Supplementary Data
American Journal of Human Genetics, Volume *88*

Population Genetic Properties of Differentiated Human Copy Number Polymorphisms

Catarina D. Campbell, Nick Sampas, Anya Tsalenko, Peter H. Sudmant, Jeffrey M. Kidd, Maika Malig, Tiffany H. Vu, Laura Vives, Peter Tsang, Laurakay Bruhn, and Evan E. Eichler

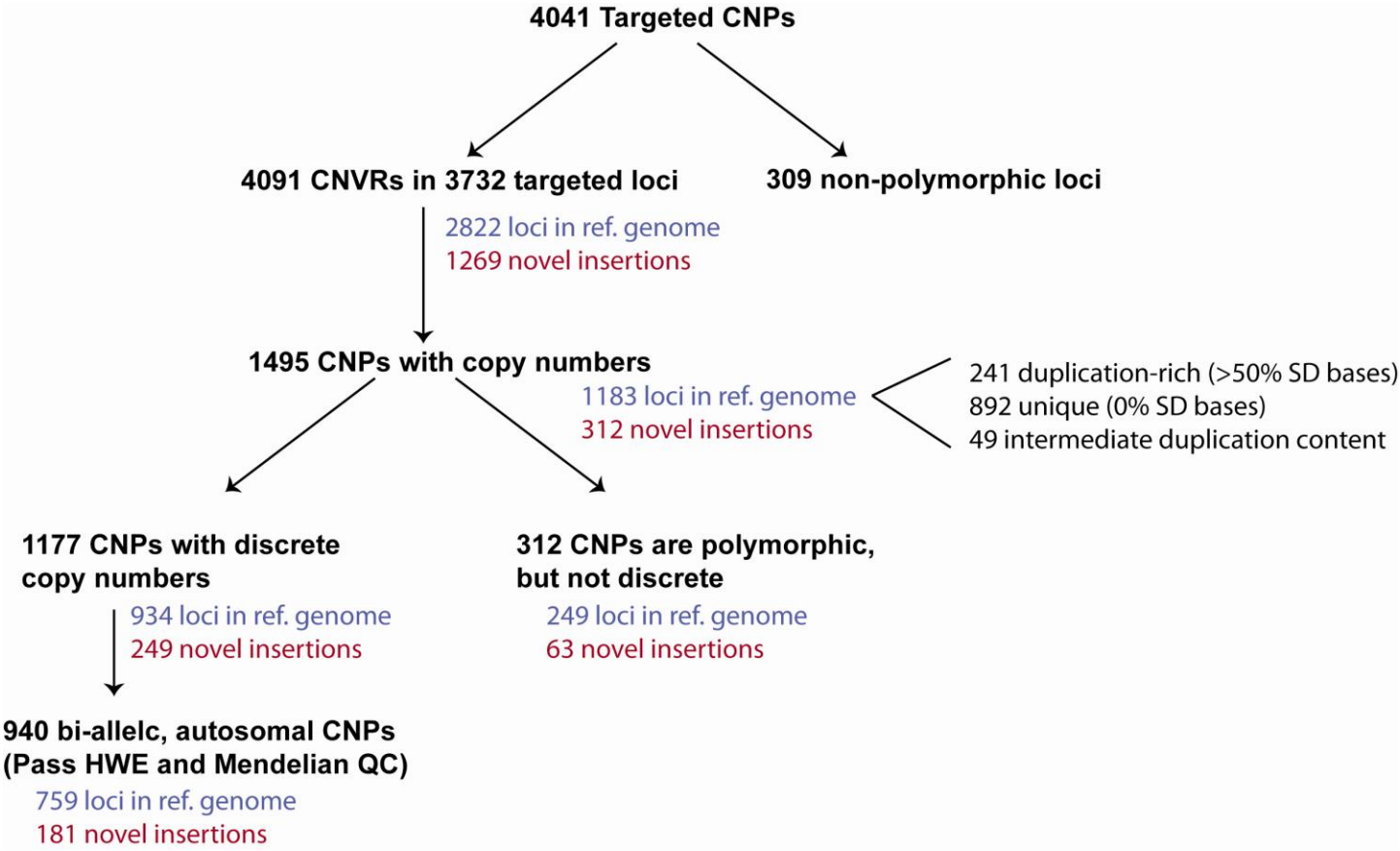

Figure S1. Classifications of CNPs Used in Analyses

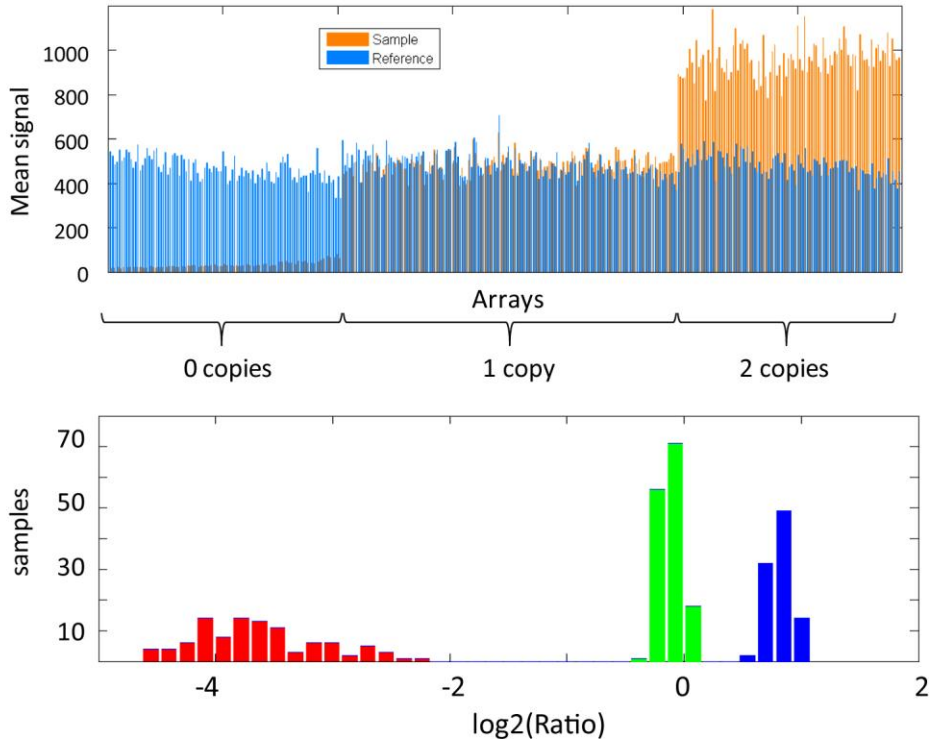

Figure S2. Copy Number Fitting Using Single Channel Intensity Data

In the top panel, the mean signal for a region for both the test and the reference samples are plotted. The reference sample has a similar mean signal for probes in this region across all hybridizations. This mean signal is used to estimate the copy number of the reference sample for this region; in this case, the reference sample has a copy number of one. Samples with the same mean signal as the reference also have one copy, samples with little to no mean signal have zero copies, and samples with double the signal of the reference have two copies. In the bottom panel, a histogram of samples based on $\log_2$ratio is shown. The three copy number genotypes form clear, discrete classes.

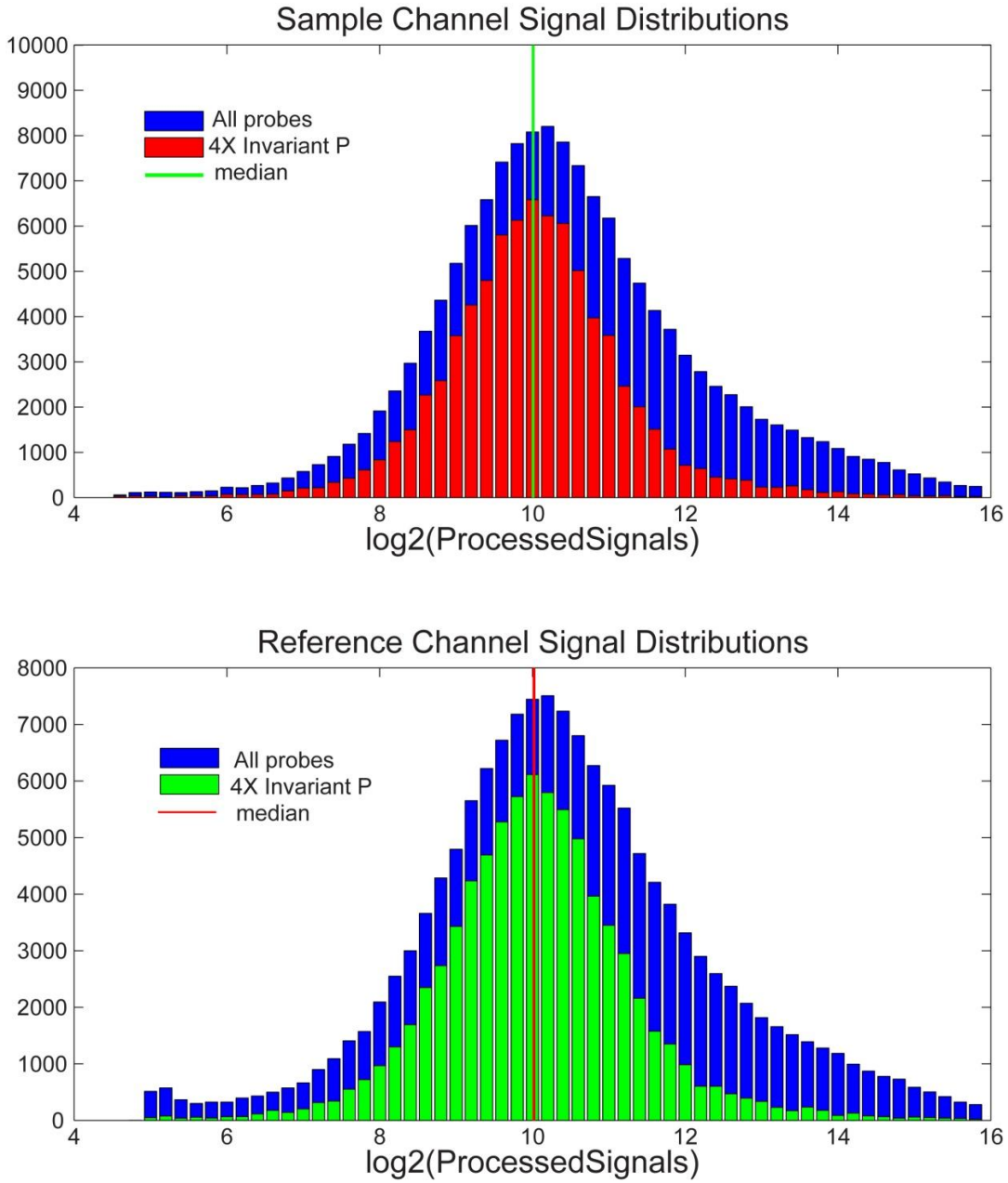

Figure S3. Distribution of Signal Intensities

The signal level associated with diploid regions was estimated by inspecting the signal distributions across all probes and all probes outside of known CNV regions. The distribution of signal intensities for all autosomal probes is highlighted in blue in the sample channel (top) and reference channel (bottom). The sample channel intensity (red-top) and reference channel intensity (green-bottom) distributions represent the probes remaining after filtering out all probes within CNVRs or within segmentally duplicated genomic regions. The histograms for filtered probes are multiplied by 4 for visual comparison. The average of the median values of the sample and reference channel is approximately 1000 counts giving a single copy value of 500 counts.

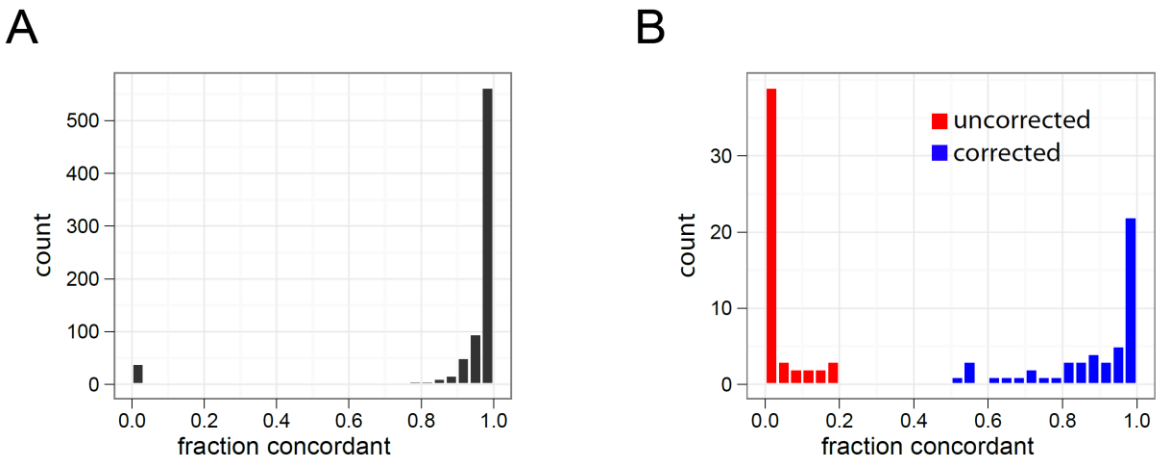

Figure S4. Microarray Copy Number Estimates Are Highly Concordant with Copy Number Estimates from Sequencing

A) Fraction of copy number estimates concordant between array-based and sequencing-based methods across 841 loci for which discrete copy numbers were estimated from array data. 709/841 (84%) of regions demonstrate >90% concordance. A small fraction of sites (51/841, 6%) display extremely low concordance; 88% of these regions overlap segmental duplications. **B)** Many of the highly discordant regions differ by an exact integer number between array- and sequencing-based copy numbers, allowing us to "correct" these copy number estimates for the true underlying baseline copy.

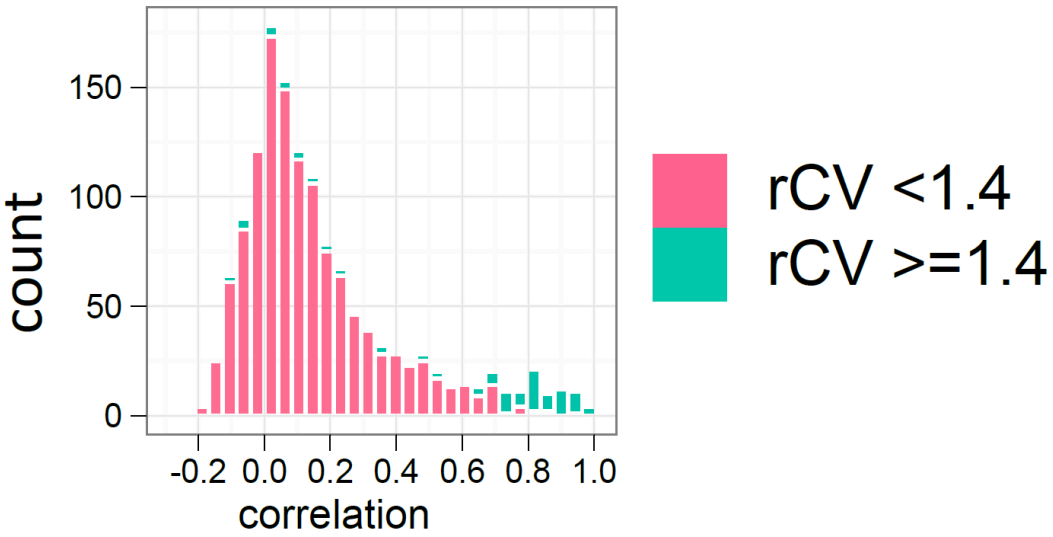

Figure S5. Distribution of Correlation Coefficients between Microarray Copy Number Estimates and Sequencing Copy Number Estimates

A histogram of correlation coefficients between estimated copy numbers and sequencing copy number estimates for regions with rCV <1.4 and regions with rCV >=1.4. Copy number estimations based on single channel intensity for genomic regions in which discrete copy number could not be assigned are highly correlated with copy number predictions made from sequencing data for array regions that have an rCV >=1.4.

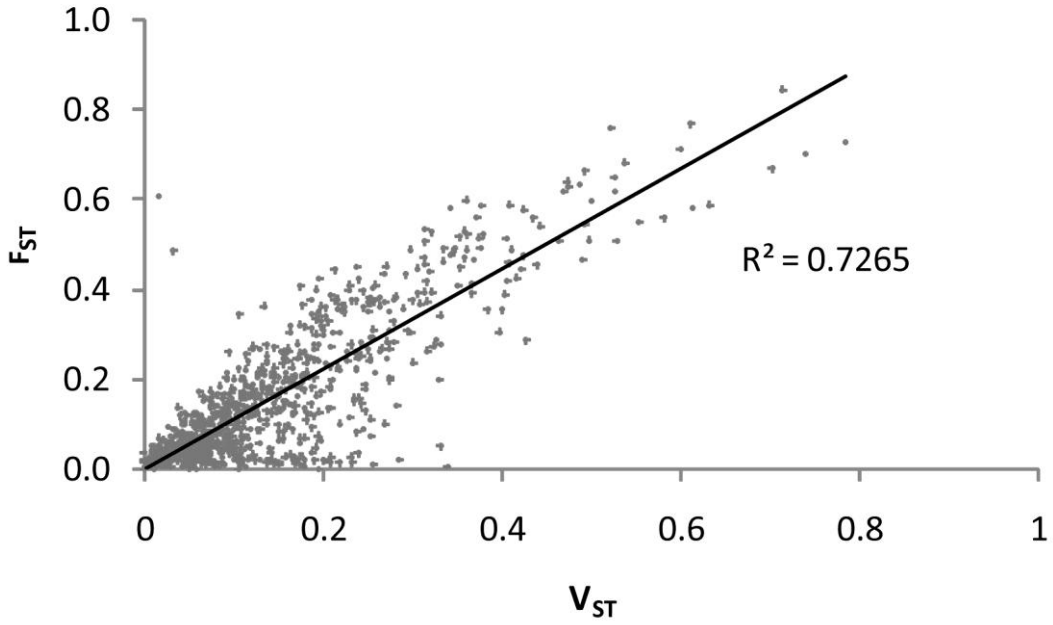

Figure S6. V_{ST} Correlated to F_{ST}

To verify that V_{ST} would be a good measure of population differentiation, we compared the maximum V_{ST} value to the maximum F_{ST} value for bi-allelic autosomal CNPs.

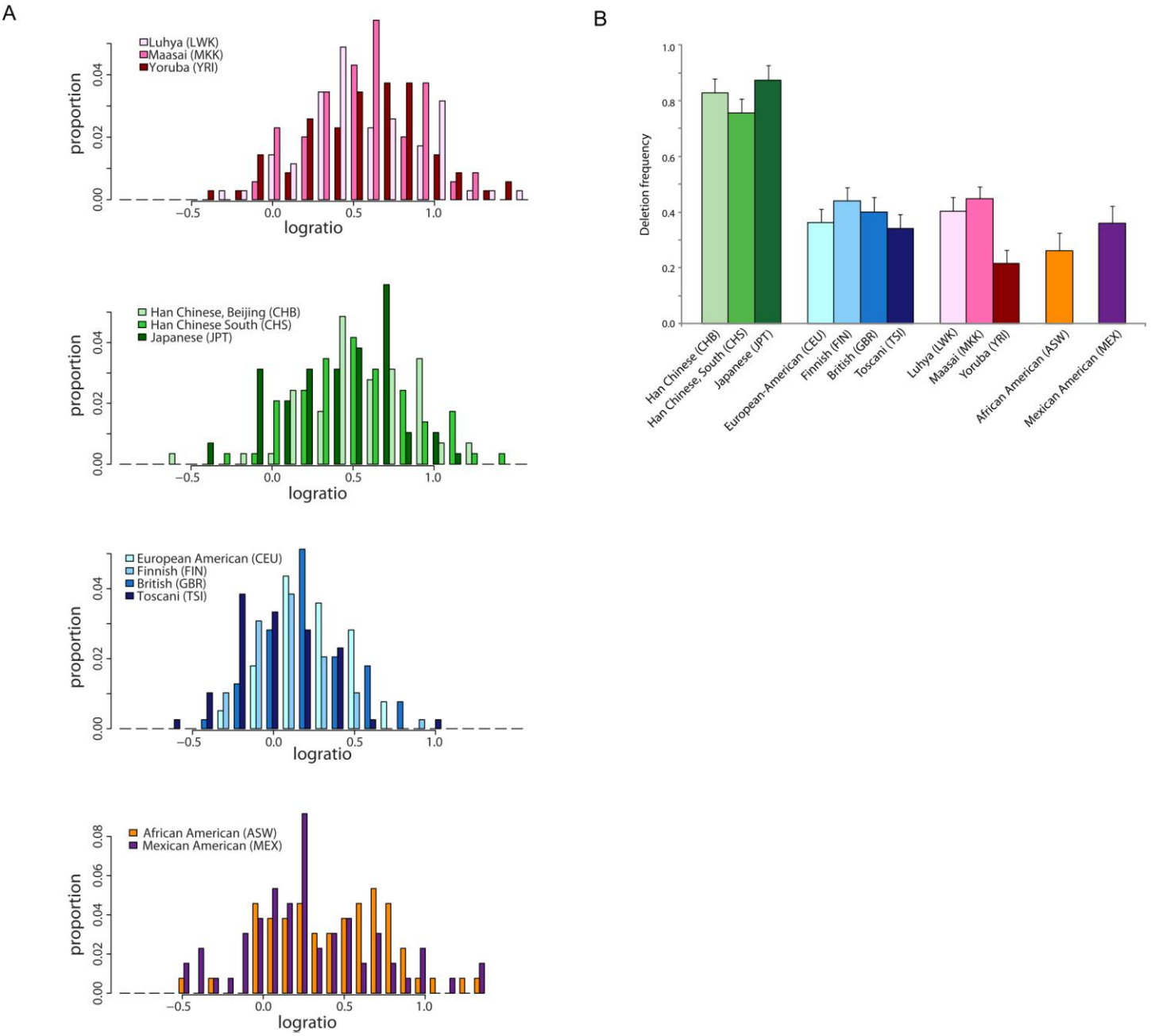

Figure S7. Reproducibility of Population Differentiation for Known Loci with Frequency Differences between Populations

A) *CCL3L1*, which is reported to have fewer copies in European individuals (Gonzales et al., 2005). Histograms of $\log_2$ratios for the unrelated individuals in each population are plotted. **B)** The deletion allele frequency for *UGT2B17* is plotted for each population. The deleted allele of this CNP has been shown to be present at a higher frequency in Asian populations (McCarroll et al., 2006). The error bars represent standard deviations.

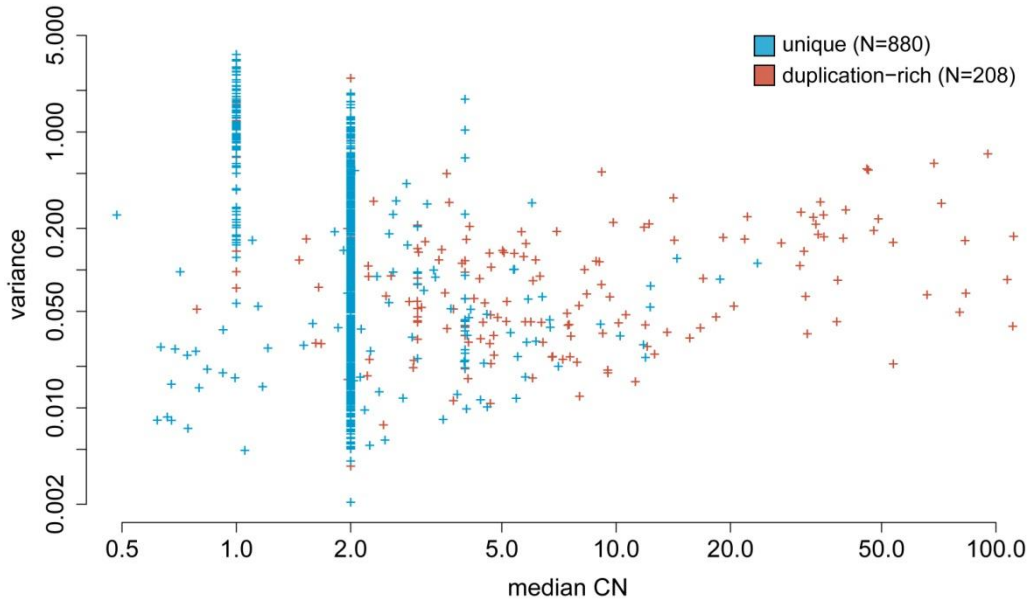

Figure S8. Variance of log_2ratios Is not Correlated to the Median Copy Number of a CNP

We compared the distribution of total variance in log_2ratios across all unrelated samples to the median copy number of all unrelated samples. We observe no correlation for CNPs in unique regions ($r^2 = 0.058$) or CNPs in segmental duplications ($r^2 = 0.029$).

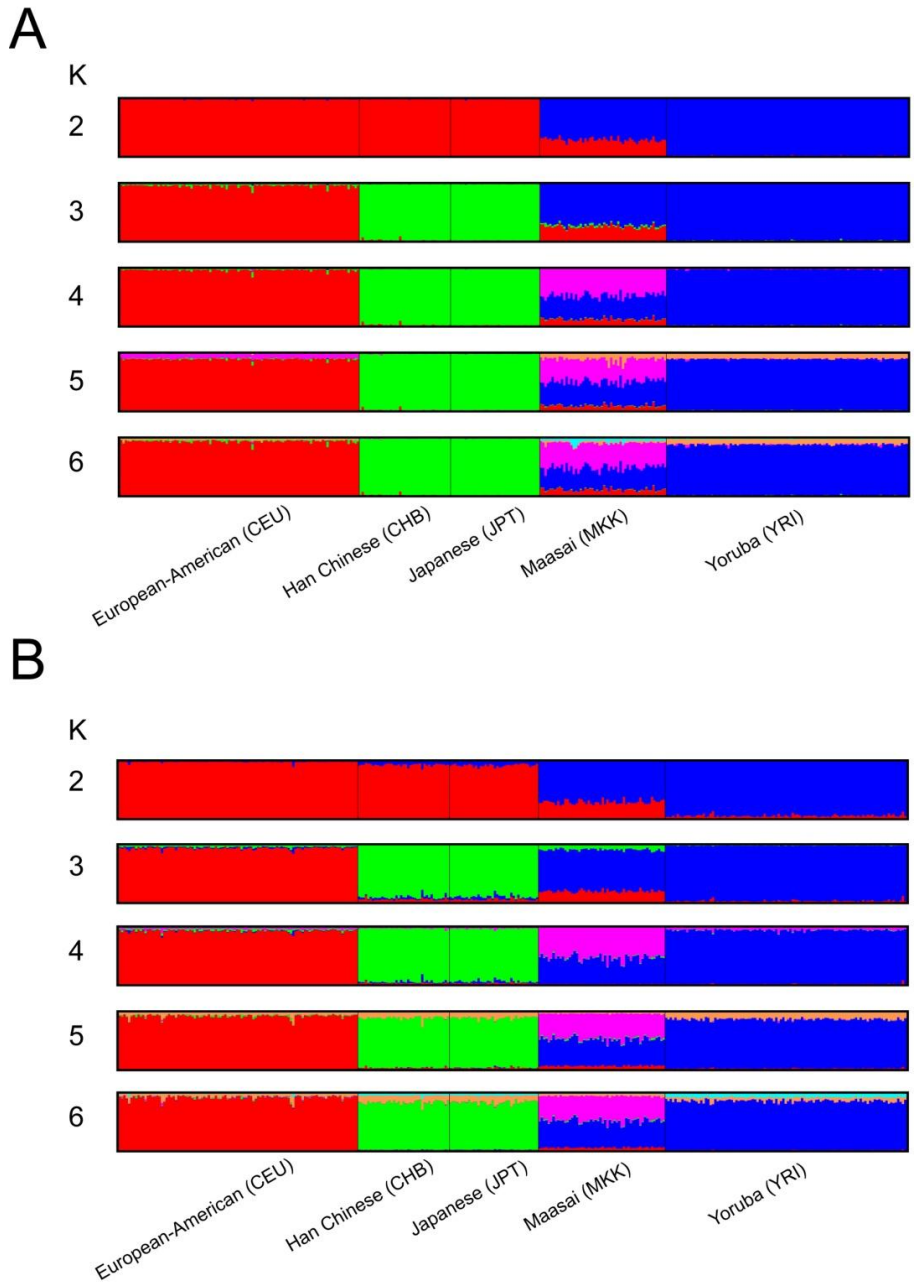

Figure S9: STRUCTURE Analysis

We performed an analysis using STRUCTURE (Pritchard *et al.*, 2000) to distinguish populations using **(A)** 940 bi-allelic CNPs or **(B)** random, autosomal SNPs that were allele frequency matched to our bi-allelic CNPs. Plots of the population membership for 336 unrelated individuals at different numbers of assumed populations (K) are shown. These 336 individuals were analyzed because there was both CNP and SNP data available. After testing multiple parameters and run lengths, we settled on the default STRUCTURE parameters with the following modifications. We used 50,000 burn-in repetitions, 50,000 MCMC repetitions after the burn-in, and we use the population information as a prior. Using these settings, we ran the program for different numbers of population clusters (K) from 1 to 6, with five iterations per value of K. The most likely value of K for our data is 4. We combined the data for all five iterations at each value of K using the program, CLUMPP, which returns the mostly likely membership of each individual in each population cluster given multiple iterations (Jakobsson et al., 2007). We used Distruct (Rosenberg, 2004) to generate the images.

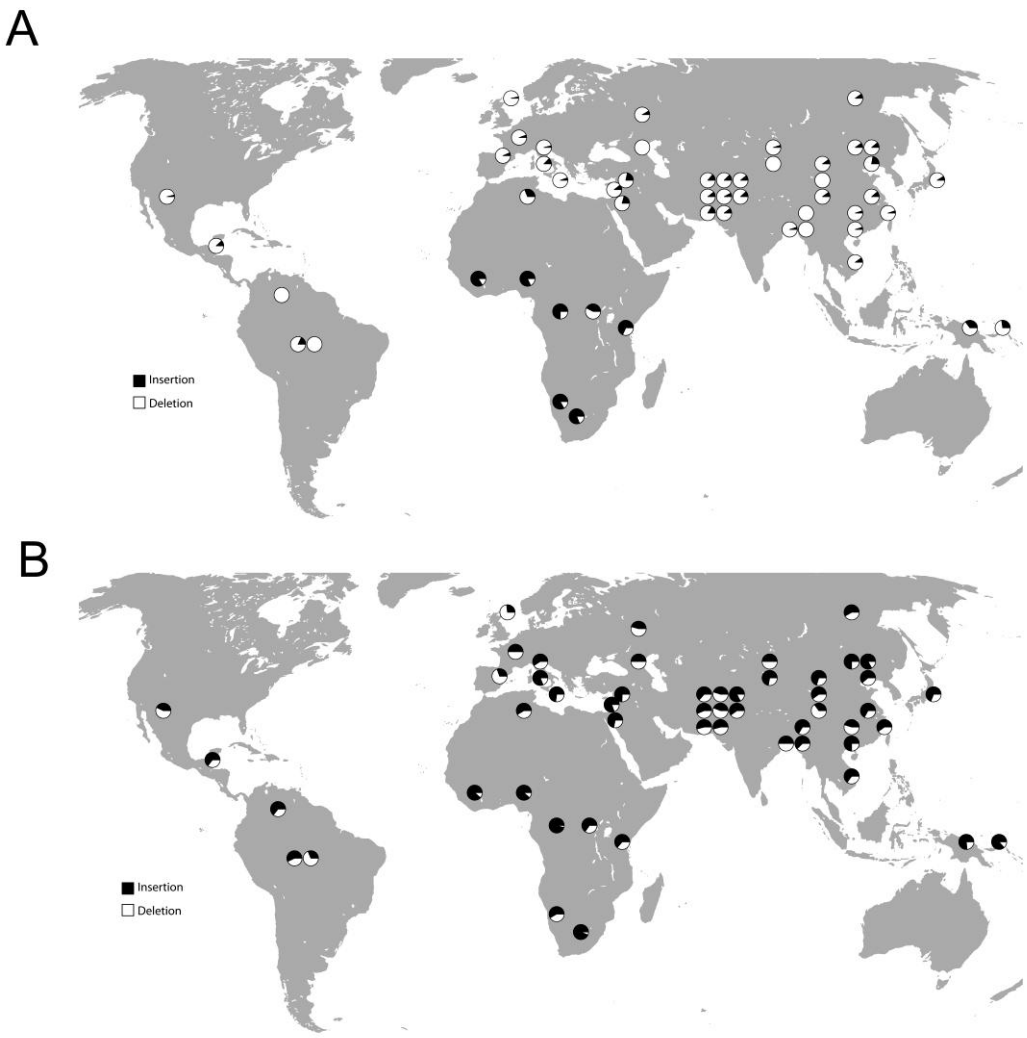

Figure S10. Worldwide Distributions of Two Novel Insertions

Pie charts of the allele frequency of the deletion (white) and insertion (black) alleles are plotted in the approximate location of each HGDP population. **A)** An insertion on chromosome 20 (novel-locus_280). **B)** An insertion in the first intron of LCT (novel-locus_687).

SUPPLEMENTARY TABLES

Table S1. Summary of Methods

Method	Purpose
Copy number determination	
ADM2 segmentation and visual data inspection	Identify and refine CNP breakpoints within targeted loci
probe clustering	Identify the most informative probes in each CNP
single copy state estimation	Determine the single channel intensity that corresponds to a single copy state
fitting integer copy number states with $\log_2$ratios	Determine whether the CNP as discrete copy number states and identify the samples in each state
fitting integer copy number states with sample channel intensities	Determine the copy number states for CNPs where the reference sample has zero copies
comparison of $\log_2$ratio and copy number of reference sample	Determine the copy numbers for CNPs that do not have discrete copy number states
Comparison of microarray copy number estimates to whole-genome sequencing data	
sequencing read-depth copy number estimation	Determine copy numbers from an orthologous technology to test the accuracy of our array-based copy number estimates
ratio of the coefficient of variation (CV)	Determine which CNPs without discrete copy numbers have accurate array-based copy number estimates
PCR and quantitative PCR assays	
PCR and quantitative PCR	Genotype specific CNPs in a larger number of individuals
Linkage disequilibrium analysis	
r^2 analysis of SNP and CNP alleles	Determine the relationship of bi-allelic CNPs to SNPs
Pearson correlation of copy number to SNP genotype	Determine the relationship of bi-allelic and multi-allelic CNPs to SNPs
multiple regression analysis	Evaluate the variables contributing to the correlation of copy number to SNP genotype
SNP haplotype phases	Test whether SNP haplotypes capture the variation of multi-allelic CNPs
Population differentiation analysis	
V_{ST}	Test for population differentiation of CNPs using array hybridization data
F_{ST}	Test for population differentiation of bi-allelic CNPs and SNPs

Table S2. Regions Targeted on the CNP Microarray
See Excel file.

Table S3. Polymorphic Novel Insertions Targeted in Microarray Design
See Excel file.

Table S4. Non-Variant Control Regions on the CNP Microarray

Chromosome	Start (hg18)	End (hg18)
chr1	23135997	23466000
chr1	32240404	32804999
chr1	205991690	206078228
chr1	207192723	208136181
chr2	400000	550000
chr2	30839541	31250000
chr2	98254331	98457914
chr2	220126146	220913126
chr3	13899362	14298853
chr3	32436354	32550000
chr3	55870149	56059460
chr3	125522206	125900000
chr4	2302567	2389651
chr4	38689025	39000512
chr5	39000512	79300000
chr5	131832406	132100000
chr5	146519693	146835222
chr6	3149820	3495198
chr6	39400000	39600000
chr6	42207525	42400000
chr7	28850000	29167069
chr7	70250000	70693285
chr7	137743285	137893285
chr8	27090341	27350000
chr8	37844489	38152357
chr8	101300000	101882000
chr9	37504916	37650000
chr9	121500978	125360267
chr10	13917075	14279792
chr10	30305099	30640237
chr10	99531094	99614661
chr11	11353927	11650000
chr11	115400000	115566825
chr12	111421663	111640806
chr12	128341073	128512417
chr13	28644654	29000000
chr14	22450000	22931651

chr14	73676171	73900000
chr15	38035103	38352997
chr15	38782403	39000000
chr16	66543817	67130253
chr16	80500000	80550000
chr17	24389874	24948320
chr17	73350000	73628135
chr18	13243662	13550000
chr18	42223274	42629843
chr19	9174321	9576249
chr19	35319187	35635517
chr20	33164714	33400332
chr20	44883862	45195752
chr21	32848897	33000000
chr22	29282126	29605446
chr22	44316588	44458127

Table S5. Integer Differences between Microarray-Based Copy Numbers and Sequencing-Based Copy Numbers

chromosome[a]	start[a]	end[a]	Difference between sequencing CN and array CN[b]
chr7	142933680	143171936	2
chr10	27678346	27680500	1
chr11	18906696	18917392	2
chr15	28393216	28462914	1
chr17	41007032	41015040	2
chr17	31440524	31520184	2
chr1	147303136	147511088	4
chr21	31354232	31356128	1
chr9	38858408	38864072	4
chr1	195005520	195068320	2
chr17	18302826	18366016	3
chr5	848743	878341	4
chr11	49667416	49709928	1
chr5	69381072	69410328	1
chr11	42926336	42927664	1
chr11	58569324	58609824	2
chr5	68857344	68890304	1
chr9	38928356	39800652	4
chr10	89027376	89065904	4
chr22	24007318	24241586	1
chr21	43794860	43796020	2
chr2	87243544	87264520	2
chr7	141416096	141440400	1

chr12	9528449	9595559	2
chr7	38359852	38363908	3
chr15	22161954	22202144	4
chr12	11396685	11433740	4
chr14	19272446	19490764	1
chr2	11312817	11317741	1
chr2	60704192	60706884	1
chr7	76168016	76386672	2
chr2	234313936	234322704	1
chr17	19440784	19478204	1
chr16	70651968	70653656	2
chrY	23941480	24027718	1
chr4	145140944	145260336	3
chr14	20432612	20481668	1
chr17	36638192	36648184	3
chr8	13643690	13645113	1
chr15	22230872	22268800	2
chr1	143670880	143792368	4

[a]Genomic position for CNPs where the copy number estimated from the microarray differ by an integer from the copy numbers estimated from sequencing read-depth. [b]Integer difference between sequencing read-depth copy number and microarray copy number. CN=copy number.

Table S6. PCR Primers and Conditions for Targeted Assays

Novel-locus_280 (Accession #: AC205876)	
Insertion forward primer	AAACCTTGCCAAATCCACAG
Insertion reverse primer	CCATTACCCTCGAAGAGCTG
Deletion forward primer	AAACCTTGCCAAATCCACAG
Deletion reverse primer	TAGCCCATGCTACCTCATCC
PCR conditions	15 ul reaction: 75 ng of DNA, 6 umoles deletion primer pair, 24 umoles insertion primer pair, and 8 ul pre-prepared master mix. Cycling conditions: 95°C for 5min, 38 cycles of 95°C for 30sec 55°C for 30sec and 72°C for 30sec, 72°C for 7min
PCR band sizes	Insertion = 164 bp; Deletion = 261 bp

Novel-locus_335 (Accession#: AC212752)	
Common forward primer	CCAGCCTAAATGTGCATCAA
Insertion reverse primer	ACTCCGCCTCAACAACAAAA
Deletion reverse primer	TGTGATTACCATGGGGCTTC
PCR conditions	15 ul reaction: 50 ng of DNA, 6 umoles of each primer, and 8 ul master mix. Cycling conditions: 95°C for 5 min, 38 cycles of 95°C for 30 sec 60°C for 30 sec and 72°C for 30 sec, 72°C for 7 min
PCR band sizes	Insertion = 185 bp; Deletion = 240 bp

Novel-locus_687 (Accession#: AC216083)	
Common forward primer	CAGGACTATGAAATGCAGAGCAGTT
Insertion reverse primer	CTCCTGGGTTAATGCCATTC
Deletion reverse primer	CCGGTGCAACTCCGTCTC
PCR conditions	15 ul reaction: 75 ng of DNA, 6 umoles insertion primer, 12 umoles deletion primer 1, 8 umoles deletion primer 2, and 9.5 ul master mix. Cycling conditions: 95°C for 5 min, 38 cycles of 95°C for 30 sec 55°C for 30 sec and 72°C for 30 sec, 72°C for 7 min
PCR band sizes	Insertion = 174 bp; Deletion = 261 bp

***OCLN* qPCR**	
Forward primer	CAGTTCGTGAAGGCAAGTTT
Reverse primer	CAACAGAAACACCCTGATCC
qPCR conditions	10 ng of DNA, .4 umoles each primer, and 5 ul SYBR green master mix.

Table S7. Copy Numbers for CNPs in the Reference Genome Assembly

See Excel file

Table S8. Copy Numbers for Novel Insertions

See Excel file

Table S9. V_{ST} Values for CNPs

See Excel file