-haplotypes are found on an *N* background (Fig. 4). It is therefore likely that associations between *8p23-inv* and *BLK/FAM167A* expression are mediated by the major A-haplotype found only on the *N* background. Moreover, the efficacy of *8p23-inv* as a predictor of *BLK-FAM167A* intergenic expression is probably attributable to the *N* background harboring both the A- and C-haplotypes. Regarding disease, risk alleles for systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA) are transmitted solely on the A-haplotype (Fig. 4), and their functional effect is putatively mediated via *BLK* expression (Hom et al. 2008; Gregersen et al. 2009). Thus, variants specific to the noninverted configuration contribute to the etiology of these autoimmune disorders.

In summary, recombination suppression induced by the inversion appears to have maintained allelic combinations that collectively influence levels of at least five transcripts, leading to distinct expression patterns associated with inversion-type. This suppression of recombination may also have contributed to preservation of a risk haplotype on the *N* allele for SLE and RA.

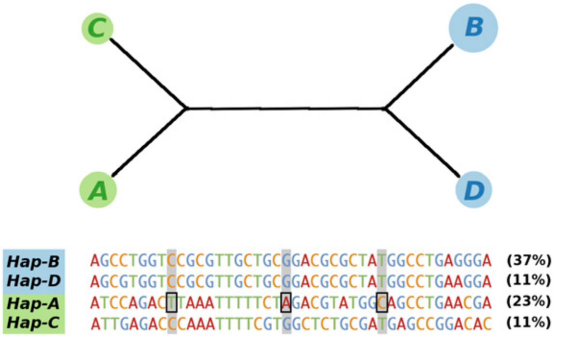
C
B
A
D
Hap-B
Hap-D
Hap-A
Hap-C
AGCCTGGTCCGCGTTGCTGCGGACGCGCTATGGCCTGAGGGA (37%)
AGCGTGGTCCGCGTTGCTGCGGACGCGCTATGGCCTGAAGGA (11%)
ATCCAGACTTAAATTTTTCTAGACGTATGGCAGCCTGAACGA (23%)
ATTGAGACCCAAATTTTCGTGGCTCTGCGATGAGCCGGACAC (11%)

Figure 4. Median-joining network representing phased HapMap CEU haplotypes influencing *BLK-FAM167A* mRNA levels. The network was constructed using 42 SNPs (Supplemental Table S5C) mapping to chr8:11,376,782–11,393,411 phased in 45 samples; *8p23-inv* heterozygotes were omitted to minimize phasing-induced errors, as were haplotypes with frequencies ≤2%. Haplotype groups are denoted by circles (with circle size proportional to haplotype frequency) and colored by whether the haplotypes derive from *II* (blue) or *NN* (green) samples. Haplotype groups are named "A–D," corresponding with previous nomenclature (Ge et al. 2009). Distances between nodes are proportional to the number of differences distinguishing each haplotype; actual haplotypes (and their frequencies in CEU founders) are provided *below* the network. Risk alleles for systemic lupus erythematosus and rheumatoid arthritis identified in four GWAS are highlighted (rs2736340, rs13277113, rs2618476) (Graham et al. 2008; Hom et al. 2008; Gregersen et al. 2009; Chung et al. 2011).

Defining the origin of *8p23-inv*

To investigate whether *8p23-inv* predates the speciation event between humans and chimpanzees, we estimated the "time to most recent common ancestor" (TMRCA) between inversion alleles (Zody et al. 2008). Overlapping allelic RPCI-11 BAC sequences (totaling 178 kb) derived from a likely African-American inversion heterozygote (Supplemental Note) were aligned to a chimpanzee genome assembly (Supplemental Table S6). Assuming the chimpanzee and human lineages diverged 4.5–6 million years ago (mya) (Locke et al. 2011), the *I*- and *N*-sequences diverged on average between 390 ± 70 and 520 ± 90 thousand years ago (kya). Including the orangutan sequence (and dating the orangutan-human lineage split at 12–16 mya) (Locke et al. 2011) yielded similar divergence dates whether comparing the RPCI-11 sequence to the chimpanzee (390 ± 100 to 520 ± 130 kya) or orangutan (420 ± 100 to 560 ± 130 kya). Similarly, aligning 2.1 Mb of the chimpanzee genome to HuRef haplotypes (a predicted inversion-heterozygote of European descent) generated comparable *8p23-inv* divergence dates of 250 ± 60 to 340 ± 80 kya.

Gene flow potentially confounds these divergence estimates by homogenizing genetic diversity over time, resulting in a seemingly recent TMRCA (Supplemental Note and Fig. S8). We therefore estimated the TMRCA between HuRef haplotypes for a ~22-kb interval with no evidence of gene flow (chr8:10848492–10870275; CEU *N* and *I* haplotypes share no polymorphic sites according to HapMap and 1000 Genome Project data), which yielded divergence dates of 315 ± 58 and 420 ± 78 kya. An alternative method designed to estimate the minimum TMRCA between inversion alleles (Andolfatto et al. 1999) indicated that *N* and *I* alleles diverged ~364–389 kya (based on HapMap CEU and YRI data). Collectively these data indicate that *8p23-inv* occurred at least 3.5 million years after the human–chimpanzee lineage split, suggesting that *8p23-inv* arose independently in the *Homo* and *Pan* lineages and is thus a recurrent event within primate lineages.

Evolution of the genomic architecture facilitating *8p23-inv*

FISH and genetic mapping have narrowed the *8p23-inv* breakpoints to two highly homologous low-copy repeat (LCR) regions known as "distal repeat" (REPD; ~1.14 Mb) and "proximal repeat" (REPP; ~635 kb) (Hollox et al. 2008). As these structures plausibly enabled inversion formation via nonallelic homologous recombination (NAHR) (Feuk 2010), we investigated their origin by analyzing the syntenic primate sequence in these locations. This identified two orangutan-derived bacterial artificial chromosomes (BACs) that span the syntenic REPP completely, each mapping to two single-copy regions within and outside the human inversion and supporting the inverted orientation as the ancestral state (AC207782 & AC212986) (Supplemental Fig. S9). Remarkably, only 1.5 kb of these BACs shows homology with human LCR sequences (Jiang et al. 2008), indicating an effective absence of LCRs at the syntenic orangutan REPP. Conversely, the orangutan REPD position contains extensive LCRs; for example, a clone that unambiguously maps to the syntenic REPD contains 26.8 kb of LCRs (AC206098) (Supplemental Fig. S9). In gorillas, however, BAC-end sequence analysis identified 11 clones with one end anchored in unique sequence and the other end mapping to REPD/REPP LCRs (Supplemental Fig. S9). These data suggest the paired REPD/REPP arrangement evolved after the emergence of orangutans in the primate lineage.

The timing of REPP formation is also reflected in the genetic divergence between LCR subunits. TMRCA analysis of a nonduplicated LCR section in orangutans that is duplicated in humans and chimpanzees dates the divergence of human REPP/REPD to ~9.5 mya (Supplemental Fig. S10A,B), suggesting REPP was formed prior to the emergence of gorillas during primate speciation. Moreover, pairwise TMRCA analyses of seven ancestral LCR subunits (Jiang et al. 2008) common to REPP/REPD exhibit two peaks in age distribution (Supplemental Fig. S10C): one at ~9 mya and another at ~800 kya (approaching the TMRCA of the *I* and *N* alleles).

Collectively, the data suggest that LCR segments were duplicatively transposed (Johnson et al. 2006) from the REPD to the REPP locus in the common ancestor of gorillas, chimpanzees, and humans, resulting in a 27-kb deletion (Supplemental Fig. S11) and the paralogous LCR loci that now bracket the human *8p23-inv*, the substrates for inversion-formation. Such a model accounts for the restriction of *8p23-inv* to the human and chimpanzee lineages and supports the inverted orientation as the ancestral state.

A breakpoint mapping to a HERV-K element is associated with *8p23-inv*

To map the inversion breakpoints at the nucleotide level, REPD and REPP were stringently reassembled into tiling paths using only finished RPCI-11 8p23 BACs (Supplemental Note). This produced two assemblies mapping to REPD (named LCR-A and -B) and two assemblies mapping to REPP (LCR-C and -D), which broadly mirror the existing reference genome assembly, except for the exclusion of non–RPCI-11 data (Supplemental Figs. S12-14). Six RPCI-11 BACs with significant homology with LCRs-A–D were identified that represent putative structural variants. These exhibited mosaic homology patterns, in which subsections of each BAC optimally aligned to different LCR haplotypes (Fig. 5A,B; Supplemental Fig. S15). Similar haplotype junction patterns were found in several clones (Supplemental Fig. S15), indicating that these are not cloning artifacts (Osoegawa et al. 2007).

The haplotype junctions fall into three broad categories (Fig. 5C): The first covers the *DEFB* locus, a site of extensive copy-number variation (Hollox et al. 2008); the second constitutes a junction between LCR-A and LCR-C; and the third represents junctions between LCR-B and LCR-D. The latter two groups map to large inverted repeats between haplotypes (Fig. 5C). Based on the premise that the inversion was created by NAHR between inverted repeats, the junctions represented in the second and third groups are reasonable candidates for inversion breakpoints.

To explore the potential breakpoints further (see also Supplemental Fig. S16 & Supplemental Note: Fosmid-End Sequences Support the LCR–Haplotype Junctions), multiple sequence alignments of the LCR-A/C and LCR-B/D BAC groups were constructed and analyzed for historical recombination events. At a Bonferroni corrected $P < 0.01$, five recombination events were detected by all seven detection methods employed (Table 1). In all cases, phylogenetic analysis of the sequences flanking the breakpoint revealed clear migration of one clade between two divergent clades (e.g., Supplemental Fig. S17A), consistent with a recombination product having formed from sequences represented by the two divergent clades. Four recombinant events were detected in multiple BAC libraries (Table 1), suggestive of common breakpoints.

To establish whether recombinant haplotypes co-segregated with inversion-type, 34 CEU founders (15x *II*, 14x *NN*, 5x *IN*) were typed for breakpoint-spanning haplotypes using the double "amplification refractory mutation system" (Lo et al. 1991), followed by PCR-product sequencing to verify haplotype specificity (Supplemental Fig. S17B). After multiple-testing correction, a single haplotype exhibited significant association with inversion-type ($P = 3.22 \times 10^{-5}$, $r_{\varphi} = 0.8$) (Table 1, event 5): The "parental" LCR-B related haplotype-group was more common in *II* samples (14/15) than *NN* samples (2/14), while the reciprocal "recombinant" LCR-D–related haplotype-group was present in all *NN* samples and five *II* samples ($P = 0.05$) (Supplemental Fig. S18).

In summary, a recombination event was identified that strongly correlates with inversion-type. The data also suggest the *N* allele derives from an ancestral *I* allele, consistent with the clinal geographic distribution of *8p23-inv* and the orientation of the region in orangutans. Moreover, the recombination event maps to an inverted pair of 9.5-kb human endogenous retrovirus elements (*HERV-K27* and *HERV-KOLD130352*) (Romano et al. 2006), members of an ancient retrovirus family frequently implicated in promoting NAHR (Jern and Coffin 2008).

Discussion

At *8p23-inv*, inversion-type and genetic substructure correlate perfectly in ancestrally diverse populations, providing a further example of inversion-mediated recombination suppression (Hoffmann and Rieseberg 2008). This feature acts as an effective surrogate for the inversion, enabling high-throughput *8p23-inv* genotyping: Crucially, our predictions perfectly correlated with all FISH-determined inversion-types ($n = 110$). The efficacy of our genotyping method (PFIDO) implicitly suggests that 8p23 inversion events were infrequent and relatively ancient. Nevertheless, given that no genetic markers were found that perfectly correlate with inversion-type, *8p23-inv* may not be an absolute recombination barrier, which corresponds with theoretical predictions for larger inversions (Andolfatto et al. 2001).

Although *8p23-inv* encompasses ~4.5 Mb, its worldwide distribution broadly reflects that of simpler genetic markers (Supplemental Fig. S6). Conceptually, the serial founder model of migration from Africa (Novembre and Di Rienzo 2009) accounts for its distribution. This implies that the inversion is not responsible for any highly penetrant adaptive/nonadaptive phenotypes and appears to be under neutral (or very weak) selection pressure. However, numerous signals of natural selection reside within *8p23-inv*

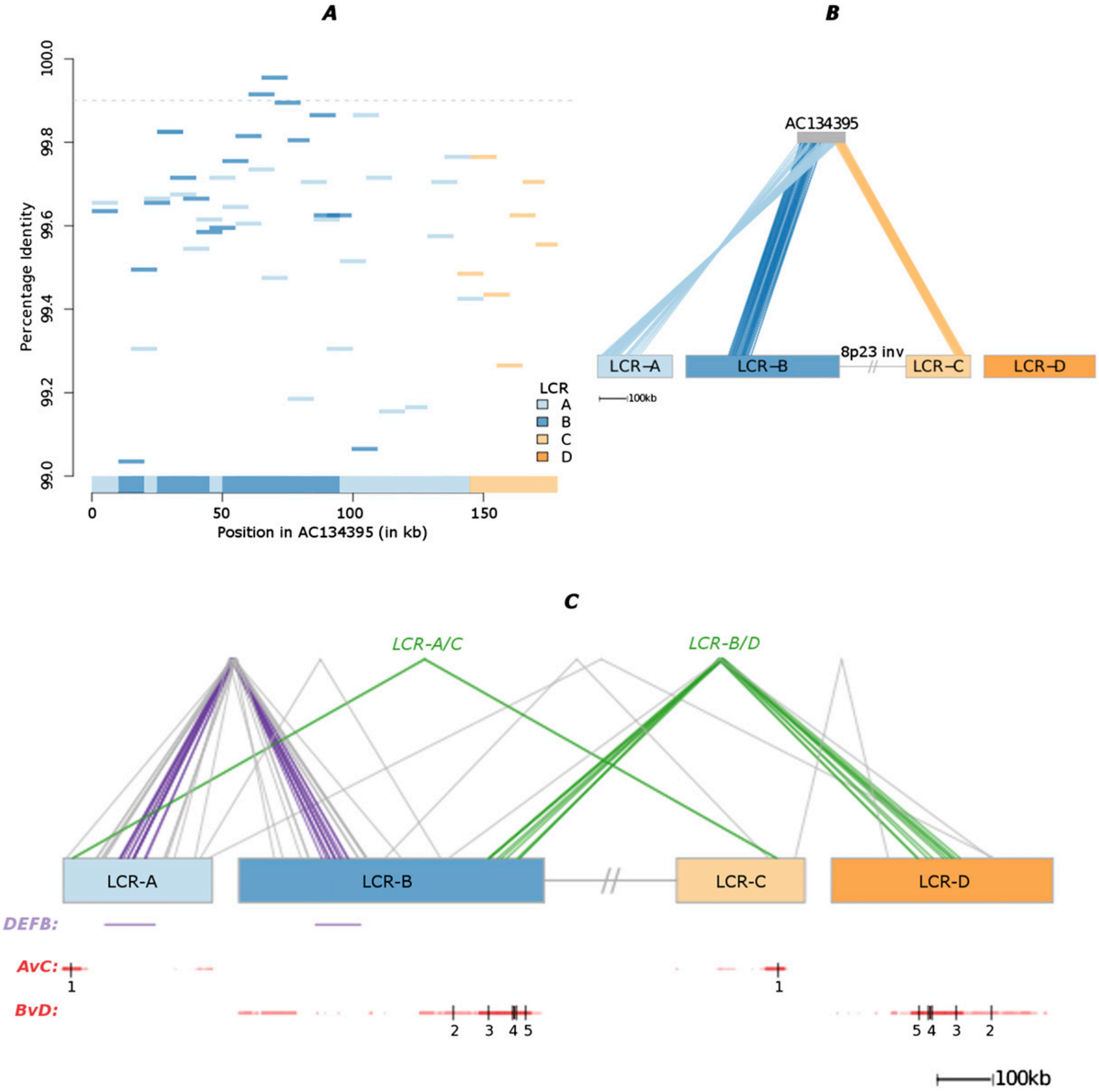

Figure 5. LCR structural diversity captured by sequenced RPCI-11 clones. (*A*) Pairwise alignments (PID > 99%) between LCRs A–D and 10-kb windows of a specified BAC. Each alignment is colored by LCR haplotype, and the dashed line represents PID > 99.9%. A "consensus" is represented on the *x*-axis, in which each 10-kb window is assigned to its most homologous LCR haplotype. In this example, an interleaved mosaic pattern between LCR-A and -B is found in the first 100 kb of the clone, followed by a transition into LCR-C toward the clone's end. (*B*) For the specified BAC, the top pairwise alignment for each 10-kb BAC window (i.e., the consensus) is mapped to its corresponding LCR haplotype (joining lines). The inversion's single-copy region (interrupted gray line) links the LCR-B and -C haplotypes. This example includes a junction between LCR-A and -C. (*C*) LCR haplotype junctions discovered in six finished RPCI-11 BACs. Each transition from one haplotype to another is represented by a joining line. (Purple/green) Those that cluster together and are observed in more than one BAC. The first group (purple, represented by AC134683, AC134395, and AC148106) covers the *DEFB* locus. The second group (green, represented by AC134683 and AC134395) constitutes a junction between LCR-A and LCR-C. The third group (green, involving AC087342, AC092766, AC105214, and AC148106) represents junctions between LCR-B and LCR-D. Notably, AC087342 is anchored by 46 kb of uniquely mapping sequence in the distal end of the inversion. The first annotation track represents the copy-number variant *DEFB* locus (purple); the second and third annotation tracks (red) indicate inverted repeats between LCR-A and -C ("AvC") or LCR-B and -D ("BvD"). Statistically significant recombinant sites (numbered 1–5) (Table 1) are marked by vertical breaks in the inverted repeats. The inversion's single-copy region is represented as in (*B*).

(e.g., Barreiro et al. 2008; Pickrell et al. 2009; Browning and Weir 2010). Whether detection of selection was confounded by *8p23-inv* remains to be determined; for example, estimates of interpopulation differentiation (Novembre and Di Rienzo 2009) may be influenced by random fluctuations of alleles "hitch-hiking" on the stratified inversion, whereas long-range LD-based neutrality tests (Barreiro and Quintana-Murci 2010) may be affected by the inversion's pronounced effect on local LD patterns (O'Reilly et al. 2008). Therefore reassessment of selection at 8p23 loci (controlled for inversion-type) is warranted.

Marked population stratification of *8p23-inv* may explain the ethnic genetic map length differences observed in the region (Wegmann et al. 2011); for example, there is an inverse correlation between African-American and Asian genetic distance for the

Table 1. Statistically significant recombination events in LCR-A/C and LCR-B/D haplotype groups

Event	1	2	3	4	5
Proximal	7,194,095	7,963,585	8,030,787	8,059,039-8,073,159	8,100,317
Distal	12,016,889	12,500,219	12,431,708	12,389,316 – 12,403,443	12,362,178
RDP	5.31×10^{-23}	2.01×10^{-19}	1.26×10^{-3}	1.72×10^{-5}	5.63×10^{-24}
GENECONV	1.20×10^{-4}	1.78×10^{-13}	5.55×10^{-3}	2.73×10^{-15}	2.29×10^{-25}
Bootscan	1.55×10^{-2}	2.01×10^{-19}	8.19×10^{-4}	3.42×10^{-5}	2.25×10^{-27}
Maxchi	7.74×10^{-8}	4.62×10^{-8}	5.58×10^{-3}	1.73×10^{-14}	9.18×10^{-20}
Chimaera	3.86×10^{-6}	3.22×10^{-8}	3.50×10^{-4}	1.69×10^{-10}	2.42×10^{-18}
SiSscan	2.12×10^{-79}	1.38×10^{-12}	2.80×10^{-4}	2.71×10^{-14}	2.69×10^{-29}
3Seq	2.60×10^{-22}	2.00×10^{-14}	6.39×10^{-5}	2.72×10^{-5}	8.74×10^{-12}
Libraries	RPCI-11, RP13, and SCb	RPCI-11 and CTB	RPCI-11 and ABC11	RPCI-11	RPCI-11, RP13, SCb, CTD1, and ABC11

Seven recombination detection methods (Martin et al. 2010) identified five recombination events (corrected *P*-values given), for which physical proximal/distal locations on chromosome 8 are reported (hg18) (see also Fig. 5C). The estimated physical breakpoint location of event 4 was broad, in contrast to the other breakpoints. No definitive "parental" sequences from which the recombinant arose were designated, possibly reflecting missing "parental" sequence data.

breakpoint spanning intervals (Jorgenson et al. 2005; He et al. 2011). Given that the *I* allele is ~4.5× more frequent in Africans than in Asians (Fig. 3A), breakpoint-flanking markers that are physically close on the (noninverted) genome reference map will frequently be separated by a considerable distance in Africans due to the inversion, resulting in (unexpectedly) long genetic map distances. In Asians, the physical distance between the same markers would generally be concordant with the reference map (with correspondingly "normal" genetic distances). Comparison of the intermarker genetic distances between Africans and Asians consequently gives an impression of increased local recombination in Africans (Jorgenson et al. 2005), whereas the converse would be expected for markers distantly separated on the genome reference map.

Like the 17q21.31 inversion (Zody et al. 2008), the *Homo* and *Pan* 8p23 inversions appear to have occurred independently as the divergence time between these lineages significantly predates that estimated for the human *I* and *N* alleles (~200–600 kya); therefore, the *Pan paniscus* inversion allele (Antonacci et al. 2009) is likely to represent an independent inversion event. Although the TMRCA analyses suffice to date *8p23-inv* relative to primate speciation events, the actual age estimates are necessarily imprecise. Not only do TMRCA estimates depend on numerous model assumptions (e.g., minimal gene flow), the construction of deep genealogies for inversions is theoretically restricted by their effective population size (Garrigan and Hammer 2006). The difficulty in precisely estimating TMRCA for inversions is exemplified by the 17q21.31 inversion, for which estimates differ greatly, from 1.9–2.7 mya (Zody et al. 2008) to ~14–108 kya (Donnelly et al. 2010); nevertheless, these estimates suffice to place the 17q21.31 inversion event in the *Homo* lineage.

The timing of REPP formation (Supplemental Fig. S10) coincides with a burst of LCR activity that occurred after the divergence of African great apes and orangutans (Marques-Bonet et al. 2009). This could explain the restriction of *8p23-inv* to the African great ape lineage; once accelerated LCR formation had generated the paired REPP/REPD arrangement, the region was susceptible to NAHR-mediated structural instability. Although a derived inversion allele was not observed in gorillas (Antonacci et al. 2009), this may be attributable to limited sampling ($n = 3$); alternatively, further REPP/REPD rearrangements may have been required prior to inversion formation (e.g., LCR expansion beyond a certain size) (Liu et al. 2011).

In humans, it is unlikely that inversion at 8p23 was a highly recurrent event; this would have eroded the gene-flow barrier between inversion haplotypes, abolishing the observed correlation between inversion-type and genetic substructure. However, a single universal inversion breakpoint was not identified (Supplemental Fig. S18) suggesting some inversion recurrence. Indeed, this might account for the unexpected number of shared polymorphisms observed between *I* and *N* alleles (Supplemental Note). Alternatively, given the structural diversity of the surrounding LCRs (Hollox et al. 2008), successive waves of gene conversion and duplication/deletion may have obscured breakpoint signals beyond reasonable recognition. In this regard, the candidate breakpoint may mark a haplotype strongly correlated with inversion status. Finally, in the absence of uninterrupted haplotype-specific LCR assemblies bridging flanking single-copy sequence, the amount of uncharacterized REPD/REPP sequence remains unclear. Therefore, to reliably resolve additional *8p23-inv* breakpoints, further positionally anchored REPP/REPD assemblies will be invaluable, a task suited to "third-generation" sequencing techniques (Schadt et al. 2010; Alkan et al. 2011b).

The single LCR–haplotype junction associated with *8p23-inv* maps to inverted HERV-K elements, which refines and validates a previously proposed *8p23-inv* breakpoint (Antonacci et al. 2009). Transposable elements such as HERV-K are common NAHR substrates (Jern and Coffin 2008); for example, ~20% of the fixed inversions distinguishing human and chimpanzee genomes are products of NAHR between *Alu* or LINE-1 elements (Lee et al. 2008). Similarly, >16% of human HERV-K elements may have mediated large-scale genome rearrangements during primate evolution (Hughes and Coffin 2001). Retrotransposons promote genomic instability through replication-fork stalling (Zaratiegui et al. 2011), and such a mechanism (mediated by paralogous HERV-K elements) may have generated a common *8p23-inv* allele.

8p23-inv exerts an indirect functional impact by inhibiting meiotic recombination, leading to the preservation of deleterious haplotypes. Indeed, SLE risk alleles were restricted to a haplotype on the derived *N* chromosome (Fig. 4), a haplotype that plausibly accounts for the strong association between *8p23-inv* and *BLK* expression (Ge et al. 2009). Again, a parallel situation exists on the 17q21.31 inversion, in which the "MAPT H1c" haplotype (associated with neurodegenerative disorders and *MAPT* expression) (Pittman et al. 2006) is restricted to the derived noninverted chromosome (Zody et al. 2008).

The inversion is also robustly associated with mRNA levels of other transcripts, particularly *PPP1R3B* whose expression levels influence serum lipid levels in rodents and humans (Gasa et al. 2002; Teslovich et al. 2010). Intriguingly, BLK is also involved in pancreatic β-cell insulin metabolism (Borowiec et al. 2009) and insulin is a key regulator of lipid metabolism (Guilherme et al. 2008), suggesting *8p23-inv* may influence lipid phenotypes via its joint association with *PPP1R3B* and *BLK* expression. Such propositions can now be formally tested using PFIDO.

Methods

All samples were collected with informed consent and approval from relevant institutional review boards. Statistical analyses were performed in R (R Development Core Team 2011) unless otherwise stated. Full descriptions of the enhanced FISH protocol and of SNP genotype data sets are provided in the Supplemental Note.

The PFIDO

The PFIDO (phase free inversion detection operator) algorithm predicts 8p23 inversion-type from diploid SNP genotype data without the need for phase inference (Supplemental Fig. S2). PFIDO first excludes SNPs missing >30% genotype data and then individuals missing >10% data. A pairwise identity-by-state distance matrix is calculated across individuals, using the *snpMatrix* package (Clayton and Leung 2007), followed by transformation with MDS. Outlier samples are identified in each dimension using the *extremevalues* package and are optionally excluded. The derived "axis" (i.e., dimension) displaying the most evidence of substructure is identified using the Shapiro-Wilk test, and the *mclust* package clusters individuals along this axis using a model-based approach (Fraley and Raftery 2006). Specifically, 18 parameterized Gaussian mixture models (1–9 component Gaussian distributions with equal or nonequal variance) are fit by maximum likelihood estimates to the univariate point distribution. The most parsimonious model is selected using the Bayesian information criterion (BIC), and the number of component Gaussian distributions in the chosen model reflects the number of estimated clusters. The conditional probability of an individual belonging to each cluster, given the mixture model parameters, is calculated for each cluster using a *z*-score. Model selection is further assessed by three clustering metrics (connectivity, the Dunn index and silhouette width) via the *clValid* package (Brock et al. 2008); the clustering solution that optimizes these measures is expected to match that selected by BIC. These features allow the user to flexibly define acceptance thresholds for the final clustering result; by selecting a *P*-value threshold, all inversion-type calls in the outcome surpass the chosen level of confidence. PFIDO functions on any R compatible platform and is freely available from http://www.whri.qmul.ac.uk/staff/Shoulders.html.

8p23-inv imputation based on tagging SNPs

Leave-n-out cross-validation analysis was performed using PLINK to impute inversion status based on "tagging" SNPs ($r^2 > 0.8$) (Supplemental Table S2). Each iteration ($n = 1000$) randomly masked one-fourth of the CEU data set for *8p23-inv* status and compared the subset's imputed inversion-types against the experimental data.

Global 8p23 inversion distribution

All populations were seeded with HapMap genotypes from an appropriate reference panel (Supplemental Table S4; Huang et al. 2009; Teo et al. 2009; International HapMap 3 Consortium 2010), serving as internal positive controls and concomitantly supporting correct cluster assignment. Following PFIDO, samples were regrouped as recommended (Donnelly et al. 2010; Xing et al. 2010) to mitigate sampling error. Six populations with sample sizes less than 10 and two populations in which inversion-allele frequency breached HWE (exact test; $P < 0.05$) were excluded.

Geographic coordinates were retrieved from the HGDP-CEPH Database (Li et al. 2008), using the mean where a range was given. Populations in the study by Xing et al. (2010) and in HapMap were geographically assigned to their sampling location, apart from CEU and CHD samples (assigned to Northern Europe and Eastern China, respectively) (Lao et al. 2008; International HapMap 3 Consortium 2010). Chinese SGVP samples were assigned to Southern China (Teo et al. 2009), whereas SGVP Malay and Indian populations were placed in their country of origin's center. Geographic distances from Addis Ababa were calculated as recommended (Handley et al. 2007) using the *geosphere* package. The correlation between geographic distance from Addis Ababa and allele frequency was calculated using the Spearman's rank correlation coefficient. F_{st} values were calculated using the *Hierfstat* package (de Meeus and Goudet 2007). Although negative F_{st} values are possible, they are biologically ill-defined (Barreiro et al. 2008), and so were set to zero.

To construct the empirical null distributions (Hancock et al. 2008), SNPs with similar allele frequency to the inversion (MAF > 0.4) in the HGDP-H952 data set were filtered to produce a set in linkage equilibrium ($r^2 < 0.2$) using PLINK. SNPs were removed within the inversion interval and gene coding regions (i.e., not defined as functionally "unknown" in the UCSC dbSNP130 Table). Rank correlation coefficients and global F_{st} values were calculated for the remaining 19,969 SNPs. Pairwise F_{st} values were calculated for a random subsample of these SNPs ($n = 1000$). Major/minor alleles were designated relative to the ancestral allele.

TMRCA analyses

The minimum TMRCA was estimated using the formula $E[T_{mrca}] = 4N_ef(1 - n_i^{-1})$ as derived for inversions according to the method of Andolfatto et al. (1999), where N_e is the effective population size (CEU = 11,418; YRI = 17,469), n_i is the number of inverted chromosomes, and f is approximated using the number of segregating sites specific to inverted and noninverted alleles, which partially accommodates minimal gene flow. A generation time of 25 yr was assumed. In a complementary analysis, sequence overlaps between finished RPCI-11 BACs were extracted. Those corresponding to alternate haplotypes (PID < 99.999%) (Supplemental Table S6) were aligned to whole-genome shotgun assemblies representing a chimpanzee ("Clint"/CH251, contig NW_001240294.1) and a Sumatran orangutan ("Susie"/ISIS no. 71, contigs NW_002882464.1, NW_002882460.1, NW_002882451.1) using MUSCLE. Assembled HuRef haplotypes (Levy et al. 2007) were similarly aligned to NW_001240294.1. Alignments <10 kb or covering LCRs were excluded. Genetic distances were calculated in MEGA4 (Tamura et al. 2007) using the Kimura 2-parameter method (complete deletion option) (Kimura 1980), and evolutionary rate equality assessed with Tajima's relative rate test (alignments with $P < 0.05$ were discarded) (Tajima 1993). Divergence times between RPCI-11 haplotypes were calculated according to the method describe by Zody et al. (2008) with the formula $T = K/2R$ (where T is time, K the Kimura 2-parameter estimate, and R the average between taxa substitution rate). Either chimpanzee or orangutan were used as an outgroup, and average divergence times were weighted by alignment length.

Analysis of primate REPD/REPP

REPD/REPP (±200 kb) were downloaded from the UCSC database (hg19) and aligned to the nr database using MegaBLAST. All nonhuman primate sequence was retained and remapped back onto the human genome assembly to confirm their mapping to

REPP/REPD. Gorilla BAC end sequence data was downloaded from the Trace Archive (ftp://ftp.ncbi.nih.gov/pub/TraceDB/; download date November 26, 2011).

Identifying recombinant sequences

"Finished" BAC sequences were fragmented into overlapping 10-kb segments and aligned to RPCI-11–specific LCR haplotypes using MegaBLAST (parameters: -p 90 -s 90 -q -3 -r 1 -W 28). Each BAC fragment's optimal alignment against each LCR haplotype was retrieved. Alignments with <99% identity were discarded, which diminished the risk of mistakenly analyzing paralogous LCRs from other cytogenetic intervals, as confirmed by 17 negative control BACs that map to paralogous LCRs on chromosomes 3, 4, 7, 11, and 12.

RPCI-11 clones with evidence of LCR mosaicism were screened for recombination events using seven algorithms implemented in the RDP3 suite (Supplemental Note; Martin et al. 2010). A Bonferroni correction was applied to the *P*-values, and $P < 0.01$ was deemed significant. Recombination events with a defined breakpoint identified by all methods and supported by phylogenetic evidence are reported.

Haplotype-specific PCR

Primers flanking putative breakpoints were designed using Primer3, positioning sequence variants that distinguish BAC haplotype groups at the primer's 3′ end (Supplemental Table S7). PCR was performed in quadruplicate with CEU DNA (Coriell Institute) using AmpliTaq Gold (Applied Biosystems; PCR parameters optimized empirically). To verify product specificity, PCR products were purified with Exonuclease I and SAP and sequenced on an ABI 3730xl using BigDye v3.1 (Applied Biosystems). Association between haplotype presence/absence and inversion-type was tested using an Fisher's exact test, followed by Bonferonni correction for the 12 haplotypes tested.

Acknowledgments

This research was supported by the Medical Research Council and a British Heart Foundation grant. We thank Emma Jones, Margaret Town, Matthew Hurles, and Panos Deloukas for helpful discussions.

References

Alkan C, Coe BP, Eichler EE. 2011a. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12:** 363–376.

Alkan C, Sajjadian S, Eichler EE. 2011b. Limitations of next-generation genome sequence assembly. *Nat Methods* **8:** 61–65.

Andolfatto P, Wall JD, Kreitman M. 1999. Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of *Drosophila melanogaster. Genetics* **153:** 1297–1311.

Andolfatto P, Depaulis F, Navarro A. 2001. Inversion polymorphisms and nucleotide variability in *Drosophila*. *Genet Res* **77:** 1–8.

Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE. 2009. Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet* **18:** 2555–2566.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: How selection shapes host defence genes. *Nat Rev Genet* **11:** 17–30.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet* **40:** 340–345.

Borowiec M, Liew CW, Thompson R, Boonyasrisawat W, Hu J, Mlynarski WM, El Khattabi I, Kim S, Marselli L, Rich SS, et al. 2009. Mutations at the *BLK* locus linked to maturity onset diabetes of the young and β-cell dysfunction. *Proc Natl Acad Sci* **106:** 14460–14465.

Bosch N, Morell M, Ponsa I, Mercader JM, Armengol L, Estivill X. 2009. Nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism. *PLoS ONE* **4:** e8269. doi: 10.1371/journal.pone.0008269.

Brock G, Pihur V, Datta S. 2008. *clValid*, an R package for cluster validation. *J Stat Softw* **25:** 1–22.

Broman K, Matsumoto N, Giglio S, Martin C, Roseberry J, Zuffardi O, Ledbetter D, Weber JL. 2003. Common long human inversion polymorphism on chromosome 8p. *In Science and statistics: A festschrift for Terry Speed. Institute of Mathematical Statistics Lecture Notes Monograph Series* (ed. D.R. Goldstein), pp. 237–245. Institute of Mathematical Statistics, Bethesda.

Browning SR, Weir BS. 2010. Population structure with localized haplotype clusters. *Genetics* **185:** 1337–1344.

Chung SA, Taylor KE, Graham RR, Nititham J, Lee AT, Ortmann WA, Jacob CO, Alarcon-Riquelme ME, Tsao BP, Harley JB, et al. 2011. Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS Genet* **7:** e1001323. doi: 10.1371/journal.pgen.1001323.

Clayton D, Leung HT. 2007. An R package for analysis of whole-genome association studies. *Hum Hered* **64:** 45–51.

Dehghan A, Dupuis J, Barbalic M, Bis JC, Eiriksdottir G, Lu C, Pellikka N, Wallaschofski H, Kettunen J, Henneman P, et al. 2011. Meta-analysis of genome-wide association studies in >80,000 subjects identifies multiple loci for C-reactive protein levels. *Circulation* **123:** 731–738.

de Meeus T, Goudet J. 2007. A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infect Genet Evol* **7:** 731–735.

Deng Y, Tsao BP. 2010. Genetic susceptibility to systemic lupus erythematosus in the genomic era. *Nat Rev Rheumatol* **6:** 683–692.

Deng L, Zhang Y, Kang J, Liu T, Zhao H, Gao Y, Li C, Pan H, Tang X, Wang D, et al. 2008. An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. *Hum Mutat* **29:** 1209–1216.

Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Mehdi SQ, Kajuna SL, Barta C, Kungulilo S, Karoma NJ, Lu RB, et al. 2010. The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am J Hum Genet* **86:** 161–171.

Feuk L. 2010. Inversion variants in the human genome: role in disease and genome architecture. *Genome Med* **2:** 11.

Fraley C, Raftery AE. 2006. *MCLUST version 3 for R: Normal mixture modeling and model-based clustering*. Technical report no. 504, Department of Statistics, University of Washington, Seattle.

Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. *Nat Rev Genet* **7:** 669–680.

Gasa R, Clark C, Yang R, DePaoli-Roach AA, Newgard CB. 2002. Reversal of diet-induced glucose intolerance by hepatic expression of a variant glycogen-targeting subunit of protein phosphatase-1. *J Biol Chem* **277:** 1524–1530.

Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KC, Gagne V, et al. 2009. Global patterns of *cis* variation in human cells revealed by high-density allelic expression analysis. *Nat Genet* **41:** 1216–1222.

Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R, et al. 2001. Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am J Hum Genet* **68:** 874–883.

Girirajan S, Eichler EE. 2010. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum Mol Genet* **19:** R176–R187.

Graham RR, Cotsapas C, Davies L, Hackett R, Lessard CJ, Leon JM, Burtt NP, Guiducci C, Parkin M, Gates C, et al. 2008. Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nat Genet* **40:** 1059–1061.

Gregersen PK, Amos CI, Lee AT, Lu Y, Remmers EF, Kastner DL, Seldin MF, Criswell LA, Plenge RM, Holers VM, et al. 2009. REL, encoding a member of the NF-κB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat Genet* **41:** 820–823.

Guilherme A, Virbasius JV, Puri V, Czech MP. 2008. Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes. *Nat Rev Mol Cell Biol* **9:** 367–377.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet* **4:** e32. doi: 10.1371/journal.pgen.0040032.

Handley LJ, Manica A, Goudet J, Balloux F. 2007. Going the distance: Human population genetics in a clinal world. *Trends Genet* **23:** 432–439.

Harley IT, Kaufman KM, Langefeld CD, Harley JB, Kelly JA. 2009. Genetic susceptibility to SLE: New insights from fine mapping and genome-wide association studies. *Nat Rev Genet* **10:** 285–290.

He C, Weeks DE, Buyske S, Abecasis GR, Stewart WC, Matise TC, Enhanced Map Consortium. 2011. Enhanced genetic maps from family-based disease studies: Population-specific comparisons. *BMC Med Genet* **12:** 15. doi: 10.1186/1471-2350-12-15.

Hoffmann AA, Rieseberg LH. 2008. Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annu Rev Ecol Evol Syst* **39:** 21–42.

Hollox EJ, Barber JC, Brookes AJ, Armour JA. 2008. Defensins and the dynamic genome: What we can learn from structural variation at human chromosome band 8p23.1. *Genome Res* **18:** 1686–1697.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: Defining, estimating and interpreting F_{ST}. *Nat Rev Genet* **10:** 639–650.

Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, Garnier S, Lee AT, Chung SA, Ferreira RC, Pant PV, et al. 2008. Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. *N Engl J Med* **358:** 900–909.

Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. 2009. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* **84:** 235–250.

Hughes JF, Coffin JM. 2001. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nat Genet* **29:** 487–489.

Iles MM. 2008. Quantification and correction of bias in tagging SNPs caused by insufficient sample size and marker density by means of haplotype-dropping. *Genet Epidemiol* **32:** 20–28.

International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467:** 52–58.

Jern P, Coffin JM. 2008. Effects of retroviruses on host genome function. *Annu Rev Genet* **42:** 709–732.

Jiang Z, Hubley R, Smit A, Eichler EE. 2008. DupMasker: A tool for annotating primate segmental duplications. *Genome Res* **18:** 1362–1368.

Johnson ME, National Institute of Health Intramural Sequencing Center Comparative Sequencing Program, Cheng Z, Morrison VA, Scherer S, Ventura M, Gibbs RA, Green ED, Eichler EE. 2006. Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad Sci* **103:** 17626–17631.

Jorgenson E, Tang H, Gadde M, Province M, Leppert M, Kardia S, Schork N, Cooper R, Rao DC, Boerwinkle E, et al. 2005. Ethnicity and human genetic linkage maps. *Am J Hum Genet* **76:** 276–290.

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477:** 203–206.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16:** 111–120.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, et al. 2008. Correlation between genetic and geographic structure in Europe. *Curr Biol* **18:** 1241–1248.

Lee J, Han K, Meyer TJ, Kim HS, Batzer MA. 2008. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS ONE* **3:** e4047. doi: 10.1371/journal.pone.0004047.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5:** e254. doi: 10.1371/journal.pbio.0050254.

Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, Glazer NL, Morrison AC, Johnson AD, Aspelund T, et al. 2009. Genome-wide association study of blood pressure and hypertension. *Nat Genet* **41:** 677–687.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100–1104.

Liu P, Lacaria M, Zhang F, Withers M, Hastings PJ, Lupski JR. 2011. Frequency of nonallelic homologous recombination is correlated with length of homology: Evidence that ectopic synapsis precedes ectopic crossing-over. *Am J Hum Genet* **89:** 580–588.

Lo YM, Patel P, Newton CR, Markham AF, Fleming KA, Wainscoat JS. 1991. Direct haplotype determination by double ARMS: Specificity, sensitivity and genetic applications. *Nucleic Acids Res* **19:** 3561–3567.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang SP, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orangutan genomes. *Nature* **469:** 529–533.

Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M. 2009. Genetic variation and recent positive selection in worldwide human populations: Evidence from nearly 1 million SNPs. *PLoS ONE* **4:** e7888. doi: 10.1371/journal.pone.0007888.

Lowry DB, Willis JH. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol* **8:** e1000500. doi: 10.1371/journal.pbio.1000500.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457:** 877–881.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* **26:** 2462–2463.

Myers AJ, Pittman AM, Zhao AS, Rohrer K, Kaleem M, Marlowe L, Lees A, Leung D, McKeith IG, Perry RH, et al. 2007. The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. *Neurobiol Dis* **25:** 561–570.

Nordmark G, Kristjansdottir G, Theander E, Appel S, Eriksson P, Vasaitis L, Kvarnstrom M, Delaleu N, Lundmark P, Lundmark A, et al. 2011. Association of EBF1, FAM167A(C8orf13)-BLK and TNFSF4 gene variants with primary Sjogren's syndrome. *Genes Immun* **12:** 100–109.

Novembre J, Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. *Nat Rev Genet* **10:** 745–755.

O'Reilly PF, Birney E, Balding DJ. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res* **18:** 1304–1313.

Osoegawa K, Vessere GM, Li Shu C, Hoskins RA, Abad JP, de Pablos B, Villasante A, de Jong PJ. 2007. BAC clones generated from sheared DNA. *Genomics* **89:** 291–299.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* **19:** 826–837.

Pittman AM, Fung HC, de Silva R. 2006. Untangling the tau gene association with neurodegenerative disorders. *Hum Mol Genet* **15:** R188–R195.

Raap AK. 1998. Advances in fluorescence in situ hybridization. *Mutat Res* **400:** 287–298.

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

Romano CM, Ramalho RF, Zanotto PM. 2006. Tempo and mode of ERV-K evolution in human and chimpanzee genomes. *Arch Virol* **151:** 2215–2228.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, et al. 2011. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **39:** D38–D51.

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Hum Mol Genet* **19:** R227–R240.

Sugawara H, Harada N, Ida T, Ishida T, Ledbetter DH, Yoshiura K, Ohta T, Kishino T, Niikawa N, Matsumoto N. 2003. Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23. *Genomics* **82:** 238–244.

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135:** 599–607.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24:** 1596–1599.

Teo YY, Sim X, Ong RT, Tan AK, Chen J, Tantoso E, Small KS, Ku CS, Lee EJ, Seielstad M, et al. 2009. Singapore Genome Variation Project: A haplotype map of three Southeast Asian populations. *Genome Res* **19:** 2154–2162.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466:** 707–713.

Wegmann D, Kessner DE, Veeramah KR, Mathias RA, Nicolae DL, Yanek LR, Sun YV, Torgerson DG, Rafaels N, Mosley T, et al. 2011. Recombination rates in admixed individuals identified by ancestry-based inference. *Nat Genet* **43:** 847–853.

Weiler KS, Wakimoto BT. 1995. Heterochromatin and gene expression in *Drosophila*. *Annu Rev Genet* **29:** 577–605.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661–678.

Xing J, Watkins WS, Shlien A, Walker E, Huff CD, Witherspoon DJ, Zhang Y, Simonson TS, Weiss RB, Schiffman JD, et al. 2010. Toward a more uniform sampling of human genetic diversity: A survey of worldwide populations by high-density genotyping. *Genomics* **96:** 199–210.

Zaratiegui M, Vaughn MW, Irvine DV, Goto D, Watt S, Bahler J, Arcangioli B, Martienssen RA. 2011. CENP-B preserves genome integrity at replication forks paused by retrotransposon LTR. *Nature* **469:** 112–115.

Zody MC, Jiang Z, Fung HC, Antonacci F, Hillier LW, Cardone MF, Graves TA, Kidd JM, Cheng Z, Abouelleil A, et al. 2008. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet* **40:** 1076–1083.

Received May 9, 2011; accepted in revised form March 6, 2012.