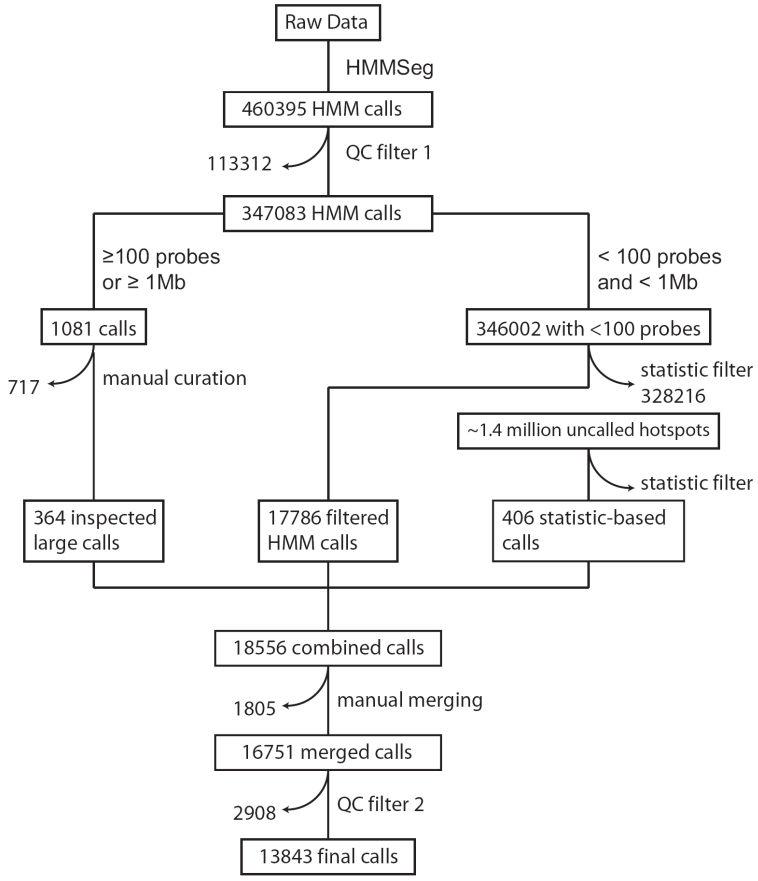
American Journal of Genetics, Volume *84*

Supplemental Data

Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease

Andy Itsara, Gregory M. Cooper, Carl Baker, Santhosh Girirajan, Jun Li, Devin Absher, Ronald M. Krauss, Richard M. Myers, Paul M. Ridker, Daniel I. Chasman, Heather Mefford, Phyllis Ying, Deborah A. Nickerson, and Evan E. Eichler



QC Filter 1	QC Filter 2
Stdev LogR≤0.25 |mean LogR|≤0.1 mean BDev<0.05	number of calls ≤25 possible artifacts in manual curation of big calls ≤2 possible artifacts in manual merging ≤2

Statistic Filter		
Homozygous Deletions	Heterozygous Deletions	Duplications
Probes ≥ 3 LogR Z-Score ≤ -4 BDev ≥ 0.1 or Probes ≥ 3 LogR Z-Score ≤-8	Probes ≥10 LogR Z-Score ≤ -1.5 ≤10% Het Calls	Probes ≥ 10 LogR Z-Score ≥ 1.5 Het BDev ≥ 0.075

Figure S1. CNV analysis Flow Chart. The number of calls discarded and remaining after each step is indicated. A detailed description of the CNV calling procedure may be found in Methods.

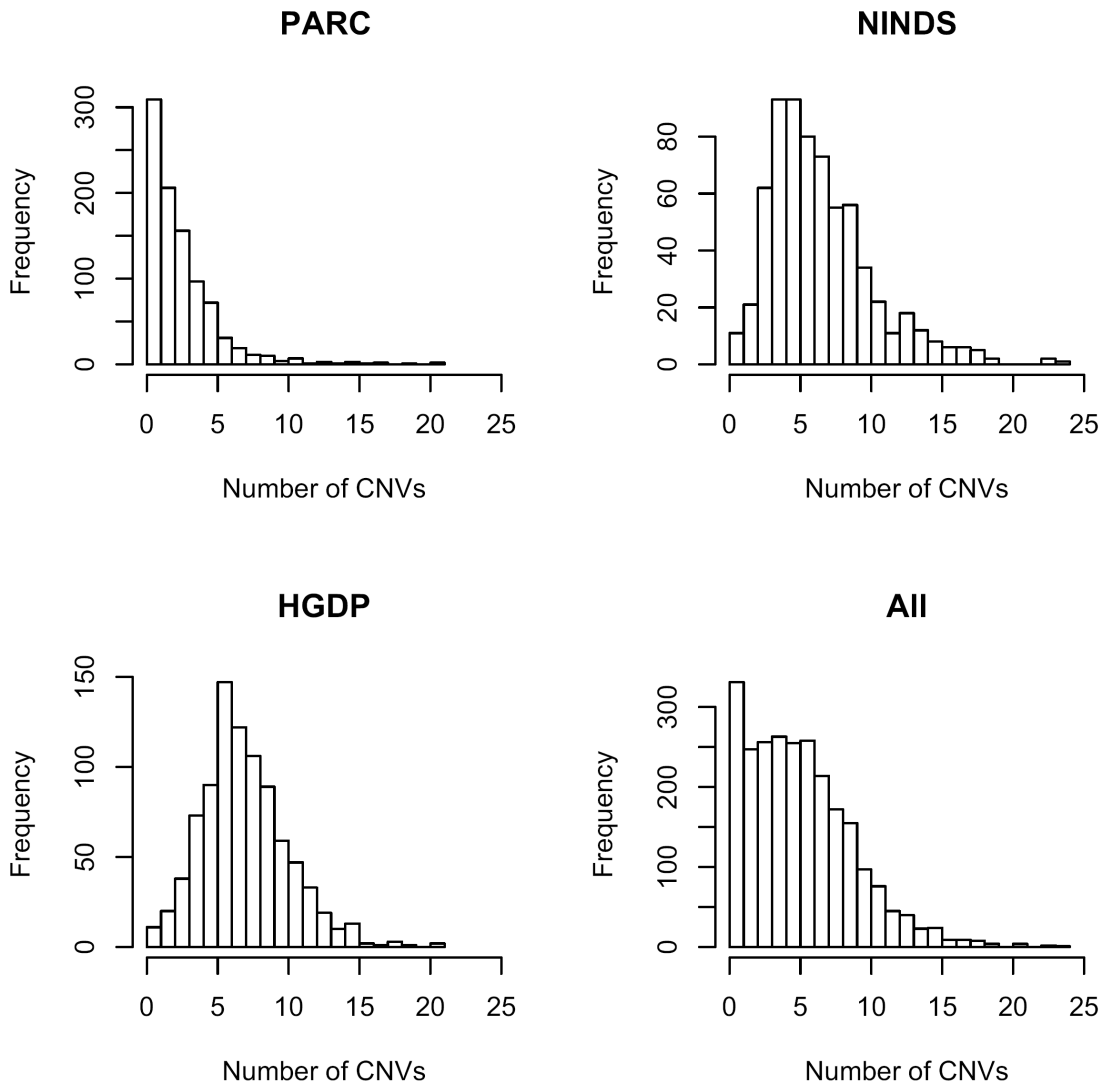


Figure S2. Distribution of CNVs per sample.
Histograms of the number of CNVs per sample over all studies and by individual study are shown, demonstrating varying sensitivity to CNVs according to probe density. PARC and HGDP samples were genotyped using Illumina 317K arrays and Illumina 650Y arrays, respectively. NINDS samples were typed using a combination of Illumina 317K + 240S and Illumina 550K arrays. As a result of quality control, all distributions appear approximately unimodal and lack apparent outliers.

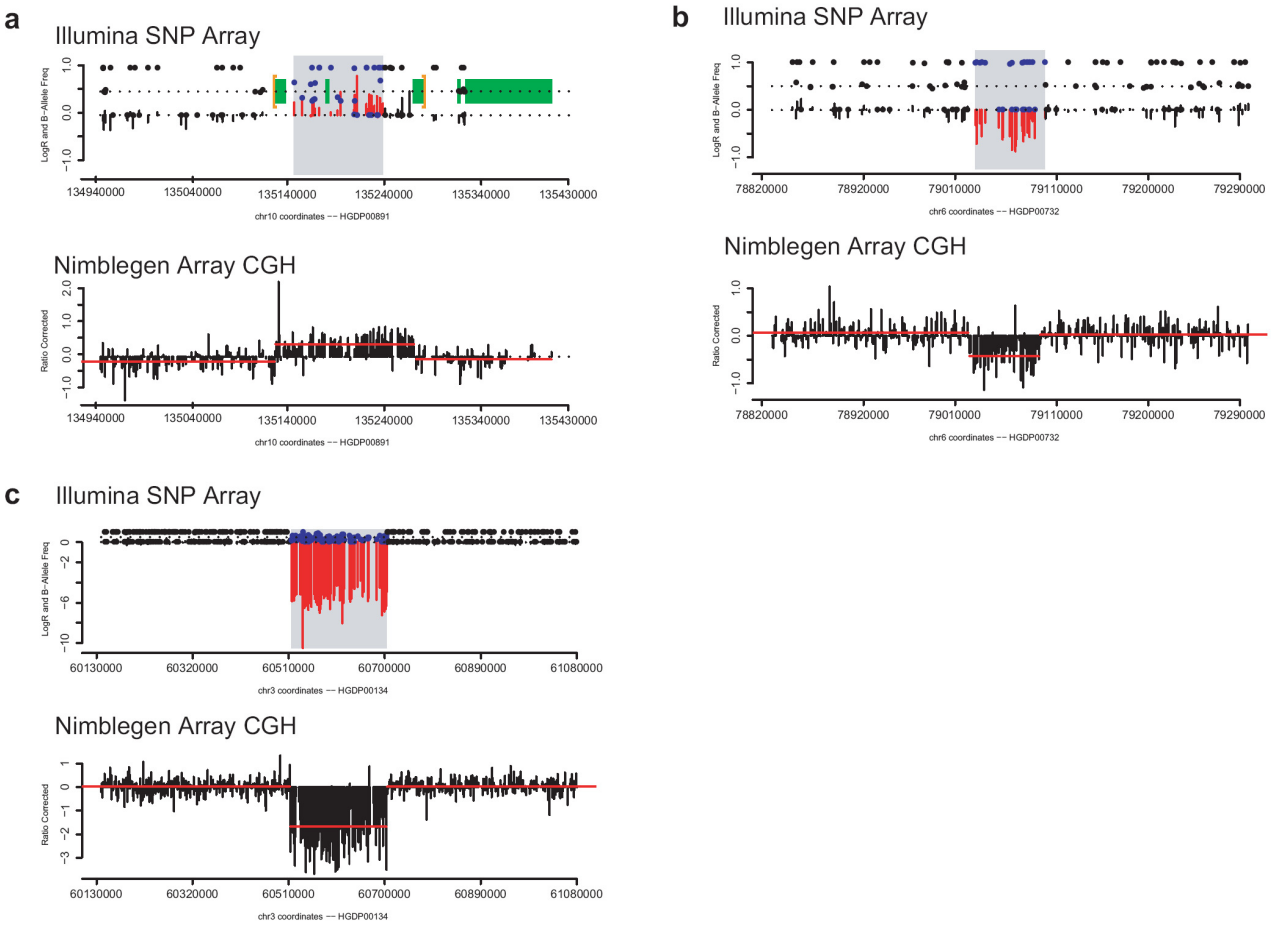


Figure S3. Validation of CNVs by Nimblegen array-CGH with SegMNT.
A homozygous deletion (a), deletion (b), and duplication (c) confirmed by SegMNT. CNVs were considered validated by SegMNT if there existed a segment in the array-CGH data with |Z-score| ≥ 1 intersecting the Illumina call.

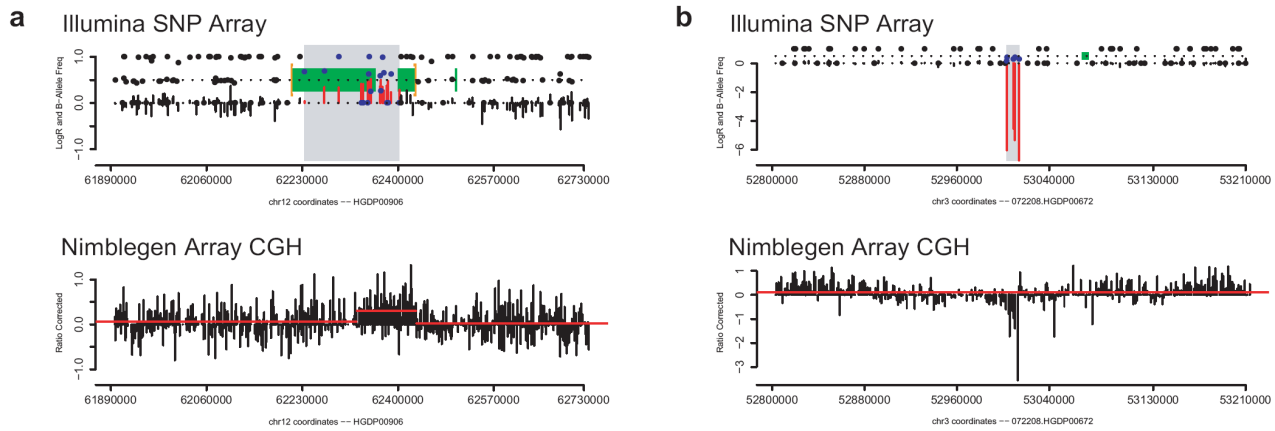


Figure S4. Manual validation of CNVs by array CGH.
In certain cases, CNV calls failed to intersect a SegMNT call with |Z-score| ≥ 1 within the validation data, but a CNV could be observed by manual inspection. Shown are an example (a) duplication and (b) homozygous deletion.

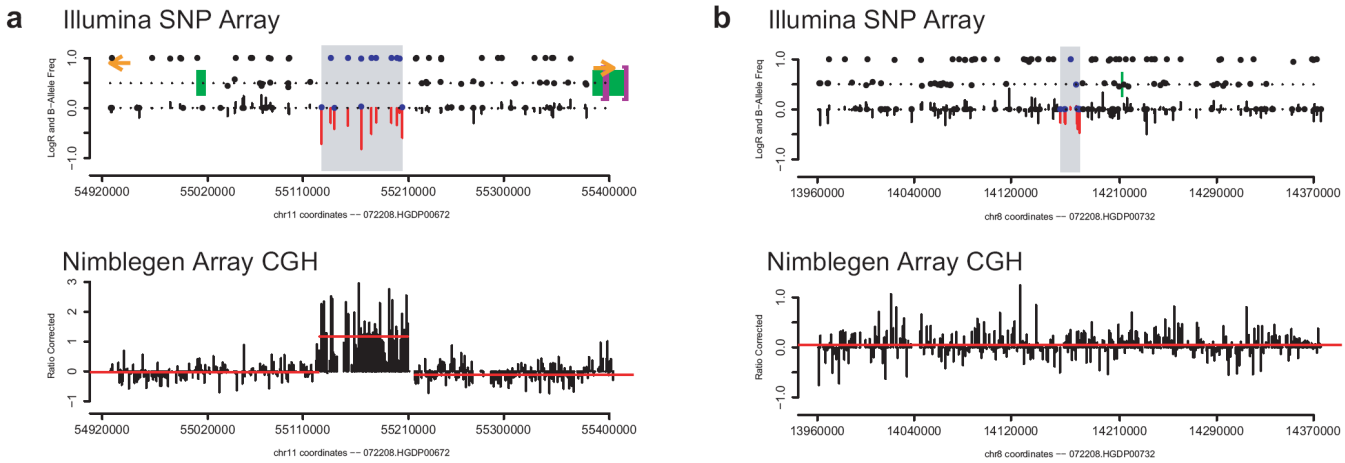


Figure S5. CNVs unable to be confirmed by array CGH. (a) Due to a homozygous deletion in the reference sample, CNVs at the chr11:55M locus could not be confirmed by array CGH, but were previously reported common CNVs.[1-4]. (b) In other cases, called CNVs were likely to be false positives.

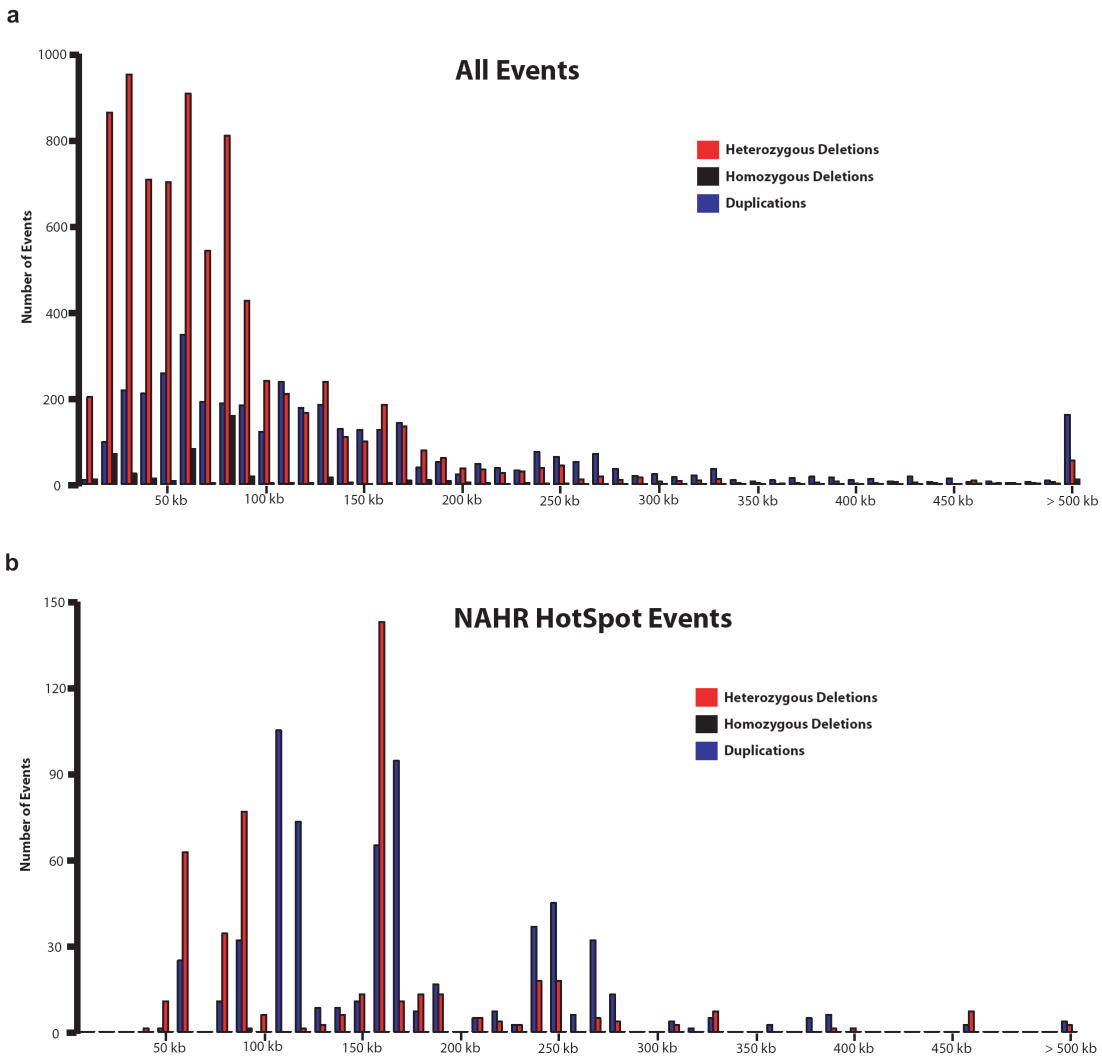


Figure S6. CNV size distribution.
(a) The number of CNVs (>10 probes) observed (y-axis) within non-overlapping 10-kb size bins (x-axis). (b) Similar to (a), only including hotspot-mediated CNVs. The distribution is dominated by polymorphic (>1%) events, corresponding to ~56% of all CNVs. CNVs in more than one individual account for ~94% of all calls.

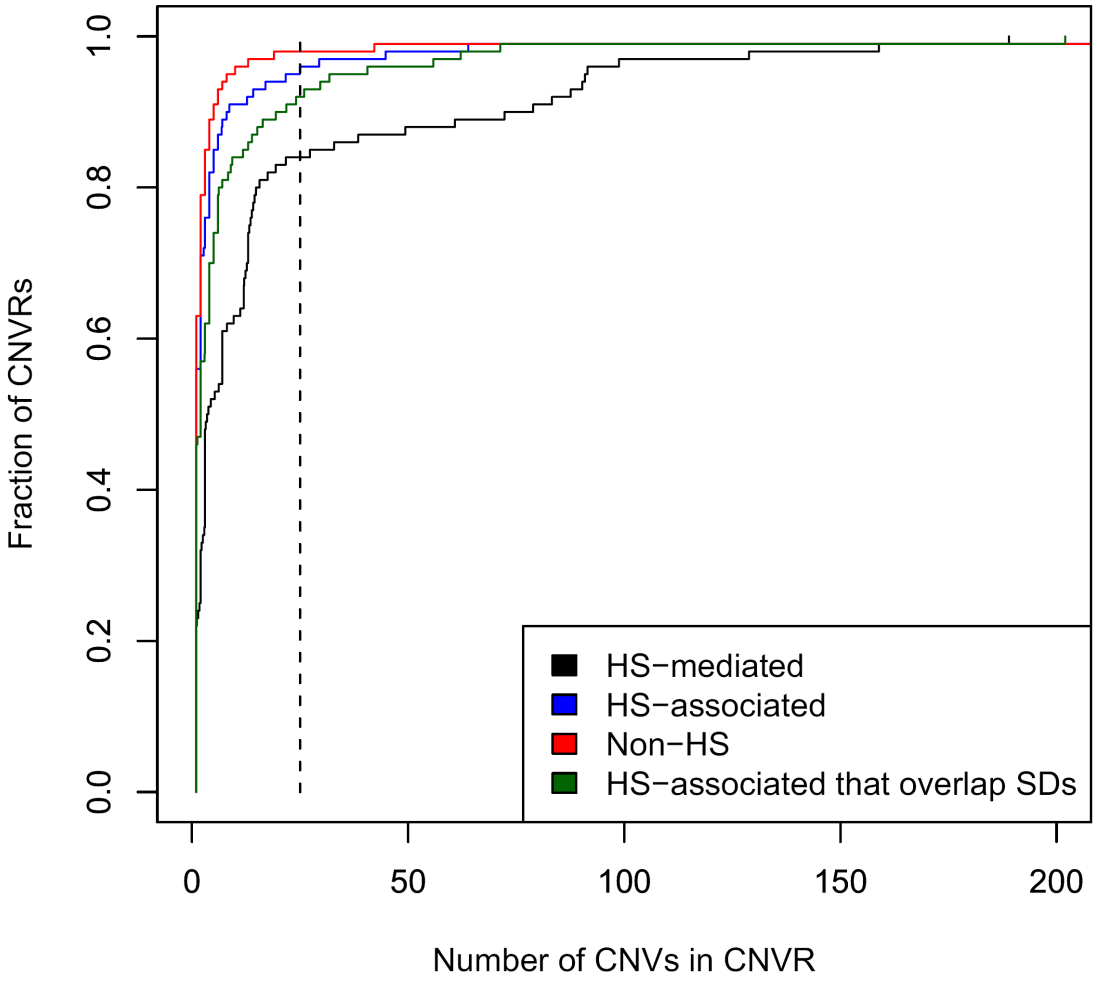


Figure S7. Cumulative Distribution of CNVR frequencies.
The fraction of CNVRs with less than a certain number of CNVs is shown for CNVRs defined as non-hotspot, hotspot-associated, hotspot-associated events overlapping segmental duplications, and hotspot-mediated events. The dotted line indicates ~1% frequency (sample not allele frequency) or 25 CNVs. Comparison of hotspot-associated events and hotspot-associated events restricted to those overlapping segmental duplications demonstrates the enrichment of segmental duplications for CNVs. CNVs assigned as hotspot-mediated are putative regions of NAHR, and demonstrate increased frequencies compared to non-hotspot or hotspot-associated events.

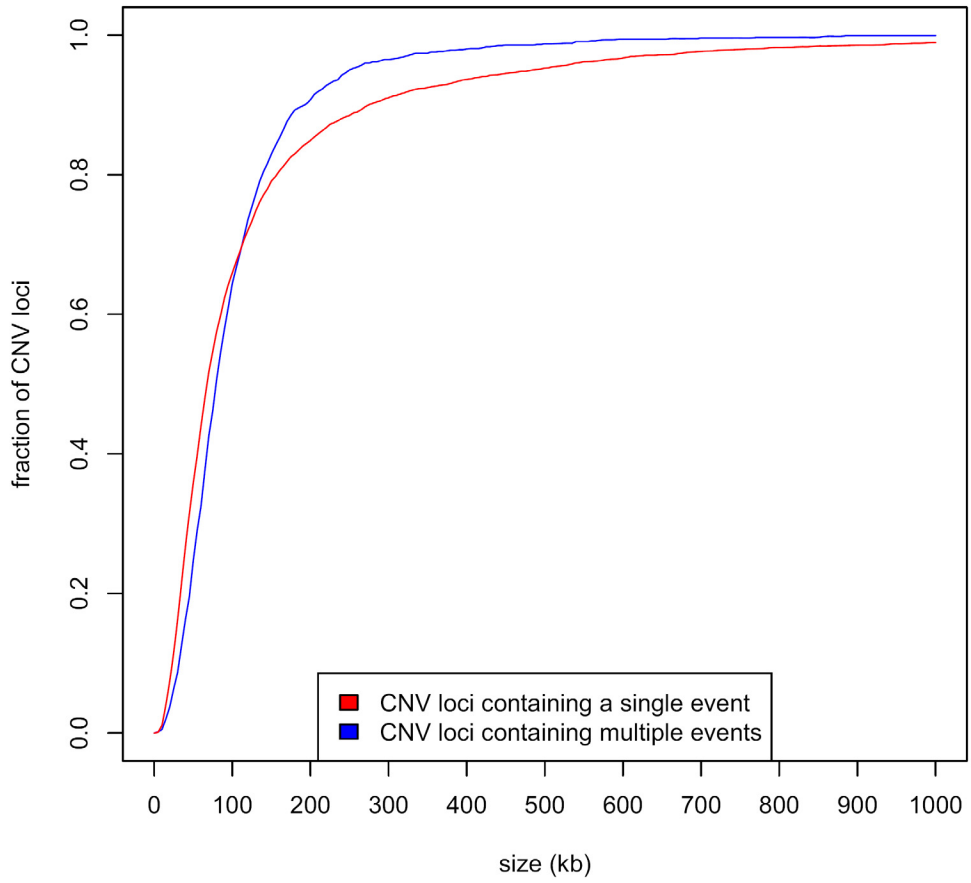


Figure S8. Cumulative distribution of CNV locus size. CNV loci have been separated into those containing a single CNV and those containing multiple events. At larger sizes (>500kb), there is a significant enrichment for events observed in a single individual ($p = 9.2 \times 10^{-8}$).

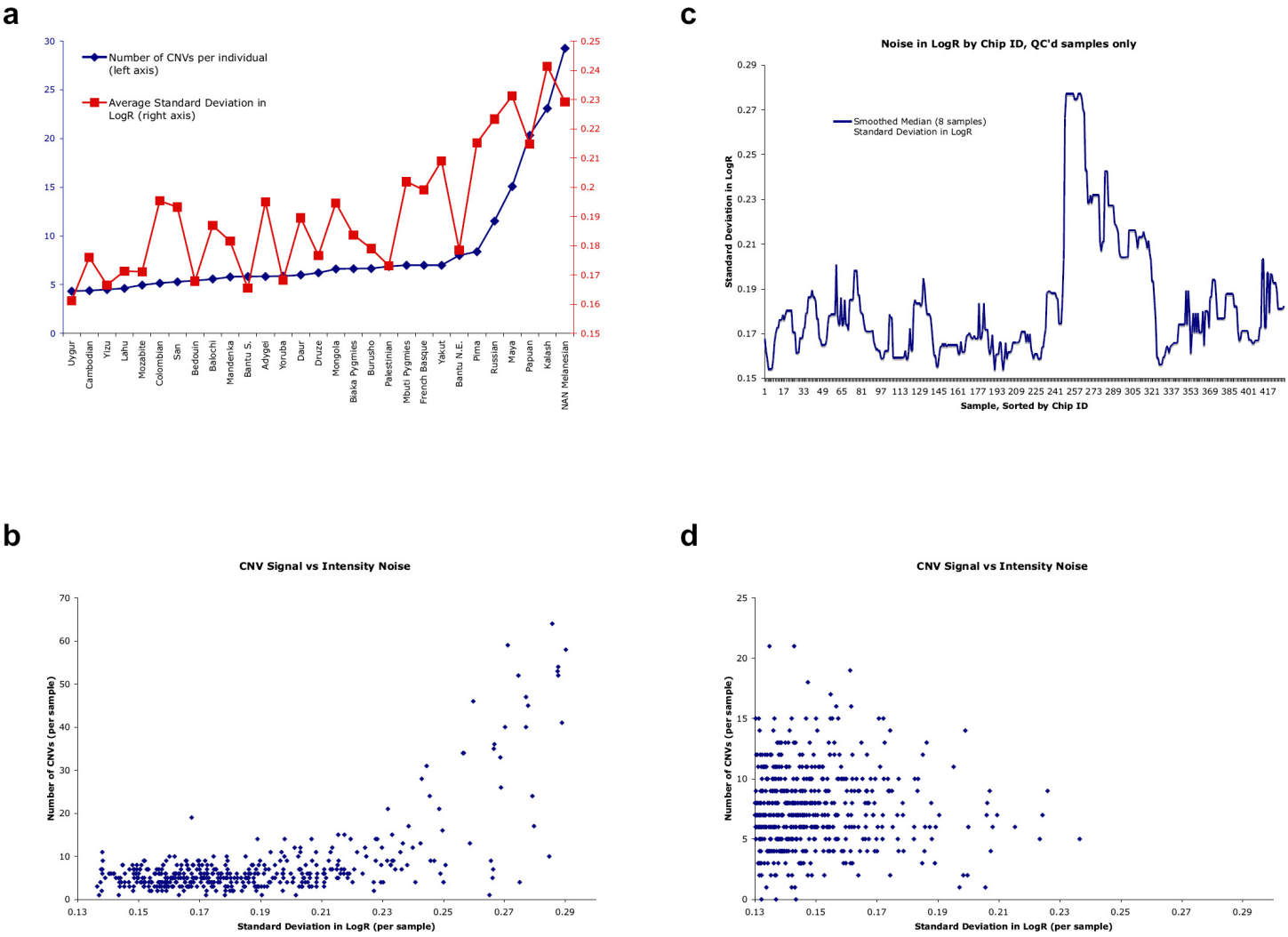


Figure S9. Relationship between hybridization/intensity noise and CNV count from SNP genotype data and CNV annotations described in Jakobsson et al. (panels a-c) or the SNP genotype data from Li et al. and the CNVs annotated here (panel d). (a) Averages in the number of annotated CNVs (blue; left y-axis) and LogR Ratio standard deviation (red; right axis) are shown for each population (x-axis) analyzed in Jakobsson et al.,[5] sorted from left to right by increasing CNV counts. Linear regression identifies a significant positive correlation between these variables ($p = 9.94 \times 10^{-7}$). (b) The number of annotated CNVs (y-axis) plotted against the standard deviation in LogR Ratio (x-axis) is shown for each sample analyzed in Jakobsson et al.; the positive correlation seen is highly significant ($p \sim 1 \times 10^{-50}$). (c) The median of the standard deviation in LogR Ratio (y-axis) is shown for sliding windows of 8 samples (same data as in (a) and (b)), sorted according to chip ID. A large spike in noise can be seen at approximately sample 250; these samples are heavily enriched for the outlier populations in panel (a). (d) Similar to (b), CNV counts (y-axis) are plotted against standard deviation in the LogR Ratio (x-axis) using the annotations generated in this study which used independently generated SNP genotype data originally described by Li et al.[6]

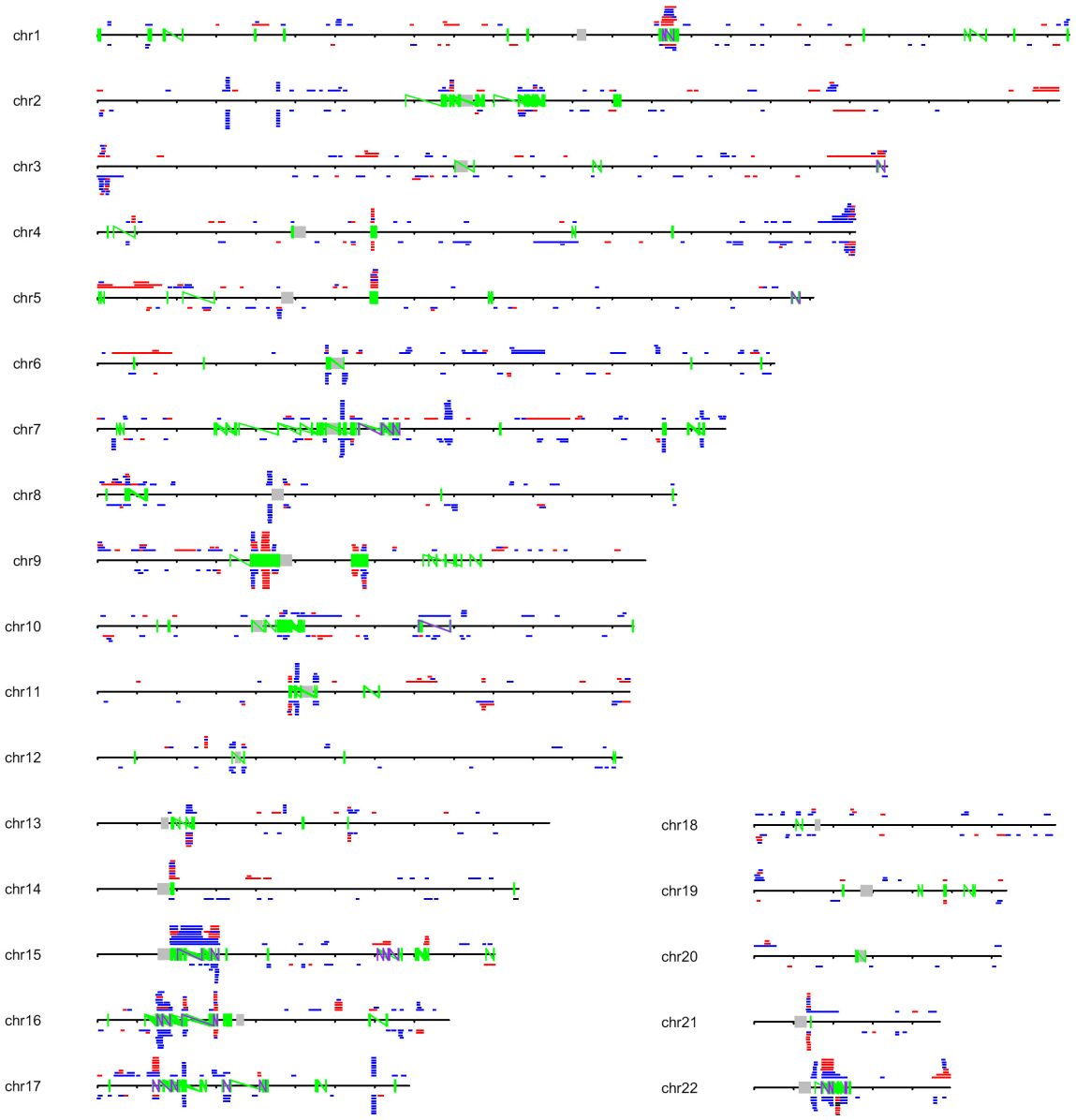


Figure S10. Map of CNVs >500kb in affected versus unaffected individuals.
For each chromosome, CNVs in affected individuals are plotted above the axis while unaffected individuals are plotted below. Duplications, deletions, and homozygous deletions are plotted blue, red, and black, respectively. Along the axis, ticks are spaced 10Mb apart, centromeres are indicated in gray, and rearrangement hotspots are indicated in green as two vertical lines connected by a diagonal. Rearrangement hotspots in which either the NAHR-mediated duplication or deletion has been associated with disease are highlighted in purple. Plotting has been cropped above eight overlapping CNVs at a given locus.

Table S1. List of All CNV Calls
See Excel table online at http://www.ajhg.org/.

Table S2. Breakdown of CNV validation via Nimblegen array CGH.

Category	Count		Citation*
overlaps SegMNT (> 1 std dev) event, no reference CNV	47		
reference has CNV, but inference of non-diploid state possible	3		
common CNV	11		
chr11:55M		7	1,2,3,4
chr15:32M		1	1,2,3,4
chr2:242M		3	2,5,6
manual inspection of aCGH data	14		
ambiguous aCGH data	3		
unconfirmed	20		
Total	98		
Number of Samples	12		

* - 1. Sebat et al., *Science* 305:525-528 (2004). 2. Redon et al., *Nature* 444:444-454 (2006). 3. Tuzun et al., *Nat Genet* 47:727-732 (2005). 4. McCarroll et al., *Nat Genet* 38:86-92 (2006). 5. Sharp et al., *Am J Hum Genet* 77:78-88 (2005). 6. Wang et al., *Genome Res* 17:1665-1674 (2007).

Table S3. Validation and Sensitivity Estimation by Array CGH.
(a) Validation of CNV calls based on Illumina SNP array analysis using conventional array CGH. For CNV calls used in later analyses in this study, a filter requiring Z-score ≥ 1.5 was used. For future reference, the subset of CNVs with Z-score ≥ 2 demonstrates a higher validation rate at the cost of sensitivity. (b) Sensitivity estimates of CNV detection by Illumina SNP array analysis. CNV calls from our analysis of Illumina SNP arrays were intersected with calls made by conventional array CGH at varying thresholds. The fraction of Nimblegen-CGH SegMNT CNV calls detected using SNP array-based analysis is shown.

Table S3A. Validation of Illumina CNVs in 12 samples via Nimblegen array-CGH.

Minimum Illumina Z-score	Num CNVs	Validated	Unvalidated	Validation Rate
1.5	98	75	23	77%
2	71	66	5	93%

Table S3B. Rates of detection of Nimblegen array-CGH CNVs.

	Minimum Nimblegen segMNT Z-score	
Minimum Illumina Z-score	1.5 (65 events)	2 (23 events)
1.5	56% (37 events)	83% (19 events)
2	52% (34 events)	78% (18 events)

Table S4. CNV relationship to segmental duplications, non-allelic homologous recombination, and rearrangement hotspots.

Category	Count	overlaps SD (fraction)	flanking SD pair within 10 kb of breakpoints	enrichment for SDs	enrichment for flanking SD pairs
all HMM-inferred CNVs	13474	3857 (0.29)	697 (0.05)	1.6	25.7
Simulation**	13474	2376 +/- 37.5* (0.18 +/- 0.003)	27.1+/- 5.7 (0.002 +/- 0.004)	1	1

SD - segmental duplication, * one standard deviation, ** 100 sets of randomly placed, size-matched controls

Table S5. CNV Counts per HGDP population

Population	Samples	CNVs	Ave CNV Count
Adygei	15	85	5.7
Balochi	24	184	7.7
BantuKenya	11	95	8.6
BantuSouthAfrica	7	58	8.3
Basque	24	159	6.6
Bedouin	44	330	7.5
BiakaPygmy	18	173	9.6
Brahui	25	201	8.0
Burusho	24	204	8.5
Cambodian	9	50	5.6
Colombian	7	49	7.0
Dai	6	48	8.0
Daur	9	51	5.7
Druze	40	287	7.2
French	24	183	7.6
Han	33	254	7.7
Han-NChina	10	52	5.2
Hazara	20	111	5.6
Hezhen	8	80	10.0
Italian	11	82	7.5
Japanese	26	185	7.1
Kalash	22	140	6.4
Karitiana	13	112	8.6
Lahu	8	55	6.9
Makrani	25	204	8.2
Mandenka	22	163	7.4
Maya	18	145	8.1
MbutiPygmy	11	80	7.3
Melanesian	9	107	11.9
Miao	9	57	6.3
Mongola	10	56	5.6
Mozabite	29	197	6.8
Naxi	7	55	7.9
Orcadian	14	86	6.1
Oroqen	9	55	6.1
Palestinian	44	337	7.7
Papuan	15	155	10.3
Pathan	22	131	6.0
Pima	13	72	5.5
Russian	24	183	7.6
San	4	29	7.3
Sardinian	27	206	7.6
She	9	59	6.6
Sindhi	21	148	7.0
Surui	8	49	6.1
Tu	9	66	7.3

Tujia	10	76	7.6
Tuscan	7	55	7.9
Uygur	10	66	6.6
Xibo	8	56	7.0
Yakut	23	163	7.1
Yi	10	62	6.2
Yoruba	21	192	9.1
Total	**886**	**6538**	**7.4**

Table S6. CNVs for which observed phenotype does not match previously reported associations

CNV coordinates (hg17)	size (Mb)	type	disease*	study**	disease locus	coordinates (MB	size (Mb)	disease state	reported disease*	overlap***	study**
chr1:143500000-145000000	1.5	loss	A	1	1q21.1	chr1:143.5-145.5	2	loss	schizophrenia, variable pediatric	0.74	2,5,6
chr3:197179156-198842299	1.7	loss	S	2	3q29	chr3:196.9-198.8	2	loss	syndromic MR +/- autism	0.84	7
chr3:197224662-198573215	1.3	loss	S	3	3q29	chr3:196.9-198.8	2	loss	syndromic MR +/- autism	0.68	7
chr3:197232320-198829110	1.6	loss	S	2	3q29	chr3:196.9-198.8	2	loss	syndromic MR +/- autism	0.81	7
chr15:21205735-26360355	5.2	gain	S	2	15q11-q13	chr15:20.2-26.4	6.2	both	loss - PW/A syndrome gain - autism	0.83	8
chr15:69601300-73890800	4.3	loss	AS	4	15q24.1-q24.2	chr15:70.7-73.4	2.7	loss	syndromic MR	1	9
chr16:15023758-16366867	1.3	loss	S	2	16p13.11	chr16:15.1-16.3	1.2	loss	MR/MCA	1	10,11
chr16:15168237-18084698	2.9	loss	S	2	16p13.11-p12.3	chr16:15.1-18.4	3.2	loss	MR/MCA	0.9	10,11
chr16:29474810-30099409	0.6	gain	S	2	16p11.2	chr16:29.3-30.2	0.9	both	autism	0.68	1,12
chr17:31895171-33318472	1.4	gain	S	2	17q12	chr17:31.8-33.4	1.6	both	loss – diabetes, renal disease gain - epilepsy	0.89	13
chr22:17014900-19786200	2.8	loss	AS	4	22q11.21	chr22:17-19.9	2.9	loss	VCFS	0.96	14

*A - autism, S - schizophrenia, AS - autism spectrum disorder, MR - mental retardation, PW/A - Prader-Willi/Angelman, MCA - multiple congenital abnormalities, VCFS - velocardiofacial syndrome

** - 1: Weiss et al. ,NEJM 358:667-675 (2008) , 2:Stone et al., Nature Epub 2008 Jul 30., 3:Walsh et al., Science Science 320:539-543 (2008), 4: Marshall et al., Am J Hum Genet 82:477-488 (2008), 5: Stefansson et al., Nature Epub 2008 Jul 30, 6: Mefford et al., in press., 7: Willatt et al., Am J Hum Genet 77:154-160 (2005), 8: Veltman et al., Psychiatr Genet 15:243-254 (2005), 9: Sharp et al., Hum Mol Genet 16:567-572 (2007), 10: Hannes et al., J Med Genet Epub 2008 Jun 14., 11: Ullman et al., Hum Mutat 28:674-82 (2007), 12: Kumar et al., Hum Mol Genet 17: 628-638 (2007), 13: Mefford et al., Am J Hum Genet 81:1057-1069 (2007), 14: McDermid et al., Am J Hum Genet 70:1077-1088 (2002).

***overlap calculated as base-pair fraction of disease locus covered by CNV

Supplementary References

1. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525-528
2. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727-732
3. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM (2006) Common deletion polymorphisms in the human genome. Nat Genet 38:86-92
4. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, et al. (2006) Global variation in copy number in the human genome. Nature 444:444-454
5. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998-1003
6. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104