Supplementary Information for

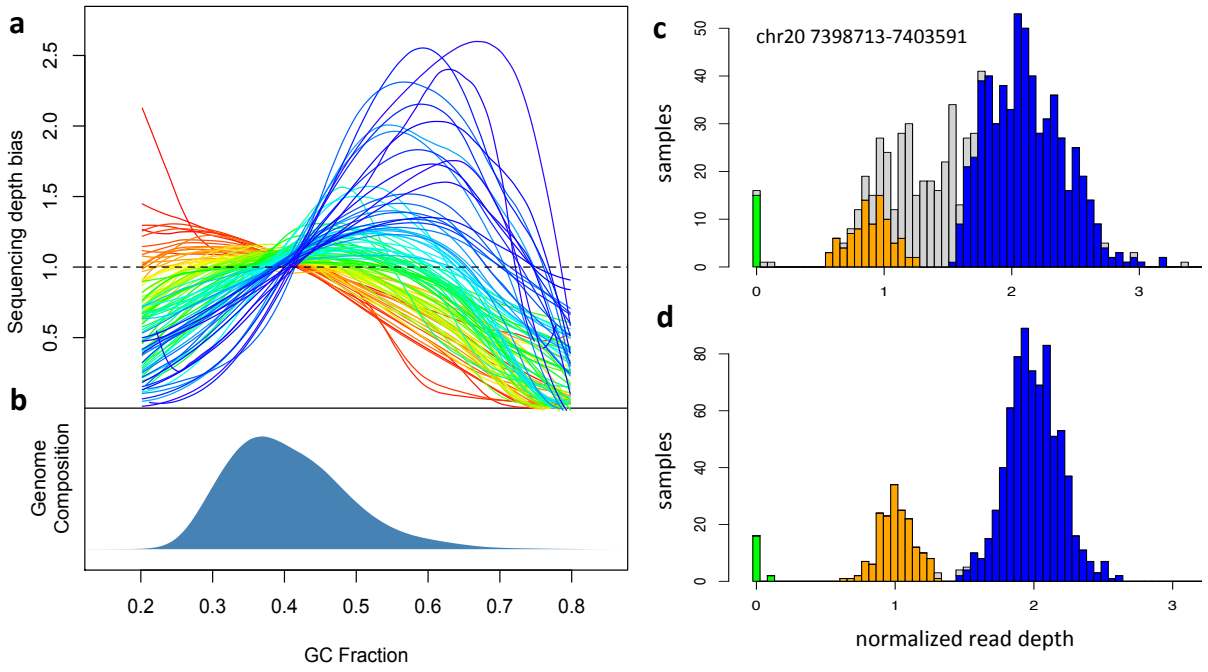
"Large multi-allelic copy number variations in humans"

RE Handsaker, V Van Doren, JR Berman, G Genovese,
S Kashin, LM Boettger, SA McCarroll

Table of Contents

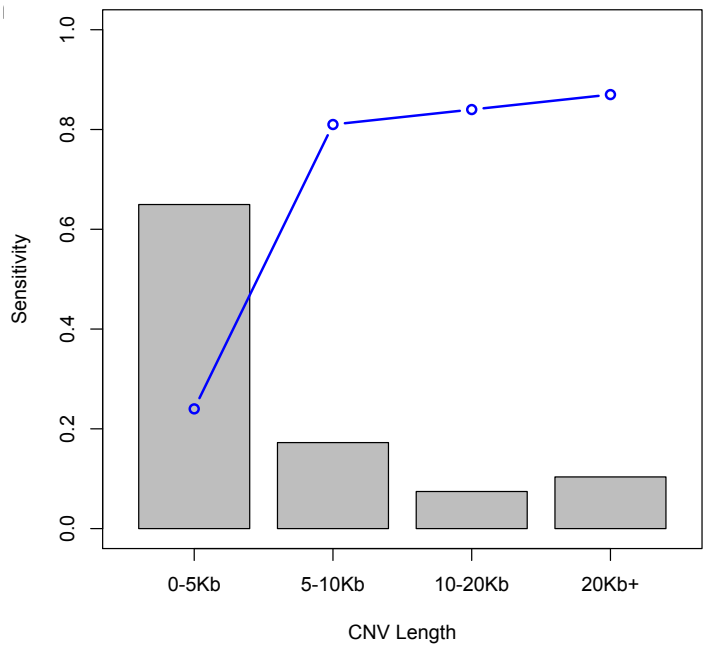
Supplementary Figures

Supplementary Figure 1



Normalization for sample-specific effects of GC content on sequence representation. **(a)** GC-bias in sequencing depth of coverage for 100 representative sequencing libraries from the 1000 Genomes Project. Colors indicate enrichment for sequences with low or high GC content. Note that different samples have different patterns of GC bias, requiring sample-specific normalization. **(b)** Distribution of GC content across the hg19 reference genome. **(c,d)** Normalization of read depth of coverage for sample-specific GC bias. At this example 5kbp locus (containing a bi-allelic deletion polymorphism), population read-depth distributions before **(c)** and after **(d)** normalization for sample-specific GC bias are shown. The post-normalization (but not the pre-normalization) read-depth distribution supports accurate genotyping.

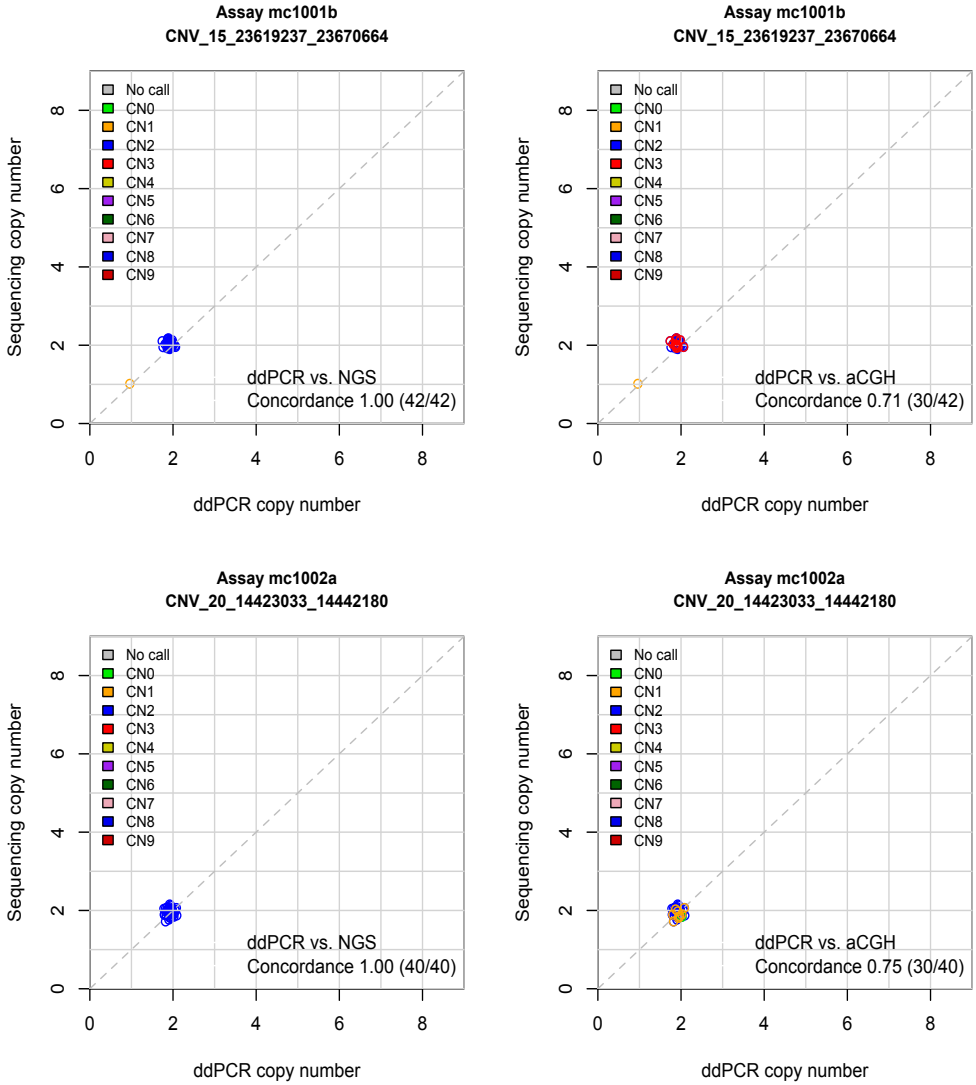
Supplementary Figure 2

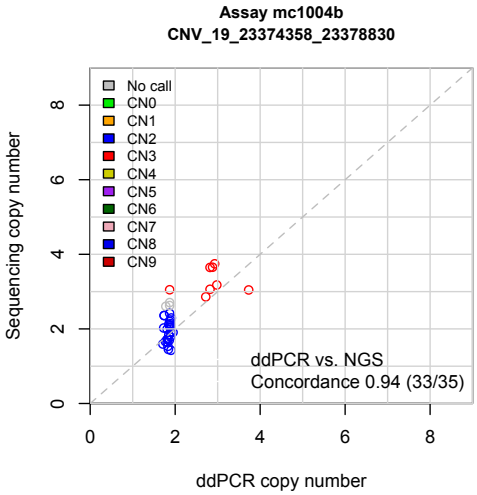


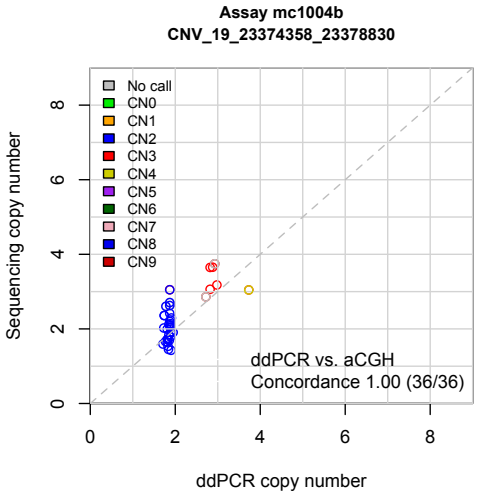
Comparison of CNVs called from sequencing data in this study to CNV calls from Conrad *et al.* using array CGH. Sensitivity measured as the fraction of CNV calls with any overlap to 4518 CNVs that were genotyped in the Conrad study and were called polymorphic in the 849 genomes analyzed here. For CNVs greater than 5kb in length, 83% were rediscovered using low-coverage sequencing data.

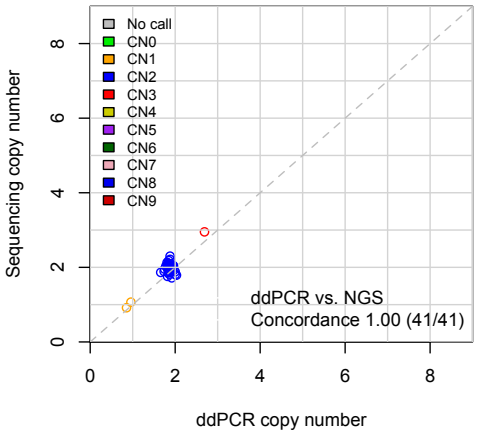
Supplementary Figure 3

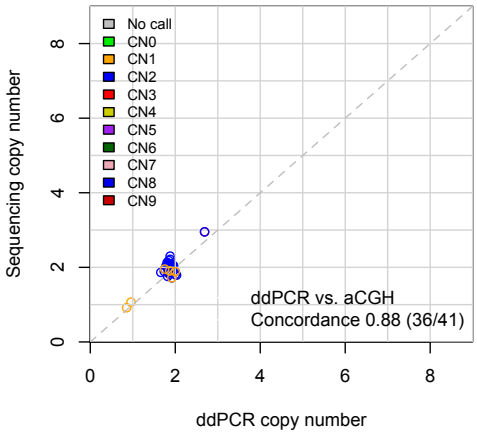
Molecular evaluation by droplet digital PCR (ddPCR) of discordant CNV genotypes determined by this study and by Conrad *et al.* Each row shows the data from the assessment of one CNV at which diploid copy-numbers calls showed discordance between the analysis of sequencing data (in this study) and earlier aCGH copy number assessment (from Conrad *et al.*). The left and right-hand panels in each row show the same measurements colored by called copy number in the sequencing data (left) and aCGH data (right). Concordance calculations at the bottom of each plot give the concordance between copy-number obtained from sequencing (left) or aCGH (right) compared to ddPCR results, at confidently called samples. Plots on subsequent pages follow the same format. Summarized results are in Supplementary Table 5.

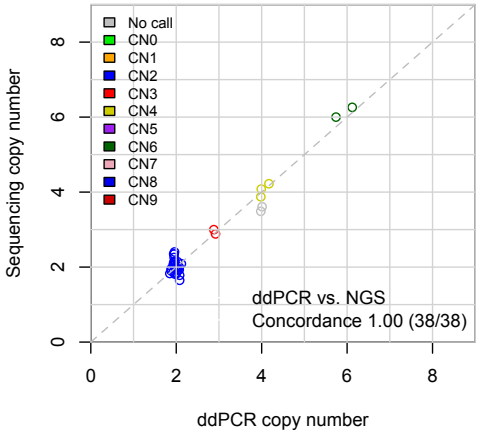


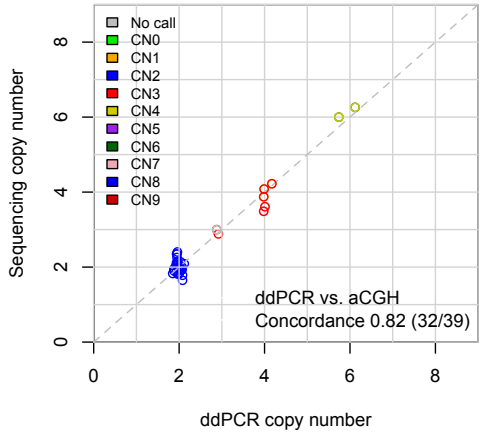

Assay mc1004b
CNV_19_23374358_23378830
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 0.94 (33/35)
Sequencing copy number
ddPCR copy number

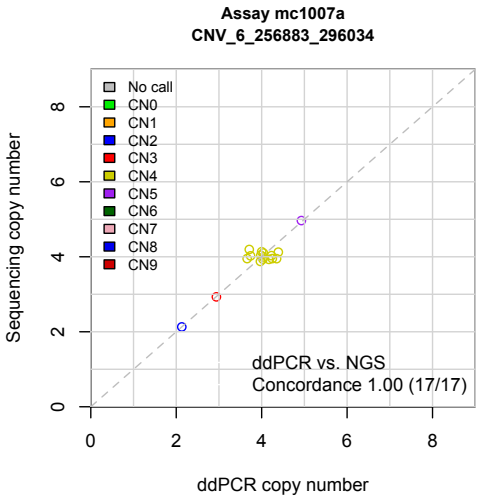

Assay mc1004b
CNV_19_23374358_23378830
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. aCGH
Concordance 1.00 (36/36)
Sequencing copy number
ddPCR copy number

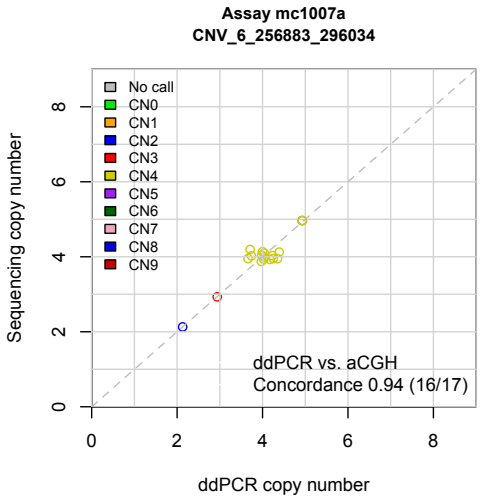

Assay mc1005a
CNV_3_63125152_63141614
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (41/41)
Sequencing copy number
ddPCR copy number

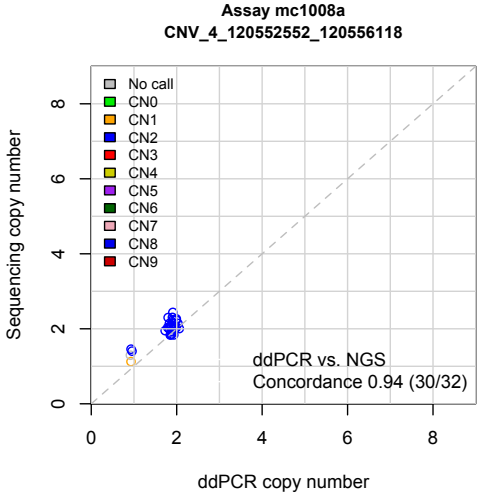

Assay mc1005a
CNV_3_63125152_63141614
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. aCGH
Concordance 0.88 (36/41)
Sequencing copy number
ddPCR copy number

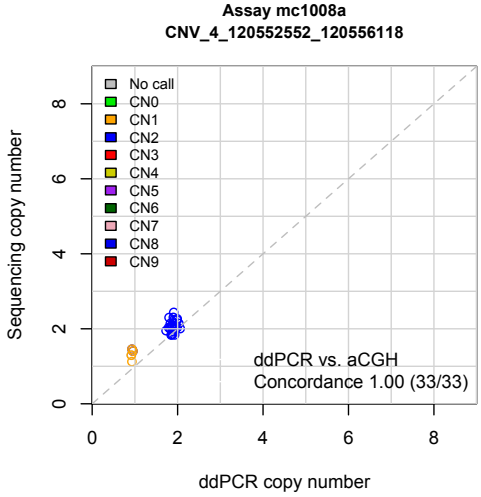

Assay mc1006a
CNV_22_18619095_18625292
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (38/38)
Sequencing copy number
ddPCR copy number

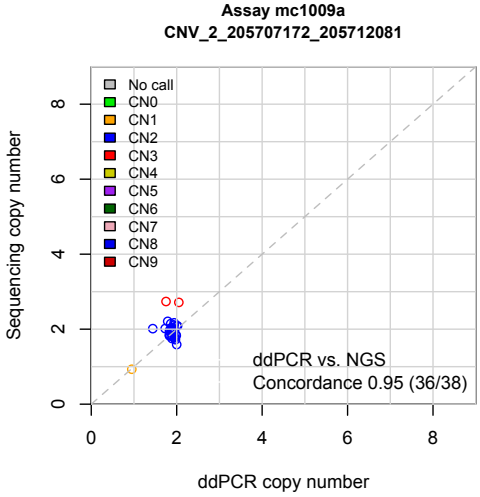

Assay mc1006a
CNV_22_18619095_18625292
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. aCGH
Concordance 0.82 (32/39)
Sequencing copy number
ddPCR copy number

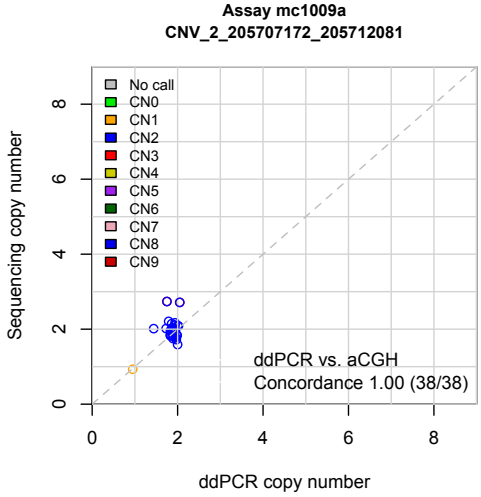

Assay mc1007a
CNV_6_256883_296034
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (17/17)
Sequencing copy number
ddPCR copy number

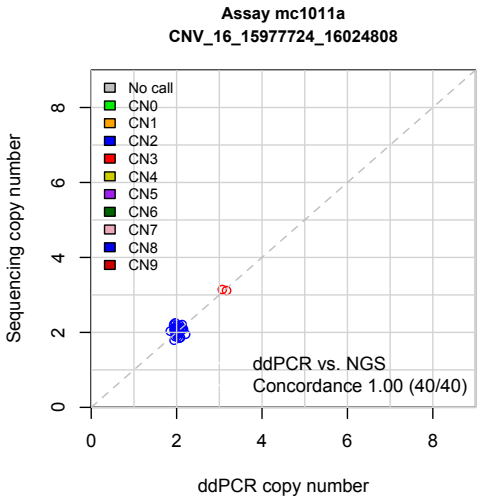

Assay mc1007a
CNV_6_256883_296034
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. aCGH
Concordance 0.94 (16/17)
Sequencing copy number
ddPCR copy number

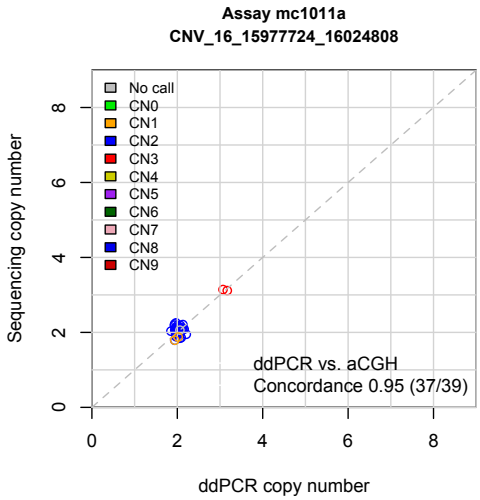

Assay mc1008a
CNV_4_120552552_120556118
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 0.94 (30/32)
Sequencing copy number
ddPCR copy number

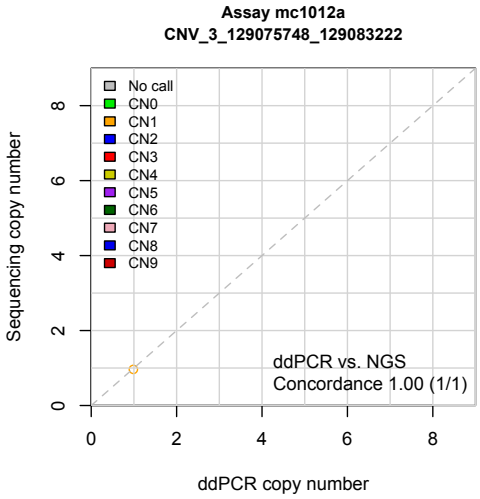

Assay mc1008a
CNV_4_120552552_120556118
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. aCGH
Concordance 1.00 (33/33)
Sequencing copy number
ddPCR copy number

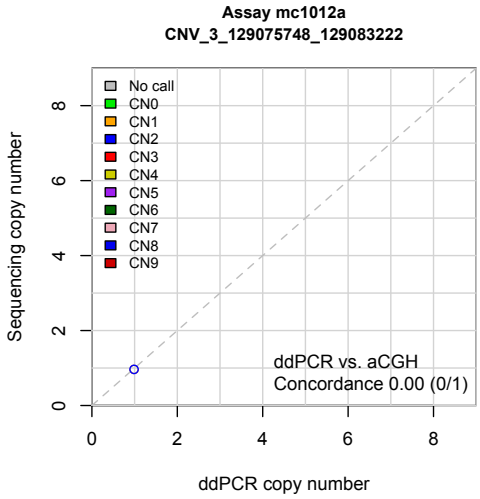

Assay mc1009a
CNV_2_205707172_205712081
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 0.95 (36/38)
Sequencing copy number
ddPCR copy number


Assay mc1009a
CNV_2_205707172_205712081
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. aCGH
Concordance 1.00 (38/38)
Sequencing copy number
ddPCR copy number


Assay mc1011a
CNV_16_15977724_16024808
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (40/40)
Sequencing copy number
ddPCR copy number

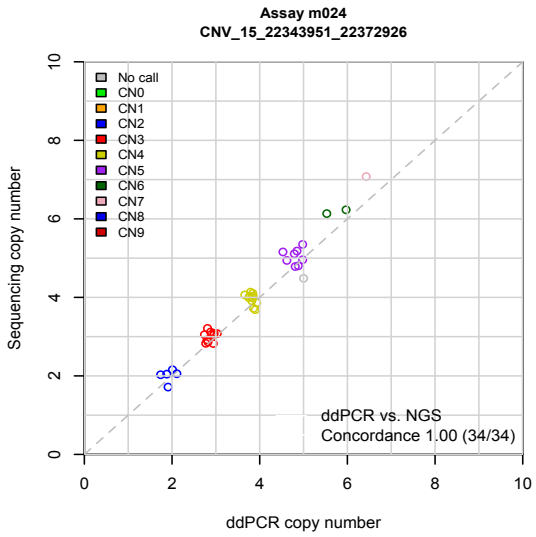

Assay mc1011a
CNV_16_15977724_16024808
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. aCGH
Concordance 0.95 (37/39)
Sequencing copy number
ddPCR copy number

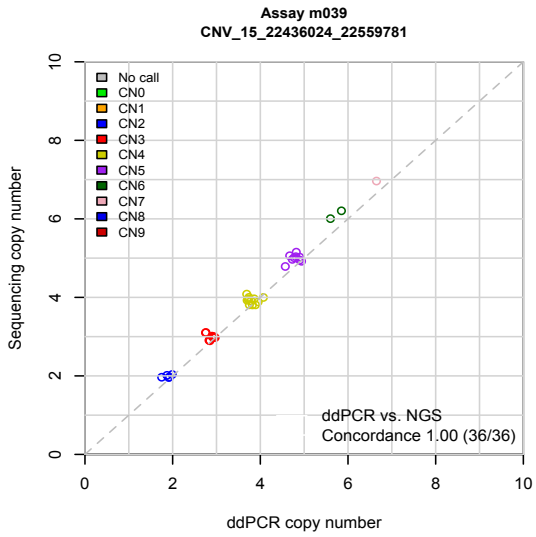

Assay mc1012a
CNV_3_129075748_129083222
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (1/1)
Sequencing copy number
ddPCR copy number

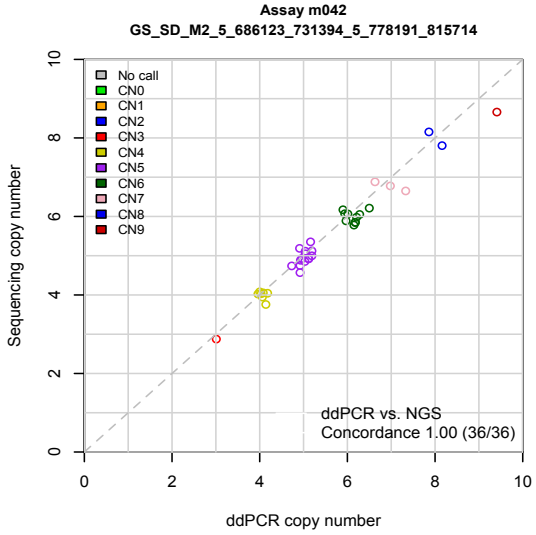

Assay mc1012a
CNV_3_129075748_129083222
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. aCGH
Concordance 0.00 (0/1)
Sequencing copy number
ddPCR copy number

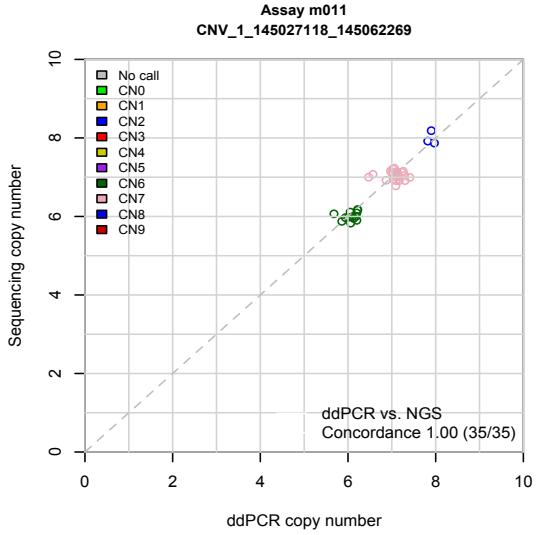
Supplementary Figure 4

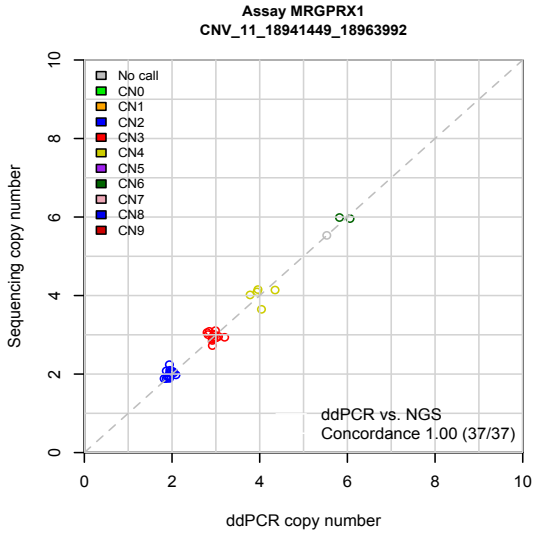
Molecular evaluation (by droplet digital PCR, ddPCR) of the accuracy of copy-number genotypes at high-copy mCNV sites. Concordance was evaluated at sites confidently called in both the sequencing data and ddPCR (**Supplementary Note**). For each site, unrounded copy-number estimates from sequencing read depth are plotted against unrounded copy-number measurements from ddPCR. The color of each circle represents the called copy-number from sequencing data (non-confident sequencing calls are shown in gray). At two sites (assays m035 and CC_SD_003), concordance was computed after a linear rescaling of the ddPCR measurements to center the clusters on integer values (**Supplementary Note**). For these two assays, plots of both the rescaled and raw data are shown. Supplementary figure includes plots for assays also shown in main Figure 3.

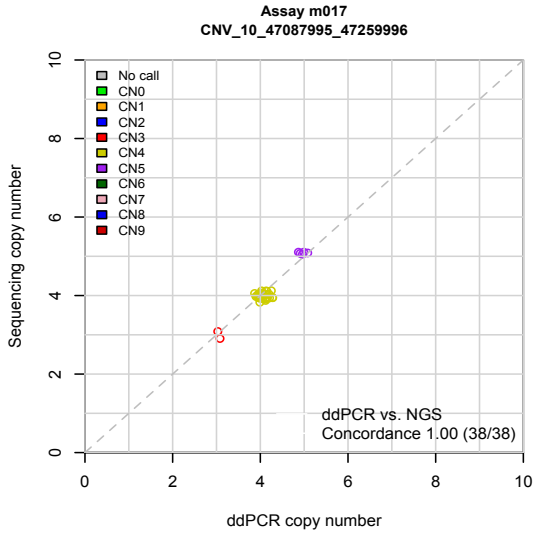

Assay m024
CNV_15_22343951_22372926
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (34/34)
Sequencing copy number
ddPCR copy number

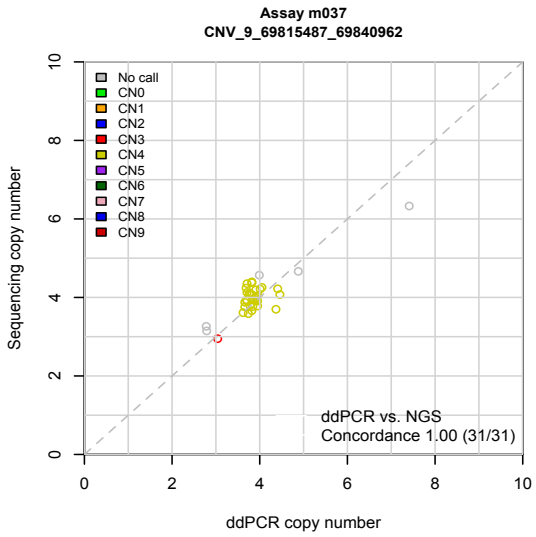

Assay m039
CNV_15_22436024_22559781
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (36/36)
Sequencing copy number
ddPCR copy number

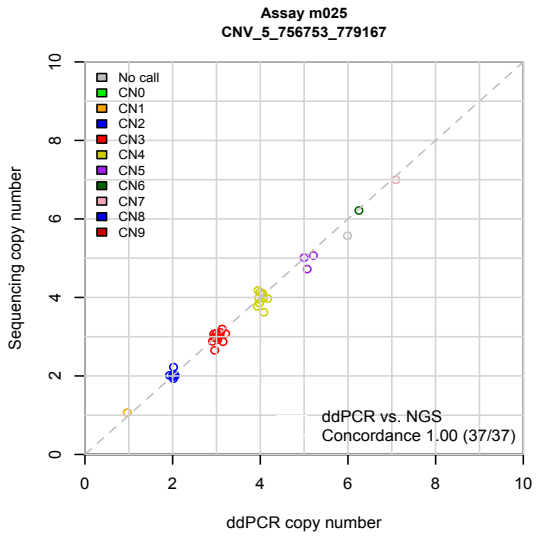

Assay m042
GS_SD_M2_5_686123_731394_5_778191_815714
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (36/36)
Sequencing copy number
ddPCR copy number

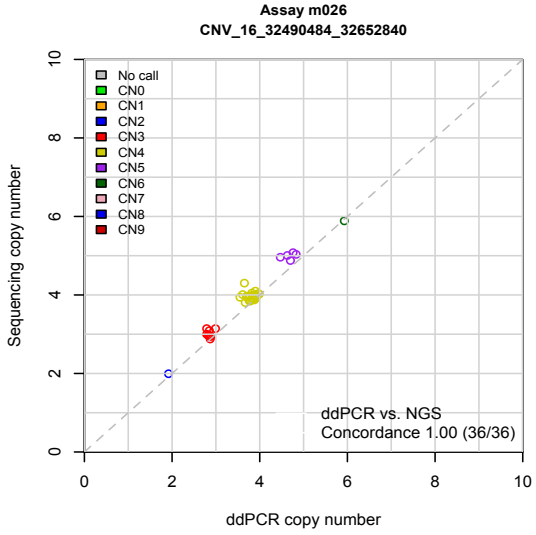

Assay m011
CNV_1_145027118_145062269
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (35/35)
Sequencing copy number
ddPCR copy number

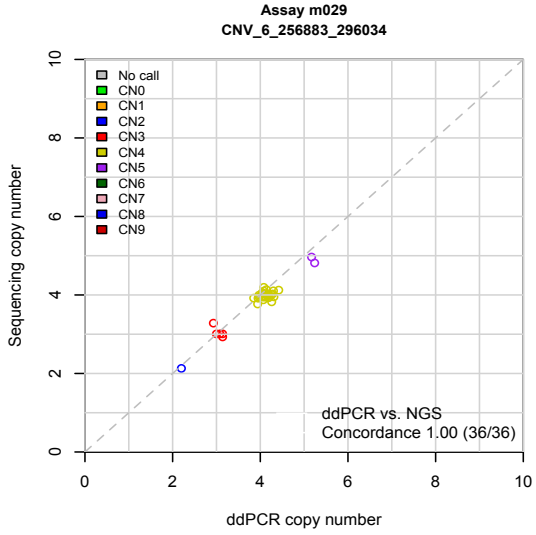

Assay MRGPRX1
CNV_11_18941449_18963992
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (37/37)
Sequencing copy number
ddPCR copy number

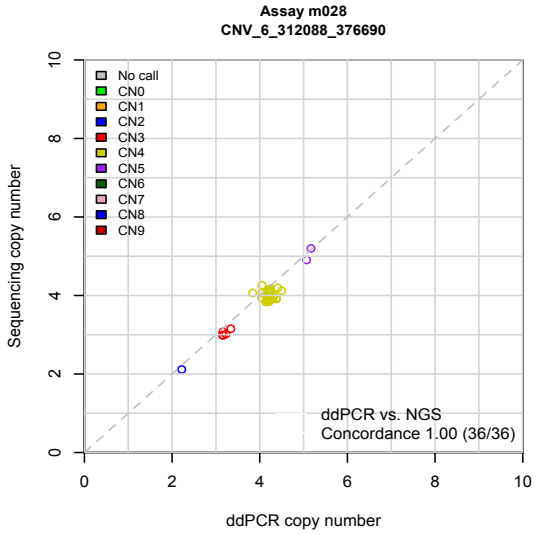

Assay m017
CNV_10_47087995_47259996
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (38/38)
Sequencing copy number
ddPCR copy number

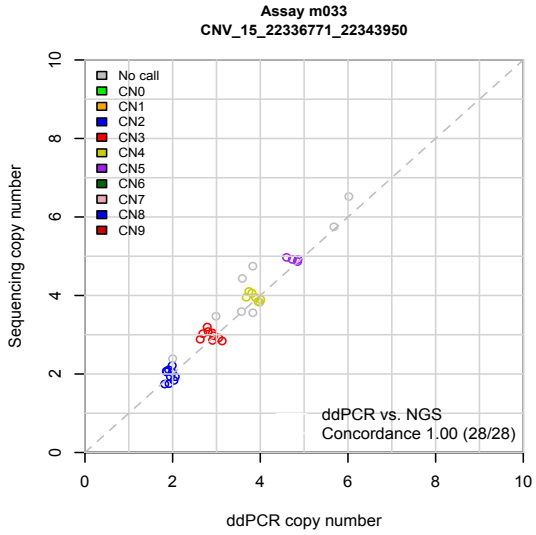

Assay m037
CNV_9_69815487_69840962
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
Sequencing copy number
ddPCR copy number
ddPCR vs. NGS
Concordance 1.00 (31/31)

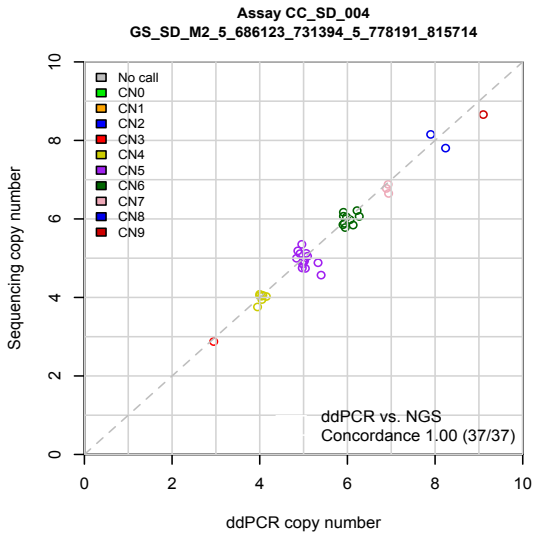

Assay m025
CNV_5_756753_779167
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
Sequencing copy number
ddPCR copy number
ddPCR vs. NGS
Concordance 1.00 (37/37)

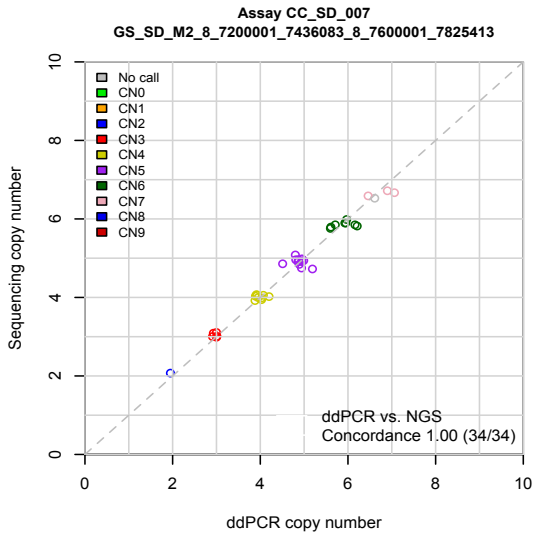

Assay m026
CNV_16_32490484_32652840
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
Sequencing copy number
ddPCR copy number
ddPCR vs. NGS
Concordance 1.00 (36/36)

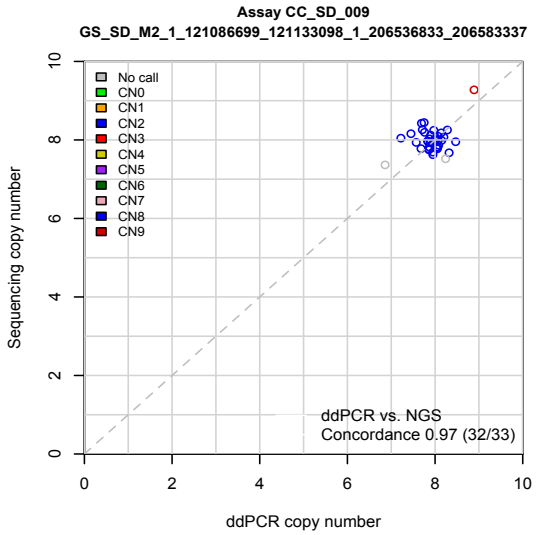

Assay m029
CNV_6_256883_296034
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
Sequencing copy number
ddPCR copy number
ddPCR vs. NGS
Concordance 1.00 (36/36)

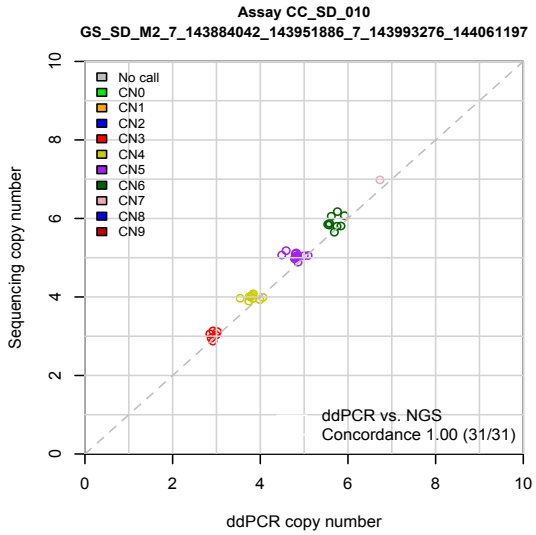

Assay m028
CNV_6_312088_376690
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
Sequencing copy number
ddPCR copy number
ddPCR vs. NGS
Concordance 1.00 (36/36)

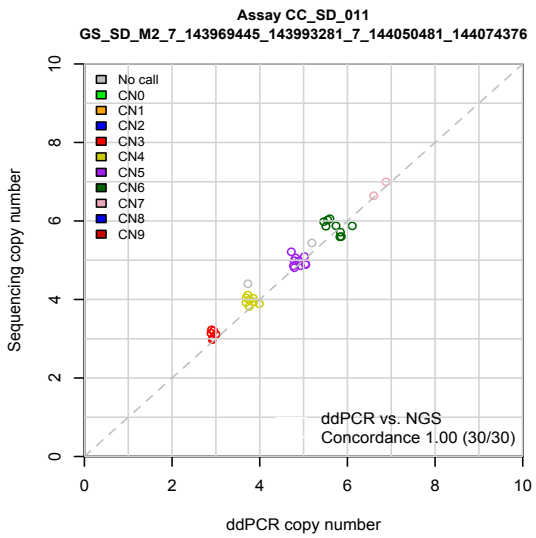

Assay m033
CNV_15_22336771_22343950
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
Sequencing copy number
ddPCR copy number
ddPCR vs. NGS
Concordance 1.00 (28/28)

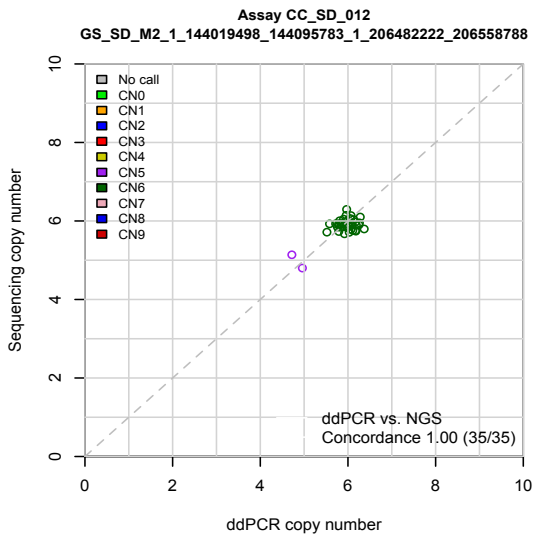

Assay CC_SD_004
GS_SD_M2_5_686123_731394_5_778191_815714
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (37/37)
Sequencing copy number
ddPCR copy number

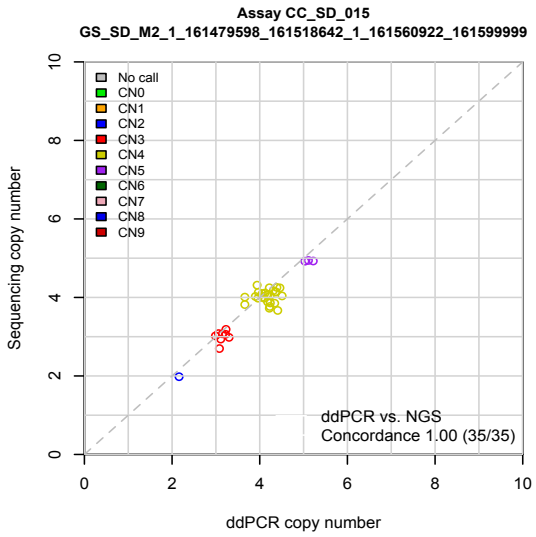

Assay CC_SD_007
GS_SD_M2_8_7200001_7436083_8_7600001_7825413
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (34/34)
Sequencing copy number
ddPCR copy number

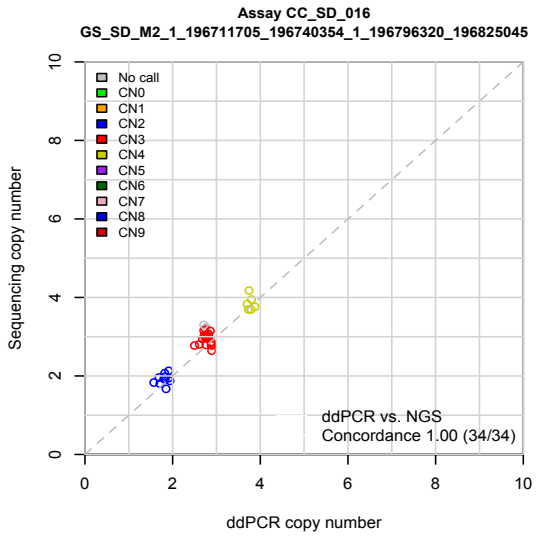

Assay CC_SD_009
GS_SD_M2_1_121086699_121133098_1_206536833_206583337
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 0.97 (32/33)
Sequencing copy number
ddPCR copy number

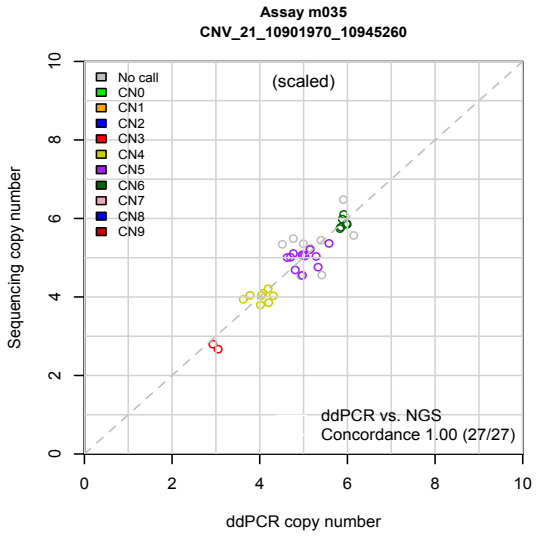

Assay CC_SD_010
GS_SD_M2_7_143884042_143951886_7_143993276_144061197
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (31/31)
Sequencing copy number
ddPCR copy number

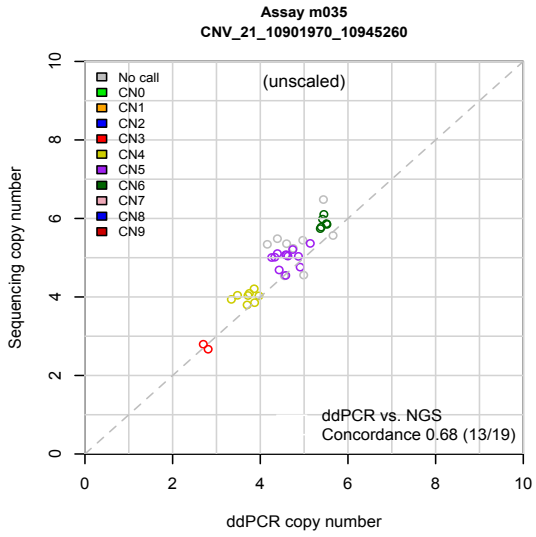

Assay CC_SD_011
GS_SD_M2_7_143969445_143993281_7_144050481_144074376
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (30/30)
Sequencing copy number
ddPCR copy number

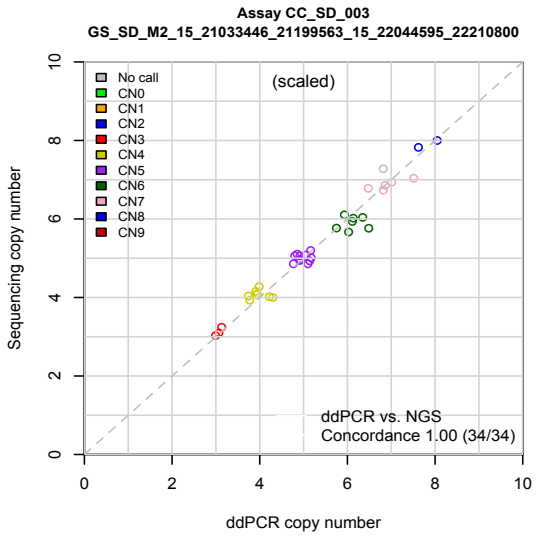

Assay CC_SD_012
GS_SD_M2_1_144019498_144095783_1_206482222_206558788
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (35/35)
Sequencing copy number
ddPCR copy number

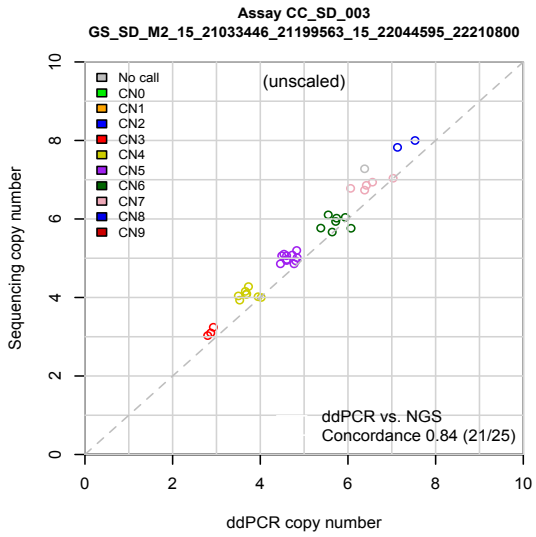

Assay CC_SD_015
GS_SD_M2_1_161479598_161518642_1_161560922_161599999
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (35/35)
Sequencing copy number
ddPCR copy number


Assay CC_SD_016
GS_SD_M2_1_196711705_196740354_1_196796320_196825045
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (34/34)
Sequencing copy number
ddPCR copy number

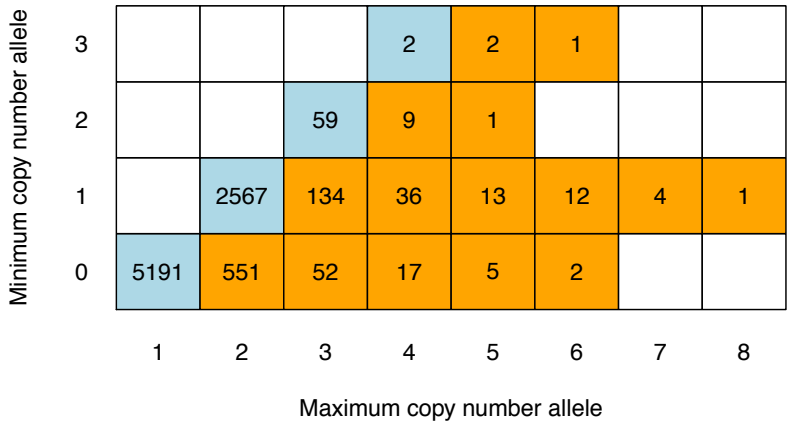

Assay m035
CNV_21_10901970_10945260
(scaled)
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (27/27)
Sequencing copy number
ddPCR copy number


Assay m035
CNV_21_10901970_10945260
(unscaled)
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 0.68 (13/19)
Sequencing copy number
ddPCR copy number


Assay CC_SD_003
GS_SD_M2_15_21033446_21199563_15_22044595_22210800
(scaled)
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 1.00 (34/34)
Sequencing copy number
ddPCR copy number

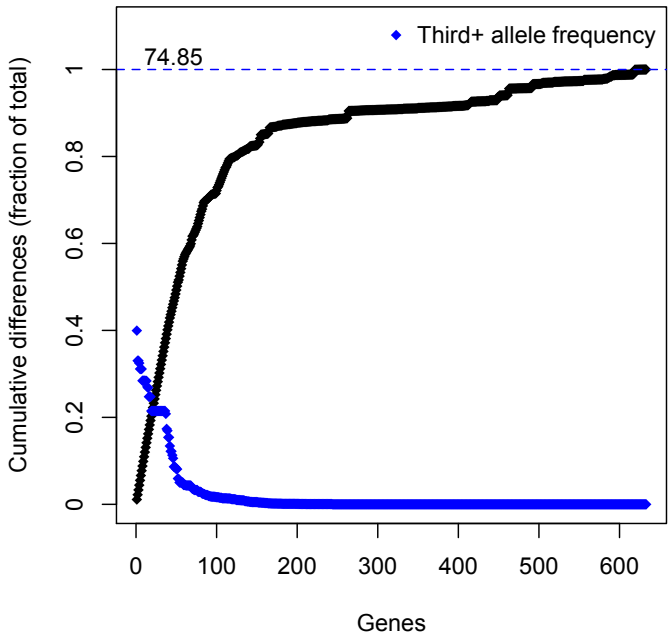

Assay CC_SD_003
GS_SD_M2_15_21033446_21199563_15_22044595_22210800
(unscaled)
No call
CN0
CN1
CN2
CN3
CN4
CN5
CN6
CN7
CN8
CN9
ddPCR vs. NGS
Concordance 0.84 (21/25)
Sequencing copy number
ddPCR copy number

Supplementary Figure 5



Range of copy-number alleles at human CNVs analyzed in this study. Blue cells indicate bi-allelic CNVs; orange cells indicate multi-allelic CNVs; numbers indicate the number of CNVs for which copy-number alleles span the range shown. For example, the most common type of mCNV locus (of which 551 were ascertained and genotyped) contains three alleles with 0, 1, and 2 copies per chromosome.

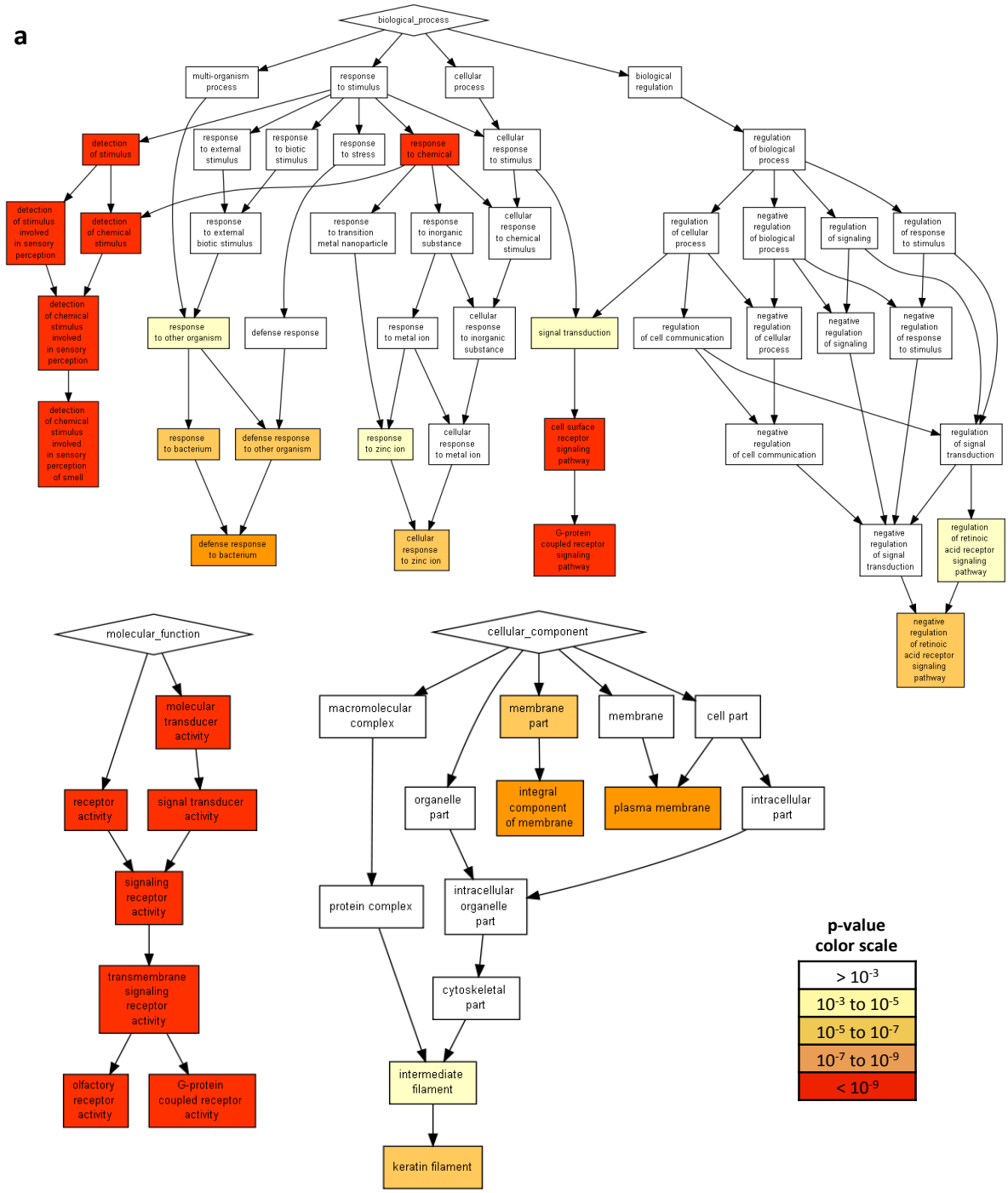
Supplementary Figure 6



Contribution of common and rare mCNVs to human gene-dosage variation. Black points show the cumulative contribution of each gene to overall gene-dosage variation. The blue diamonds indicate "third+ allele frequency", the estimated total frequency of all alleles beyond the two most common copy-number alleles of that CNV; here the genes are ordered by decreasing "third+ allele frequency". The blue dashed line is the average number of genes that differ in copy number between any two individuals (74.85). Some 100-200 high-frequency mCNVs (each with three or more common alleles) account for most of the gene dosage variations among humans. Note that our analysis includes only those CNVs for which we could infer integer genotypes with high confidence, generally limiting analysis to CNV for which copy number ranges within 0 to 12; these numbers are therefore likely to be a lower bound on the overall contribution of high-frequency mCNVs to gene dosage variation in humans.

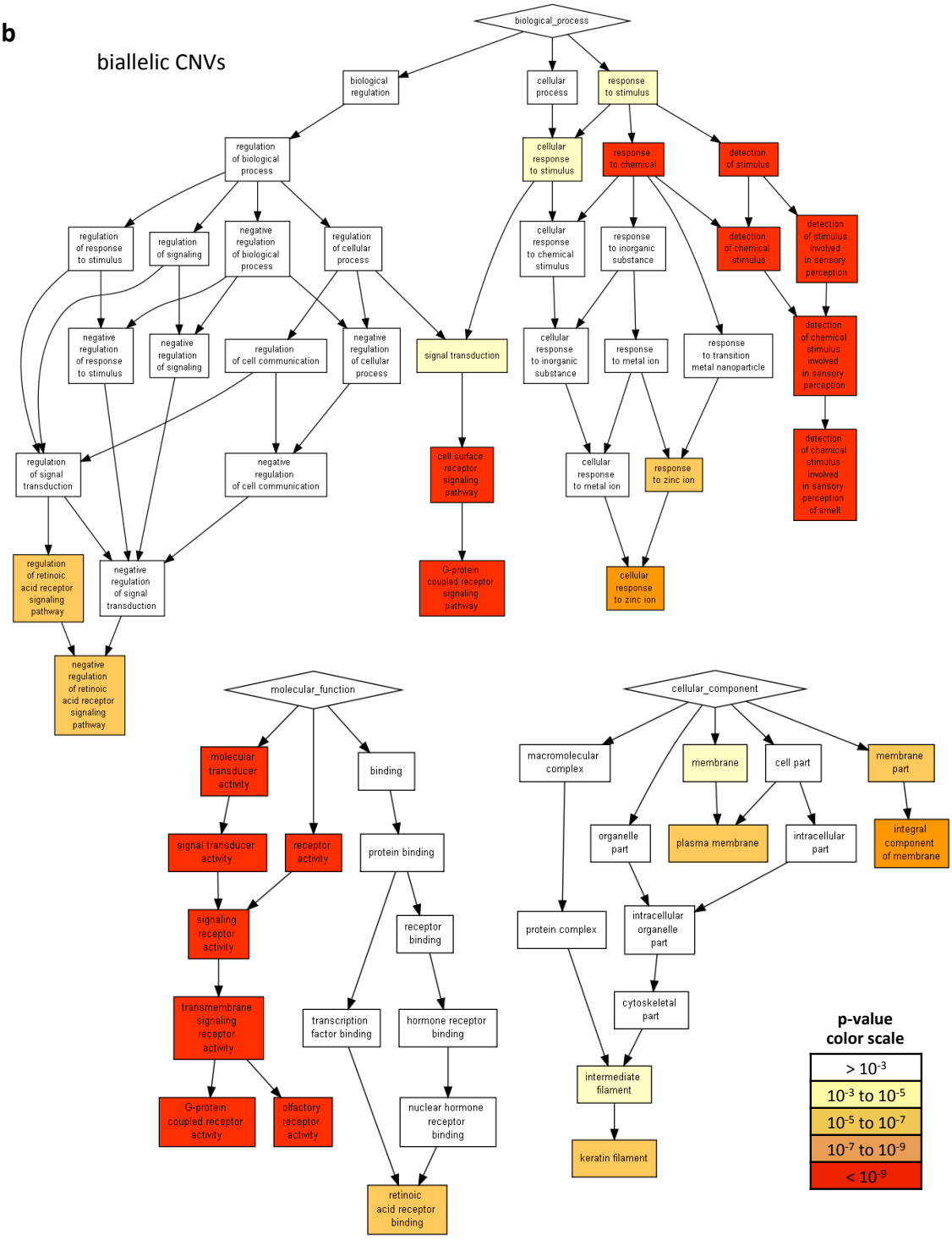
Supplementary Figure 7

Enrichment of gene ontology categories for genic CNVs. Graphical presentation of GO enrichment results using visualization from GOrilla (http://cbl-gorilla.cs.technion.ac.il). Enrichment analysis is shown for all ascertained CNVs **(a)** and then separately for the biallelic **(b)** and multi-allelic **(c)** subsets for all three GO ontologies (biological process, molecular function and cellular component). Visualizations were generated using a p-value threshold of 10^{-5}.



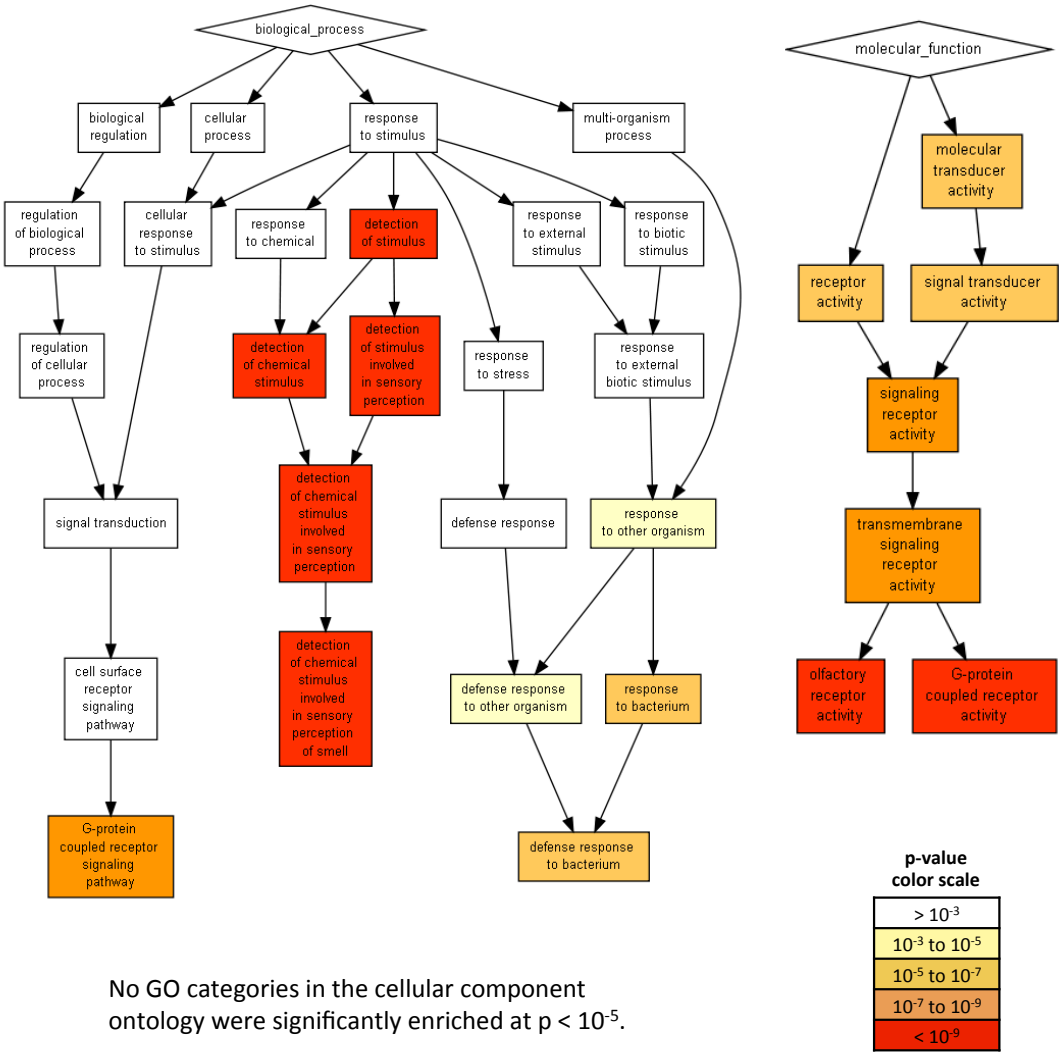
b

biallelic CNVs



c

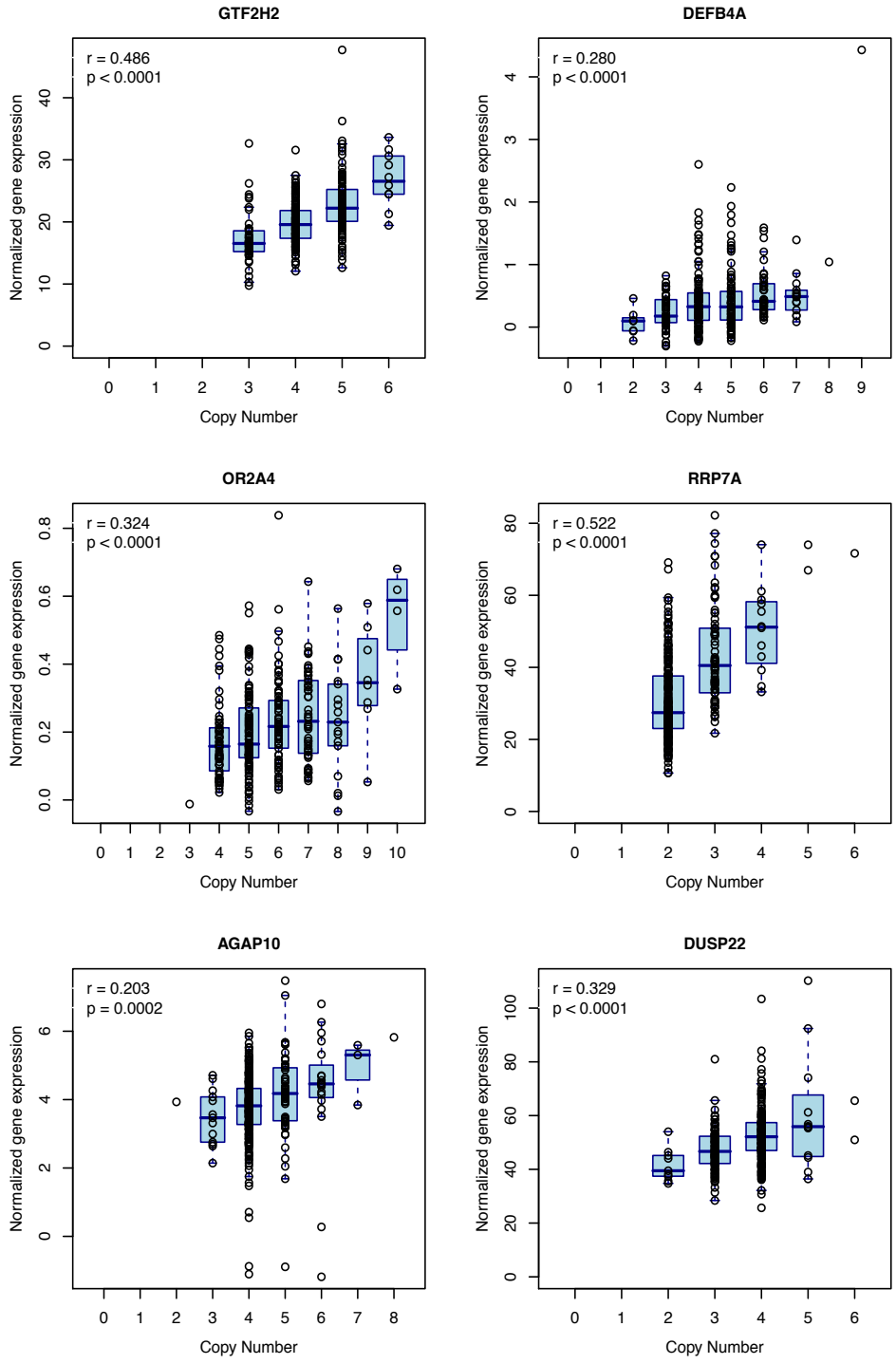
multi-allelic CNVs

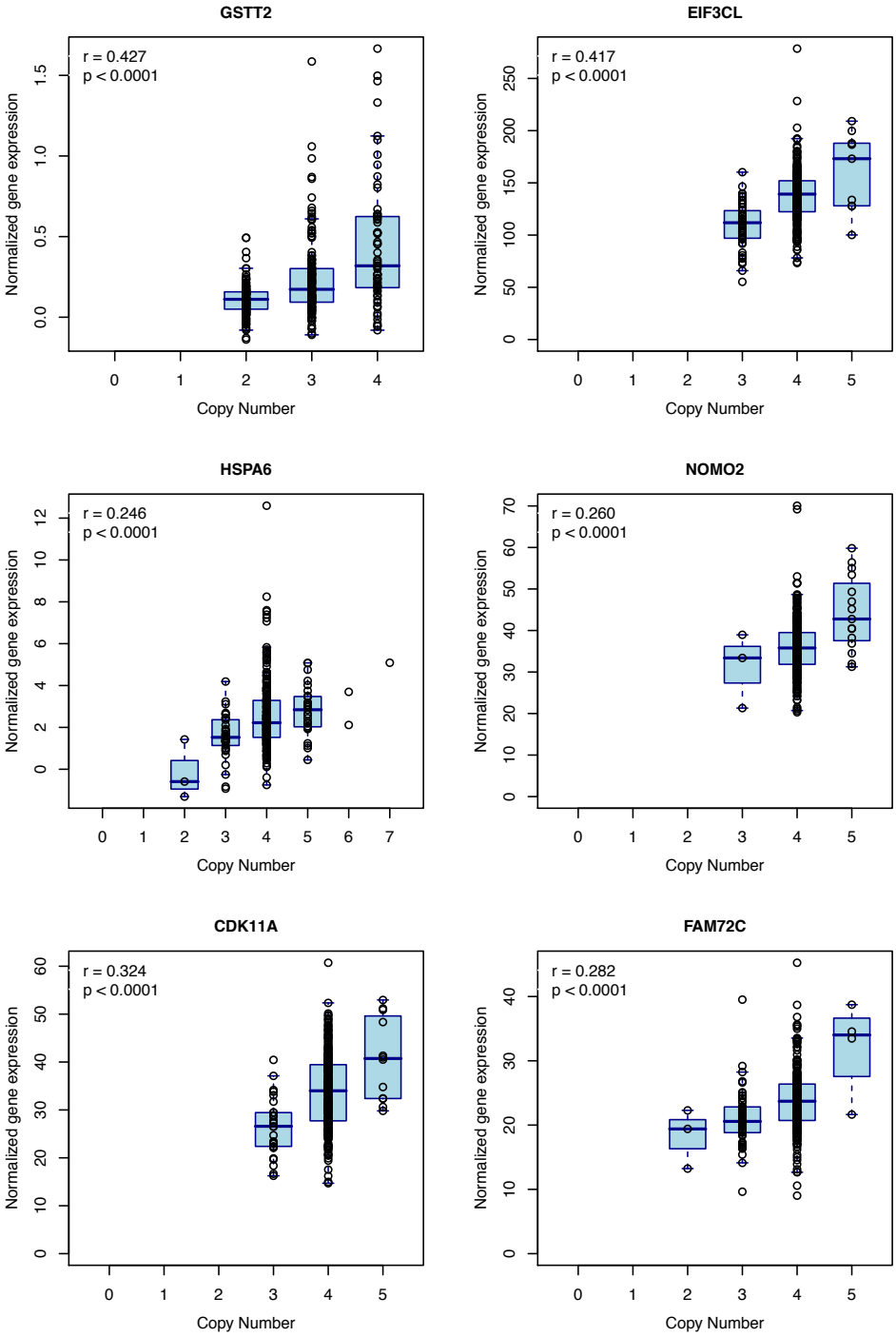


No GO categories in the cellular component ontology were significantly enriched at $p < 10^{-5}$.

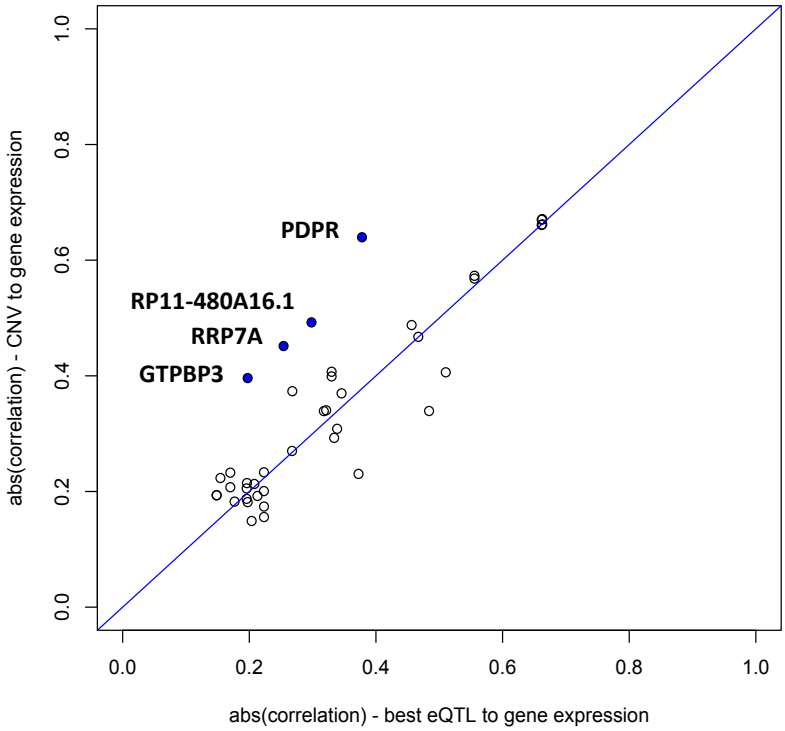
Supplementary Figure 8

Relationships at 12 mCNV loci between gene copy number and mRNA expression. Correlation is measured in lymphoblastoid cell lines from 310 individuals. Four such loci were shown in Figure 4; the panels below show additional examples of genes with significant correlations between gene dosage and normalized gene expression measured from RNA-seq.



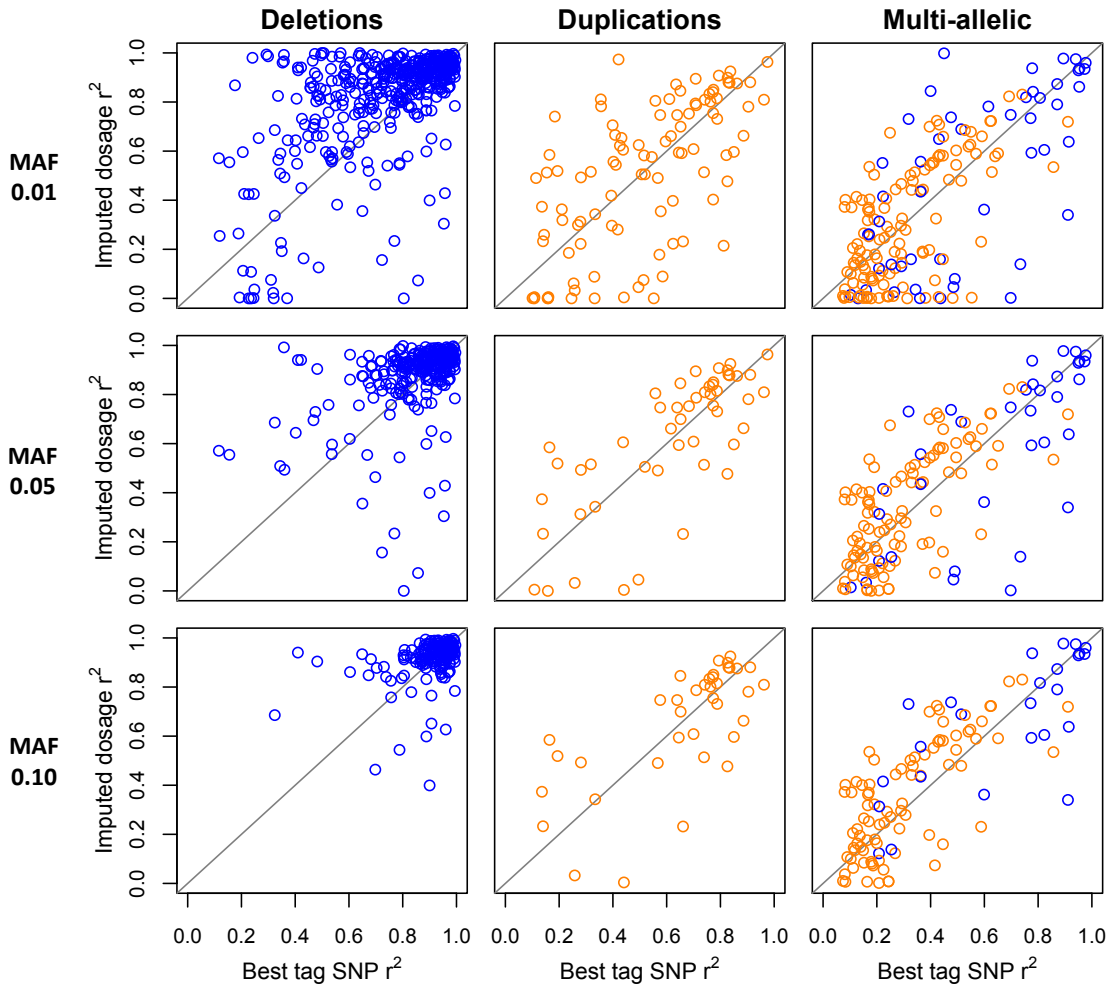

GSTT2
r = 0.427
p < 0.0001
Normalized gene expression
Copy Number
EIF3CL
r = 0.417
p < 0.0001
Normalized gene expression
Copy Number
HSPA6
r = 0.246
p < 0.0001
Normalized gene expression
Copy Number
NOMO2
r = 0.260
p < 0.0001
Normalized gene expression
Copy Number
CDK11A
r = 0.324
p < 0.0001
Normalized gene expression
Copy Number
FAM72C
r = 0.282
p < 0.0001
Normalized gene expression
Copy Number

Supplementary Figure 9



Comparison of CNV and eQTL effects on gene expression for CNVs in strong LD with the eQTLs. For the genes listed in Supplementary Table 10, most show comparable magnitude of correlation between the CNV dosage and gene expression and the correlation for the best eQTL. Of the four cases with the largest positive differences (blue points), the CNV overlaps the coding sequence of the gene in three, suggesting the effect on gene expression may be due to direct interaction as opposed to an effect on gene regulation. In the fourth case (RP11-480A16.1), analysis with blat suggests the possible presence of an unannotated copy of the lincRNA occurring within the CNV (data not shown).

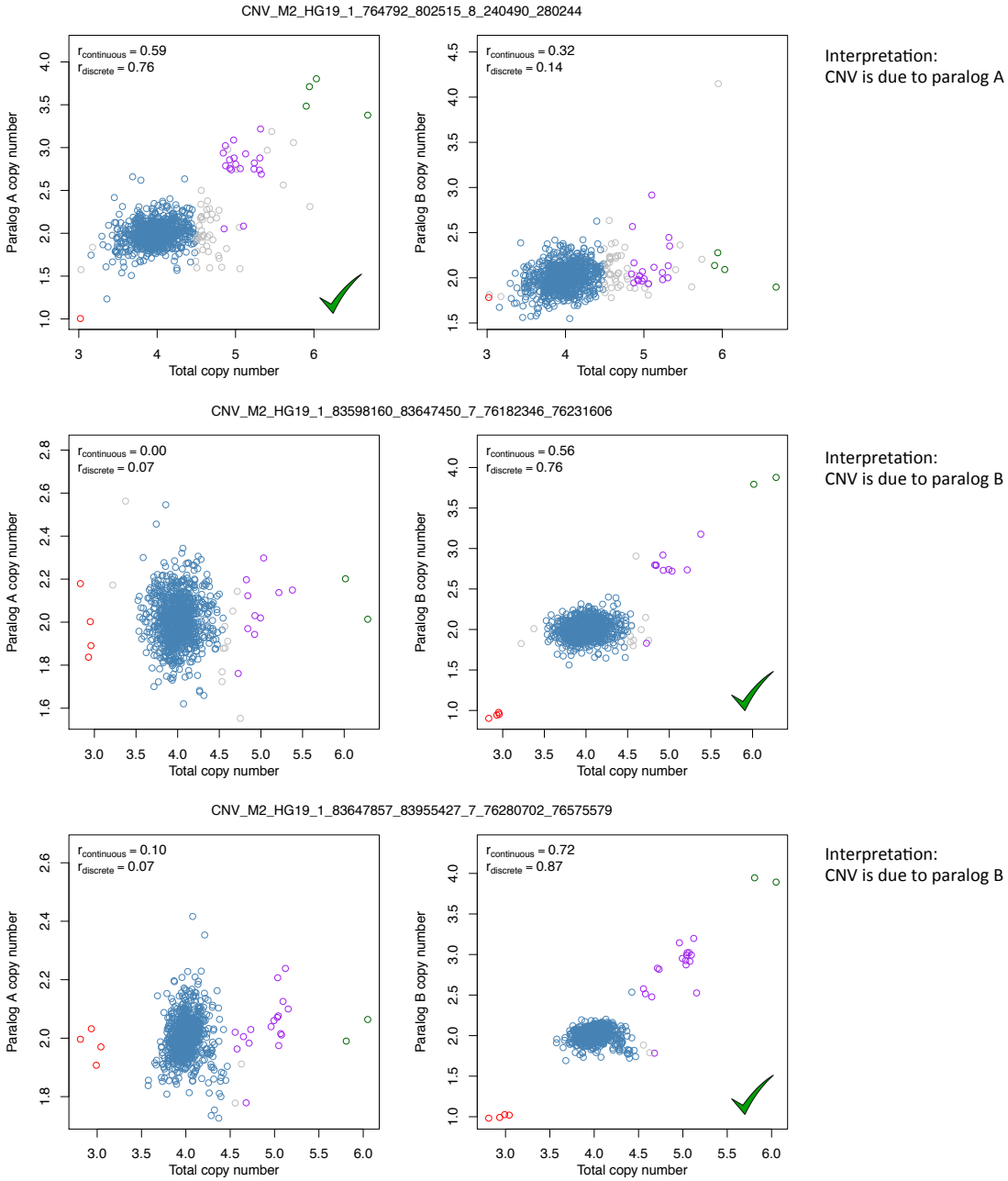
Supplementary Figure 10



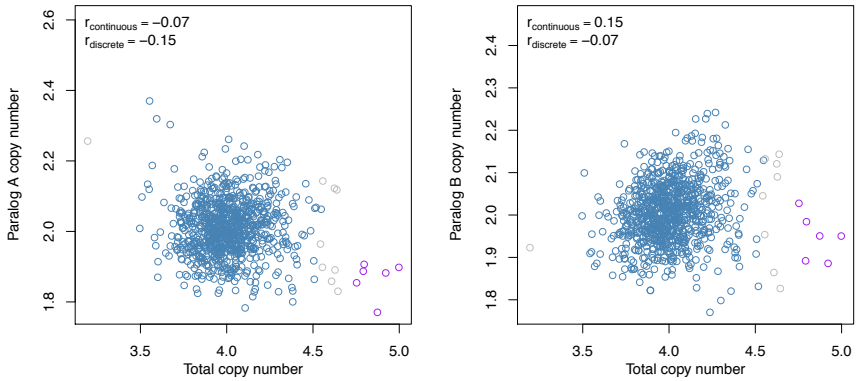
Imputability of deletions, biallelic duplications, and multi-allelic CNVs, assessed using leave-out trials. For each CNV with non-modal AF > 1%, we conducted a set of leave-out trials in which we withheld 10 samples at a time and imputed their allelic copy-number state based on the CNV genotypes from the other samples and flanking SNP genotypes for all samples. Imputed dosage correlation (r^2) between the imputed and directly genotyped (from WGS data) diploid copy-number in the EUR continental population is compared to the correlation between copy number and the most strongly correlated local SNP ("best tag SNP") in the European (EUR) population samples from the 1000 Genomes Project. The data is grouped by CNV category in each column and filtered by different CNV minor allele frequency thresholds in each row (top: AF > 1%, middle: AF > 5%, bottom: AF > 10%). Only sites with a significant tag SNP ($p < 10^{-3}$) are shown (**Supplementary Note**). Colors indicate whether the average copy number of each multi-allelic CNV is less than two (blue) or greater than two (orange). Note that mCNVs with higher average copy number tend to be less imputable.

Supplementary Figure 11

Resolving CNVs in segmentally duplicated regions. Each row shows a pair of plots for a CNV region that is in a segmental duplication where the duplicated copies are present on two different reference chromosomes. Read depth at positions with paralog-specific-differences (PSDs) can be used to estimate which paralog is contributing to the total copy number. For non-tandem CNVs, we used the paralog with the strongest correlation between the continuous PSD read depth and discrete total copy number ($r_{discrete}$). Colors indicate discrete copy number of samples called at 95% confidence for total copy number.

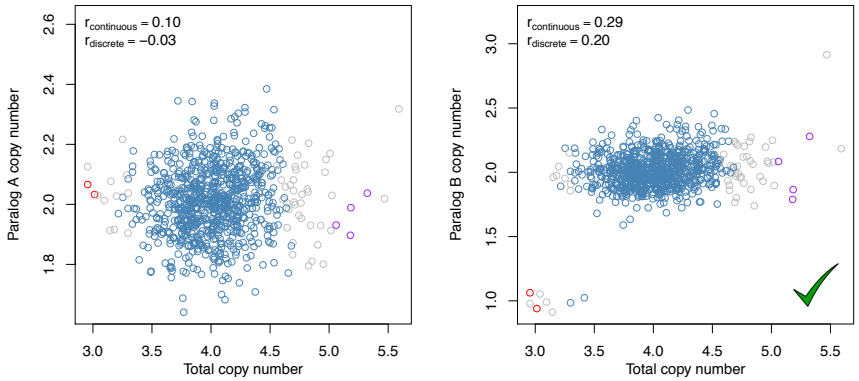


CNV_M2_HG19_1_242413205_242528828_10_38520100_38637504



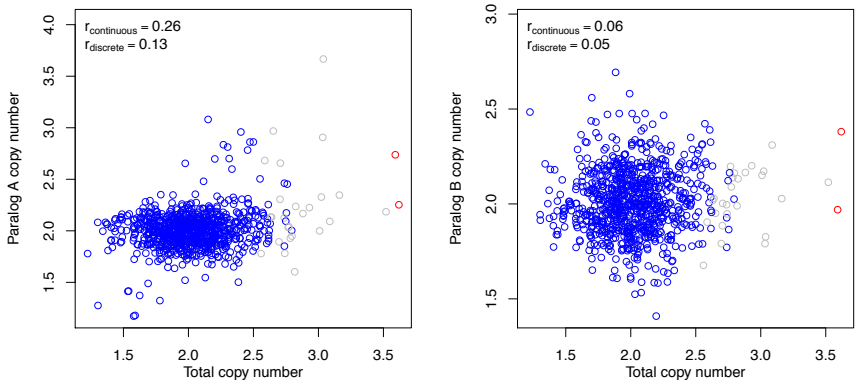
Interpretation:
CNV source is unclear

CNV_M2_HG19_2_132660995_132699803_21_14593409_14627482



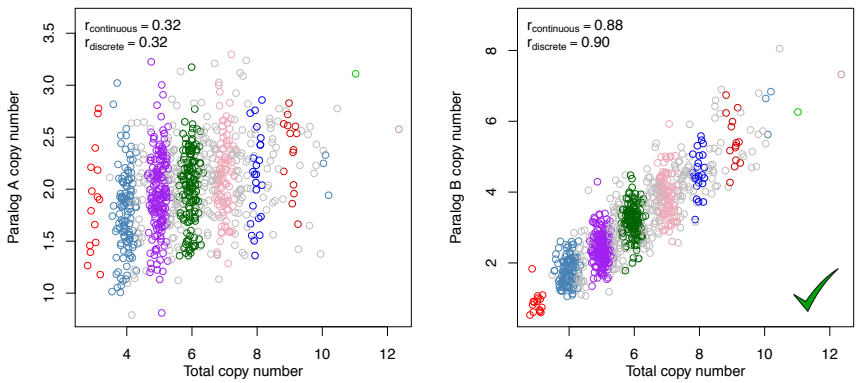
Interpretation:
CNV is due to paralog B

CNV_M2_HG19_2_159703462_159734019_3_125415459_125446156



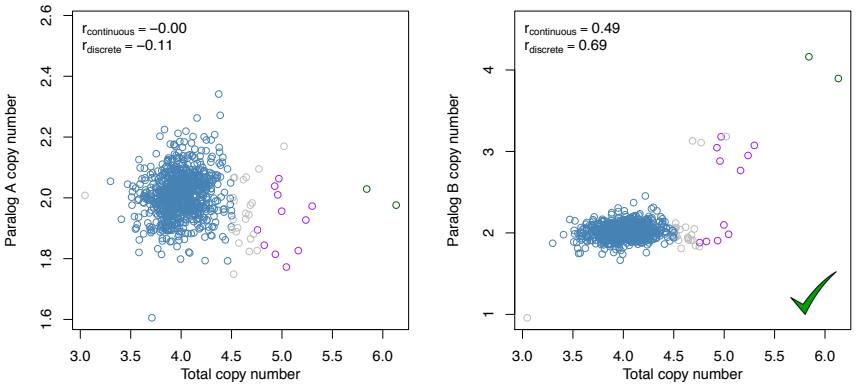
Interpretation:
CNV source is unclear

CNV_M2_HG19_6_132019328_132035259_7_143953514_143969444



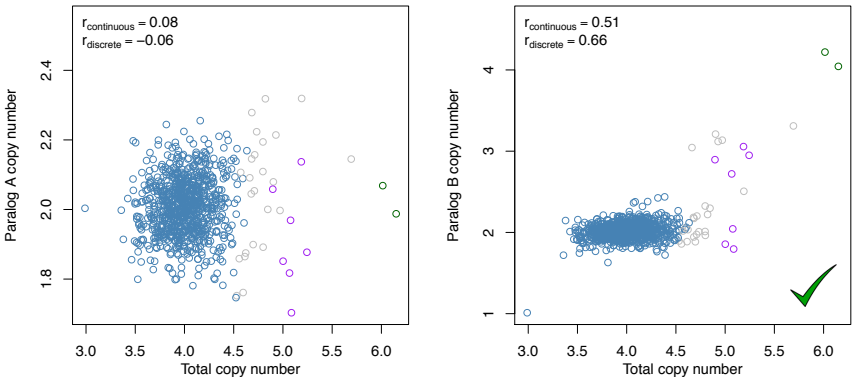
Interpretation:
CNV is due to paralog B

CNV_M2_HG19_7_35139142_35281183_12_63971080_64119245



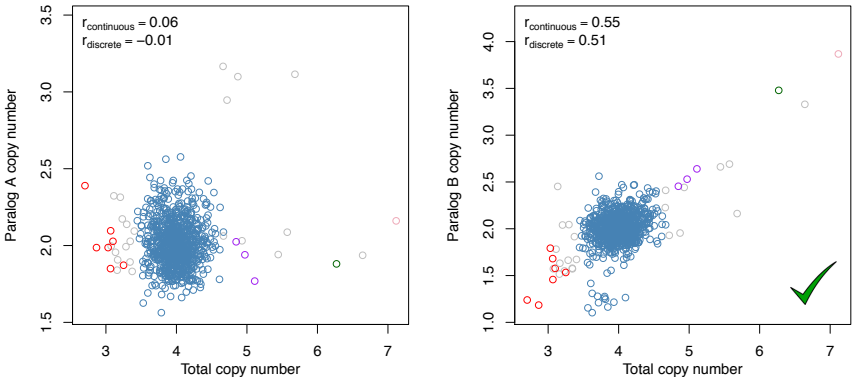
Interpretation:
CNV is due primarily to paralog B

CNV_M2_HG19_7_102815781_102929218_12_63954365_64072240



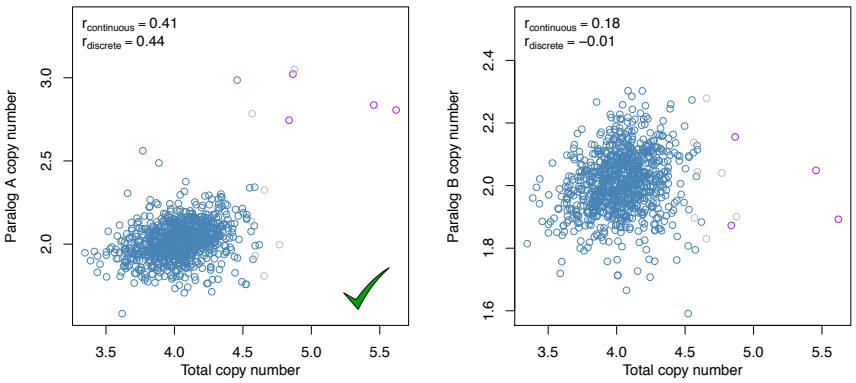
Interpretation:
CNV is due primarily to paralog B

CNV_M2_HG19_10_60001_130311_18_14415_84190



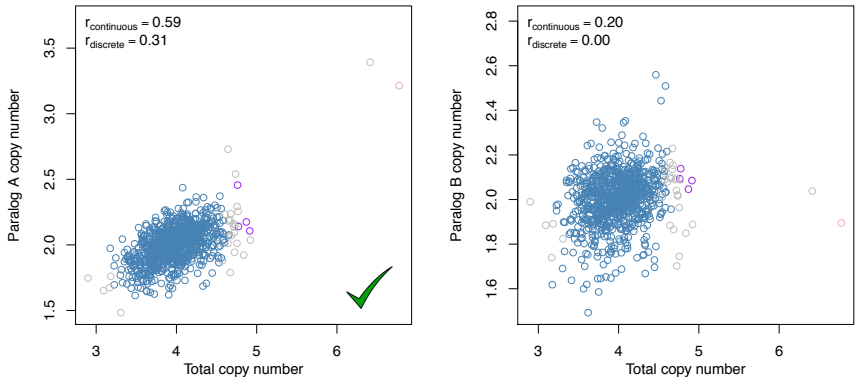
Interpretation:
CNV is due primarily to paralog B

CNV_M2_HG19_13_19167974_19275982_18_14358135_14464819



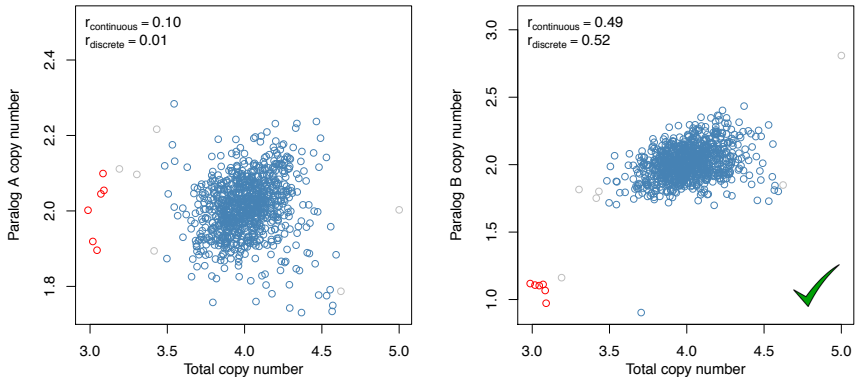
Interpretation:
CNV is due to paralog A

CNV_M2_HG19_13_19301588_19448886_18_14185210_14353419



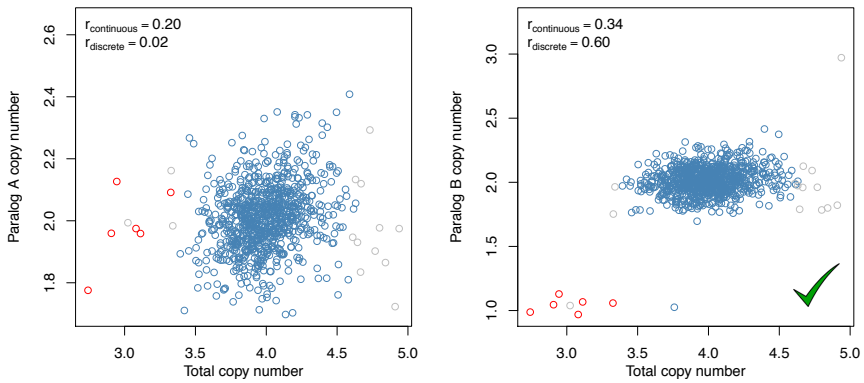
Interpretation:
CNV is due to paralog A

CNV_M2_HG19_18_14358135_14728624_21_14801588_15174050



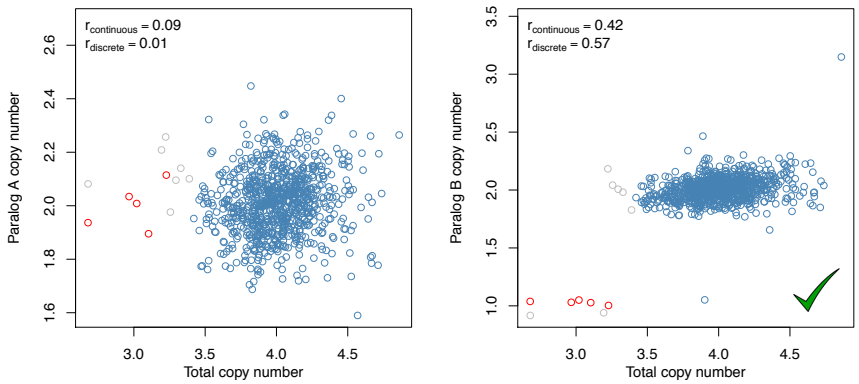
Interpretation:
CNV is due to paralog B

CNV_M2_HG19_18_14807068_14897137_21_14714527_14801577



Interpretation:
CNV is due to paralog B

CNV_M2_HG19_18_15026971_15101188_21_14640424_14714507



Interpretation:
CNV is due to paralog B

Supplementary Tables

Supplementary Table 1

Samples and populations analyzed from the 1000 Genomes Project

Population Code	Population grouping used in this study	Samples	Description
ASW	ADM	45	Americans of African Ancestry in SW USA
CLM	ADM	49	Colombians from Medellin, Colombia
MXL	ADM	54	Mexican Ancestry from Los Angeles USA
PUR	ADM	52	Puerto Ricans from Puerto Rico
LWK	AFR	69	Luhya in Webuye, Kenya
YRI	AFR	71	Yoruba in Ibadan, Nigera
CHB	ASN	78	Han Chinese in Bejing, China
CHS	ASN	91	Southern Han Chinese
JPT	ASN	69	Japanese in Tokyo, Japan
CEU	EUR	66	Utah Residents (CEPH) with Northern and Western European ancestry
FIN	EUR	62	Finnish in Finland
GBR	EUR	54	British in England and Scotland
IBS	EUR	6	Iberian population in Spain
TSI	EUR	83	Toscani in Italia

Supplementary Table 2

Frequency of CNVs based on observed range of diploid copy number

		Max Copy Number													
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Min Copy Number	**0**	613	86	15	10	11	3	1	-	-	2	-	-	-	-
	1	4168	897	79	17	10	3	9	1	-	-	-	-	-	-
	2	-	2254	216	61	26	17	8	7	2	6	2	2	1	-
	3	-	-	5	35	24	3	-	1	1	2	1	-	-	1
	4	-	-	-	28	17	6	1	-	-	-	-	-	-	-
	5	-	-	-	-	-	2	1	-	-	-	-	-	-	-
	6	-	-	-	-	-	-	2	2	-	-	-	-	-	-

Supplementary Table 3

Estimated per-site false discovery rate (FDR) using array intensity data

Sites longer than 20 kilobases

Site Category	Sites	Affymetrix 6.0			Omni 2.5		
		Sites with probes	%	Est. FDR	Sites with probes	%	Est. FDR
CN <= 2 only	982	953	97.0%	0.000	975	99.3%	0.002
CN >= 2 only	1165	1129	96.9%	0.011	1148	98.5%	0.010
Mixed	309	297	96.1%	0.000	303	98.1%	0.000
All (20Kb+)	2456	2379	96.9%	0.005	2426	98.8%	0.006

Sites longer than 10 kilobases

Site Category	Sites	Affymetrix 6.0		
		Sites with probes	%	Est. FDR
CN <= 2 only	2058	2019	98.1%	0.002
CN >= 2 only	1736	1691	97.4%	0.017
Mixed	562	533	94.8%	0.000
All (10Kb+)	4356	4243	97.4%	0.008

All sites *

Site Category	Sites	Affymetrix 6.0		
		Sites with probes	%	Est. FDR
CN <= 2 only	4780	4490	93.9%	0.014
CN >= 2 only	2530	2401	94.9%	0.058
Mixed	1137	1039	91.4%	0.012
All	8447	7930	93.9%	0.027

* Sites based on annotated segmental duplications in the hg19 reference are not included

Supplementary Table 4

Genotype concordance of sequencing-based genotypes to aCGH-based genotypes

Overall concordance	0.9992
Concordance at non-homref calls	0.9896

		Conrad aCGH copy number genotype						
		No call	CN0	CN1	CN2	CN3	CN4	CN5
Genome STRiP copy number genotype	No call	3	4	161	147	21	5	-
	CN0	6	2097	1	-	-	-	-
	CN1	127	1	12762	28	-	-	-
	CN2	735	1	46	186839	35	1	-
	CN3	24	-	1	14	675	6	-
	CN4	12	-	-	-	20	179	-
	CN5	5	-	-	-	1	4	8
	CN6	-	-	-	-	-	2	2
	CN7	-	-	-	-	-	1	-
	CN8	-	-	-	-	-	-	-
	CN9	-	-	-	-	-	-	-
	CN10	-	-	-	-	-	-	-
	CN11	-	-	-	-	-	1	-
Discordance by aCGH genotype			0.0010	0.0037	0.0002	0.0766	0.0773	0.2000

Supplementary Table 5

ddPCR evaluation of discordant genotypes between sequencing data and Conrad aCGH data

Genome STRiP Genotype Accuracy (at discordant genotypes) 0.8571
Conrad Genotype Accuracy (at discordant genotypes) 0.1429

ASSAYID	SITEID	Genotypes Concordant GS vs Conrad	Genotypes Discordant GS vs Conrad	At Concordant Genotypes (GS vs Conrad)			At Discordant Genotypes (GS vs Conrad)				DDPCR Call Rate
				DD Conc	DD Disc	DD Nocall	DD = GS	DD = Conrad	DD = Neither	DD Nocall	
mc1001b	CNV_15_23619237_23670664	30	12	30	0	0	12	0	0	0	1.000
mc1002a	CNV_20_14423033_14442180	30	10	30	0	0	10	0	0	0	1.000
mc1004b	CNV_19_23374358_23378830	31	2	31	0	0	0	2	0	0	1.000
mc1005a	CNV_3_63125152_63141614	36	5	36	0	0	5	0	0	0	1.000
mc1006a	CNV_22_18619095_18625292	32	5	32	0	0	5	0	0	0	1.000
mc1007a	CNV_6_256883_296034	36	2	16	0	20	1	0	0	1	0.447
mc1008a	CNV_4_120552552_120556118	36	2	30	0	6	0	2	0	0	0.842
mc1009a	CNV_2_205707172_205712081	38	2	36	1	1	0	2	0	0	0.975
mc1011a	CNV_16_15977724_16024808	37	2	37	0	0	2	0	0	0	1.000
mc1012a	CNV_3_129075748_129083222	0	1	0	0	0	1	0	0	0	1.000
Totals		306	43	278	1	27	36	6	0	1	

Conrad Accuracy (overall) * 0.8847
Genome STRiP Accuracy (overall) * 0.9782
ddPCR discordance rate 0.0036

* Overall accuracy calculated at the YRI2 genotypes from these 10 discordant sites where genotype calls are made by all 3 methods (Genome STRiP, Conrad, ddPCR)
ddPCR discordance rate calculated from sites where Genome STRiP and Conrad agree but ddPCR does not
ddPCR genotype calls are calculated by rounding to the nearest integer

Supplementary Table 6

Genotype validation - concordance between copy number calls from sequencing and ddPCR

Overall concordance:	0.9986
Sequencing call rate:	0.9574
ddPCR call rate:	0.9298

ddPCR copy number \ Sequencing copy number	No call	CN0	CN1	CN2	CN3	CN4	CN5	CN6	CN7	CN8	CN9
No call	6	-	-	1	2	10	10	15	9	3	-
CN0	-	-	-	-	-	-	-	-	-	-	-
CN1	-	-	1	-	-	-	-	-	-	-	-
CN2	1	-	-	57	-	-	-	-	-	-	-
CN3	5	-	-	-	128	-	-	-	-	-	-
CN4	5	-	-	-	-	259	-	-	-	-	-
CN5	8	-	-	-	-	-	105	-	-	-	-
CN6	5	-	-	-	-	-	-	92	-	-	-
CN7	3	-	-	-	-	-	-	-	31	1	-
CN8	1	-	-	-	-	-	-	-	-	38	-
CN9	-	-	-	-	-	-	-	-	-	-	2

Supplementary Table 7

CNV impact on genes by coding sequence overlap

Genes with CNVs overlapping coding sequence

				Average gene impact difference between two individuals				
Category	# CNVs	Genic CNVs	Genes	Singletons	AAF < 1%	AAF < 5%	AAF > 5%	**Overall**
Deletions	4781	476	546	0.83	1.89	3.27	12.08	**15.35**
Duplications	2522	703	1018	1.45	3.80	6.88	4.06	**10.94**
Multi-allelic	1356	270	449	n/a	1.46	5.61	102.38	**107.99**
All	8659	1449	2013	2.28	7.15	15.76	118.52	**134.28**

Supplementary Table 8

Impact on genes (requiring minimum 2 observations per copy number class)

Impact on whole genes

				Average gene dosage difference between two individuals				
Category	# CNVs	Genic CNVs	Genes	Singletons	AAF < 1%	AAF < 5%	AAF > 5%	**Overall**
Deletions	2794	40	50	n/a	0.36	0.73	5.04	**5.77**
Duplications	1369	109	178	n/a	1.24	2.52	2.42	**4.95**
Multi-allelic	451	83	162	n/a	0.14	1.55	61.91	**63.46**
All	4614	232	390	0.00	1.74	4.80	69.37	**74.18**

Impact on gene CDS

				Average gene impact difference between two individuals				
Category	# CNVs	Genic CNVs	Genes	Singletons	AAF < 1%	AAF < 5%	AAF > 5%	**Overall**
Deletions	2794	178	211	n/a	1.21	2.93	12.89	**15.82**
Duplications	1369	311	448	n/a	2.66	6.62	6.60	**13.22**
Multi-allelic	451	159	290	n/a	0.36	3.28	99.03	**102.31**
All	4614	648	949	0.00	4.23	12.83	118.52	**131.35**

Supplementary Table 9

Gene set enrichment

Gene set enrichment as computed by Amigo2 for genes completely overlapped by CNVs ascertained in this study. Three tables are listed below, one for enrichment based on all genic CNVs, one for enrichment based on the bi-allelic subset of these CNVs and one for the multi-allelic subset. In all three tables, only results with a p-value less than 10^{-5} are listed.

Table 9a: All genic CNVs

Term	Ontology	Background frequency	Sample frequency	Expected	+/-	P-value
olfactory receptor activity (GO:0004984)	Molecular function	376	100	9.04E+00	+	1.38E-68
detection of chemical stimulus involved in sensory perception (GO:0050907)	Biological process	419	104	1.01E+01	+	2.76E-68
detection of chemical stimulus involved in sensory perception of smell (GO:0050911)	Biological process	376	100	9.04E+00	+	3.00E-68
sensory perception of chemical stimulus (GO:0007606)	Biological process	467	106	1.12E+01	+	6.48E-66
sensory perception of smell (GO:0007608)	Biological process	403	100	9.69E+00	+	1.81E-65
detection of stimulus involved in sensory perception (GO:0050906)	Biological process	463	105	1.11E+01	+	3.29E-65
detection of chemical stimulus (GO:0009593)	Biological process	454	104	1.09E+01	+	5.87E-65
detection of stimulus (GO:0051606)	Biological process	642	108	1.54E+01	+	1.33E-54
G-protein coupled receptor activity (GO:0004930)	Molecular function	806	112	1.94E+01	+	5.00E-49
sensory perception (GO:0007600)	Biological process	847	107	2.04E+01	+	1.55E-42
G-protein coupled receptor signaling pathway (GO:0007186)	Biological process	1072	116	2.58E+01	+	5.24E-40
transmembrane signaling receptor activity (GO:0004888)	Molecular function	1172	116	2.82E+01	+	1.07E-36
neurological system process (GO:0050877)	Biological process	1118	110	2.69E+01	+	5.02E-34
signaling receptor activity (GO:0038023)	Molecular function	1269	116	3.05E+01	+	1.62E-33
molecular transducer activity (GO:0060089)	Molecular function	1571	123	3.78E+01	+	1.94E-29
signal transducer activity (GO:0004871)	Molecular function	1571	123	3.78E+01	+	1.94E-29
receptor activity (GO:0004872)	Molecular function	1488	117	3.58E+01	+	5.73E-28
system process (GO:0003008)	Biological process	1562	119	3.75E+01	+	5.83E-27
cell surface receptor signaling pathway (GO:0007166)	Biological process	2707	139	6.51E+01	+	5.75E-16
cellular metabolic process (GO:0044237)	Biological process	8281	113	1.99E+02	-	1.43E-13
response to chemical (GO:0042221)	Biological process	3427	153	8.24E+01	+	1.50E-12
primary metabolic process (GO:0044238)	Biological process	8428	120	2.03E+02	-	3.00E-12
organic substance metabolic process (GO:0071704)	Biological process	8680	126	2.09E+02	-	4.75E-12
metabolic process (GO:0008152)	Biological process	9570	146	2.30E+02	-	7.11E-12
intrinsic component of membrane (GO:0031224)	Cellular component	5305	200	1.28E+02	+	1.27E-10
binding (GO:0005488)	Molecular function	12375	217	2.97E+02	-	1.47E-10
cellular macromolecule metabolic process (GO:0044260)	Biological process	6150	79	1.48E+02	-	2.43E-10
protein binding (GO:0005515)	Molecular function	8230	123	1.98E+02	-	2.59E-10
integral component of membrane (GO:0016021)	Cellular component	5188	194	1.25E+02	+	7.69E-10
macromolecule metabolic process (GO:0043170)	Biological process	6833	95	1.64E+02	-	1.40E-09
intracellular part (GO:0044424)	Cellular component	12509	225	3.01E+02	-	1.89E-09
anion binding (GO:0043168)	Molecular function	2519	18	6.05E+01	-	2.34E-09
membrane part (GO:0044425)	Cellular component	6175	219	1.48E+02	+	2.42E-09
cellular aromatic compound metabolic process (GO:0006725)	Biological process	4737	55	1.14E+02	-	2.93E-09
organic cyclic compound binding (GO:0097159)	Molecular function	5719	75	1.37E+02	-	3.70E-09
small molecule binding (GO:0036094)	Molecular function	2544	19	6.11E+01	-	5.20E-09
nitrogen compound metabolic process (GO:0006807)	Biological process	5347	68	1.29E+02	-	9.75E-09
heterocyclic compound binding (GO:1901363)	Molecular function	5650	75	1.36E+02	-	1.03E-08
nucleus (GO:0005634)	Cellular component	6207	87	1.49E+02	-	1.39E-08
cellular nitrogen compound metabolic process (GO:0034641)	Biological process	4934	61	1.19E+02	-	2.01E-08
nuclear lumen (GO:0031981)	Cellular component	2767	24	6.65E+01	-	2.25E-08
nucleobase-containing compound metabolic process (GO:0006139)	Biological process	4550	54	1.09E+02	-	2.40E-08
organelle (GO:0043226)	Cellular component	11824	213	2.84E+02	-	2.62E-08
anatomical structure development (GO:0048856)	Biological process	4047	45	9.73E+01	-	2.86E-08
nervous system development (GO:0007399)	Biological process	1865	10	4.48E+01	-	2.91E-08
developmental process (GO:0032502)	Biological process	4634	56	1.11E+02	-	3.44E-08
heterocycle metabolic process (GO:0046483)	Biological process	4731	58	1.14E+02	-	3.97E-08
cell periphery (GO:0071944)	Cellular component	4522	169	1.09E+02	+	4.25E-08
intracellular organelle (GO:0043229)	Cellular component	10900	192	2.62E+02	-	4.41E-08
plasma membrane (GO:0005886)	Cellular component	4428	166	1.06E+02	+	5.27E-08
organic cyclic compound metabolic process (GO:1901360)	Biological process	4960	63	1.19E+02	-	6.61E-08
single-organism developmental process (GO:0044767)	Biological process	4587	56	1.10E+02	-	7.06E-08
intracellular (GO:0005622)	Cellular component	12662	236	3.04E+02	-	1.12E-07
multicellular organismal development (GO:0007275)	Biological process	4080	48	9.81E+01	-	2.27E-07
nucleoside phosphate binding (GO:1901265)	Molecular function	2271	18	5.46E+01	-	2.44E-07
nucleotide binding (GO:0000166)	Molecular function	2270	18	5.46E+01	-	2.48E-07
carbohydrate derivative binding (GO:0097367)	Molecular function	2138	16	5.14E+01	-	2.51E-07
nuclear part (GO:0044428)	Cellular component	3097	32	7.44E+01	-	2.90E-07
membrane-bounded organelle (GO:0043227)	Cellular component	10892	196	2.62E+02	-	4.23E-07
intracellular membrane-bounded organelle (GO:0043231)	Cellular component	9853	172	2.37E+02	-	4.60E-07
cellular component organization or biogenesis (GO:0071840)	Biological process	4354	54	1.05E+02	-	4.70E-07
cytosol (GO:0005829)	Cellular component	2632	25	6.33E+01	-	6.70E-07
intracellular organelle lumen (GO:0070013)	Cellular component	3348	38	8.05E+01	-	1.20E-06
cell projection (GO:0042995)	Cellular component	1395	7	3.35E+01	-	1.36E-06
cellular component organization (GO:0016043)	Biological process	4248	54	1.02E+02	-	2.20E-06
organelle lumen (GO:0043233)	Cellular component	3405	40	8.18E+01	-	2.69E-06
enzyme binding (GO:0019899)	Molecular function	1289	6	3.10E+01	-	3.35E-06
cytoplasm (GO:0005737)	Cellular component	9554	169	2.30E+02	-	3.40E-06
RNA metabolic process (GO:0016070)	Biological process	3180	35	7.64E+01	-	3.57E-06
system development (GO:0048731)	Biological process	3490	41	8.39E+01	-	5.18E-06
membrane-enclosed lumen (GO:0031974)	Cellular component	3462	42	8.32E+01	-	5.77E-06
purine nucleotide binding (GO:0017076)	Molecular function	1825	14	4.39E+01	-	6.81E-06
ribonucleotide binding (GO:0032553)	Molecular function	1820	14	4.37E+01	-	7.45E-06
defense response to other organism (GO:0098542)	Biological process	319	27	7.67E+00	+	8.11E-06
purine ribonucleotide binding (GO:0032555)	Molecular function	1805	14	4.34E+01	-	9.76E-06

Table 9b: All bi-allelic CNVs

Term	Ontology	Background frequency	Sample frequency	Expected	+/-	P-value
detection of chemical stimulus involved in sensory perception (GO:0050907)	Biological process	419	79	6.82E+00	+	6.76E-56
olfactory receptor activity (GO:0004984)	Molecular function	376	75	6.12E+00	+	5.91E-55
sensory perception of chemical stimulus (GO:0007606)	Biological process	467	81	7.60E+00	+	1.08E-54
detection of chemical stimulus involved in sensory perception of smell (GO:0050911)	Biological process	376	75	6.12E+00	+	1.43E-54
detection of stimulus involved in sensory perception (GO:0050906)	Biological process	463	80	7.54E+00	+	7.69E-54
detection of chemical stimulus (GO:0009593)	Biological process	454	79	7.39E+00	+	2.45E-53
sensory perception of smell (GO:0007608)	Biological process	403	75	6.56E+00	+	1.83E-52
detection of stimulus (GO:0051606)	Biological process	642	82	1.05E+01	+	1.85E-45
G-protein coupled receptor activity (GO:0004930)	Molecular function	806	84	1.31E+01	+	1.86E-40
sensory perception (GO:0007600)	Biological process	847	82	1.38E+01	+	9.96E-37
G-protein coupled receptor signaling pathway (GO:0007186)	Biological process	1072	87	1.75E+01	+	1.61E-33
neurological system process (GO:0050877)	Biological process	1118	85	1.82E+01	+	1.23E-30
transmembrane signaling receptor activity (GO:0004888)	Molecular function	1172	85	1.91E+01	+	1.40E-29
signaling receptor activity (GO:0038023)	Molecular function	1269	85	2.07E+01	+	3.43E-27
system process (GO:0003008)	Biological process	1562	93	2.54E+01	+	6.39E-26
molecular transducer activity (GO:0060089)	Molecular function	1571	89	2.56E+01	+	1.61E-23
signal transducer activity (GO:0004871)	Molecular function	1571	89	2.56E+01	+	1.61E-23
receptor activity (GO:0004872)	Molecular function	1488	86	2.42E+01	+	3.43E-23
cell surface receptor signaling pathway (GO:0007166)	Biological process	2707	101	4.41E+01	+	1.18E-13
response to chemical (GO:0042221)	Biological process	3427	110	5.58E+01	+	1.02E-10
metabolic process (GO:0008152)	Biological process	9570	95	1.56E+02	-	3.60E-09
integral component of membrane (GO:0016021)	Cellular component	5188	139	8.45E+01	+	6.11E-09
cellular metabolic process (GO:0044237)	Biological process	8281	77	1.35E+02	-	6.67E-09
intrinsic component of membrane (GO:0031224)	Cellular component	5305	141	8.64E+01	+	7.21E-09
organic substance metabolic process (GO:0071704)	Biological process	8680	84	1.41E+02	-	1.80E-08
primary metabolic process (GO:0044238)	Biological process	8428	81	1.37E+02	-	2.94E-08
membrane part (GO:0044425)	Cellular component	6175	155	1.01E+02	+	3.39E-08
cell periphery (GO:0071944)	Cellular component	4522	123	7.36E+01	+	6.05E-08
plasma membrane (GO:0005886)	Cellular component	4428	121	7.21E+01	+	6.80E-08
negative regulation of retinoic acid receptor signaling pathway (GO:0048387)	Biological process	30	9	4.88E-01	+	4.90E-07
anion binding (GO:0043168)	Molecular function	2519	11	4.10E+01	-	5.59E-07
intracellular part (GO:0044424)	Cellular component	12509	151	2.04E+02	-	8.97E-07
binding (GO:0005488)	Molecular function	12375	149	2.02E+02	-	1.22E-06
small molecule binding (GO:0036094)	Molecular function	2544	12	4.14E+01	-	1.55E-06
nuclear part (GO:0044428)	Cellular component	3097	18	5.04E+01	-	1.63E-06
organic cyclic compound binding (GO:0097159)	Molecular function	5719	50	9.31E+01	-	1.81E-06
regulation of retinoic acid receptor signaling pathway (GO:0048385)	Biological process	35	9	5.70E-01	+	1.83E-06
nuclear lumen (GO:0031981)	Cellular component	2767	15	4.51E+01	-	3.10E-06
retinoic acid receptor binding (GO:0042974)	Molecular function	42	9	6.84E-01	+	3.53E-06
heterocyclic compound binding (GO:1901363)	Molecular function	5650	50	9.20E+01	-	3.68E-06
carbohydrate derivative binding (GO:0097367)	Molecular function	2138	9	3.48E+01	-	5.96E-06
intracellular (GO:0005622)	Cellular component	12662	157	2.06E+02	-	6.93E-06

Table 9c: All multi-allelic CNVs

Term	Ontology	Background frequency	Sample frequency	Expected	+/-	P-value
olfactory receptor activity (GO:0004984)	Molecular function	376	25	2.91E+00	+	2.00E-14
detection of chemical stimulus involved in sensory perception of smell (GO:0050911)	Biological process	376	25	2.91E+00	+	2.74E-14
sensory perception of smell (GO:0007608)	Biological process	403	25	3.12E+00	+	1.30E-13
detection of chemical stimulus involved in sensory perception (GO:0050907)	Biological process	419	25	3.25E+00	+	3.11E-13
detection of chemical stimulus (GO:0009593)	Biological process	454	25	3.52E+00	+	1.84E-12
detection of stimulus involved in sensory perception (GO:0050906)	Biological process	463	25	3.59E+00	+	2.84E-12
sensory perception of chemical stimulus (GO:0007606)	Biological process	467	25	3.62E+00	+	3.44E-12
detection of stimulus (GO:0051606)	Biological process	642	26	4.98E+00	+	5.26E-10
G-protein coupled receptor activity (GO:0004930)	Molecular function	806	28	6.25E+00	+	1.91E-09
transmembrane signaling receptor activity (GO:0004888)	Molecular function	1172	31	9.08E+00	+	1.15E-07
G-protein coupled receptor signaling pathway (GO:0007186)	Biological process	1072	29	8.31E+00	+	3.43E-07
signaling receptor activity (GO:0038023)	Molecular function	1269	31	9.84E+00	+	7.23E-07
sensory perception (GO:0007600)	Biological process	847	25	6.57E+00	+	8.49E-07
cell differentiation (GO:0030154)	Biological process	2793	2	2.17E+01	-	2.14E-06
molecular transducer activity (GO:0060089)	Molecular function	1571	34	1.22E+01	+	2.46E-06
signal transducer activity (GO:0004871)	Molecular function	1571	34	1.22E+01	+	2.46E-06
cellular developmental process (GO:0048869)	Biological process	2915	3	2.26E+01	-	7.09E-06
anatomical structure development (GO:0048856)	Biological process	4047	8	3.14E+01	-	8.11E-06

Supplementary Table 10

Gene expression eQTLs intersecting CNV proxy SNPs

The best proxy SNP for each CNV (with a minimum r2 of 0.3) is matched against gene expression eQTLs from the Geuvadis study. The results were then filtered for intersections where the CNV proxy p-value (-log10(p)) was at least 70% of the best eQTL and the correlation p-value between the CNV and gene expression was less than 0.01. For CNVs with multiple equivalent proxy SNPs, we report the SNP with the lowest eQTL p-value. Distance between gene and CNV are based on the outermost coordinates of both; a distance of zero implies partial or complete overlap. Locus numbers were assigned manually to eliminate likely redundant loci. Gene overlap indicates whether the CNVs at the locus fully contain at least one gene or partially contain at least one gene.

Locus	Gene	CNV	Category	Copy Number	Population	Best SNP	r2	eQTL Gene		Distance	eQTL -log10(p)	eQTL -log10(p)	r	|r|
ID	overlap			Range		proxy to CNV	(SNP to CNV)	ID	Symbol	Gene to CNV	(CNV proxy SNP)	(eQTL peak SNP)	CNV to Gene	eQTL to Gene
1	Full	CNV_M1_HG19_1_25593922_25661196	mCNV	0-3	EUR	rs72660908	0.962	ENSG00000187010.13	RHD	0	30.70	31.92	0.568	0.556
		CNV_M2_HG19_1_25594516_25655519_1_25688914_25751819	mCNV	2-5	EUR	rs72660908	0.604	ENSG00000187010.13	RHD	0	30.70	31.92	0.568	0.556
2		CNV_M2_HG19_1_248584240_248623089_1_248795556_248834181	DUP	4-5	YRI	rs9724898	0.361	ENSG00000196539.2	OR2T3	13537	6.34	6.34	0.373	0.268
		CNV_M1_HG19_1_248620631_248635914	mCNV	1-5	EUR	rs61834536	0.497	ENSG00000196539.2	OR2T3	712	8.80	10.92	0.373	0.268
3	Partial	CNV_M1_HG19_3_37978283_37986832	DEL	0-2	EUR	rs7629707	0.635	ENSG00000144677.10	CTDSPL	0	5.40	5.70	-0.233	0.170
4	Partial	CNV_M1_HG19_3_136021069_136026053	DEL	0-2	EUR	rs1279949	0.936	ENSG00000114054.9	PCCB	0	11.13	11.21	-0.270	0.268
5	Partial	CNV_M1_HG19_3_191064583_191071561	DEL	0-2	EUR	rs76898988	1.000	ENSG00000152492.9	CCDC50	0	6.72	6.72	-0.183	0.177
6		CNV_M1_HG19_3_195421487_195446004	mCNV	2-9	EUR	rs13303068	0.416	ENSG00000260261.1	RP11-480A16.1	230054	13.36	16.22	0.492	0.298
7	Partial	CNV_M1_HG19_4_69382540_69430421	DEL	0-2	EUR	rs149896834	0.966	ENSG00000196620.4	UGT2B15	81926	66.77	68.12	0.671	0.663
		CNV_M1_HG19_4_69382540_69430421	DEL	0-2	EUR	rs149896834	0.966	ENSG00000197888.2	UGT2B17	0	66.24	68.12	0.670	0.662
		CNV_M1_HG19_4_69382540_69430421	DEL	0-2	YRI	rs148518713	0.916	ENSG00000196620.4	UGT2B15	81926	8.70	11.14	0.671	0.663
		CNV_M1_HG19_4_69382540_69430421	DEL	0-2	YRI	rs148518713	0.916	ENSG00000197888.2	UGT2B17	0	8.40	11.08	0.670	0.662
		CNV_M1_HG19_4_69435168_69484713	DEL	0-2	EUR	rs149896834	0.966	ENSG00000196620.4	UGT2B15	27634	66.77	68.12	0.671	0.663
		CNV_M1_HG19_4_69435168_69484713	DEL	0-2	EUR	rs149896834	0.966	ENSG00000197888.2	UGT2B17	922	66.24	68.12	0.670	0.662
		CNV_M1_HG19_4_69435168_69484713	DEL	0-2	YRI	rs148518713	0.889	ENSG00000196620.4	UGT2B15	27634	8.70	11.14	0.671	0.663
		CNV_M1_HG19_4_69435168_69484713	DEL	0-2	YRI	rs148518713	0.889	ENSG00000197888.2	UGT2B17	922	8.40	11.08	0.670	0.662
8		CNV_M1_HG19_5_178109202_178113340	DEL	0-2	EUR	rs12187838	0.789	ENSG00000169131.6	ZNF354A	25252	8.52	11.08	-0.192	0.213
		CNV_M1_HG19_5_178348230_178353193	DEL	0-2	EUR	rs186159699	0.709	ENSG00000178338.6	ZNF354B	33106	5.58	5.98	-0.149	0.204
9		CNV_M1_HG19_5_180375492_180428287	mCNV	0-3	EUR	rs72494581	0.977	ENSG00000165810.11	BTNL9	38937	11.32	12.40	0.223	0.154
10		CNV_M1_HG19_6_29865905_29897945	DEL	0-2	EUR	rs115988571	1.000	ENSG00000204632.7	HLA-G	67002	12.78	14.60	0.340	0.321
11	Partial	CNV_M1_HG19_7_75664249_75667815	DEL	0-2	EUR	rs7798298	0.986	ENSG00000127952.11	STYXL1	0	15.33	16.64	-0.467	0.467
12	Partial	CNV_M2_HG19_9_69088571_69278385_9_70841062_71031684	mCNV	3-6	EUR	rs139206165	0.339	ENSG00000231242.1	RP11-87H9.3	0	8.43	9.38	0.187	0.196
		CNV_M1_HG19_9_69815487_69840962	mCNV	3-7	EUR	rs2794909	0.455	ENSG00000231242.1	RP11-87H9.3	706693	7.52	9.38	0.187	0.196
13		CNV_M2_HG19_12_10569234_10584672_12_10584673_10600097	mCNV	2-5	EUR	rs2246809	0.925	ENSG00000245648.1	RP11-277P12.20	18128	9.76	10.20	0.233	0.223
		CNV_M1_HG19_12_10581412_10586874	mCNV	0-3	EUR	rs2246809	0.982	ENSG00000245648.1	RP11-277P12.20	30306	9.76	10.20	0.233	0.223
14		CNV_M1_HG19_15_45154128_45160343	mCNV	0-3	YRI	rs146286069	0.315	ENSG00000140263.9	SORD	154958	6.54	7.11	-0.230	0.373
15		CNV_M1_HG19_16_19945435_19967702	DEL	0-2	EUR	rs11074418	1.000	ENSG00000167191.6	GPRC5B	47945	6.76	7.11	-0.207	0.170
16	Partial	CNV_M1_HG19_16_70183673_70198386	mCNV	1-8	EUR	.	0.462	ENSG00000090857.8	PDPR	0	9.32	11.68	0.640	0.378
17		CNV_M1_HG19_17_16710818_16724595	mCNV	0-4	EUR	rs678522	0.310	ENSG00000170160.10	CCDC144A	3050	5.31	5.31	0.182	0.197
18	Full	CNV_M2_HG19_17_43593498_43631279_17_45090987_45129812	DUP	2-4	EUR	rs114279117	0.861	ENSG00000120071.8	KANSL1	476002	6.35	7.27	0.308	0.339
		CNV_M2_HG19_17_43593498_43631279_17_45090987_45129812	DUP	2-4	EUR	rs114279117	0.861	ENSG00000176681.9	LRRC37A	675826	21.19	28.59	-0.370	0.346
		CNV_M2_HG19_17_43593498_43631279_17_45090987_45129812	DUP	2-4	EUR	rs114279117	0.861	ENSG00000214425.1	AC091132.3	0	20.62	28.09	0.339	0.484
		CNV_M1_HG19_17_44227967_44263121	mCNV	2-5	EUR	rs145827285	0.447	ENSG00000120071.8	KANSL1	0	7.26	7.27	0.308	0.339
19	Partial	CNV_M1_HG19_19_17443159_17449780	DUP	2-4	EUR	rs7259703	0.482	ENSG00000130299.10	GTPBP3	0	7.62	7.62	0.396	0.197
20		CNV_M1_HG19_19_20595496_20604163	DEL	0-2	EUR	rs148344613	0.779	ENSG00000237440.2	ZNF737	116635	6.44	6.44	0.194	0.148
		CNV_M1_HG19_19_20642556_20703930	DEL	0-2	EUR	rs148344613	0.898	ENSG00000237440.2	ZNF737	16868	6.44	6.44	0.194	0.148
21	Partial	CNV_M1_HG19_19_36840814_36846644	DEL	0-2	EUR	rs12462725	0.757	ENSG00000142065.7	ZFP14	0	5.56	5.93	0.213	0.208
		CNV_M1_HG19_19_36840814_36846644	DEL	0-2	EUR	rs62113149	0.757	ENSG00000181007.6	ZFP82	27948	13.55	13.91	0.293	0.334
22		CNV_M1_HG19_20_44203923_44207351	DEL	0-2	EUR	rs73129045	0.402	ENSG00000101473.11	ACOT8	263008	5.41	5.48	0.215	0.196
23	Partial	CNV_M1_HG19_21_44969557_44974112	mCNV	1-3	EUR	rs230646	0.784	ENSG00000160207.4	HSF2BP	0	26.57	35.18	0.406	0.510
24	Partial	CNV_M1_HG19_22_39359540_39386346	mCNV	0-3	EUR	rs12628403	0.907	ENSG00000128383.8	APOBEC3A	351	22.50	23.25	0.339	0.318
		CNV_M1_HG19_22_39359540_39386346	mCNV	0-3	EUR	rs12628403	0.907	ENSG00000179750.11	APOBEC3B	0	31.29	32.90	0.488	0.456
25		CNV_M2_HG19_22_42896180_42902137_22_42949740_42955460	mCNV	2-8	EUR	rs5751297	0.618	ENSG00000189306.6	RRP7A	3836	6.81	8.85	0.452	0.254

Supplementary Table 11

Candidate dispersed duplications identified by long-range LD

A duplication can in principle be "dispersed" in the sense that extra copies of a genomic locus are at genomic sites distant from the original sequence. We sought to identify potential dispersed duplications by looking for CNVs for which our genotypes exhibited linkage disequilibrium with SNPs on other chromosomes or at long genomic distances on the same chromosomes. We note that such an analysis is complicated by many potential issues, not least the possibility that the SNPs themselves may be mismapped due to segmental duplications. The list below should therefore be considered preliminary and a set of candidates rather than validated dispersed duplications. Four known examples of dispersed duplications (three experimentally validated, marked with asterisks below) served as positive controls for the analysis; all were re-discovered in this analysis. Most of the CNVs below were found in complex regions with extensive segmental duplication; while such architecture could contribute to dispersed duplication, it could also contribute to mismapping of nearby SNPs (note that many of these CNVs also showed some LD to local SNPs). Other forms of analysis would be required to confirm these as dispersed duplications. These 15 dispersed-duplication candidates represent 2.2% of all of the duplications in our data set with a duplication allele count of at least 15.

CNVID	Category	Copy Number Range	Frequency	Genes	Best association p-value -log10(p)		Local chromosome	Distant chromosome	Distant locus
					Local to CNV	Distant locus			
CNV_M2_HG19_1_144019498_144095783_1_206482222_206558788 *	mCNV	4-7	0.051	SRGAP2	15.1	91.4	1	1	148515254-148594051
CNV_M1_HG19_2_37958363_37970514 (Note 1)	mCNV	2-5	0.443	-	153.2	208.3	2	4	49537270-49633309
CNV_M1_HG19_2_37971788_38002047 (Note 1)	mCNV	2-5	0.446	-	153.9	249	2	4	49537270-49633309
CNV_M1_HG19_6_256883_296034 *	mCNV	2-6	0.414	DUSP22	16.5	108.2	6	16	33430008-33529083
CNV_M1_HG19_7_51435268_51477367	Dup	2-3	0.009	-	21	127.7	7	7	56241903-56341629
CNV_M1_HG19_7_65261544_65275079	Dup	2-4	0.500	-	146.5	324	7	7	55792718-55892500
CNV_M1_HG19_8_124082676_124086058	Dup	2-3	0.010	WDR67	6.2	49.5	8	11	119320100-119383544
CNV_M1_HG19_9_86509134_86518206	Dup	2-3	0.050	KIF27	31.4	239.4	9	9	88413820-88511940
CNV_M1_HG19_9_141071532_141092563	Dup	2-4	0.236	-	175.9	299.1	9	16	80212-175186
CNV_M1_HG19_10_81270579_81292860	Dup	2-3	0.020	EIF5AL1	12.7	46.6	10	2	89331340-89413194
CNV_M1_HG19_12_31312361_31407812	Dup	2-3	0.017	-	207.3	252.1	12	12	9528725-9623175
CNV_M1_HG19_12_124495889_124500138 *	Dup	2-4	0.057	ZNF664	NA	324	12	2	3913680-4008688
CNV_M1_HG19_16_34498796_34570427	mCNV	1-4	0.043	-	324	324	16	16	46698387-46797523
CNV_M1_HG19_16_34587447_34685602	mCNV	1-4	0.046	-	324	324	16	16	46698387-46797523
CNV_M1_HG19_16_34715349_34760389	Dup	2-4	0.043	-	324	324	16	16	46698387-46797523

Note 1: This CNV shows long-range LD to an isolated contig on chromosome 4 of build 37 of the human reference, but admixture analysis suggests that this duplication (and possibly duplications at the locus on chr4) both localize to chromosome 22.

Supplementary Table 12

CNVs with paralogs on multiple chromosomes

For 15 CNVs in segmentally duplicated regions with the paralogs on different chromosomes, we evaluated whether the copy number variation arises mostly from one paralog or the other. Analysis of read depth at sites of paralog-specific differences (PSDs) suggests that one paralog is contributing most of the variation in 13 out of 15 sites (green). For four of these sites, imputability of the CNV is better at the site of the correct paralog (blue). At one CNV (chr21:14801588-15174050, blue hatching) found mostly in Europeans, imputation dosage r^2 for paralog B is 0.264 in EUR (data not shown).

Paralog A			Paralog B				Correlation of paralog to total copy number (est.)		Imputed dosage r2 in leave-out trials	
Chrom	Start	End	Chrom	Start	End	Frequency	Paralog A	Paralog B	Paralog A	Paralog B
1	764792	802515	8	240490	280244	0.043	0.757	0.144	0.257	0.004
1	83598160	83647450	7	76182346	76231606	0.017	0.068	0.758	0.000	0.028
1	83647857	83955427	7	76280702	76575579	0.015	0.074	0.866	0.000	0.134
1	242413205	242528828	10	38520100	38637504	0.008	-0.153	-0.068	0.000	0.000
2	132660995	132699803	21	14593409	14627482	0.043	-0.034	0.196	0.000	0.000
2	159703462	159734019	3	125415459	125446156	0.023	0.134	0.048	0.000	0.000
6	132019328	132035259	7	143953514	143969444	0.615	0.320	0.905	0.050	0.114
7	35139142	35281183	12	63971080	64119245	0.020	-0.109	0.686	0.008	0.000
7	102815781	102929218	12	63954365	64072240	0.022	-0.062	0.662	0.000	0.000
10	60001	130311	18	14415	84190	0.025	-0.012	0.508	0.108	0.000
13	19167974	19275982	18	14358135	14464819	0.011	0.443	-0.011	0.000	0.000
13	19301588	19448886	18	14185210	14353419	0.026	0.308	0.001	0.000	0.000
18	14358135	14728624	21	14801588	15174050	0.008	0.008	0.518	0.000	0.000
18	14807068	14897137	21	14714527	14801577	0.016	0.023	0.604	0.000	0.054
18	15026971	15101188	21	14640424	14714507	0.009	0.009	0.568	0.000	0.000

Supplementary Table 13

Droplet digital PCR assays

Legend

SegDup: Is the assay targeting a region where the reference genome has two copies

Conrad: Yes if this assay was used for resolving discordant genotypes between sequencing and aCGH, no if this assay was used for sequencing concordance analysis

Assay ID	Data Location	CNV Region	SegDup	Conrad	Assay Notes	Commercial ID
mc1001b	Supp Fig 3; Supp Table 6	CNV_M1_HG19_15_23619237_23670664	N	Y	good	
mc1002a	Supp Fig 3; Supp Table 6	CNV_M1_HG19_20_14423033_14442180	N	Y	good	
mc1004b	Supp Fig 3; Supp Table 6	CNV_M1_HG19_19_23374358_23378830	N	Y	good	
mc1005a	Supp Fig 3; Supp Table 6	CNV_M1_HG19_3_63125152_63141614	N	Y	good	
mc1006a	Supp Fig 3; Supp Table 6	CNV_M1_HG19_22_18619095_18625292	N	Y	good	
mc1007a	Supp Fig 3; Supp Table 6	CNV_M1_HG19_6_256883_296034	N	Y	good	
mc1008a	Supp Fig 3; Supp Table 6	CNV_M1_HG19_4_120552552_120556118	N	Y	good	
mc1009a	Supp Fig 3; Supp Table 6	CNV_M1_HG19_2_205707172_205712081	N	Y	good	
mc1011a	Supp Fig 3; Supp Table 6	CNV_M1_HG19_16_15977724_16024808	N	Y	good	
mc1012a	Supp Fig 3; Supp Table 6	CNV_M1_HG19_3_129075748_129083222	N	Y	good	
CC_SD_003	Supp Fig 4	CNV_M2_HG19_15_21033446_21199563_15_22044595_22210800	Y	N	good	
CC_SD_004	Fig 3d, Supp Fig 4	CNV_M2_HG19_5_686123_731394_5_778191_815714	Y	N	good	
CC_SD_007	Fig 3e, Supp Fig 4	CNV_M2_HG19_8_7200001_7436083_8_7600001_7825413	Y	N	good	
CC_SD_009	Supp Fig 4	CNV_M2_HG19_1_121086699_121133098_1_206536833_206583337	Y	N	good; some variance around CN8	
CC_SD_010	Fig 3f, Supp Fig 4	CNV_M2_HG19_7_143884042_143951886_7_143993276_144061197	Y	N	good	
CC_SD_011	Supp Fig 4	CNV_M2_HG19_7_143969445_143993281_7_144050481_144074376	Y	N	good	
CC_SD_012	Supp Fig 4	CNV_M2_HG19_1_144019498_144095783_1_206482222_206558788	Y	N	good; high negatives	
CC_SD_015	Supp Fig 4	CNV_M2_HG19_1_161479598_161518642_1_161560922_161599999	Y	N	good	
CC_SD_016	Supp Fig 4	CNV_M2_HG19_1_196711705_196740354_1_196796320_196825045	Y	N	good	
m011	Fig 3c, Supp Fig 4	CNV_M1_HG19_1_145027118_145062269	N	N	good	
m017	Fig 3b, Supp Fig 4	CNV_M1_HG19_10_47087995_47259996	N	N	good	
m024	Supp Fig 4	CNV_M1_HG19_15_22343951_22372926	N	N	good	
m025	Supp Fig 4	CNV_M1_HG19_5_756753_779167	N	N	good	
m026	Supp Fig 4	CNV_M1_HG19_16_32490484_32652840	N	N	good	
m028	Supp Fig 4	CNV_M1_HG19_6_312088_376690	N	N	good	
m029	Supp Fig 4	CNV_M1_HG19_6_256883_296034	N	N	good	
m033	Supp Fig 4	CNV_M1_HG19_15_22336771_22343950	N	N	good	
m035	Supp Fig 4	CNV_M1_HG19_21_10901970_10945260	N	N	secondary clusters	
m037	Supp Fig 4	CNV_M1_HG19_9_69815487_69840962	N	N	rainy	
m039	Fig 3a, Supp Fig 4	CNV_M1_HG19_15_22436024_22559781	N	N	good	
m042	Supp Fig 4	CNV_M2_HG19_5_686123_731394_5_778191_815714	Y	N	rainy	
MRGPRX1	Supp Fig 4	CNV_M1_HG19_11_18941449_18963992	N	N	Validated, commercialized assay	dHsaCP1000474
TERT	Reference Assay				Validated, commercialized assay	dHsaCP1000100

Assay ID	CNV Region	F Primer Seq	R Primer Seq	Probe Seq	dMIQE (hg19)
mc1001b	CNV_M1_HG19_15_23619237_23670664	GTGTTTGCAGGGAGACATCA	CCGGTTGACATTGACTCTCC	TCCTCTTCTCTTGCGCTTTCTGAATG	
mc1002a	CNV_M1_HG19_20_14423033_14442180	TGGTCCTAGAAGCTGAATACCT	CCTGTAGTTCTCATTTCTTACTCTT	ACCTCCCCACAAGAGATGGCA	
mc1004b	CNV_M1_HG19_19_23374358_23378830	ACTTTTAGTGAGCTCAGCAGA	TGGAGCCCTGGATGGTTTAT	ACCATGGGTTCAAACTCTACTGACCA	
mc1005a	CNV_M1_HG19_3_63125152_63141614	TTGGGGTTTCTGGTTCTCTTC	AGGCAGCTCACAGAATTAGG	TGAGTCCCAGGTTTCTGAAGCTCT	
mc1006a	CNV_M1_HG19_22_18619095_18625292	TCCAGTTAACCTGTGGAGCT	CTCCGGATCCGACCTTGTTA	AGAAGCTCTCACTCACAACTGGGA	
mc1007a	CNV_M1_HG19_6_256883_296034	TGACACTGCGAGCACTCA	TATGGCATGTTACAGGGAGGA	CTCCCGCGCGCCTCCTCC	
mc1008a	CNV_M1_HG19_4_120552552_120556118	CAAGGGATTGGGGAACAGA	AGTTACACATTTTGGCTGAATCT	AGTGTCTCAAACAGTTCTAGTCTCTGGG	
mc1009a	CNV_M1_HG19_2_205707172_205712081	TTAGACAAGCCTTCCCCAAAG	AATGGTAGTGTCCTCCTCCA	TTTGGCAGTCAGCTTCATAGTTGGA	
mc1011a	CNV_M1_HG19_16_15977724_16024808	GACAATGATGGTCGGGGTTC	GCAAGGATCCGAGCTGAAG	ACGGTCATCATCTAGGGCATTGAAAA	
mc1012a	CNV_M1_HG19_3_129075748_129083222	AGTTTCTCCTAAGCTGCGTG	TCGTATCATCCCATTTGCATCT	CCACTGGCACTCATTAAAACCCACC	
CC_SD_003	CNV_M2_HG19_15_21033446_21199563_15_22044595_22210800	AGACACTGCCCACTAGCCT	TCATCTCAACCTCCTCAACGT	ACACCACGGTCATATGCCCATGG	
CC_SD_004	CNV_M2_HG19_5_686123_731394_5_778191_815714	GTTCCTGTTCTGGGGTTACTGTT	GTGCAAACTCAAGGGCTGG	ACATCAAGAGACAAACGACAAGCAGC	
CC_SD_007	CNV_M2_HG19_8_7200001_7436083_8_7600001_7825413	AGCCTGACAGACATACGTTGC	AACTCTGGGGAAGACACACT	TGCACATCCTAAGCAAAATCTAGCAGT	
CC_SD_009	CNV_M2_HG19_1_121086699_121133098_1_206536833_206583337	ACCTTTTTCACCGAAATACTGATCTC	GCAAAGGTTGTGGCTGATACAG	AGGGGAAGGGAGGGGTTACTG	
CC_SD_010	CNV_M2_HG19_7_143884042_143951886_7_143993276_144061197	CATCCTGCCCGTCAAGAGG	AGCAAGAAGGAACTGTGACTCC	TCCTCCTCCGTCTCTGACCCT	
CC_SD_011	CNV_M2_HG19_7_143969445_143993281_7_144050481_144074376	ATGCCTATGGTCCAACTCAGG	TTCAGTTCAGGCTTCTAGGGC	AGTGAGTGTGTGGGACAGCTCT	
CC_SD_012	CNV_M2_HG19_1_144019498_144095783_1_206482222_206558788	ACACACAAGCCACAAGACCA	GGCTTTCCTGCTCCTAAACTG	ATGTGGGTTGTGGCTGATGCA	
CC_SD_015	CNV_M2_HG19_1_161479598_161518642_1_161560922_161599999	GATGGTTTGGTGAGTGCCCT	AAGAATTGTGCACAGTGATGCT	AGCCTTCCATTTTCCTGGGACAC	
CC_SD_016	CNV_M2_HG19_1_196711705_196740354_1_196796320_196825045	TTGTCAGTATATGCTCCAGCTTCAT	GTTATTCGCTTGTTACCCTCAAGTT	AGTTGAGTACCAATGCCAGAACTTGT	
m011	CNV_M1_HG19_1_145027118_145062269	AGAGCAAGCTCTTCCAGACA	GTTCTGAAGGCTGGACCAAG	CTGAATCTGCCAGCACCTTGAT	
m017	CNV_M1_HG19_10_47087995_47259996	TGAAGTAGAGATCTGCCAGCAA	CACCTGCCATGTACTGGATATT	AGCCAGCATCAGGACCGA	
m024	CNV_M1_HG19_15_22343951_22372926	GCTTGTTCTTCCTTGAAGTGAC	GCCACTACTCATCACACACA	ACACACACACACACACACACACA	
m025	CNV_M1_HG19_5_756753_779167	TGATGCTGTGGTCAAGGAGA	ACGGAGTCTCTTGGTGTGTT	ACCAGCAGCTGCCAGGAA	
m026	CNV_M1_HG19_16_32490484_32652840	GTCCACTGCCTGGTTACATAC	CCAGACACTTCTTCTGTAAGGTC	ACTGTGTGCTCAGTCTGCTCT	
m028	CNV_M1_HG19_6_312088_376690	GAGCTGGTACGTTGAGTGC	TGGATTATATGTGGACAGCAAGC	TCGCCACCAGGAAGCCA	
m029	CNV_M1_HG19_6_256883_296034	TGGTGTTACATTGTGCTTAGGA	GGATGAAGAAGGAACATTGAGGA	AACTCTCCAGGACCTGAGCAG	
m033	CNV_M1_HG19_15_22336771_22343950	CCAATCAGAGTGGCTGTTATGAA	ACAACTCTGTACAAGCACTCC	ACAGGCTCACCAGCATCTGT	
m035	CNV_M1_HG19_21_10901970_10945260	AGTTGTGTTCCAAGAGCCAA	CCTAACGTGTAATACCTGAGCTT	TCCATGAGTAGCTTGACCAGCA	
m037	CNV_M1_HG19_9_69815487_69840962	AACCTGGTACCTGAGATGGA	CTCAGAACCAGAAGCAGAGG	TCCAAGCATCAGCAACGCA	
m039	CNV_M1_HG19_15_22436024_22559781	CCACTGTCTCCAGTCCACA	TACAGTGACTCCTGGAGCTG	CAGCAGCCATCATGTGACCTT	
m042	CNV_M2_HG19_5_686123_731394_5_778191_815714	CACGGCAACAGGACAGC	CCATCCTCTCTCTGGTCCTC	CATGCCTGGCAGAGACACG	
MRGPRX1	CNV_M1_HG19_11_18941449_18963992				chr11:18956210-18956332
TERT					chr5:1258717-1258839

Supplementary Note

Sequencing data and population cohorts

The sequencing data analyzed in this study consists of whole-genome shotgun sequencing from 849 individuals sequenced as part of Phase 1 of the 1000 Genomes Project. Where sequencing was available from multiple platforms, we used only the Illumina sequencing data. Sequencing coverage ranged from 2.0x to 20.6x across the 849 individuals (median 4.8x). Read lengths ranged from 36bp to 108bp.

The sequenced individuals were sampled from 14 populations (**Supplementary Table 1**). These individuals were further grouped into 4 continental populations (AFR: Africa, EUR: Europe, ASN: Asian, ADM: Admixed Americas). In contrast to the continental groupings used in the 1000 Genomes Project, in this study we analyze the ASW cohort (Americans of African ancestry), which exhibit European and African admixture, with the ADM (Admixed Americas) group instead of with the other African populations. To minimize confusion, we refer to the cohort that includes the admixed populations as ADM (instead of AMR) and refrain from using the AFR notation where the context would be unclear.

In total, we called CNVs utilizing the data from a total of 946 individuals from 1000 Genomes Phase 1. Of these individuals, 97 were filtered during QC as described below.

CNV calling

To discover and genotype the CNVs for this study, we further developed the Genome STRiP software (reaching internal version 1.04.1383, public version 2.0). Variants were called using two different pipelines and then merged. The first pipeline targets CNVs affecting sequences that are unique on the human genome reference. The second pipeline targets CNVs for which the human genome reference contains multiple copies of the CNV-affected segment. We developed new features and modules of Genome STRiP linked together into pipelines for each specific phase of discovery, genotyping and analysis, as described in the following sections.

CNV genotyping

The genotyping methods in Genome STRiP were used during CNV discovery, as described below, to generate a set of genotyped copy-number-variable regions mapped to the human reference genome. Previous versions of Genome STRiP (Handsaker, 2011) contained support for genotyping deletion variants but not duplications or multi-allelic CNVs. In this study, we developed a greatly enhanced version of the read-depth genotyping method from our previous work.

Normalization of read depth signal

As in our previous work, read depth of coverage is measured by counting sequenced DNA fragments overlapping a genomic interval of interest. To facilitate normalization, each fragment is counted as a point event, arbitrarily assigned to the coordinate that is the midpoint of the left-most read (for paired-end sequencing). Reads are only counted if they map to locations which should be uniquely alignable based on the structure of the reference genome. In addition, for CNV analysis, we applied a "low complexity mask" that masked genome coordinates falling in regions of low-complexity sequence (as categorized by the RepeatMasker tracks from the UCSC browser).

The raw read counts are normalized to correct for sequencing bias as a function of GC-content in each sequencing library. For each library, we estimate the GC-bias by binning the read counts by the GC-fraction of the reference sequencing in a 400bp window centered on each counted read. We compute the enrichment/depletion of read counts as a function of GC-fraction compared to a genome-wide average and normalize the expected read depth in each library by this factor. The GC-bias estimates are computed over a subset of the autosome excluding regions of potential copy-number variation, repeats and segmental duplication.

Genotyping mixture model

To genotype a CNV interval, Genome STRiP fits a constrained Gaussian mixture model to the read depth signal across the reference interval, using data from all available DNA samples. For deletion variants, a mixture of three Gaussians is used, corresponding to diploid copy number classes of 0, 1 and 2. For CNV genotyping, we fit a mixture of multiple Gaussians, corresponding to diploid copy number classes from zero to a site-specific maximum. The maximum copy number class modeled for each site is the maximum read depth signal from any of the samples at that site, plus one (but is never less than two times the reference genome copy number, plus one).

An advantage of the constrained Gaussian mixture model used in Genome STRiP is that in practice the mixture weights can be allowed to go to zero without adversely affecting the model fit. This eliminates the need to test and compare many different models with different numbers of copy number classes.

Assignment of absolute copy number

A key problem for CNV calling algorithms that use clustering is to accurately estimate the correct absolute copy number for each cluster. The constrained mixture model used in Genome STRiP is advantageous in this respect, especially when large numbers of genomes can be called together. The means of the copy number classes are required to scale as integer multiples (with a scaling parameter *m1* fitted from the data). Thus, the model is sensitive to the ratio between adjacent clusters, which can help to distinguish a cluster of CN 2-4 from a cluster of CN 4-6, for example. To avoid over-fitting, we reject models with a value of *m1* that is too high or too low (by default, requiring $0.5 <= m1 <= 2.0$).

Using copy number parity

An enhancement to the Genome STRiP genotyping model provides additional accuracy at calling absolute copy-number, especially at sites with high copy number.

This enhancement is based on the mathematical observation that for autosomal loci in a population in Hardy-Weinberg equilibrium, the number of individuals with even diploid copy numbers should not be less than the number of individuals with odd diploid copy numbers. (Intuitively, this is a generalization to mCNVs of the observation in SNP genotypes that the frequency of a heterozygote class should not exceed the combined frequencies of the homozygote classes.) At most sites, the allele frequencies will fall into a range where incorrectly shifted copy number assignments will cause the site to deviate strongly from HWE.

For example, a low to moderate frequency CN 2-4 site will have a very different observed distribution than a CN 1-3 site or a CN 3-5 site at a similar frequency. Using this information increases the effective separation between reasonable cluster assignments by a factor of two at most CNV sites.

We exploit this by performing a simple parity test on copy number. As we fit the mixture model, we estimate the number of even and odd copy numbers observed in the population. If the fraction of even copy-numbers falls below a threshold parameter (default 0.4), we shift the cluster assignments either up or down by one (the direction is determined by the current estimate of the *m1* model parameter) and restart the expectation-maximization loop. For small populations, family studies or highly-stratified populations, this optimization can be explicitly disabled by setting the parity threshold parameter to zero.

Genotyping both unique and duplicated sequences

Previous versions of Genome STRiP analyzed read depth only at positions on the reference that are sufficiently unique that the input reads should be able to align uniquely (based on read length and the repetitive structure of the reference genome). To genotype CNVs that are present in multiple copies on the reference genome, we extended our method to utilize positions that are non-unique on the reference by considering reads in reference k-mers that are not globally unique, but are present only within the paralogous CNV regions. The normalized read counts from such positions can be used to estimate the total copy number of a non-unique segment on the reference genome.

The read counts from both the unique and non-unique positions can be utilized together to estimate both the total copy number and the paralog-specific copy number. In this case, the unique positions represent paralog-specific differences (or paralog-specific variation, PSVs) that differentiate one paralog from the other, based on the reference genome.

CNV discovery set 1

In the first discovery set, we ascertained polymorphic CNVs with one copy of the CNV in the reference genome using a CNV discovery pipeline implemented on top of Genome STRiP. The CNV pipeline consists of the following major stages:

Seed windows

The pipeline begins by dividing the reference genome into overlapping seed windows, each window containing the same amount of uniquely-alignable sequence and thereby giving roughly equal power for variant calling. For this data set, we utilized windows with 5 Kb of uniquely-alignable sequence overlapping by 2.5 Kb. For this data set, we defined uniquely-alignable sequence as 36-mers from the reference which could be aligned uniquely back to the reference genome using bwa. For the reference genome, we used the 1000 Genomes Phase 1 reference, which is based on hg19.

Each of the initial seed windows was genotyped using Genome STRiP with default parameter settings. We found that for CNV calling we obtained better results by augmenting the usual alignability mask (based on 36-mers) with an additional alignability mask that filters out regions of low sequence complexity. This low-complexity mask was constructed from the intervals on the reference genome identified as low complexity repeat by RepeatMasker as downloaded from the UCSC browser. Specifically, all RepeatMasker annotated regions with a repeat class of “Low_complexity”, “Simple_repeat” or “Satellite” were masked.

After initial genotyping, windows that pass Genome STRiP default quality filters are retained (in this data set, 99.95% of all seed windows passed these quality filters). Seed windows are promoted for further processing by ignoring the DUPLICATE filter and eliminating any windows classified as NONVARIANT.

Seed window merging

Among the seed windows passing the first stage of the pipeline, all overlapping or adjacent windows are compared to detect windows with concordant genotypes and these windows are then merged. Concordant windows are detected by using the Genome STRiP Redundancy annotator and requiring that the merged windows have a duplicate score >= 0 (i.e. based on the computed copy-number likelihoods, the copy-number genotype of every sample is more likely to be the same than different).

The seed windows and merged windows are re-genotyped together and the Genome STRiP Redundancy annotator is used to perform duplicate elimination by selecting between the merged and unmerged seed windows at each site.

Sample filtering

Based on all called sites across the genome, all confidently called copy-numbers are evaluated to determine the number of variant sites called in each sample. Samples with

more than the median number of variant sites plus 3 median absolute deviations (MAD) are dropped from further analysis, as are all sites that are called variant only in these samples.

In this data set, 97 samples were filtered from the original input set of 946 samples to yield the final set of 849 samples used in this analysis.

Boundary refinement

After sample filtering, a hill-climbing algorithm was used to determine the best boundaries for each CNV segment detected in the pipeline. The boundary refinement algorithm alternately varies the left and right boundary of the CNV segment over a range of input values to converge on the best set of boundaries. The boundary increments are initially large (10% of the size of the input boundaries) and then gradually halved until they reach a target boundary precision (in this data set, 200bp) or a minimum interval length (in this data set, 2500bp).

Boundary refinement is performed separately on each input window. After refinement, two nearby or overlapping windows may predict the same or similar CNV boundaries. The refined intervals are all genotyped together and duplicate intervals are removed using the Genome STRiP Redundancy annotator.

Adjacent site merging

Compatible calls that were adjacent but not overlapping were checked and merged. Calls separated by less than 1Mb were evaluated for compatible copy-number genotypes using the Genome STRiP Redundancy annotator with a score threshold of zero (calls are compatible if every sample is more likely to have the same copy number genotype than a different copy number). When compatible calls were found, these were merged into a union interval, which was then genotyped and compared to the constituent calls using the Redundancy annotator. Merged calls were retained in favor of the constituent calls if the total posterior likelihood was greater than that of the constituent calls.

Filtering and site selection

The CNV calls resulting from boundary refinement were critically evaluated using the intensity signals from two SNP genotyping arrays: the Illumina Omni 2.5 and the Affymetrix 6.0 arrays. We estimated the false discovery rate (FDR) of the CNVs using an Intensity Rank Sum (IRS) test, as described below.

Guided by the IRS test results on the Omni 2.5 array, we establishing the following filtering criteria: site length >= 3Kb, density of uniquely alignable bases >= 25%, genotype call rate >= 80%, at least one sample called non-reference at 95% confidence. After applying these filters, we found that the estimated FDR from the IRS test using the Omni 2.5 array was 1.4% for sites over 20Kb in length. All sites with length greater than 20Kb were retained and in addition sites shorter than 20Kb were retained if they contained at least one array probe and either IRS p-value was less than 0.01.

Post-phasing site filtering

After phasing each CNV with beagle (see below), we removed an additional 647 sites from this discovery set that were monomorphic in the discovery cohort after phasing.

CNV discovery set 2

In the second discovery set, we ascertained polymorphic CNVs with multiple copies of the CNV in the reference genome. In this pipeline, rather than using seed windows, the sites are seeded using the segmental duplication annotation track from the UCSC genome browser. For each annotated segmental duplication, we prospectively genotype the segments to assess total copy number across the two segments. To determine total copy number, we applied the same genotyping model used in Discovery Set 1, but consider only sequence positions with equivalent alignability across the two segments. For this data set, we assumed two positions had equal alignability if 101bp windows centered on each position were identical to each other but distinct from all other 101-mers on the reference genome.

Filtering and site selection

Sites were initially genotyped and filtered using the default Genome STRiP parameters. Evaluation of the passing sites using the IRS test and the Omni 2.5 array indicated a small number of reliable array probes in these segmentally duplicated regions. We manually reviewed histograms of read depth (as in Figure 2) for each site and chose filtering parameters to select sites with clean histograms and good separation between the clusters. The filtering parameters used in this data set were sites with copy-number call rate > 50%, alignable length >= 1000bp, density (fraction of alignable bases) > 0.05 and cluster separation >= 3.0. Cluster separation is measured as the mean Mahalanobis distance between the copy number 1 and copy number 2 clusters.

Post-phasing site filtering

After phasing each CNV with beagle (see below), we removed an additional 49 sites from this discovery set that were monomorphic in the discovery cohort after phasing.

Discovery set merging

The final set of genotyped CNV loci was constructed as the union of the two discovery sets. No removal of redundant sites was performed as the two discovery sets provide different information about a site (discovery set 1 predicts the unique or paralog-specific copy number at a site whereas discovery set 2 predicts the total copy number across multiple copies on the reference). Some genomic loci are covered by calls from both discovery sets.

For congruity in the final call set, only genotypes from the 849 samples passing QC for the first discovery set were used in genotyping the merged set of CNV loci.

Intensity rank sum (IRS) test

The IRS test estimates a false discovery rate for a set of putative CNV calls, by utilizing the distribution of a test statistic (across all calls) derived from the relative probe-level intensities of the same probe(s) between samples expected to have different copy number levels. We drew upon (in an integrated analysis) the SNP probe intensities from two different SNP arrays: the Omni 2.5 and the Affymetrix 6.0. Data sets for both arrays were obtained from the 1000 Genomes Project web site, where they are publicly available.

The IRS test works in the following way. First, a matrix of normalized probe intensities was created for each sample and SNP probe set by summing intensity values for the A and B allele probes for each SNP (for Affymetrix 6.0 copy-number probes, the individual probe intensity was used). For each genotyped CNV site, up to two tests are performed: One test based on samples with predicted copy number less than the reference copy number (i.e. the expected copy number, generally copy number 2, for an individual that is homozygous for the allele represented by the reference genome), and one test for samples with predicted copy number greater than the reference copy number. For the first test, the samples are divided into two subsets, those with reference copy number and those with copy number less than the reference copy number (samples with copy number above this value are not used in the first test). For each probe underneath the CNV, the samples are first ranked according to the probe intensities with ties broken randomly. Then using the ranks at each probe, the samples are re-ranked across all probes, with ties broken randomly. A rank-sum test is performed to test whether the samples predicted to have copy number below the reference copy number have lower ranks than the samples with reference copy number. The second test is symmetrical to the first test, comparing the subset of samples with copy number above the reference copy number to the samples with reference copy number.

For each CNV, these tests yield either one or two p-values depending on the range of copy number genotypes at that CNV. Empirical evaluation of the null distribution of these p-values by randomizing the assignment of samples to each category shows that it is symmetrical and almost uniform (though slightly overdispersed). The false discovery rate of a set of CNVs was estimated by dividing the putative CNVs into three subsets: (a) CNVs with observed copy numbers either at or below the reference copy number (b) CNVs with observed copy numbers either at or above the reference coyp number and (c) CNVs with observed copy numbers both above and below the reference copy number. Subsets (a) and (b) have one p-value while subset (c) has two p-values. For subsets (a) and (b), we estimate the FDR of these subset as two times the fraction of sites with p-value > 0.5. For subset (c) with two p-values, we estimate the FDR of this subset as four times the fraction of sites having both p-values > 0.5. An overall FDR for the call set as a whole is calculated as the weighted sum of the FDRs of the three subsets (a-c).

An implementation of this test is available as the IntensityRankSum annotator module in the Genome STRiP software.

Phasing of copy number alleles

Generating genotype likelihoods

Prior to phasing, we convert the diploid copy number calls to genotype likelihoods for haploid copy number alleles. First, we determine the set of potential haploid copy number alleles we will emit for each site. As part of the genotyping model described previously, Genome STRiP computes a set of diploid copy-number likelihoods (CNLs) for each sample for each copy number state from zero up to a site-dependent maximum value (all copy numbers above this value have negligible likelihood in every sample). From this set of copy-number likelihoods, we estimated the frequency spectrum of the haploid copy-number alleles using an expectation maximization (EM) algorithm (see G. Abecasis, http://www.sph.umich.edu/csg/abecasis/class/666.08.pdf). The frequencies were estimated separately in each population and summed. For each site, we computed the set of haploid copy number alleles as the list of consecutive integers from the minimum to the maximum haploid copy numbers with an estimated allele frequency greater than 0.001.

Given the set of haploid copy number alleles for each site and the copy-number likelihoods (CNLs) for each sample, we computed for each sample the set of potential allele combinations summing to each potential diploid copy number and distributed the likelihood for that diploid copy number over the set of potential allele combinations. In addition, we used the population-specific estimated allele frequency for each haploid copy number as a prior on the genotype likelihoods. This improved phasing and imputation performance compared to a uniform prior where, for example, a diploid copy number of 2 would be equally likely to be 1+1 or 0+2, even if 0 was a rarely observed allele.

Phasing copy number variants

We used the beagle software (beagle4, version r1128) and the 1000 Genomes reference panel provided on the beagle web site (http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/) to phase the CNVs. For phasing and for imputation, we used only bi-allelic SNPs as flanking reference panel markers and used up to 1000 markers within 500Kb on each flank. We also excluded any markers underneath the target CNV and within 1Kb of the estimated breakpoints on each side of the target CNV to account for potential inaccuracy in the breakpoint localization. Each CNV was phased in an independent run.

For CNVs from Discovery Set 2 with multiple reference segments on the genome, we treated them differently depending on whether they were nearby (separated by less than 100Kb) or widely spaced. For CNVs where the segments were nearby, the two segments were combined into a single interval for phasing and imputation. For those CNVs that were widely spaced (separated by more than 100Kb or on different chromosomes), we attempted to discern whether most of the copy number variation was arising primarily in one paralog compared to the other. We did this by measuring the read depth at any positions containing paralog-specific variations (PSVs) that distinguish the two paralogs based on the reference genome and then comparing the correlation (Pearson's *r*) between the continuous read

depth estimate for each paralog and the total (discrete) copy number across the two paralogs. We used the interval for the paralog with the strongest correlation for phasing and imputation. For CNVs with multiple reference segments where one segments was in an unplaced reference contig, we used the interval with chromosomal coordinates in all cases.

Droplet digital PCR experiments

We used droplet digital PCR (ddPCR, Bio-Rad Laboratories) to evaluate copy number genotypes originally obtained via analysis of sequencing data from the 1000 Genomes project. We screened a subset of 96 genomic DNA samples from the YRI2 (Yoruba in Ibidan, Nigeria) cohort of the Coriell 1000 Genomes Project samples (Cat#: HAPMAPPT04) for 32 multi-allelic CNV and segmental duplication target sites (Supplementary Table 13). Primer/probe ddPCR assays for each target were custom designed by Bio-Rad Laboratories based on a set of target genomic regions ascertained from sequencing data analysis. At CNV sites in segmental duplications on the reference genome, the assays were designed to identically duplicated sequence to measure total copy number across the duplicated regions. Assays that failed quality control (see below) were not pursued. All target assays utilized FAM-labeled probes. The target assay MRGPRX1 (Assay ID: dHsaCP1000474, Bio-Rad Laboratories) was used. Reference assay targeted the TERT locus (Assay ID: dHsaCP1000100, Bio-Rad Laboratories), and was labeled in HEX.

Droplet digital PCR workflow

250 ng DNA was digested with the restriction enzyme HindIII (2.5 U/rxn, New England Biolabs, Cat#: R0104) in NEB Buffer #4 (Cat#: B7004) for 1 hour at 37°C in a 50 µl volume and then diluted 1:2 with molecular biology grade water to a final 100 µl volume. A 20 µl mixture comprised of 1) 1 µl of each 20x primer/probe mix 2) 10 µl of Bio-Rad ddPCR Supermix for Probes (no dUTP) (Bio-Rad Laboratories, Cat#: 186-3023), and 3) 8 µl digested, diluted DNA (20 ng/rxn) was emulsified with Bio-Rad Droplet Generator Oil (Cat#: 186-3005) using a Bio-Rad QX200 Droplet Generator according to the manufacturer's instructions. The droplets were then transferred to a 96 well reaction plate (Eppendorf, Cat#: 0030128) and heat-sealed with pierce-able sealing foil sheets (Bio-Rad, Cat#: 181-4040). PCR amplification was performed using a Bio-Rad C1000 thermal cycler with the following cycling parameters: 10 minutes at 95°C, 40 cycles consisting of a 30 second denaturation at 94°C and a 1 minute extension at 60°C, followed by 10 minutes at 98°C and a hold at 8°C. All steps had a ramp rate of 2 °C/s, except for the final hold at 8°C, which had a ramp rate of 1 °C/s.

Immediately following PCR amplification, droplets were analyzed using a Bio-Rad QX200 droplet reader in which droplets from each well are aspirated, streamed toward a detector and aligned single-file for two-color detection. Fluorescence data for each well were analyzed using QuantaSoft software, version 1.4. Thresholds were determined manually for each experiment. Droplet positivity was determined by fluorescence intensity; only droplets above a minimum amplitude threshold were counted as positive.

Quality control on droplet digital PCR assays

The ddPCR assays were assessed first for technical quality. Assays that did not achieve good amplification, did not form clear clusters or had excessive "rain" (droplets whose positive/negative assignment was ambiguous) were not used. Assays were further assessed as to whether they were measuring the same copy number variant ascertained from the sequencing data; a few assays required redesign due to imprecise boundary estimates from the low coverage sequencing data, the presence of paralogous sequence variants within the multiple copies or due to non-specific binding to paralogous sites elsewhere on the reference genome.

Droplet digital PCR concordance analysis

To evaluate CNV sites where the copy numbers obtained from sequencing data were in disagreement with the copy numbers obtained from array CGH in Conrad 2011, we used ddPCR assays for 10 sites (**Supplementary Table 4**) where the CNV locations obtained from sequencing data had 80% or greater reciprocal overlap with the CNV locations ascertained in the Conrad study and where there were large numbers of discordant copy numbers in the YRI2 samples. Concordance analysis was performed by comparing 95% confident copy number calls from the sequencing data to integer copy number calls from ddPCR obtained by rounding the ddPCR copy number estimate to the nearest integer.

To evaluate copy number genotyping accuracy from sequencing data, we chose CNV sites with a wide range of high copy number in the YRI2 samples. Concordance was analyzed at 21 CNV sites, using 21 ddPCR assays on 38 samples in common between the YRI2 samples and the 1000 Genomes analysis cohort. Concordance was assessed by comparing 95% confident copy number calls from the sequencing data to integer copy number calls from ddPCR obtained by rounding the ddPCR copy number estimate to the nearest integer. In cases where the ddPCR copy number estimates were not with +/- 0.4 of an integer value, they were considered no-calls for concordance analysis.

At two sites (assays m035 and CC_SD_003), we applied a linear scaling factor to the ddPCR data (but not to the sequencing data) to overcome apparent artifacts causing the ddPCR copy-number estimates to not cluster at integral values (the ddPCR data clustered, but the cluster centers seemed to be inflated linearly with copy number). The scaling factor was estimated separately at each site as the mean of the ratio between the continuous ddPCR copy-number estimate and the normalized copy-number estimate from sequencing read depth for each sample. Applying this scaling factor to other sites yielded no change in genotype concordance. Both the scaled and unscaled data are shown in **Supplementary Figure 4**.

Comparison to CNVs from aCGH (Conrad, 2011)

Site-level sensitivity

From the genotyped CNVs ascertained by aCGH, the subset was computed that were polymorphic in the overlapping set of 205 samples between the two studies. Only autosomal CNVs were considered. Reciprocal overlap was computed between the CNVs ascertained from sequencing data and CNVs from the aCGH study. For reporting sensitivity, we considered CNVs from the aCGH study that were overlapped by any amount by at least one CNV ascertained from sequencing data.

Genotype concordance

Genotype concordance with the aCGH CNVs was assessed at sites where the two CNV calls had at least 80% reciprocal overlap. Genotype concordance was computed as exact agreement between the integer copy number assessments from the two methods at these sites. Concordance was measured on 95% confident calls from sequencing data (any genotype less than 95% confident was treated as missing data) and all reported calls from the aCGH data. Nineteen sites from the aCGH data set were excluded from the analysis based on manual review (these sites suffered from diagnosable problems such as copy number shifts caused by the reference sample not being copy number two or mis-clustering that collapsed two adjacent copy number clusters in the aCGH analysis). Discrepancies at additional sites were evaluated using ddPCR (Supplementary Table 3).

Excluded aCGH CNV sites

Affy6_73	CNVR1847_full	CNVR4227.2	CNVR6769_full
CNVR116_full	CNVR2561.1	CNVR5600_full	CNVR7540.1
CNVR1539.1	CNVR3613.1	CNVR5917.1	CNVR7702.1
CNVR1574_full	CNVR3734.1	CNVR6297.12	CNVR7708.1
CNVR1815.1	CNVR3773.1	CNVR6703.6	

Genic overlap of CNVs (Table 1)

Classification of CNVs

CNVs were classified based on 95% confident copy-number genotypes. CNVs were classified as deletions if all samples had a diploid copy number that is either 0, 1 or 2. CNVs were classified as (bi-allelic) duplications if all samples had a diploid copy number that is 2 or greater and the range of observed copy-number could be explained by only two alleles. All other CNVs were classified as multi-allelic. These classifications were made independently of whether the CNV was present in one or two copies on the human genome reference.

Classification of gene overlap

Genic overlap was computed using the transcript annotations from Gencode v17. CNVs were classified as containing a gene if the entire gene model, including all exons and annotated UTRs from all isoforms were contained within the predicted CNV region (Table 1). CNVs were classified as partially overlapping a gene if they overlapped any annotated CDS sequence from any gene isoform (**Supplementary Table 7**).

Classification of CNVs by allele count / AAF was based on 95% confident genotype calls (**Supplementary Table 8**).

Differential gene dosage per individual

For each CNV containing a gene, we estimated the allele frequency spectrum using an expectation-maximization (EM) algorithm (see, e.g., G. Abecasis, http://www.sph.umich.edu/csg/abecasis/class/666.08.pdf). The allele frequencies were estimated separately in each population and combined.

In cases where multiple CNVs overlapped the same gene, we chose a single CNV and used the allele frequencies for this CNV to estimate differential gene dosage. CNVs from discovery set 1 (represented by a single segment on the reference genome and thus where copy number is paralog-specific) were used in preference to CNVs from discovery set 2. In other cases, we used the CNV that minimized the estimated number of differences per individual for that gene. For CNVs from discovery set 2 (with multiple segments on the reference genome), we assumed that the CNV affects the dosage of the genes in both reference segments.

When analyzing genes partially overlapped by CNVs (**Supplementary Tables 6 and 7**), the same procedure was used except that all CNVs overlapping the gene (fully or partially) were considered and as above the CNV with the minimum contribution to differential gene dosage was assigned to each gene to achieve a conservative estimate.

Analysis of sensitivity to genotyping error

Although our validation analyses indicate our genotyping error rate is low, we sought to evaluate the potential effect of any residual genotyping error on the analysis of per-individual differences in gene dosage. We performed a second analysis where we removed any copy-number states that were not observed in at least 2 individuals (**Supplementary Table 8**). The conclusions were not significantly affected.

In addition, the level of per-individual gene dosage variation is strongly determined by the highest-frequency mCNVs (**Supplementary Figure 6**). This suggests that the magnitude of gene-dosage variation will have limited sensitivity to modest genotyping error (for example, incorrectly assigning high copy numbers to be one more or less than the true copy number), as most of the high-frequency mCNV sites are strongly multi-allelic and exhibit a wide range of copy number states.

Gene ontology category enrichment

GO category enrichment was evaluated using CNVs that fully overlapped a gene model from Gencode v17. The gene symbols for each gene were used to test for enrichment for different gene ontology categories using the Amigo2 term enrichment service (http://amigo2.berkeleybop.org/amigo). We tested separately the set of all genic CNVs and the subsets that were bi-allelic and the subsets that were multi-allelic.

GO category enrichment was also evaluated using the GOrilla web tool (http://cbl-gorilla.cs.technion.ac.il/) which was used to produce the graphical summaries plotted in Supplementary Fig. 7. GOrilla and Amigo2 generated substantially similar results, although GOrilla does not automatically calculate term depletion. Using inverse gene lists (all autosomal protein coding genes from Gencode v17 not overlapped by CNVs) showed slight enrichment for metabolic processes, consistent with results from Amigo2 (data not shown).

Effect of gene dosage on gene expression

RNA sequencing data was downloaded from the Geuvadis project web site at ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/.

For CNVs that fully overlapped genes, we computed the Pearson's correlation coefficient between the CNV integer copy number and the normalized gene expression quantitation from the RNA sequencing data (GD462.GeneQuantRPKM.50FN.samplename.resk10.norm.txt.gz).

P-values for gene expression were calculated using 10,000 random permutations and a one-sided test with Pearson's correlation coefficient as the test statistic. In each trial, we randomly permuted the mapping between genes and CNVs. To control for potential interactions between genes and nearby but not overlapping CNVs, we generated permutations where genes were always assigned to CNVs on a different chromosome.

Intersection between eQTLs and CNV proxy SNPs

To evaluate the potential for CNVs to influence gene expression indirectly, for example through regulatory elements, we searched for candidate interactions using the intersection between CNV tag SNPs and eQTLs found in Lappalainen, 2013. The best proxy SNP (or SNPs, when there were equally good proxies) for each CNV was intersected with all gene expression eQTLs (FDR 5%) for each gene. Multiple intersections were ranked based eQTL p-value (and then distance). Distance was computed for each intersection based on the minimum separation between the CNV and the gene (a distance of zero implies full or partial overlap between the CNV and gene). Intersections were computed separately in the EUR and YRI populations.

The list of intersections was filtered by requiring that the eQTL corresponding to the CNV proxy SNP had an eQTL p-value (measured as –log10(p)) that was at least 70% of the best eQTL for the gene (measured as –log10(p)) and requiring that the direct correlation between the CNV dosage and gene expression had a p-value (calculated by permutation testing) $p < 0.01$.

Due to effects of local LD and potential redundancy in the CNV calls, the resulting list of intersections was manually annotated to group them into distinct genomic loci. Each locus was annotated as to whether there was at least one case of either full or partial overlap between a CNV and gene.

For RP11-480A16.1, a lincRNA that showed strong correlation between CNV and gene expression, despite a separation of 230Kb, we looked for possible additional explanations for the association. Using blat, we observed two potential paralogs for this lincRNA (both with < 2% divergence) occurring within the CNV. Neither locus is annotated as a transcribed gene.

Imputation

Imputation utilized the allelic copy-number likelihoods calculated for each sample as described above. Imputation was carried out using the beagle software (beagle4, version r1128) and the 1000 Genomes reference panel provided on the beagle web site (http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/).

Leave-out trials

To assess imputation accuracy, we performed a series of trials in which we withheld the CNV copy-number genotypes for 10 samples at a time and then used the flanking SNPs and the CNV copy-number genotypes for the remaining samples to impute the copy number alleles for the 10 withheld samples. At each site, this was repeated for distinct sets of 10 samples (without regard to genotype) to collectively impute all 849 samples (the last trial withheld only 9 samples). We compared the set of imputed CNV alleles in these trials to the original phased copy-number genotypes using multiple metrics, including phased haplotype concordance, unphased genotype concordance and diploid copy number agreement. For diploid copy number agreement, we measured both concordance and dosage correlation using Pearson's r. In our analysis, we report Pearson's r^2 between the called and imputed diploid integer copy number as the imputed dosage r^2.

Taggability of CNVs by SNPs

As context for evaluating the potential imputability of the CNVs in our study, we also measured how well tagged each CNV is by individual flanking SNPs. We evaluated SNPs from 1000 Genomes Phase 1 in the same samples that were within 100Kb of each CNV, not using any SNPs that fell underneath the estimated CNV boundaries. For each SNP, we calculated the Pearson's correlation coefficient between the SNP genotype (encoded as 0,1,2) and the unrounded copy number estimate for that CNV in the same samples. We used

the unrounded copy number (as opposed to discrete integer copy numbers) to provide a more direct comparison to the information available to imputation, which consists of genotype likelihoods for each allelic copy-number combination.

To evaluate the significance of the best tag SNP r^2 for each CNV, we calculated a p-value using 1000 random trials for each CNV. In each trial, we randomly permuted the CNV genotypes and then recomputed the maximum r^2 value across the same set of eligible tag SNPs at the CNV locus. We set the significance threshold for these tag SNP r^2 values at 10^{-3}. (In practice, almost all r^2 values greater than 0.1 and involving common (maf>0.10) variants were confirmed as significant.)

Candidate dispersed duplications by long-range LD

All CNVs with a duplication allele count (the number of individuals with a duplication allele) of at least 5 were evaluated for genome-wide association to SNPs using the plink software (version 1.90b2i). We tested for association using a linear regression of the CNV dosage as a quantitative trait against the 1000 Genomes SNPs and using as covariates the first 10 principle components (to control for population stratification) and sample gender. CNV dosage was computed as the discrete integer total copy number for each sample, with genotypes below 95% confidence set to missing.

The plink association results were post-processed to remove all associations with $p < 10^{-6}$. To find consistent haplotype association signals (as opposed to isolated SNPs), we ranked all 100Kb windows across the genome by the sum of the –log10(p) values from any associated SNPs and the top scoring window was considered the best genome-wide association locus. We performed a similar analysis using a 1Mb window around the CNV locus itself (or loci, for CNVs with multiple reference segments). In both cases, SNPs underneath the CNV itself were excluded. CNVs were considered to be dispersed duplication candidates if they met the following criteria: (a) the best genome-wide association window was further than 1Mb from the CNV (or on a different chromosome) and (b) the best single SNP in the best genome-wide association window had –log10(p) > 100 or the best single SNP in the best genome-wide association window had –log10(p) > 20 and this value was greater than 3 times the value for the best local SNP and (c) at least 15 individuals were carriers of duplication alleles (to increase the power of the association test).

When we observed very strong long-range LD, we considered a site as a candidate dispersed duplication even when there was also evidence for strong local LD. It should be noted that inaccuracies in CNV boundary determination (for example, dividing a CNV region into multiple segments that are highly correlated by not identical in different individuals) could be one potential cause for highly associated local SNPs even for a truly dispersed duplication allele.