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

Maximilian P.A. Salm, Stuart D. Horswell, Claire E. Hutchison, Helen E. Speedy, Xia Yang, Liming Liang, Eric E. Schadt, William O. Cookson, Anthony S. Wierzbicki, Rossi P. Naoumova, and Carol C. Shoulders

Supplemental Note

Table of Contents

Optimal SNPs for PFIDO

Gene Flow within *8p23-inv*

Gene Expression and *8p23-inv*

8p23-inv in Genome Assemblies

8p23 LCR re-assembly

Fosmid-end sequences support the LCR-haplotype junctions

Supplementary Methods

References

Optimal SNPs for PFIDO

Currently, high-throughput genotyping commonly focuses on a subset of HapMap tag-SNPs in a test sample (Wellcome Trust Case Control Consortium 2007). To characterize the SNP data requirements for PFIDO, the algorithm was run iteratively in cumulative 25-SNP steps, comparing inversion prediction accuracy and the three internal measures of cluster validity. Specifically, FISH-derived inversion status was coded as a bialleleic SNP, and LD between this marker and unphased HapMap SNP genotypes mapping to the inversion were calculated in PLINK (Purcell et al. 2007). PFIDO was then run iteratively in cumulative 25-SNP steps through the r^2-ranked SNP list, comparing the outcomes by number of correctly predicted inversion-types and three internal measures of cluster validity (C/D/S), selecting the input SNPs that optimized all four metrics as defined by the *BruteAggreg* function (Pihur et al. 2009).

In the 54 CEU founders (Fig. 1), 802 SNPs are optimal for inversion-type prediction (Supplemental Fig. S3 & Table S2). Additional SNPs appear to reduce clustering quality, potentially by adding noise to the dataset. Although accurate inversion-type classification is possible with as few as 50 SNPs, the underlying clustering appears less robust (Supplemental Fig. S3). Increasing sample size to 110 by including HapMap Phase III CEU founders allows effective PFIDO function with 123 SNPs (Supplemental Fig. S3 & Table S2), indicating that a sample size increase improves algorithm performance. Importantly, the specific SNP subset input is relatively flexible in content: of 4600 randomly sampled 123-SNP sets, 74% achieved perfect inversion-type classification at $p < 0.05$. Moreover, the data suggest that the SNP genotyping burden is reduced by a sample size increase.

Optimal PFIDO performance with YRI samples was achieved by including HapMap phase III samples, and reducing the number of input SNPs from 6506 to 2039 (Supplemental Table S2). To explore whether SNP ascertainment bias (Rosenberg et al. 2010) influenced PFIDO in this sample, genotypes for an extra 4267 SNPs (MAF > 0.01 and in HWE ($p > 0.001$)) were retrieved from the 1000 genome project (1000 Genomes Project Consortium et al. 2010) (pilot 1, version 23/06/2010). However, no significant improvement in cluster resolution was found (data not shown), suggesting that PFIDO's performance may not benefit from further YRI SNP sampling. These data also indicate that genetic differentiation between YRI *8p23-inv* alleles is less pronounced than in the CEU/JPT samples, which is supported by population-specific multilocus F_{st} estimates between phased *I* and *N* HapMap SNP haplotypes (F_{st} = 0.39 & 0.16 in CEU & YRI respectively). Therefore, it appears that the inversion "LD signal" is stronger in the CEU than the YRI, although it is still detectable in the YRI by increasing sample size and reducing noise from uninformative SNPs.

Gene Flow within *8p23-inv*

Genetic maps frequently exhibit suppressed recombination in the inversion region (Jorgenson et al. 2005; He et al. 2011; Wegmann et al. 2011). Theoretically, reduced recombination permits the accumulation of inter-marker LD and consequently structured genetic variation; it is the detection of discrete genetic variation substructure that forms the basis of PFIDO. The paucity of effective *8p23-inv* tagging SNPs, however, suggests the historic transfer of genetic variation (i.e. "gene flow") between inversion-types.

To characterise allele sharing between inversion-types, we explored HapMap genotype data in 40 *II* and 17 *NN* CEU samples. Under the simplifying assumptions that *8p23-inv* formation occurred once during human evolution (i.e. there was a single ancestral founder inversion) and that the inversion prohibits recombination between *I* and *N* haplotypes, genetic polymorphisms would be expected to segregate on either *I* or *N* haplotypes but not both. Of the 5425 SNPs analysed, 1473 were only polymorphic in *II* samples, while 596 were only polymorphic in *NN* samples. However, 3356 SNPs segregated in both *II* and *NN* samples. Even though a proportion of these are likely attributable to recurrent mutation (Hodgkinson et al. 2011) and genotyping error, the large proportion of shared SNPs suggests that the inversion was not a unique event and/or there has been limited gene flow. Nevertheless, compared to 10,000 randomly selected SNPs, significantly fewer polymorphisms are shared between inversion-types ($p < 2.2 \times 10^{-16}$, chi-squared test).

Furthermore, shared polymorphisms exhibit significantly higher F_{st} values than 10,000 randomly selected SNPs (mean $F_{st} = 0.2$, $p < 2.2 \times 10^{-16}$, Wilcoxon rank sum test), indicative of marked genetic differentiation between inversion-types and contrary to a hypothesis of substantial gene flow of these shared polymorphisms.

Viable recombinant haplotypes can be generated in inversion heterozygotes if an inversion loop is formed and an even number of crossovers occur within the loop. To explore this avenue of potential gene flow in *8p23-inv*, we first determined contemporary recombination rates relative to PFIDO-predicted inversion-type. Specifically, meiotic recombination events were identified within *8p23-inv* in two-generation families with at least two children from a densely genotyped European cohort (Dixon et al. 2007; Moffatt et al. 2007) using a previously implemented method (Chowdhury et al. 2009). To diminish the influence of cryptic genotyping errors, each recombination event was supported by ≥5 consecutive "informative" SNPs (a SNP where one parent is homozygous in one parent but heterozygous in the other parent). Parental crossover events for 116 nuclear families were mapped at high-resolution using 449 inversion-specific SNPs with an algorithm that detects "parental phase switches" (Chowdhury et al. 2009). Nineteen crossovers were detected within the inversion for 264 meioses, which equates to a genetic distance of 7.20cM and is equivalent to the inversion's length on the Marshfield genetic map (7.19cM; Broman et al. 1998). However, no crossover events were detected in inversion heterozygotes ($n = 100$, $p = 0.003$, fisher's exact test, H_0; for *II/IN/NN*, observed rate = expected rate). Therefore no double recombinants mediating inter-inversion gene flow were detected at the level of a single generation.

An alternative to identifying recombinants in pedigrees is to identify ancestral "admixture" events between *I* and *N* chromosomes, allowing recombination events to be identified over multiple generations (Hinch et al. 2011). To implement this, we applied HAPMIX (Price et al. 2009b) to estimate the number of *N*- or *I*-derived alleles at each SNP in 1378 *8p23-inv* heterozygotes (Supplemental Fig. S8A). We examined the chr8: 4Mb – 15Mb interval using 529 *II* samples and 529 *NN* samples as "reference" populations, the HapMap genetic map, a prior hypothesis of 50% *I* ancestry and 6 generations since admixture per individual. Only confidently inferred cross-over events were retained as described (Hinch et al. 2011), except that instead of removing switch points flanked by blocks < 2cM we required switch points to be separated by >500kb. This procedure mapped 45 double crossover events (and no single crossovers) to the inversion interval. The estimated crossover positions largely map around a central segment of the inversion (Supplemental Fig. 8B), supporting theoretical predictions of gene flow due to double crossovers in inversion loops (Navarro et al. 1997). It seems noteworthy that gene flow appears to be asymmetric, with 42 of the recombinant segments being of *N* ancestry (Supplemental Fig. 8C). Given that double recombinants were only detected in ~3% of the sample, and that recombinants are inferred over multiple generations, it is reasonable to conclude that double recombinants are rare events; this would account for the shared polymorphisms between *I* and *N* haplotypes even though genetic substructure between haplotypes is generally preserved.

Gene Expression and *8p23-inv*

To investigate the influence of *8p23-inv* on local gene expression, five gene expression datasets were re-analysed with respect to inversion status. In addition to providing inter-study replication, each individual dataset offers distinct advantages. The first dataset comprises expression levels derived from 165 HapMap LCLs representing 60 YRI, 60 CEU, and 45 JPT samples (Stranger et al. 2007), and thus allows inter-population comparisons. The second dataset focuses on allele-specific expression in 53 CEU LCLs (Ge et al. 2009) and allows direct assessment of the inversion's effect on expression in *cis*. The third (Dixon et al. 2007; Moffatt et al. 2007) and fourth datasets (W.O. Cookson & M. Moffatt, unpublished) were generated from LCLs representing large British cohorts (n=395 & n=550, respectively) and are therefore statistically well powered to detect expression quantitative trait loci (eQTLs). Finally, the well-powered fifth dataset (n=372) is derived from post-mortem liver samples, allowing eQTL analysis in another tissue (Schadt et al. 2008). Although some concordance between studies is expected, perfect overlap is not, due to the multiple technical differences between studies (Draghici et al. 2006).

For the analyses, inversion status was assigned by PFIDO ($p < 0.05$), using SNP genotype data merged with HapMap CEU SNP data (rel. 27). Association analyses were restricted to transcripts on, or within 1Mb, of the inversion. The Holm step-down procedure (Manly et al. 2004) was applied to assess significance at $p < 0.05$.

The HapMap LCL dataset (n=270; GSE6536) comprised normalized $\log_2$ transformed mRNA expression values (Stranger et al. 2007). mRNAs with median $\log_2$ expression values < 6.4 were discarded (Johnston et al. 2008). For each population sample, linear regression was performed between normalized expression values and inversion status in unrelated samples using the *lm* function in R, assuming an additive genetic model. P-values were determined using ANOVA, testing the null hypothesis of no association.

The alleleic expression (AE) dataset derives from unspliced primary transcripts prepared from 53 unrelated HapMap CEU LCL samples (Ge et al. 2009). The data (http://www.genomequebec.mcgill.ca/publications/pastinen/) represents quantitative measurements for expressed SNPs, normalized to quantitative measurements for the same SNPs in the sample's corresponding genomic DNA. PFIDO-defined inversion-types were assigned to specific chromosomes via hierarchical clustering of pairwise distance between phased HapMap (rel.27) *8p23-inv* regions using the ape package (Paradis et al. 2004). Sixty inversion-localized "AE informative windows" (i.e. a collection of SNPs representing an expressed region) were defined and analysed as described (Ge et al. 2009).

A third dataset (Dixon et al. 2007; Moffatt et al. 2007) (GSE8052) comprising 778 SNP genotypes mapping to the inversion in addition to normalized gene-expression level data for 395 European LCLs was analyzed for association as described (Dixon et al. 2007).

A post-mortem liver dataset (Schadt et al. 2008) (GSE9588), representing 372 Caucasians, comprised 1421 genotyped SNPs with corresponding gene-expression levels for 94 transcripts mapping to the 8p23 inversion, normalized for age, sex and sample collection site. Association analysis was performed as described (Schadt et al. 2008).

The final dataset represents 950 individuals from 320 families of British descent (W.O. Cookson & M. Moffatt, unpublished data): subjects with asthma (n=347) and/or atopic dermatitis (n=487; 259 with both diseases) were genotyped using the Illumina Sentrix HumanHap300 BeadChip. Of the 314,552 annotated SNPs (UCSC genome browser; hg18), 8,345 were excluded due to low genotyping success rate (< 95%) or deviation from HWE ($P<1x10^{-6}$): this left 306,207 SNPs (296,533,535 genotypes; 99.1% call rate) available for further analyses. Genotype calls with non-Mendelian inheritance patterns were removed (Wigginton et al. 2005a). RNA was prepared from LCLs representing 550 samples as described (Dixon et al. 2007), and hybridized to Illumina Human-6 BeadChips; these samples represented atopic dermatitis probands and their siblings, of which 496 had been genotyped. Expression values were estimated using the Illumina BeadStudio and bead summary data were used for downstream analysis. 16,487 probes were retained for analysis, after excluding 30,806 probes (out of 47,293) called as "absent" (detection score < 0.95) in more than 80% arrays. The data were then normalized using quantile normalization (Bolstad et al. 2003) and analyzed for association as described (Dixon et al. 2007).

In all four European LCL-derived datasets a highly significant and consistent association was found between inversion-type and *BLK* transcript abundance ($p_{min} = 7.12x10^{-9}$; Supplemental Table S5). Given that *BLK* is not expressed in the liver (Su et al. 2004), no association between *BLK* transcript levels and inversion status was expected in the liver samples, nor was any found (p = 0.89). The number of *N* alleles correlated with decreased *BLK* expression level (Supplemental Table S5) but this relationship may be population-specific as it was not evident in the YRI or JPT samples.

Inversion-type was also significantly associated with *PPP1R3B* mRNA levels in both LCL and liver datasets, with *N* allele dosage positively correlating with transcript abundance (Supplemental Table S5 and Fig. S7). This trend is consistent across populations, being observed in CEU, YRI and JPT samples (Supplemental Fig. S7), although the association signal is lost in the JPT following multiple-testing correction suggesting that this sample is statistically underpowered to detect this association robustly.

Other replicated inversion-eQTLs include *XKR6*, *FAM167A* and the first intron of *CTSB* (Supplemental Table S5). The inversion's association with levels of *XKR6*-related transcripts (AB073660 & AJ305312) is in fact the strongest found in the large British LCL datasets (Dixon et al. 2007) ($p = 3.4x10^{-14}$ & $p = 2.2x10^{-18}$). *FAM167A* mRNA levels were significantly positively correlated with *N* allele dosage, and this trend is mirrored in the CEU, YRI and JPT samples, although the associations do not retain statistical significance after multiple-testing correction in this dataset (Stranger et al. 2007) (data not shown). Finally, the *CTSB* inversion-eQTL was only present in CEU LCLs (Stranger et al. 2007; Ge et al. 2009), exhibiting a positive correlation between *N* allele dosage and mRNA levels (Supplemental Table S5).

8p23-inv in Genome Assemblies

Determining the 8p23 inversion-type represented in reference genome assemblies could aid in inversion breakpoint refinement within the flanking LCRs (Zody et al. 2008). PFIDO can enable this: for example, the HuRef (Levy et al. 2007) donor is predicted to be *IN* ($p=3.84x10^{-9}$, HuRef SNP genotype data retrieved from the UCSC genome browser). However, the HuRef-derived assemblies are generally lacking in LCR coverage (Levy et al. 2007) due to the production strategy used (random shotgun sequencing and *de novo* assembly). In fact, LCRs can currently only be resolved reliably using very high quality sequence (i.e. with an error rate significantly lower than the polymorphism rate) from haplotype-specific large-insert clones (Eichler et al. 2004; International Human Genome Sequencing Consortium 2004).

To this end, we focused on "finished" sequence data (≤ 1 error in 10,000 bases (International Human Genome Sequencing Consortium 2004)) from the redundantly sequenced RPCI-11 BAC library, which represents a single anonymous male (Osoegawa et al. 2001) of probable African American ancestry (Green et al. 2010), and whose sequence constitutes ~70% of the inversion reference assembly (NT_077531, hg18). Notably FISH-based inversion-typing is precluded by the absence of a transformed donor cell line (Osoegawa et al. 2001), necessitating an alternative approach.

The RPCI-11 BACs encompassing *8p23-inv* and its surrounding regions (from 5.5-17.1 Mb) are predominantly (96%) of African ancestry (Green et al. 2010) and so inversion-type was predicted relative to HapMap YRI samples. Insufficient SNP genotype data (n=209) existed for direct PFIDO use, but BAC haplotypes could be assigned to PFIDO-predicted *II* or *NN* categories using HAPMIX (Price et al. 2009a). Specifically, dbSNP fasta entries were aligned to "finished" (International Human Genome Sequencing Consortium 2004) inversion-mapping BACs using MegaBLAST (Johnson et al. 2008). After parsing the alignments in Biopython (Cock et al. 2009) to retrieve BAC-specific SNP alleles, the BAC haplotypes were individually analysed by HAPMIX, using phased YRI haplotypes (HapMap rel. 27) as a reference with related recombination rates. The two "parental" populations comprised either *II* or *NN* individuals (as determined by PFIDO), and we assumed a 50% prior probability of ancestry from either group.

Three clones were allocated to the *I* (AC011008, AC090790, AC023385) or *N* (AC022239, AC025857, AC069185) categories with >90% confidence. Furthermore, in four cases this confidence exceeded 99% (AC011008, AC022239, AC025857, AC069185). This suggests that the RPCI-11 donor was heterozygous for *8p23-inv*. It is noteworthy that PFIDO analysis with HapMap CEU samples as a reference yields the same outcome (*IN*; p = 5.92 x10^{-6}). Therefore, investigation of RPCI-11 sequence to reveal precise inversion breakpoints is feasible.

8p23 LCR re-assembly

The inversion is flanked by two highly homologous LCRs (REPD and REPP) that harbor multiple copy-number variants. This structural diversity, compounded with extensive stretches of near-perfect homology between LCRs has complicated genome assembly in the region (Taudien et al. 2004), potentially explaining the assembly gaps that currently interrupt both REPD and REPP (Zhang et al. 2009). However, using high-quality BAC sequence from a single inversion heterozygote (RPCI-11) reduces the complexity of 8p23 LCR assembly and diminishes the risk of paralogue-mediated mis-assembly. To identify all BACs eligible for re-assembly (i.e. of likely REPD/REPP origin), we applied a strategy based on “feature similarity” that clustered BACs by LCR sub-unit homology. First, duplicons with >95% identity to REPP/REPD were identified in a database representing ancestral duplicons (Jiang et al. 2007); these were individually re-aligned using BLAST (Altschul et al. 1990) to all “finished” RPCI-11 BACs, to retrieve all RPCI-11 sequences of a specific duplicon. The following algorithm was subsequently applied to each duplicon sequence set:

1. Create a multiple sequence alignment (MSA) for the duplicon set using MUSCLE (Edgar 2004).
2. Transform the MSA into a Neighbour-Joining tree (based on Kimura-2 parameter genetic distance)
3. Retrieve the inter-tip branch distances
4. Apply hierarchical clustering to the inter-tip branch distances and bootstrap (x1000, using the *pvclust* package (Suzuki et al. 2006))
5. Record all clusters with an “approximately unbiased” p-value < 0.01.

Having repeated this operation for all duplicons, the pairwise co-occurrence of BACs within each significant recorded cluster was calculated; a crude multiple-testing adjustment was applied by randomly removing 1% of all pairwise co-occurrences. This data was represented as a network (Supplemental Fig. S12), which exhibits two clear communities; all those previously assigned to chromosome 8 by the BAC submitters group together while those assigned to other chromosomes form a second group.

Using the RPCI-11 BACs from the “chromosome 8 group”, REPD and REPP were stringently re-assembled into tiling paths. First, pairwise alignments of RPCI-11 clone sequences were constructed using nucmer (–maxmatch) in the MUMmer package (Kurtz et al. 2004). Lower-quality alignments (mismatch > 0.5% and length < 5000 bp) were removed, before retrieving the best overlaps (defined as a function of sequence similarity and alignment length) for each BAC sequence, effectively organizing library-specific BACs into region-specific tiling paths. Sequence contents of clone versions were verified for completeness by BLASTing successive accessioned versions against the used version. This identified three partial submissions (AC087203.12, AC068353.34 & AC105214.6), in which only the first 105.3-158.2kb were submitted as the remainders overlap other sequenced clones; un-submitted sequence from previous accession versions was not used to extend these sequences.

Tiling paths were subjected to a further quality control procedure. After masking sequence intervals annotated as "unsure" (Schmutz et al. 2004) using the SeqinR package (Charif et al. 2005), but leaving low-complexity regions and high-copy repeats un-masked, each BAC sequence overlap was stringently re-aligned using a 0.02% mismatch threshold (MegaBLAST –p 99.98 –s 90). By only allowing a <0.02% sequence mismatch, which is below the reported ~0.1% SNP rate (International Human Genome Sequencing Consortium 2004), the resulting assemblies are expected to faithfully represent the underlying physically contiguous sequence (Zody et al. 2008).

This produced two assemblies mapping to REPD (named LCR-A & B) and two assemblies mapping to REPP (LCR-C & D; Supplemental Fig. S13). These broadly mirror the existing reference genome assembly, except for the exclusion of non-RPCI-11 data (Supplemental Fig. S13). Other RPCI-11 clones also support sections of the LCR haplotype assemblies: for example AC130367 corroborates LCR-A, whilst AC131269 and AC144950 support LCR-B. Seven out of nine tiling-path BAC overlaps matched perfectly (Supplemental Fig. S13v), with the single nucleotide mismatch between AC068020 and AC105233 in LCR-B attributable to lower confidence sequence ("single clone coverage"), and the imperfect alignment between AC087203 & AC130352 in LCR-D attributable to a single di-nucleotide insertion in an $(AT)_{18}$ repeat. Given that this is the only sequence difference in the 75,155 bp overlap, this is likely to be an artifact; comparison with draft sequence (AC138201) suggests the error is in AC130352.

To be confident in the chromosomal location of LCR haplotypes, assembly should extend from positionally unambiguous sequence (Bailey et al. 2006). LCRs-B, -C and -D are all anchored in single-copy sequence (Supplemental Fig. S13). Using HAPMIX as described in the previous section and 1000 Genome Project data from YRI *II* and *NN* samples as a reference, LCR-B is likely to represent an *I* background while LCR-C is likely to represent an *N* background. No sequence was found that links LCR-A to a single-copy sequence although it is most similar to LCRs B-D at the genome-wide level (PID > 99.5%), with highest similarity to LCR-B (Supplemental Fig. S14), which suggests that it is allocated to the correct cytogenetic interval. Consistent with previous reports (Sugawara et al. 2003), the four RPCI-11 LCR haplotypes exhibit significant homology to one another, containing multiple large inverted repeats that could sponsor inversion formation through NAHR (Supplemental Fig. S14).

Fosmid-end sequences support the LCR-haplotype junctions

As a secondary screen for human BACs containing recombination events, fosmid end-sequence pairs (ESPs) generated from 15 ancestrally diverse HapMap samples (Kidd et al. 2010) of known *8p23-inv* genotype (Antonacci et al. 2009) (Supplemental Table S1) were aligned to all finished human BACs; paired-ends discordant in orientation when mapped to a reference assembly (thus potentially spanning an inversion breakpoint) were predicted to align in their correct orientations to breakpoint-harbouring BACs. Specifically, fosmid paired-end sequence data were retrieved from the NCBI trace archive (http://www.ncbi.nlm.nih.gov/Traces), and aligned to the RPCI-11 LCR haplotypes (MegaBlast –p 98 –s 90, alignment length > 400bp). End-sequence pairs (ESPs) with unique map positions that aligned to the same strand of the reference were re-aligned to all human-derived BACs in the nr/nt database (MegaBLAST –q -4 –r 1 –e 10). This approach mitigates mistaken inferences resulting from alignment to potentially mis-assembled references, a particular problem in LCR regions (Eichler et al. 2004; Feuk 2010). Moreover, to reduce false positive signals arising from LCR-mediated ambiguity in mapping (Feuk 2010), only end-sequences with a unique “best-hit” in the reference were considered.

Of the 625 ESPs discordant in orientation when mapped to the RPCI-11 LCR haplotypes, 197 ESPs could be re-assigned to single BACs in the *nr/nt* database with equal or higher alignment confidence than against the reference (Supplemental Fig. S16), bringing the paired-ends into their expected orientation and alignment span (median = 39468 bp, IQR = 4624 bp). 126 of the 197 resolved ESPs map to previously identified BACs in the LCR-A/C and LCR-B/D junction groups (Supplemental Fig. S16). Moreover, all ESPs (n=110) with unique placements but not in the same BAC map to at least one of the junction group BACs. In summary, the majority of resolved ESPs support the LCR-A/C and LCR-B/D junction groups as representing breakpoint-harbouring clones.

Under the simplest model of a single ancestral inversion breakpoint, one would expect *II*- and *NN*-derived ESPs to map to non-overlapping regions in each BAC. However, most resolved *II*- and *NN*-derived ESPs appear to map to overlapping regions (Supplemental Fig. S16). Given that the *II*-derived fosmid libraries represent a YRI and a CEU sample (GM19240/ABC10 & GM12878/ABC12) while the *NN*-derived ESPs principally represent a JPT sample (GM18956/ABC9), these ESP-bounded regions may reflect population-specific *8p23-inv* breakpoints. Alternatively, the resolved ESPs may be indicative of common uncharacterised LCR conformations in the region; the presence of multiple LCR-haplotype junctions within single samples attests to the region’s diversity (e.g. LCRs –A/B, -A/C, -A/D, -B/B and –B/D in GM12878/ABC12; Supplemental Fig. S16). This pervasive structural variation may in turn complicate delineation of the *8p23-inv* breakpoints.

Supplementary Methods

FISH analysis

To enhance FISH mapping resolution, elongated metaphase chromosomes were prepared from lymphoblastoid cell lines (LCLs; Supplemental Table S1). Briefly, 48h after seeding a culture, Chromosome Resolution Additive (Procell Reagents) was added for 30 mins, followed by 7.5 μg/ml ethidium bromide (Sigma) for 50 min (Ikeuchi 1984), and 0.07μg/ml of Colcemid (Gibco BRL) for 10 min (Yunis 1981). Cells were incubated for 10 min at room temperature (RT) with 75mM KCl, recovered by centrifugation and re-suspended in fresh fixative (methanol:acetic acid, 3:1, vol/vol) and slides prepared (MacLeod et al. 2007). Chromosome morphology was enhanced by a final 30s drying stage in an improvised humidity chamber (adapted from (Henegariu et al. 2001)). Following overnight drying at RT, preparations were artificially "aged" in 2xSSC buffer (Invitrogen) for 2.5hrs at RT, 2xSSC for 1.5hrs at 37°C, and dehydrated through an ethanol series (adapted from (Bayani et al. 2008)). Prior to hybridization, preparations were incubated with RNaseA (Sigma) and pepsin (Sigma) (Henegariu et al. 2001).

The orthodox 8p inversion-assay probes (Giglio et al. 2001), RP11-589N15 and RP11-399J23, and a reference probe (RP11-73M19), were obtained from BACPAC (http://bacpac.chori.org/) and purified using the Large Construct MaxiPrep Kit (Qiagen). Probe identities were confirmed by PCR or BAC end sequencing. Probes were labeled by nick translation with dUTPs SpectrumGreen (RP11-589N15), SpectrumOrange (RP11-399J23) or SpectrumGreen and SpectrumOrange (RP11-73M19) according to the manufacturer's protocol (Vysis).

Chromosome preparations were hybridized with 33.3ng of each probe and 1μg unlabelled C0t-1 DNA (Invitrogen) in 10μl of LSI/WCP buffer (Vysis; 2x SSC, 50% formamide, 10% dextran sulfate) according to standard protocols (MacLeod et al. 2007). Post-hybridization washes were performed at 45°C; 50% deionised formamide/2xSSC (5 min x 3), then 1xSSC wash (5 min x 3) and finally 4xSSC/0.1%Tween20 (5 min x 3).

Chromosome spreads were counterstained with DAPI according to the manufacturer's instructions (Vectashield, Vector Laboratories). Digital images were obtained using an upright fluorescence microscope (DM RB, Leica) attached to a CoolSNAPHQ CCD camera and processed with MetaMorph software (Molecular Devices). Inversion status was scored from 5-10 chromosome spreads to accommodate potential mosaicism, although none was observed. The significance of association between FISH-derived and PFIDO-derived inversion-types was assessed using Fisher's exact test, with H_0: no correlation between FISH results and PFIDO classification.

SNP genotype data

HapMap

SNP genotypes for 7 HapMap populations were retrieved from the HapMap website (International HapMap 3 Consortium et al. 2010) (http://hapmap.org; HapMap Public Release (rel.) #27). All analyses aimed at investigating genetic substructure used only founder genotypes, to avoid bias introduced by IBD genotypes. Seven samples with reported cryptic relatedness (NA19192, NA18913, NA19092, NA12155, NA06993, NA18987 & NA18992) were excluded as recommended (International HapMap Consortium 2005). SNP inclusion criteria were: MAF > 1%; Hardy-Weinberg Equilibrium (HWE; two-sided selome p-value > 0.001, exact test (Wigginton et al. 2005b)); and location within the inversion interval (chr8: 8,137,473-11,802,039 (hg18)), excluding local segmental duplications.

Family Data

SNPs (n=236, Supplemental Table S3) from an ongoing candidate gene association mapping study were genotyped in 1,748 white British individuals (Naoumova et al. 2003) using the GoldenGate (Illumina Inc.) and MassARRAY iPLEX Gold (Sequenom) platforms. No discrepant genotype calls were found between duplicate samples. Pedcheck (O'Connell et al. 1998) and PEDSTATS (Wigginton et al. 2005a) were used to isolate and remove 55 genotype calls with non-Mendelian inheritance patterns. Two SNPs breaching HWE ($p < 0.001$) were also removed. For PFIDO analysis, the dataset was segregated into 7 subsets of unrelated individuals of $n > 100$, and each subset was seeded with HapMap CEU genotypes (rel.27), identifying and matching potential allele switches using the test.allele.switch function in the snpMatrix package (Clayton et al. 2007).

Worldwide Population Datasets

Supplemental Table S4 summarizes the datasets used. This included genotype data from the Human Genome Diversity Panel (HGDP)-CEPH (http://www.cephb.fr/), focusing on the H952 population subset (which excludes known first- and second-degree relatives (Rosenberg 2006)) and the Illumina 650K platform generated genotypes (Li et al. 2008). Three samples of ambiguous origin (Biswas et al. 2009) (HGDP00980, HGDP00770 & HGDP00621) and one with evidence of cryptic relatedness (Price et al. 2009a) (HGDP01281) were excluded. Additionally, unrelated samples representing 13 populations (n=296) genotyped with the Affymetrix 6.0 array and 23 populations (n=344) genotyped with the Affymetrix 250k NspI array (Xing et al. 2010) were included (http://jorde-lab.genetics.utah.edu/). Three reputedly admixed populations from this dataset were excluded (Xing et al. 2010). Singapore Genome Variation Project samples (SGVP; http://www.nus-cme.org.sg/SGVP/; n = 268) representing Chinese, Malay and Indian populations genotyped on Affymetrix 6.0 and Illumina 1M arrays were also added (Teo et al. 2009).

The Wellcome Trust Case-Control Consortium data (Wellcome Trust Case Control Consortium 2007) (see www.wtccc.org.uk for a full list of contributing investigators; funded by Wellcome Trust award 076113) were generated using the 500K Affymetrix chip with samples from the UK Blood Service Control Group (“NBS”; n=1500) and the 1958 British Birth (“58C”; n=1504), Type 2 Diabetes (n=1999) and Coronary Artery Disease (n=1988) cohorts. Using PLINK (Purcell et al. 2007) (v1.06), samples with evidence of non-European ancestry were identified and excluded as recommended (Wellcome Trust Case Control

Consortium 2007). To mitigate potential batch effects, cohorts were combined prior to analyses using PLINK.

1000 Genome Project data

Nucleotide variant data mapping to the inversion interval was retrieved for 261 samples of European ancestry (CEU, TSI, GBR and FIN) from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20100804/ALL.2of4intersection.20100804.genotypes.vcf.gz using Tabix (v0.2.3, (Li 2011)). Using VCFtools (v0.1.6; (Danecek et al. 2011)), sites were first filtered by quality (options= --hwe 0.001 --geno 0.9 –maf 0.01 --minDP 10 --minGQ 40 --remove-filtered-all) before the calculation of summary statistics for *II* and *NN* samples.

Identifying Recombination Events in Sequence Alignments using RDP3

RPCI-11 clones (with evidence of LCR mosaicism) were queried against human clones in the nr/nt database using MegaBLAST (-e 1 –q -4 –r 1 –W 256), and resulting flat-query anchored files parsed in R to produce multiple sequence alignments. “Low-scoring” alignment segments (as defined in CLUSTALX) were manually edited to avoid false positive recombination signals: editing primarily involved mis-aligned di/tri-nucleotide repeats. The recombination detection algorithms RDP (Martin et al. 2000) (no reference, window-size=60), GENECONV (Padidam et al. 1999), Bootscan (Martin et al. 2005) (window-size=2000, 200 bootstrap replicates, cutoff=95%), Maxchi (Smith 1992) (120 variable sites/window), Chimaera (Posada et al. 2001) (120 variable sites/window), SisScan (Gibbs et al. 2000) (window-size=2000) and 3Seq (Boni et al. 2007) were used, with default algorithm settings unless stated. To focus analyses on the RPCI-11 library, all clones from other libraries were masked in the initial analysis. Short recombinant tracts (<5kb), indicative of gene conversion or double-crossover events (Chen et al. 2007), were excluded.

References for Supplemental Note

1000 Genomes Project Consortium, Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061-1073.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403-410.

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

Bayani, J. and Squire, J. A. 2008. *Molecular biomethods handbook.* Humana; Springer distributor, Totowa, N.J.; London.

Biswas, S., Scheinfeldt, L. B., and Akey, J. M. 2009. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am. J. Hum. Genet.* **84:** 641-650.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19:** 185-193.

Boni, M. F., Posada, D., and Feldman, M. W. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176:** 1035-1047.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., and Weber, J. L. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63:** 861-869.

Charif, D., Thioulouse, J., Lobry, J. R., and Perriere, G. 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics* **21:** 545-547.

Chen, J. M., Cooper, D. N., Chuzhanova, N., Ferec, C., and Patrinos, G. P. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8:** 762-775.

Chowdhury, R., Bois, P. R., Feingold, E., Sherman, S. L., and Cheung, V. G. 2009. Genetic analysis of variation in human meiotic recombination. *PLoS Genet.* **5:** e1000648.

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

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25:** 1422-1423.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27:** 2156-2158.

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., et al. 2007. A genome-wide association study of global gene expression. *Nat. Genet.* **39:** 1202-1207.

Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* **22:** 101-109.

Eichler, E. E., Clark, R. A., and She, X. 2004. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* **5:** 345-354.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792-1797.

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

Ge, B., Pokholok, D. K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D. J., Le, J., Koka, V., Lam, K. C., 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.

Gibbs, M. J., Armstrong, J. S., and Gibbs, A. J. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16:** 573-582.

Giglio, S., Broman, K. W., 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:** 874-83.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., et al. 2010. A draft sequence of the Neandertal genome. *Science* **328:** 710-722.

He, C., Weeks, D. E., Buyske, S., Abecasis, G. R., Stewart, W. C., Matise, T. C., and Enhanced Map Consortium. 2011. Enhanced genetic maps from family-based disease studies: population-specific comparisons. *BMC Med. Genet.* **12:** 15.

Henegariu, O., Heerema, N. A., Lowe Wright, L., Bray-Ward, P., Ward, D. C., and Vance, G. H. 2001. Improvements in cytogenetic slide preparation: controlled chromosome spreading, chemical aging and gradual denaturing. *Cytometry* **43:** 101-109.

Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., Chen, G. K., Wang, K., Buxbaum, S. G., Akylbekova, E. L., et al. 2011. The landscape of recombination in African Americans. *Nature* **476:** 170-175.

Hodgkinson, A. and Eyre-Walker, A. 2011. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12:** 756-766.

Ikeuchi, T. 1984. Inhibitory effect of ethidium bromide on mitotic chromosome condensation and its application to high-resolution chromosome banding. *Cytogenet. Cell Genet.* **38:** 56-61.

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467:** 52-58.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299-1320.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-945.

Jiang, Z., Tang, H., Ventura, M., Cardone, M. F., Marques-Bonet, T., She, X., Pevzner, P. A., and Eichler, E. E. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39:** 1361-1368.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T. L. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36:** W5-9.

Johnston, C. M., Lovell, F. L., Leongamornlert, D. A., Stranger, B. E., Dermitzakis, E. T., and Ross, M. T. 2008. Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. *PLoS Genet.* **4:** e9.

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

Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., Kallicki, J., Kaul, R., Wilson, R. K., and Eichler, E. E. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143:** 837-847.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* **5:** R12.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol.* **5:** e254.

Li, H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27:** 718-719.

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.

MacLeod, R. A., Kaufmann, M., and Drexler, H. G. 2007. Cytogenetic harvesting of commonly used tumor cell lines. *Nat. Protoc.* **2:** 372-382.

Manly, K. F., Nettleton, D., and Hwang, J. T. 2004. Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res.* **14:** 997-1001.

Martin, D. and Rybicki, E. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16:** 562-563.

Martin, D. P., Posada, D., Crandall, K. A., and Williamson, C. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res. Hum. Retroviruses* **21:** 98-102.

Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E., et al. 2007. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448:** 470-473.

Naoumova, R. P., Bonney, S. A., Eichenbaum-Voline, S., Patel, H. N., Jones, B., Jones, E. L., Amey, J., Colilla, S., Neuwirth, C. K., Allotey, R., et al. 2003. Confirmed locus on chromosome 11p and candidate loci on 6q and 8p for the triglyceride and cholesterol traits of combined hyperlipidemia. *Arterioscler Thromb Vasc Biol* **23:** 2070-7.

Navarro, A., Betran, E., Barbadilla, A., and Ruiz, A. 1997. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics* **146:** 695-709.

O'Connell, J. R. and Weeks, D. E. 1998. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.* **63:** 259-266.

Osoegawa, K., Mammoser, A. G., Wu, C., Frengen, E., Zeng, C., Catanese, J. J., and de Jong, P. J. 2001. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11:** 483-496.

Padidam, M., Sawyer, S., and Fauquet, C. M. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265:** 218-225.

Paradis, E., Claude, J., and Strimmer, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20:** 289-290.

Pihur, V., Datta, S., and Datta, S. 2009. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* **10:** 62.

Posada, D. and Crandall, K. A. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. U. S. A.* **98:** 13757-13762.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. 2009a. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5:** e1000519.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. 2009b. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5:** e1000519.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81:** 559-575.

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.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. 2010. Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11:** 356-366.

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6:** e107.

Schmutz, J., Grimwood, J., and Myers, R. M. 2004. Quality assessment of finished BAC sequences. *Methods Mol. Biol.* **255:** 343-349.

Smith, J. M. 1992. Analyzing the mosaic structure of genes. *J. Mol. Evol.* **34:** 126-129.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., et al. 2007. Population genomics of human gene expression. *Nat. Genet.* **39:** 1217-1224.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **101:** 6062-6067.

Suzuki, R. and Shimodaira, H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22:** 1540-1542.

Taudien, S., Galgoczy, P., Huse, K., Reichwald, K., Schilhabel, M., Szafranski, K., Shimizu, A., Asakawa, S., Frankish, A., Loncarevic, I. F., et al. 2004. Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence. *BMC Genomics* **5:** 92.

Teo, Y. Y., Sim, X., Ong, R. T., Tan, A. K., Chen, J., Tantoso, E., Small, K. S., Ku, C. S., Lee, E. J., Seielstad, M., et al. 2009. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19:** 2154-2162.

Wegmann, D., Kessner, D. E., Veeramah, K. R., Mathias, R. A., Nicolae, D. L., Yanek, L. R., Sun, Y. V., Torgerson, D. G., Rafaels, N., Mosley, T., et al. 2011. Recombination rates in admixed individuals identified by ancestry-based inference. *Nat. Genet.* **43:** 847-853.

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.

Wigginton, J. E. and Abecasis, G. R. 2005a. PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* **21:** 3445-3447.

Wigginton, J. E., Cutler, D. J., and Abecasis, G. R. 2005b. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76:** 887-893.

Xing, J., Watkins, W. S., Shlien, A., Walker, E., Huff, C. D., Witherspoon, D. J., Zhang, Y., Simonson, T. S., Weiss, R. B., Schiffman, J. D., 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.

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. 2009. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10:** 451-481.

Yunis, J. J. 1981. Mid-prophase human chromosomes. The attainment of 2000 bands. *Hum. Genet.* **56:** 293-298.

Zody, M. C., Jiang, Z., Fung, H. C., Antonacci, F., Hillier, L. W., Cardone, M. F., Graves, T. A., Kidd, J. M., Cheng, Z., Abouelleil, A., et al. 2008. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40:** 1076-1083.