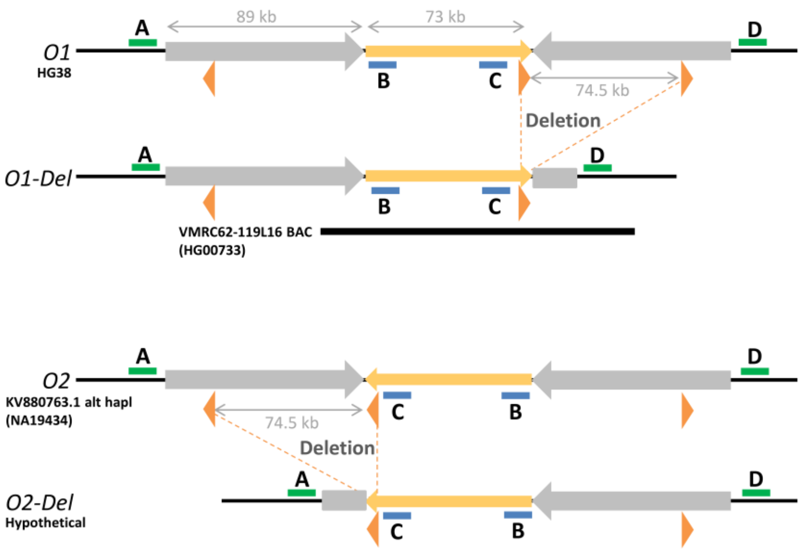
SUPPLEMENTAL MATERIAL

Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR

Marta Puig, Jon Lerga-Jaso, Carla Giner-Delgado, Sarai Pacheco, David Izquierdo, Alejandra Delprat, Magdalena Gayà-Vidal, Jack F. Regan, George Karlin-Neumann, Mario Cáceres

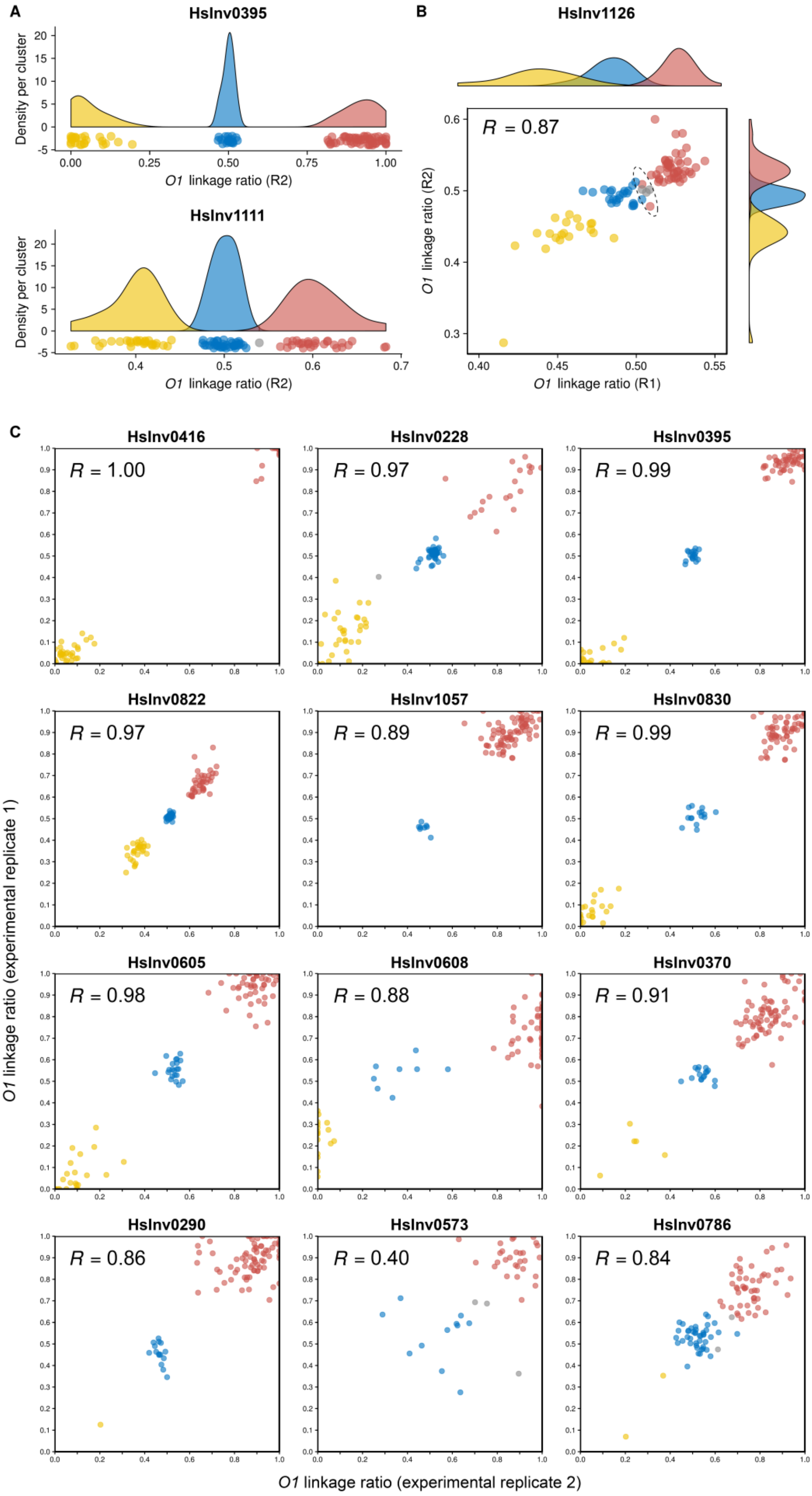
CONTENTS

SUPPLEMENTAL FIGURES

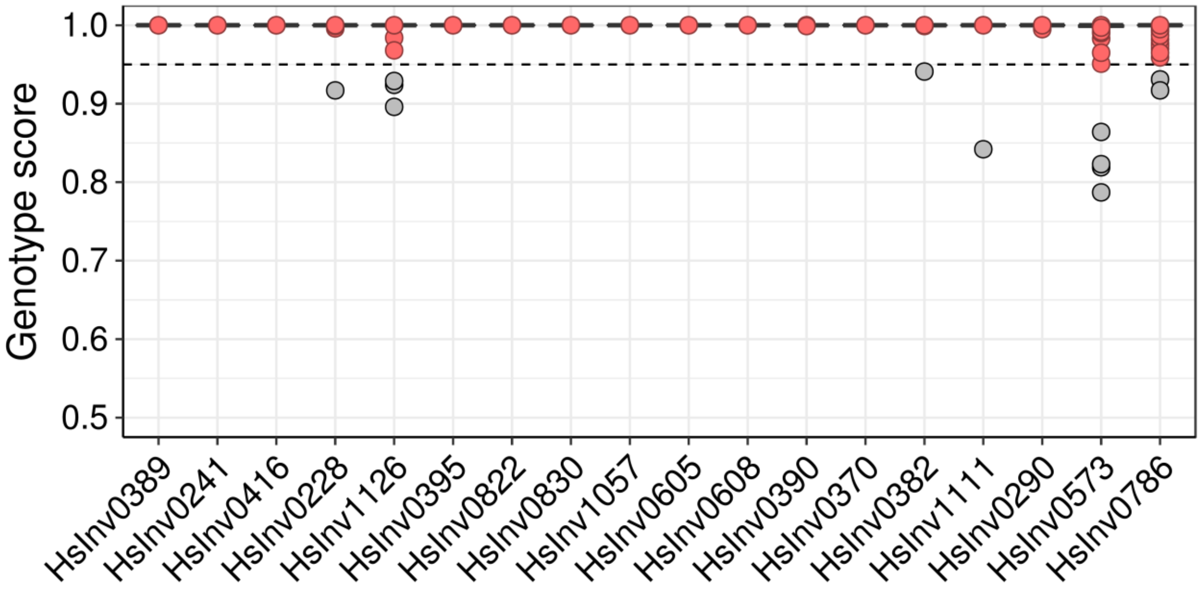


AB	CD	AC	BD
99.7	123.7	145.1	169.2
99.7	49.3	145.1	94.7
168.9	145.2	123.4	99.7
94.4	145.2	48.9	99.7

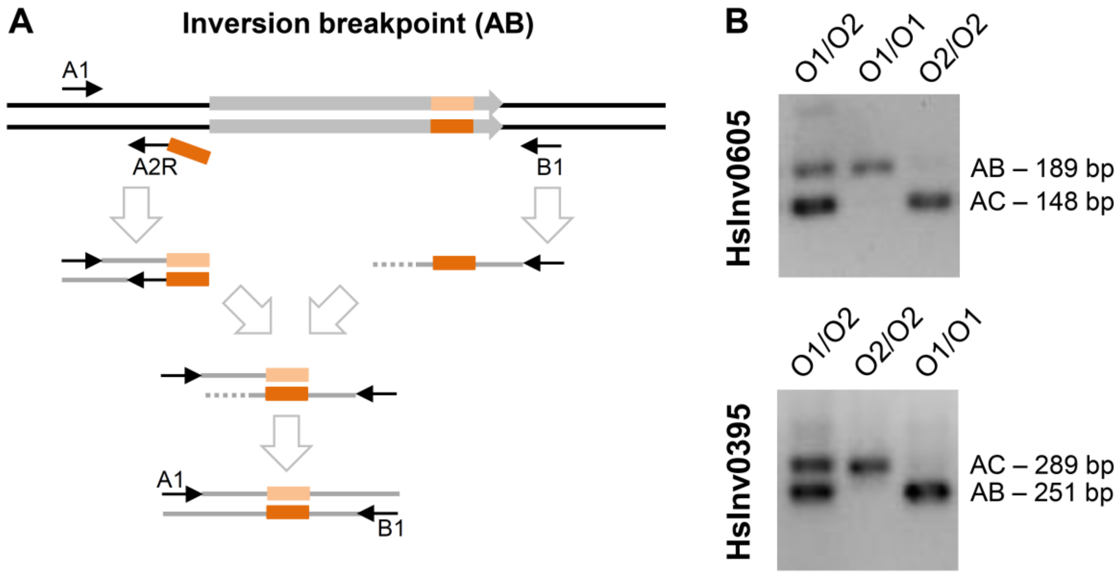
Supplemental Figure S1. ddPCR analysis of structural variation at the HsInv0233 inversion region. Grey arrows indicate inverted segmental duplications (SDs) and the yellow arrow the orientation of the inverted sequence (*O1* or *O2*). ddPCR amplicons are represented as green (outside the inversion) and blue (inside the inversion) bars labeled according to their position, and orange triangles correspond to the 11.6-kb repeats that mediate a 74.5-kb deletion affecting most of one of the SDs. The source of the sequences is indicated below the name of each conformation and in *O1-Del* a black bar below indicates the region included in the BAC clone that supports this structure. The distances in kb between each pair of amplicons are represented to the right of each structure, with combinations expected to have lower linkages shaded in grey and those affected by the deletion shown in orange and red. As a consequence of the deletion, the BD distance is similar in an *O1-Del* and *O2* chromosome (and the same happens with AB in the hypothetical *O2-Del* and *O1* conformation). Since the measured linkage is an average of the two chromosomes of an individual, this makes it impossible to interpret the results without additional information and prevents accurate genotyping of the inversion orientation.


A
HsInv0395
Density per cluster
20
15
10
5
0
-5
0.00
0.25
0.50
0.75
1.00
O1 linkage ratio (R2)
HsInv1111
Density per cluster
0.4
0.5
0.6
0.7
O1 linkage ratio (R2)
B
HsInv1126
R = 0.87
O1 linkage ratio (R2)
0.6
0.5
0.4
0.3
0.40
0.45
0.50
0.55
O1 linkage ratio (R1)
C
HsInv0416
R = 1.00
HsInv0228
R = 0.97
HsInv0395
R = 0.99
HsInv0822
R = 0.97
HsInv1057
R = 0.89
HsInv0830
R = 0.99
HsInv0605
R = 0.98
HsInv0608
R = 0.88
HsInv0370
R = 0.91
HsInv0290
R = 0.86
HsInv0573
R = 0.40
HsInv0786
R = 0.84
O1 linkage ratio (experimental replicate 1)
O1 linkage ratio (experimental replicate 2)

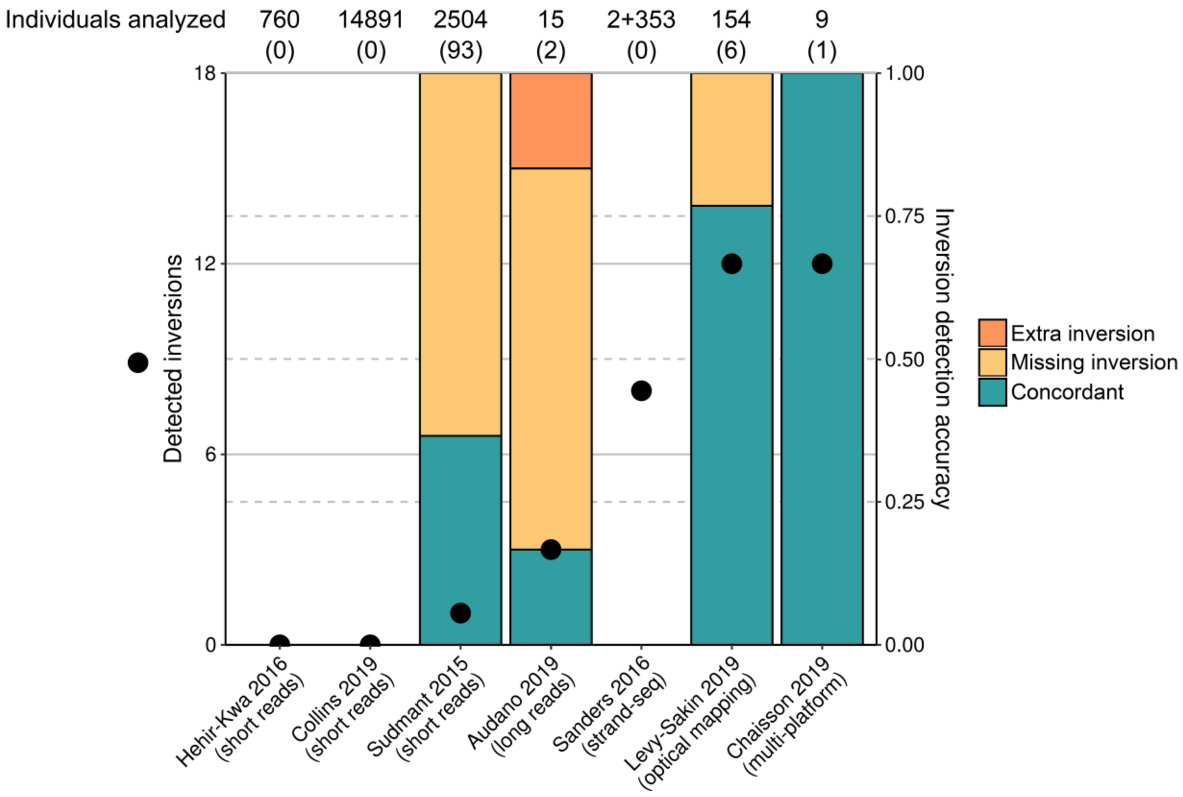
Supplemental Figure S2. Inversion genotype calling by clustering of *O1* linkage ratios. (*A*/*B*) Clustering density plots for two inversions that show separated genotype groups, HsInv0395 and HsInv1111 (*A*), and HsInv1126 that has overlapping clusters (*B*). Dots indicate the *O1* linkage ratio for each sample and colors mark genotype groups: *O1/O1*, red; *O1/O2*, blue; *O2/O2*, yellow. Grey dots are samples without a clear genotype. R1 and R2 represent two randomly selected *O1* linkage ratio replicate values for the different samples. For HsInv1126, dots inside the ellipse correspond to samples included in both *O1/O1* and *O1/O2* clusters with similar probability that were not genotyped (with the exception of two males in red recovered in the male-specific clustering analysis for Chr X inversions). (*C*) Correlation between the *O1* linkage ratios of two replicates for the rest of inversions in which a single breakpoint was analyzed, showing the clustering of the samples in the three genotype groups (represented as in *A* and *B*). The correlation coefficient (*R*) is indicated within each graph and values are always above 0.85, except for the two inversions with amplicons separated by more than 145 kb (HsInv0786, *R* = 0.84, and HsInv0573, *R* = 0.40).



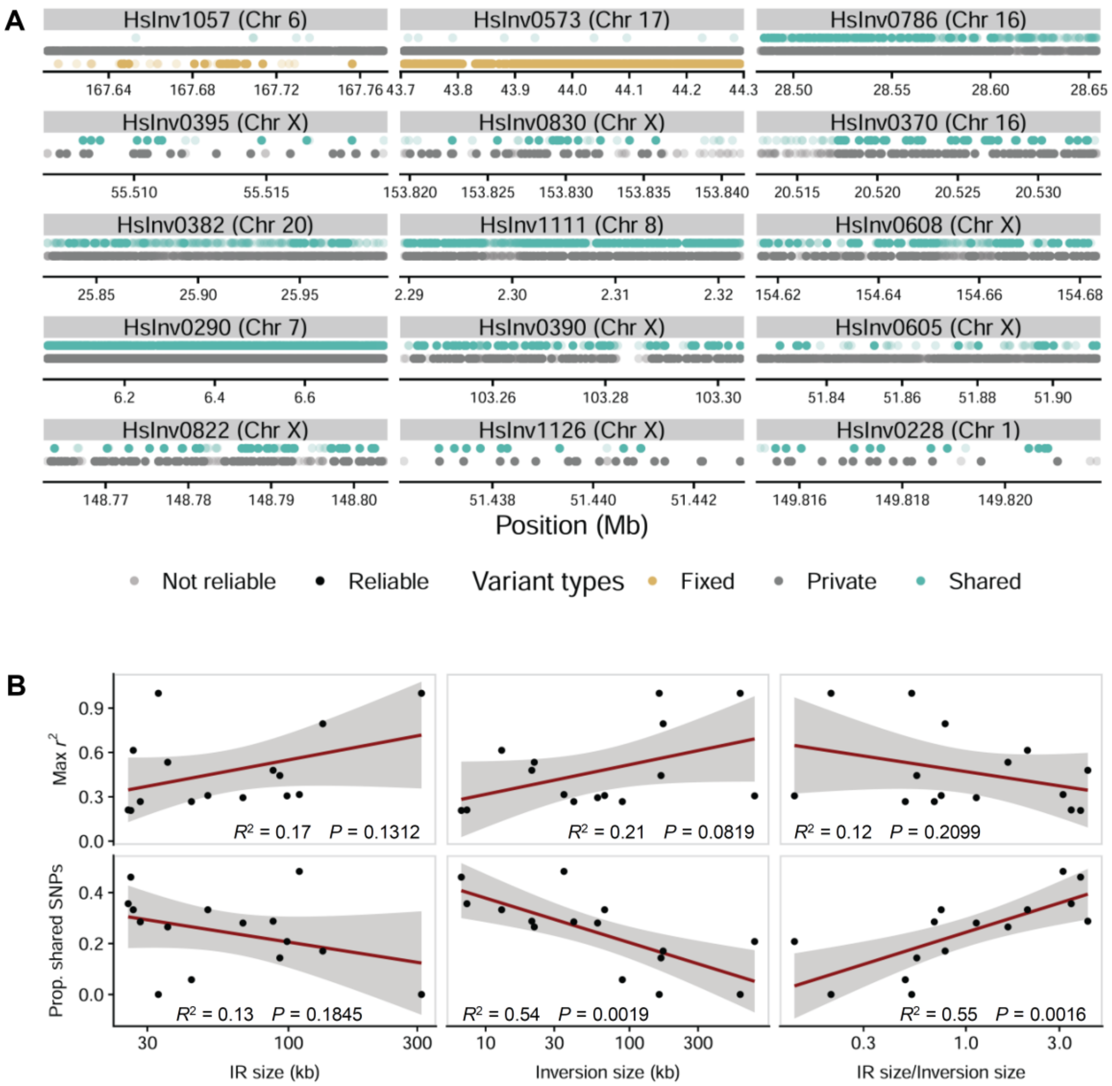
Supplemental Figure S3. Summary of ddPCR inversion genotype scores. Genotype scores correspond to the percentage of times that a sample is included within the most likely cluster and inversions are ordered according to the distance between amplicons. For virtually all the samples, genotypes tend to be very reliable with a score of 1, indicating that they are always clustered in the same genotype group (Supplemental Table S3). Genotypes with scores below 0.95 (dashed line) were considered unreliable and were not taken into account (grey dots). The inversions that accumulate more samples with low scores are HsInv1126, where the IRs are very close together and it was not possible to digest the DNA, and the two with the longest IRs at the breakpoints. However, for inversion 17q21 (HsInv0573) ddPCR genotypes match perfectly those predicted from tag SNPs.



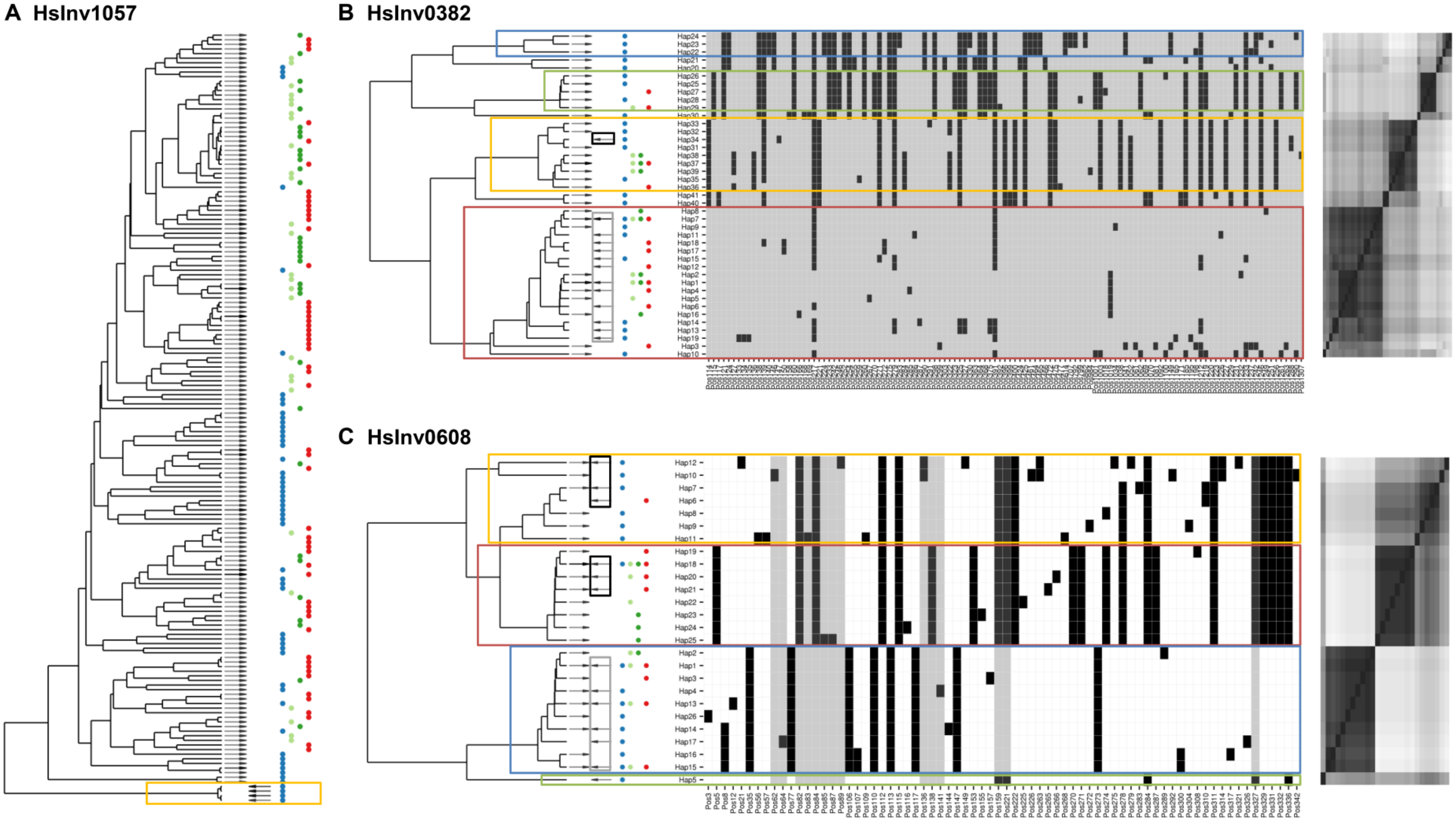
Supplemental Figure S4. Inversion genotyping by haplotype-fusion PCR (HF-PCR). (*A*) Summary of the HF-PCR strategy to genotype the AB breakpoint of an inversion (Turner et al. 2006). A double-stranded amplicon located outside an inversion is amplified with two primers A1 and A2R (black arrows), one of which has a 5' extension (orange rectangle) with a sequence found within the inverted repeat (grey arrow) at the inversion breakpoint. In the same reaction, a single-stranded product is linearly amplified at the other side of the breakpoint with primer B1. This single-stranded product contains the sequence able to hybridize with amplicon A (orange rectangle) and an AB fusion product containing sequences from both sides of the breakpoint is amplified once primer A2R runs out. Since the PCR reaction takes place in an emulsion, only one DNA template molecule is expected to be found within a single droplet and the fusion product will indicate the presence in the sample of an AB junction. By adding a primer C1 able to amplify the other end of the inverted sequence, we can detect both AB and AC breakpoints and genotype the inversion. (*B*) Examples of HF-PCR results. After reamplification with nested primers, the three genotypes can be clearly distinguished for the two inversions analyzed in this work by visualizing the fusion products in an agarose gel.



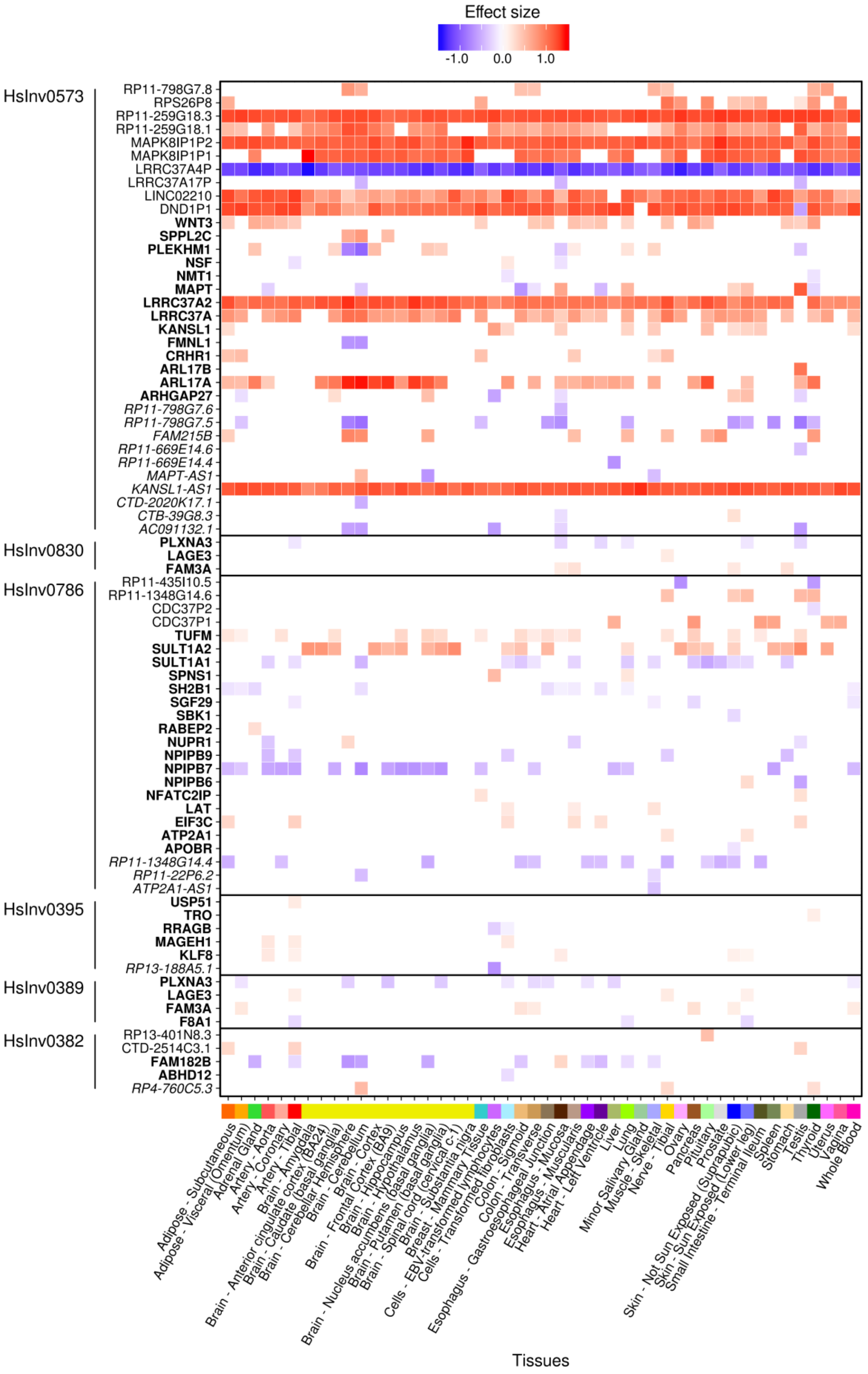
Supplemental Figure S5. Comparison of detection of 18 inversions genotyped by ddPCR with other genome-wide techniques. The total number of individuals analyzed in each study and those in common with this work and for whom genotypes can be compared (within parenthesis) are shown on top of the graph. Sanders et al. (2016) included samples of two individuals plus a pool of cells from 353 individuals. Black dots represent the number of inversions detected in each study. HsInv0233 and HsInv0012, not genotyped by ddPCR, are not included, although they were detected by Levy-Sakin et al. (2019) and Chaisson et al. (2019) (Supplemental Table S1). Only optical mapping and the multi-platform approach (including short and long reads, optical mapping and strand-seq) are able to detect a substantial part of the inversions flanked by large IRs analyzed here. Bars indicate the proportion of concordant genotype calls among the comparable individuals (green) and that of different types of inversion detection errors (orange and yellow). Audano et al. (2019) and Levy-Sakin et al. (2019) just provide information about the presence or absence of the inversion in each sample, while Sudmant et al. (2015) and Chaisson et al. (2019) provide genotypes. Only the multi-platform approach detects correctly the 12 inversions identified in the individual in common. In all cases, discrepant ddPCR results are very reliable according to the genotype scores, suggesting that the other techniques tend to be too conservative and miss the inverted allele in many samples. For example, it had already been found that the inverted allele of HsInv0241, genotyped previously by iPCR, was missed in most samples in 1000GP data (Giner-Delgado et al. 2019). Optical mapping is the single technique that genotypes better as it is able to go across longer repeats.



Supplemental Figure S6. LD and SNP distribution in the regions of 15 newly genotyped inversions. (*A*) SNPs within each of the inversions are classified as fixed between orientations (yellow), private to one orientation (grey) or shared between orientations (green), and color intensity indicates their reliability according to the 1000GP Phase 3 strict accessibility mask and their location outside segmental duplications. The presence of shared variants polymorphic in both orientations along the entire inversion length is difficult to explain by gene conversion events and it can be thus considered a sign of recurrence. (*B*) Correlation of the maximum LD with 1000GP variants (1 Mb at each side of the inversion) in all analyzed individuals (r^2) (top) and the proportion of shared SNPs within the inverted region (bottom) with IR and inversion size measures for the 15 inversions. Correlation (R^2) and *P* values were calculated with the lm R function (Maechler et al. 2018) and are shown within each graph. Longer inversions tend to have higher LD with other variants and a lower proportion of shared SNPs, whereas the opposite is found for the ratio between the IR and inversion size (IR/Inv ratio), consistent with the results of the recurrence model in which this last variable explains a significant part of the variance in the number of inversion events.


A HsInv1057
B HsInv0382
C HsInv0608
Hap24
Hap23
Hap22
Hap21
Hap20
Hap26
Hap25
Hap27
Hap28
Hap29
Hap30
Hap33
Hap32
Hap34
Hap31
Hap38
Hap37
Hap39
Hap35
Hap36
Hap41
Hap40
Hap8
Hap7
Hap9
Hap11
Hap18
Hap17
Hap15
Hap12
Hap2
Hap1
Hap4
Hap5
Hap6
Hap16
Hap14
Hap13
Hap19
Hap3
Hap10
Hap12
Hap10
Hap7
Hap6
Hap8
Hap9
Hap11
Hap19
Hap18
Hap20
Hap21
Hap22
Hap23
Hap24
Hap25
Hap2
Hap1
Hap3
Hap4
Hap13
Hap26
Hap14
Hap17
Hap16
Hap15
Hap5
Pos3
Pos5
Pos8
Pos12
Pos21
Pos35
Pos56
Pos57
Pos62
Pos64
Pos77
Pos82
Pos83
Pos84
Pos85
Pos87
Pos89
Pos106
Pos107
Pos109
Pos110
Pos112
Pos113
Pos115
Pos116
Pos117
Pos136
Pos138
Pos141
Pos144
Pos147
Pos149
Pos153
Pos155
Pos157
Pos159
Pos221
Pos222
Pos225
Pos226
Pos263
Pos265
Pos266
Pos268
Pos270
Pos271
Pos272
Pos273
Pos274
Pos275
Pos278
Pos279
Pos283
Pos284
Pos287
Pos289
Pos292
Pos300
Pos304
Pos308
Pos310
Pos311
Pos314
Pos317
Pos321
Pos326
Pos327
Pos329
Pos331
Pos332
Pos336
Pos342

Supplemental Figure S7. Estimation of the number of inversion events from inverted region haplotypes. Each inversion was analyzed using integrated haplotype plots (iHPlots) (Giner-Delgado et al. 2019), with the tree indicating the relationship between the different haplotypes, the rightwards (*O1*) and leftwards (*O2*) arrows the orientations observed for each haplotype, and dots the populations where each haplotype has been found (blue for YRI, light and dark green for CHB and JPT, and red for CEU). Inverted region haplotypes are represented by the variable positions (see Methods for variant selection) with different colors indicating the two alleles (white, ancestral; grey, hg19 reference; black, derived/alternative), and colored boxes showing the main differentiated haplotype clusters. For the unique inversion (*A*) only the first part of the plot is shown with a yellow box highlighting the single group of tightly-clustered inverted haplotypes. In the recurrent inversions (*B*, *C*), the orientation of the haplotypes of what we considered the original inversion event are included in a grey box and any additional inversion event in a black box (one in HsInv0382 and two in HsInv0608).


Effect size
-1.0
0.0
1.0
HsInv0573
RP11-798G7.8
RPS26P8
RP11-259G18.3
RP11-259G18.1
MAPK8IP1P2
MAPK8IP1P1
LRRC37A4P
LRRC37A17P
LINC02210
DND1P1
WNT3
SPPL2C
PLEKHM1
NSF
NMT1
MAPT
LRRC37A2
LRRC37A
KANSL1
FMNL1
CRHR1
ARL17B
ARL17A
ARHGAP27
RP11-798G7.6
RP11-798G7.5
FAM215B
RP11-669E14.6
RP11-669E14.4
MAPT-AS1
KANSL1-AS1
CTD-2020K17.1
CTB-39G8.3
AC091132.1
HsInv0830
PLXNA3
LAGE3
FAM3A
HsInv0786
RP11-435I10.5
RP11-1348G14.6
CDC37P2
CDC37P1
TUFM
SULT1A2
SULT1A1
SPNS1
SH2B1
SGF29
SBK1
RABEP2
NUPR1
NPIPB9
NPIPB7
NPIPB6
NFATC2IP
LAT
EIF3C
ATP2A1
APOBR
RP11-1348G14.4
RP11-22P6.2
ATP2A1-AS1
HsInv0395
USP51
TRO
RRAGB
MAGEH1
KLF8
RP13-188A5.1
HsInv0389
PLXNA3
LAGE3
FAM3A
F8A1
HsInv0382
RP13-401N8.3
CTD-2514C3.1
FAM182B
ABHD12
RP4-760C5.3
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cells - Transformed fibroblasts
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Heart - Atrial Appendage
Heart - Left Ventricle
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
Tissues

Supplemental Figure S8. Summary of gene expression changes associated to inversions across GTEx tissues. Inversion effects were estimated using FAPI (Kwan et al. 2016) through LD patterns with eQTLs in GTEx Analysis Release v7 (see Supplemental Table S10). The direction and strength of the beta effect is indicated in blue or red representing, respectively, a lower or higher expression associated to the *O2* orientation. Gene names in bold correspond to protein coding genes, in italics to non-coding RNA genes, and the rest are classified as pseudogenes in GENCODE version 26.

SUPPLEMENTAL METHODS

High-molecular-weight DNA isolation

To obtain high-molecular-weight DNA, the cell pellet was resuspended in extraction buffer (10 mM Tris-HCl pH 8, 10 mM EDTA pH 8, 150 nM NaCl, 0.5% SDS) and incubated overnight with slow rotation at 37 °C with RNase cocktail (Invitrogen) and 100 µg/ml Proteinase K (Invitrogen). Four purification steps with one volume of TE-equilibrated phenol pH 7.9 (twice), phenol:chloroform:isoamyl alcohol pH 7.9, and chloroform:isoamyl alcohol were performed by mixing by rotation until an emulsion was formed, and then centrifuging at 5,000 x g for 10-15 min to separate organic and aqueous phases. Finally, DNA was precipitated by adding 0.1 volumes of 3 M sodium acetate and 2 volumes of absolute ethanol, centrifuged, washed with 70% ethanol, and resuspended in 100-300 µl of water.

Genotype clustering

In order to determine reliably the genotypes of each inversion, we calculated the Euclidian distance between individuals (stats::dist R function) (R Core Team 2017) using two randomly-selected *O1* linkage ratios per sample scaled to normal scores (base::scale R function). All replicates of every sample were used, except those based on a total linkage <7.5% (or <15% if only one measurement was available), low droplet couts (<10,000 droplets) or altered amplicon ratios due to deletions or duplications. Also, in HsInv0382, where a deletion increases the linkage in one of the breakpoints, samples genotyped only by one breakpoint were excluded. Since in some cases there is a variable number of measurements, this process was repeated 200,000 times and a mean pairwise distance between individuals was obtained. Next, we performed a hierarchical clustering analysis (ward.D implemented method) on this similarity matrix to determine group membership (stats::hclust, stats::cutree R functions) (R Core Team 2017). Clustering was run to find two or three clusters that were defined taking into account that heterozygotes should be centered around 0.5. For Chr X inversions, we repeated the analysis only with males clustered into *O1* or *O2* and, if these genotypes were more robust, they were the ones used. Finally, we tried to recover samples without a clear genotype in an extra clustering step by selecting proportionally more often those *O1* linkage ratios based on a higher total linkage when calculating Euclidean distances and repeating the bootstrapping to obtain the final genotype clusters. In the successive rounds, genotype score was calculated by clustering two thirds of the samples selected at random 10,000 times (ensuring that at least three samples of each of the clusters defined with all the individuals were present, with the exception of HsInv0786 and HsInv0290 in which there are only two and one *O2*/*O2* homozygotes) and determining the percentage of inclusion of each sample in its most common cluster. Individual genotypes were assigned the best score of the three clustering rounds (Supplemental Table S3).

Haplotype fusion PCR (HF-PCR)

For HF-PCR, to separate both inversion breakpoints, first 250 ng of genomic DNA were digested overnight at 37 °C in a volume of 20 µl including 5 U of SwaI for HsInv395 or SalI for HsInv605 and 1x buffer, the restriction enzyme was then heat inactivated at 65 °C for 15 min, and 25 ng of DNA were used as template. Emulsion PCR reactions were performed in 25 µl in 96-well plates for 40 cycles as previously described (Turner and Hurles 2009). The main differences were that we used SOLiD^{TM} EZ^{TM} Bead Emulsifier Oil Kit (Applied Biosystems) to form the emulsions, and that after amplification we carefully transferred the emulsion to a fresh plate, added 50 µl of 1x Phusion HF buffer (Thermo Scientific) to increase volume, centrifuged for 5 minutes at maximum speed and recovered the aqueous phase containing the amplification products. Next, we did a 30-cycle reamplification step with 1 µl of a 1/10 dilution of the previous PCR, 1.5 U Taq DNA polymerase (Roche), and 200-400 nM of each of the three nested primers in a 25 µl total volume (Turner et al. 2006). Sequences of all primers can be found in Supplemental Table S15. Finally, 10 µl of the PCR reaction were loaded into a 3% agarose gel for visualization.

Analysis of inversion frequency

Frequency differences between populations were measured with Weir and Cockerham's F_{ST} estimator implemented in vcftools (v0.1.15) (Danecek et al. 2011), using the 92 samples common to the 1000GP and only females for Chr X inversions or paired male chromosomes for the Chr Y inversion. F_{ST} values of each inversion were compared with an empirical distribution from 10,000 genome-wide biallelic 1000GP SNPs polymorphic in at least two of the populations, matched by chromosome type (autosome or Chr X) and excluding those SNPs overlapping inversion regions ±100 kb. Correlation between MAF and the logarithm of the physical and genetic lengths of inversions was measured with a linear model implemented in robustbase::lmrob R function (Maechler et al. 2018), including data from 45 inversions in a larger sample of the same populations (Giner-Delgado et al. 2019). Inversion physical length corresponds to the distance between IRs and genetic length was interpolated from Bhérer at al. (2017) high-resolution recombination map, using the female map for Chr X and the sex average map for autosomes. No genetic length was available for Chr Y inversions and HsInv0608 (Chr X), which falls outside the last marker in the map.

Calculation of inversion mutation rate

To calculate the mutation rate in Chr Y inversion HsInv0416, we estimated a number of 30,931.1 generations for all branches involved in the phylogeny that relates the 48 analyzed males (Poznik et al. 2016), including a C-T branch split time of 76,000 years, a total number of mutations of 5,591, and an average number of mutations of all branches of 549.5, plus a generation time of 25 years (Repping et al. 2006). Considering that four independent inversion events were detected, this results in an inversion mutation rate of 1.29 x 10^{-4} inversions per generation. The mutation rate for other inversions was extrapolated from the average of the values of the two Chr Y inversions (HsInv0416 and HsInv0832) based on the predicted number of inversion events per chromosome according to the model.

Gene expression analysis

For gene expression analysis in lymphoblastoid cell lines, Geuvadis RNA-seq reads (EMBL-EBI ArrayExpress experiment E-GEUV-1) (Lappalainen et al. 2013) were aligned against the human reference genome GRCh38.p10 (excluding patches and alternative haplotypes) with STAR v2.4.2a (Dobin et al. 2013). We estimated gene expression levels as reads per kilobase per million mapped reads (RPKM) based on GENCODE version 26 annotations (Harrow et al. 2012) and quantified transcript expression with RSEM v1.2.31 (Li and Dewey 2011), filtering out non-expressed genes and transcripts with <0.1 RPKM in >80% of the samples. RPKM values were normalized by quantile transformation across all samples and expression of each gene/transcript was adjusted to a standard normal distribution by rank-based inverse normal transformation. Association with the expression of 418 genes and 2,044 transcripts was calculated for all biallelic variants with MAF >0.05 (including the inversion) within 1 Mb at either side of the transcription start site through linear regressions implemented in FastQTL (Ongen et al. 2016). Since technical or biological confounders reduce the power to find associations, we adjusted expression values by the top three 1000GP genotyping principal components (corresponding to population structure), sequencing laboratory, gender, and an optimal number of PEER (probabilistic estimation of expression residuals) components (Stegle et al. 2012) for eQTL finding (for genes and transcripts, respectively, 12 and 15 for the experimental and 25 and 30 for the imputed set).

In order to estimate inversion gene-expression effects in other tissues, first we randomly took three samples of 30 experimentally genotyped individuals following ethnic proportions of GTEx donors (25 individuals from CEU, 4 YRI and 1 EAS) per inversion-gene pair and tissue to calculate LD patterns between each inversion and neighbouring SNPs, which were subsequently used to impute the corresponding inversion association *P* values from GTEx V7 release eQTL *P* values (The GTEx Consortium 2017). If any *P* value was lower than the genome-wide empirical threshold defined by GTEx for each gene and tissue (The GTEx Consortium 2017), we generated 30 samples of 30 individuals to calculate statistical significance more accurately. The eQTL *P* value of the inversion was defined as the median of permuted *P* values and the association confidence interval as the 25th and 75th percentiles, since the small number of individuals for LD calculation can produce extreme *P* values. In addition, we filtered out those associations with estimated *P* value lower than GTEx significance threshold or with a confidence interval spanning more than two orders of magnitude. Effect sizes were calculated as a function of MAF and *P* value, whereas direction was determined through LD with eQTLs using PLINK v1.9 *--r2 in-phase* option (Purcell et al. 2007). LD was estimated as the median LD of permutations as explained above. The conservative nature of this analysis is represented by HsInv0389, in which only four of the 16 genes previously associated to the inversion based directly on the LD with GTEx eQTLs from a larger number of samples were identified.

GWAS enrichment analysis

To determine enrichment of GWAS signals within the inversions we used a similar approach as in Giner-Delgado et al. (2019). First, we translated GWAS Catalog (http://www.ebi.ac.uk/gwas/) [release 2018-06-25, v1.0] (MacArthur et al. 2017) coordinates to hg19 using Ensembl REST API (Yates et al. 2016) and grouped together the signals associated with SNPs in high LD (r^2 ≥ 0.8) in 1000GP data and corresponding exactly to the same phenotype, resulting in 67,035 non-redundant SNP-trait associations. Then, we created a background distribution of each inversion with 1,500 random genomic regions of the same size than the inverted segment to calculate the enrichment *P* values. We excluded from permutations Chr Y, gaps, and the major histocompatibility complex region (chr6:28,477,797-33,448,354), known to harbor a vast number of associations. In addition, we tested that the GWAS enrichment was not biased by the allele SNP frequencies by selecting 150 random regions per inversion with comparable patterns of common variants (number of 1000GP loci with global MAF > 0.05 per kb ±20%) and without this criteria, which showed very similar results (R^2 = 0.99). To explore which inversions were driving the enrichment, we repeated the analysis for each inversion independently using a one-tailed permutation test (to account for inversions with zero GWAS signals).

SUPPLEMENTAL REFERENCES

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the major structural variant alleles of the human genome. *Cell* **176**: 663–675.

Bhérer C, Campbell CL, Auton A. 2017. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat Commun* **8**: 14994.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784.

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Khera A V., Francioli LC, Gauthier LD, Wang H, Watts NA, et al. 2019. An open resource of structural variation for medical and population genetics. *bioRxiv* doi.org/10.1101/578674.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Giner-Delgado C, Villatoro S, Lerga-Jaso J, Gaya-Vidal M, Oliva M, Castellano D, Pantano L, Bitarello B, Izquierdo D, Noguera I, et al. 2019. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat Commun* **10**: 4222.

The GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204–213.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, et al. 2016. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**: 12989.

Kwan JS, Li M-X, Deng J-E, Sham PC. 2016. FAPI: Fast and accurate P-value Imputation for genome-wide association study. *Eur J Hum Genet* **24**: 761–766.

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–511.

Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KHY, et al. 2019. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat Commun* **10**: 1025.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**: D896–901.

Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller, Manuel Conceicao ELT, di Palma MA. 2018. robustbase: Basic Robust Statistics R package version 0.93-2.

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**: 1479–1485.

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet* **48**: 593–599.

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

R Core Team. 2017. R: A Language and Environment for Statistical Computing. *R Found Stat Comput*.

Repping S, van Daalen SKM, Brown LG, Korver CM, Lange J, Marszalek JD, Pyntikova T, van der Veen F, Skaletsky H, Page DC, et al. 2006. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat Genet* **38**: 463–467.

Sanders AD, Hills M, Porubský D, Guryev V, Falconer E, Lansdorp PM. 2016. Characterizing polymorphic inversions in human genomes by single cell sequencing. *Genome Res* **26**: 1575–1587.

Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**: 500–507.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Turner DJ, Hurles ME. 2009. High-throughput haplotype determination over long distances by haplotype fusion PCR and ligation haplotyping. *Nat Protoc* **4**: 1771–1783.

Turner DJ, Shendure J, Porreca G, Church G, Green P, Tyler-Smith C, Hurles ME. 2006. Assaying chromosomal inversions by single-molecule haplotyping. *Nat Methods* **3**: 439–445.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucleic Acids Res* **44**: D710–716.