Supplementary Material for
EnsembleCNV: An ensemble machine learning algorithm to identify and genotype copy number variation using SNP array data

Zhongyang Zhang,† Haoxiang Cheng,† Xiumei Hong, Antonio F. Di Narzo, Oscar Franzen, Shouneng Peng, Arno Ruusalepp, Jason C. Kovacic, Johan LM Bjorkegren, Xiaobin Wang, and Ke Hao*

† These authors contributed equally to this work.
* To whom correspondence should be addressed. Tel: +1 (212) 659-8534; Fax: +1 (646) 537-8660; Email: ke.hao@mssm.edu

Supplementary Results

Supplementary Methods

Supplementary Figures

Supplementary Tables

Supplementary Results

Batch effect in FA dataset

While we did not identify batch effects in the STARNET dataset, we did find clear evidence that there were batches in the FA dataset. We tried to alleviate batch effects in CNV calling at the initial data processing step with Genome Studio (Figure 1). In the FA dataset, we identified 5 batches by two types of signals. First, before any CNV analysis was performed, we extracted LRR values at 100,000 randomly drawn probes across all samples and applied PCA to the LRR matrix. This visualization revealed 3 major clusters along with a few outliers (Figure S2A). Second, while generating CNV calls, PennCNV, QuantiSNP and iPattern also provided summary statistics of CNV results at the sample-level, such as standard deviations (SD) of LRR, SD of BAF, wave factor in LRR, BAF drift, and the number of CNVs detected per sample, which are highly correlated among themselves and between methods (Figure S1). We also used PCA to visualize this information (Figure S2C-E). The 3 major clusters from PCA of LRR were confirmed as three batches and one of the batches further consisted of three sub-batches. Therefore, we identified 5 batches, 3 of which were enriched for sample plates used for grouping and preparation of DNA samples (Figure S2C-E). Except for the largest batch, we re-processed the other four smaller batches with Genome Studio (see Methods; Figure 1) and re-did the CNV calling using the three methods. After these processing steps, the clusters in the original PCA plots of LRR and summary statistics were eliminated (Figure S2B and F). Importantly, the variability in sample-wise QC statistics stratified by the 5 batches was dramatically reduced (Figure S3).

Supplementary Methods

Remarks on fitting likelihood model for CNV re-genotyping

Before we fitted the local likelihood model for each CNVR, we selected CNVRs with common frequency (e.g. 10%~30%) and adequate sample size (e.g. ≥ 20) for each CN status based on the results from steps (a) and (c) of ensembleCNV. We used LRR and BAF signals of initial CNV calls in these selected CNVRs to estimate CNVR-specific parameters and take their average weighted by associated sample sizes to obtain a global estimate of the parameters.

To fit the GMM of LRR data, when the number of samples for each CN status was adequate (e.g., ≥1%) relative to the total sample size, we used the mean (with respect to location of the normal CN mode) and standard deviation of the LRR values from these samples for each CN status as the initial values of the EM algorithm; otherwise, we drew adequate random LRR values with the estimated global parameters (see above) as auxiliary samples for the CN statuses lack of adequate samples to estimate parameter values. In practice, the augmented data with random samples was able to improve the stability of the GMM fitting for CNVRs with rare variant alleles.

For some common copy number polymorphisms (CNPs), the normal CN component (CN = 2) in GMM may not be dominant and the initial CNV calls from the three selected methods may misclassify the actual CN status. As shown in Figure S6, the large proportion of homozygous deletions (CN = 0) implies that a large component of heterozygous deletions, which is misclassified as normal component, is expected. In such cases, we re-assigned CNs to the components identified by the GMM such that the proportions of different CN statuses are compatible. For unrelated samples from the same genetic population, the CNV genotype should approximately follow Hardy-Weinberg equilibrium (HWE).

Supplementary Figures

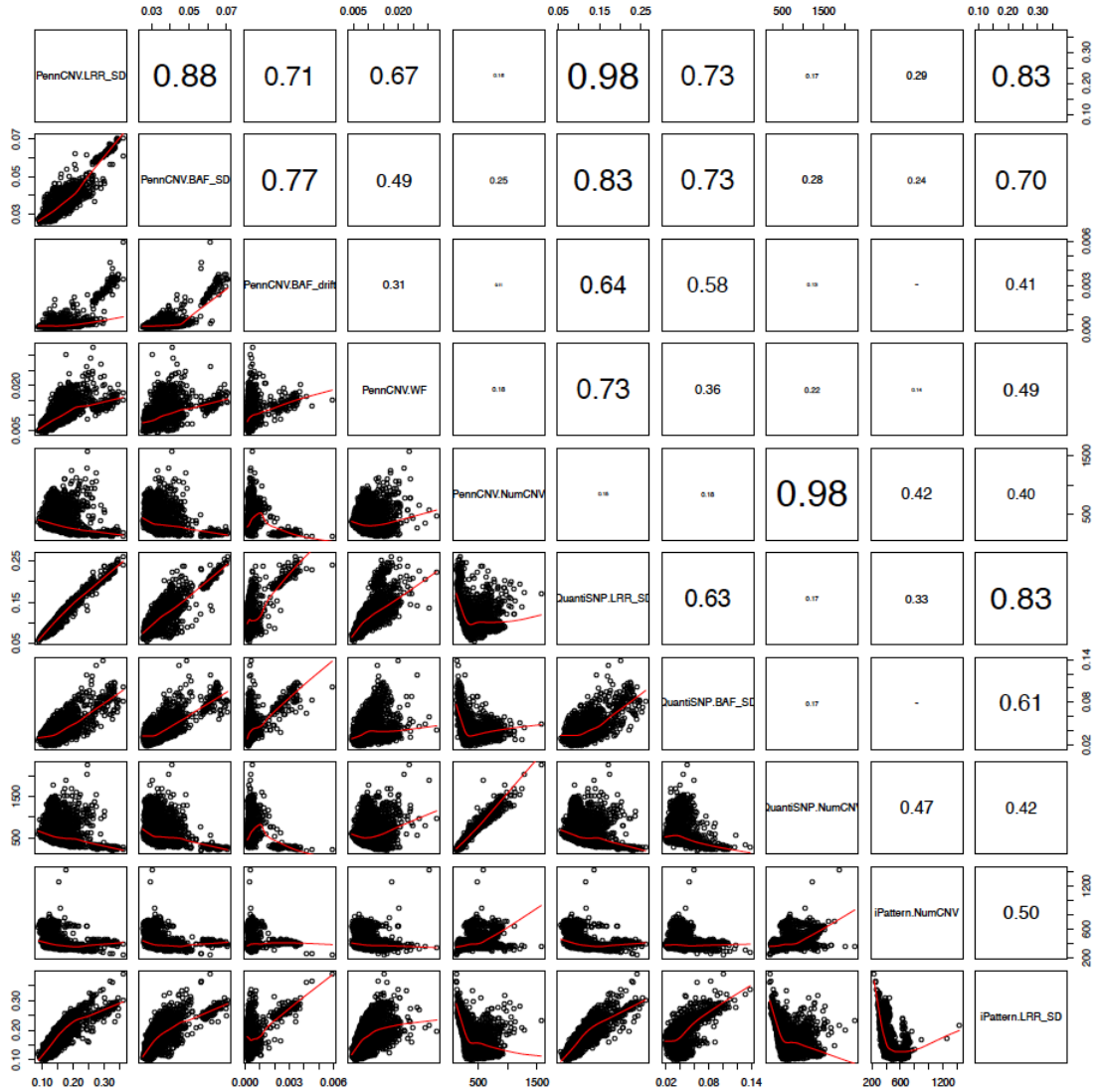

Figure S1: QC statistics of CNV results from PennCNV, QuantiSNP and iPattern on the FA dataset

The lower left half shows the pair-wise scatter plots of the 10 statistics, indicated in the diagonal; the upper right half shows the pair-wise correlations of the 10 statistics, with the font size being proportional to the Pearson correlation coefficient.

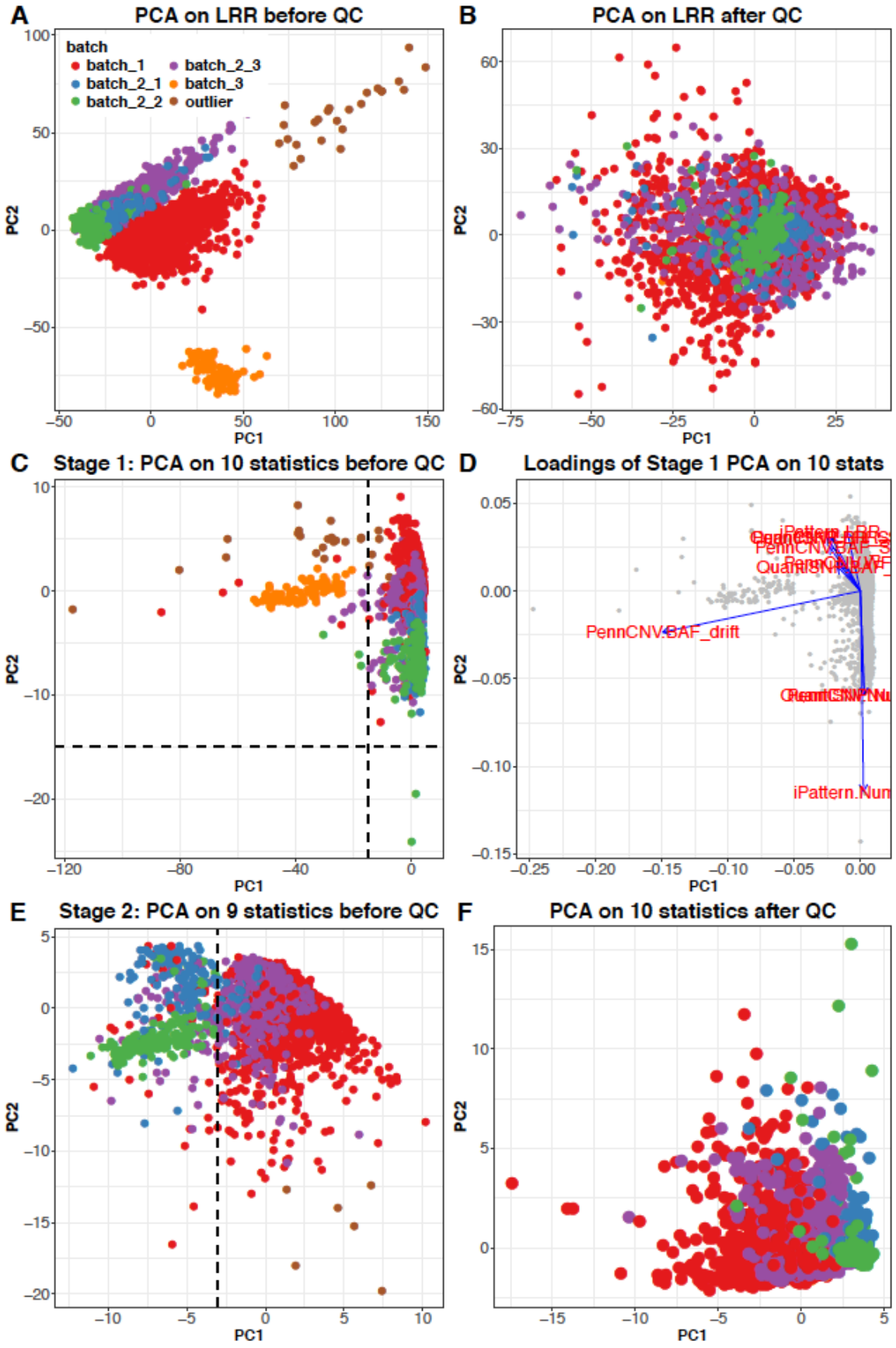
A
PCA on LRR before QC
batch
batch_1
batch_2_1
batch_2_2
batch_2_3
batch_3
outlier
PC1
PC2
B
PCA on LRR after QC
C
Stage 1: PCA on 10 statistics before QC
D
Loadings of Stage 1 PCA on 10 stats
PennCNV.BAF_drift
iPattern.Num
E
Stage 2: PCA on 9 statistics before QC
F
PCA on 10 statistics after QC

Figure S2: Batch effects in the FA dataset

PCA results on the LRR values of 100,000 randomly selected probes are displayed on the first two PCs (**A**) before and (**B**) after sample-level QC. Three major clusters and several outliers are clearly identified before QC. After QC, the batch effect is eliminated and the clusters are well mixed together. (**C**) The first round of PCA on the 10 sample-level QC statistics. A major cluster (batch_3) is clearly seen, and we further identified that samples in batch_3 were enriched in one sample plate (XW-FA-P04). (**D**) The load of the 10 QC statistics on the first two PCs. It is clear that BAF drift reported by PennCNV dominates the first PC and separates batch_3 from the remaining samples. (**E**) After excluding the BAF drift statistic and batch_3, the second round of PCA on the remaining 9 statistics and samples was performed. Three sub-clusters from batch_2 are clearly seen, and we further identified that samples in batch_2_1 and batch_2_2 were each enriched in two sample plates, respectively (batch_2_1 in XW_FA_P30 and XW_FA_P31; batch_2_2 in XW_FA_P32 and XW_FA_P33). (**F**) After sample-level QC, PCA on the 10 statistics was performed. All five batches are well mixed together, consistent with the results shown in (**B**).

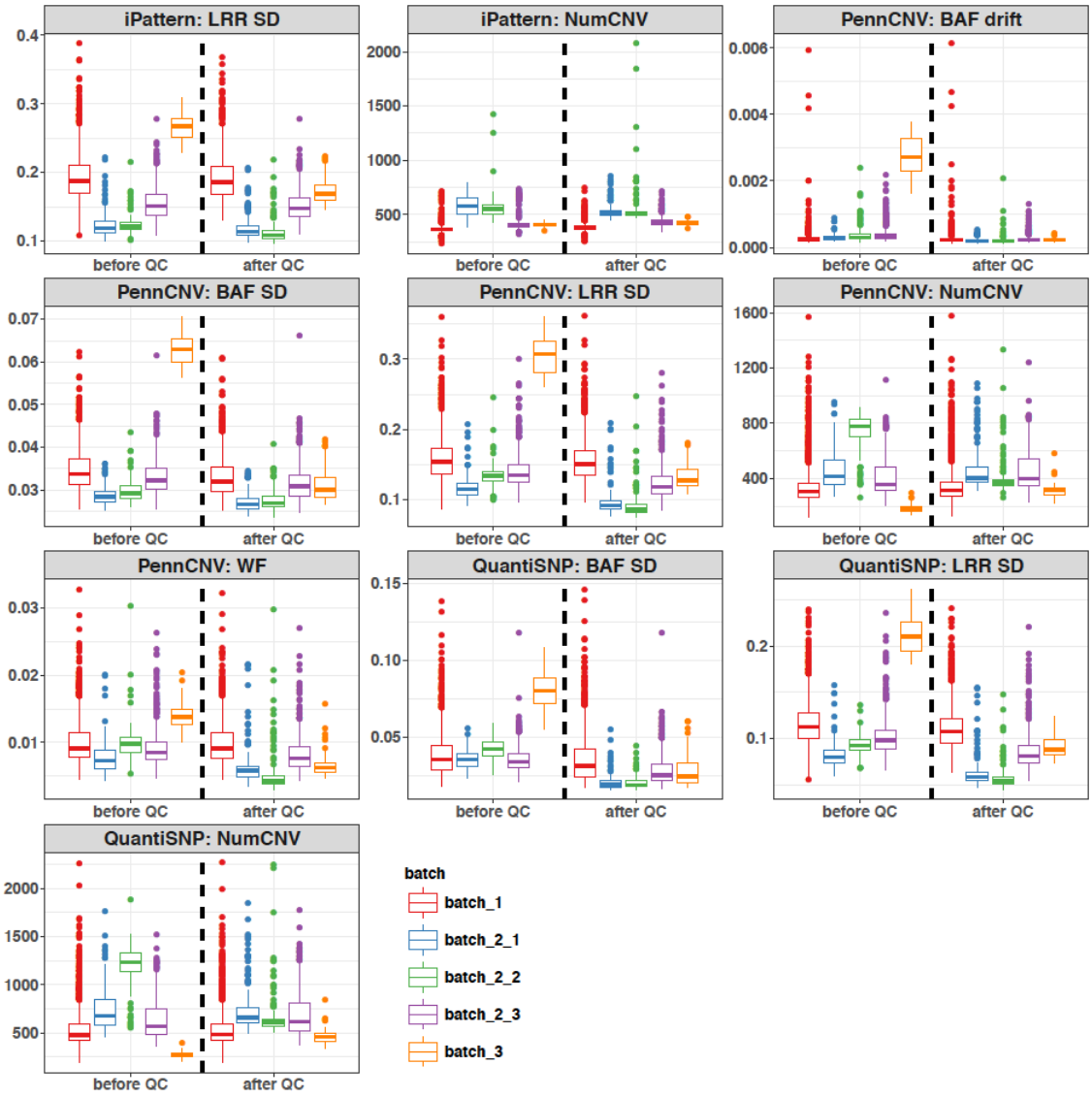

Figure S3: Sample-level QC statistics before and after the removal of batch effects

Each of the 10 panels shows the boxplots of each of the 10 statistics (Figure S1) grouped by the identified five batches (Figure S2) before and after the removal of batch effects (Figure S2). The variability in the CNV QC statistics between batches was largely alleviated after the sample-level QC.

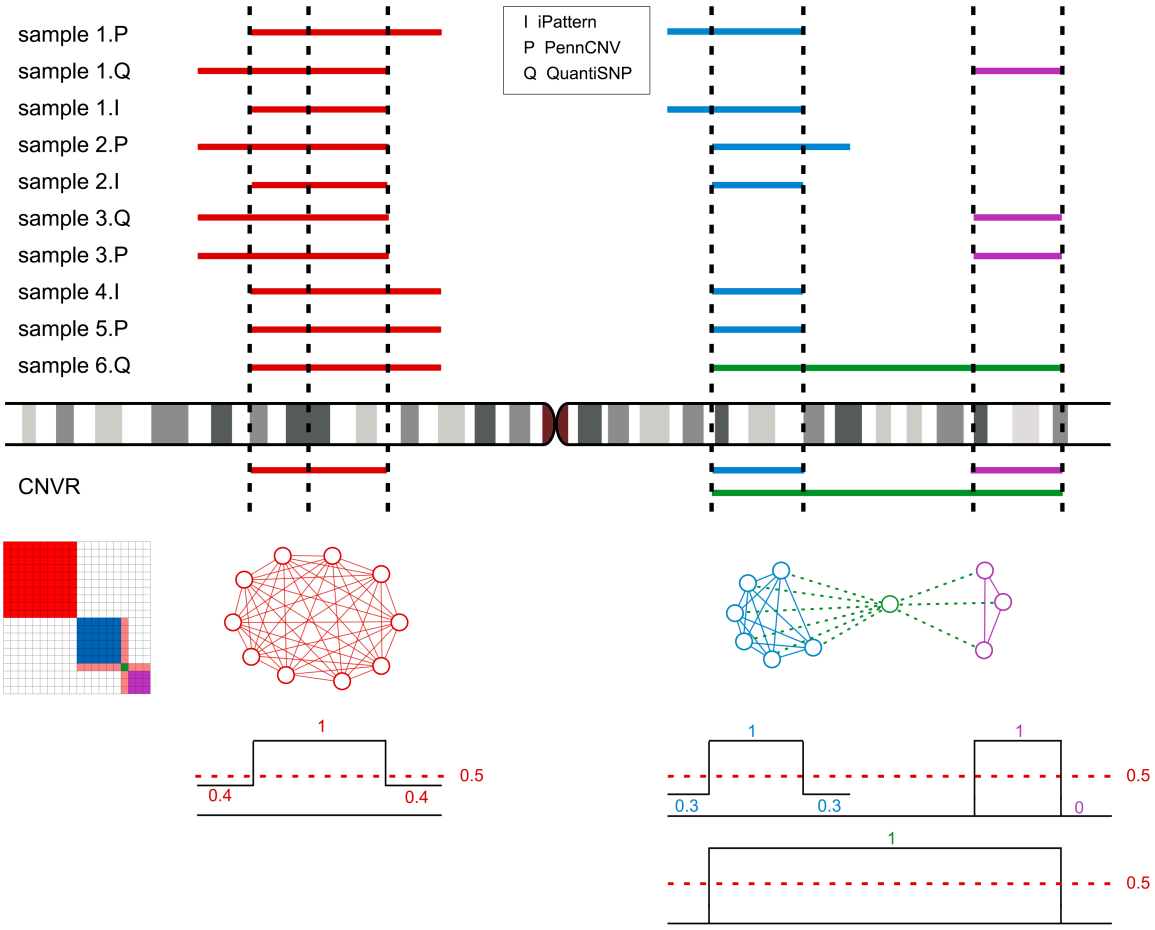

Figure S4: Intuitive diagram of the CNVR construction algorithm

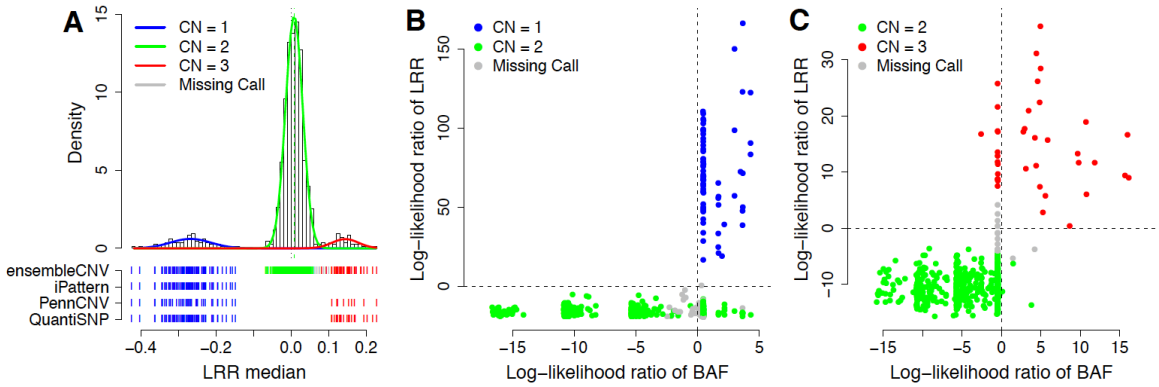

Figure S5: Example of CNV re-genotyping with local likelihood model fitted for each CNVR

(**A**) A histogram of LRR medians across probes within the CNVR is shown in the upper panel, superimposed by fitted GMM; the dotted black line indicates the location 0; the green dashed line indicates the median location of normal (CN = 2) peak. The CNVs called by iPattern, PennCNV, and QuantiSNP as well as the CNV genotype (including "no call") by ensembleCNV are marked in the lower panel. The log-likelihood ratio in LRR and BAF is shown for (**B**) CN=1 vs. CN=2 and (**C**) CN=3 vs. CN=2.

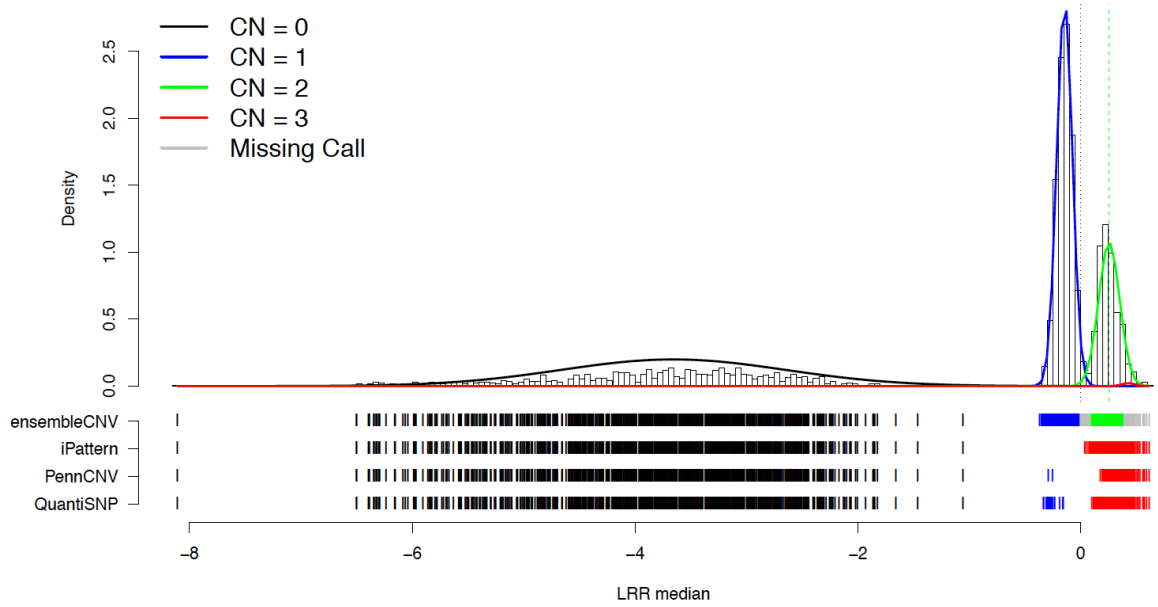

Figure S6: Example CNVR with deletion polymorphism

The histogram of LRR medians across 15 probes within the CNVR is shown in the upper panel, superimposed by fitted GMM; the dotted black line indicates the location 0; the green dashed line indicates the median location of the normal (CN = 2) peak. The CNVs called by iPattern, PennCNV, and QuantiSNP as well as the CNV genotype (including "no call") by ensembleCNV are marked in the lower panel. The initial duplication calls (CN = 3) from iPattern, PennCNV and QuantiSNP were misclassified and were adjusted by ensembleCNV to normal CN (CN = 2) in the re-genotyping step.

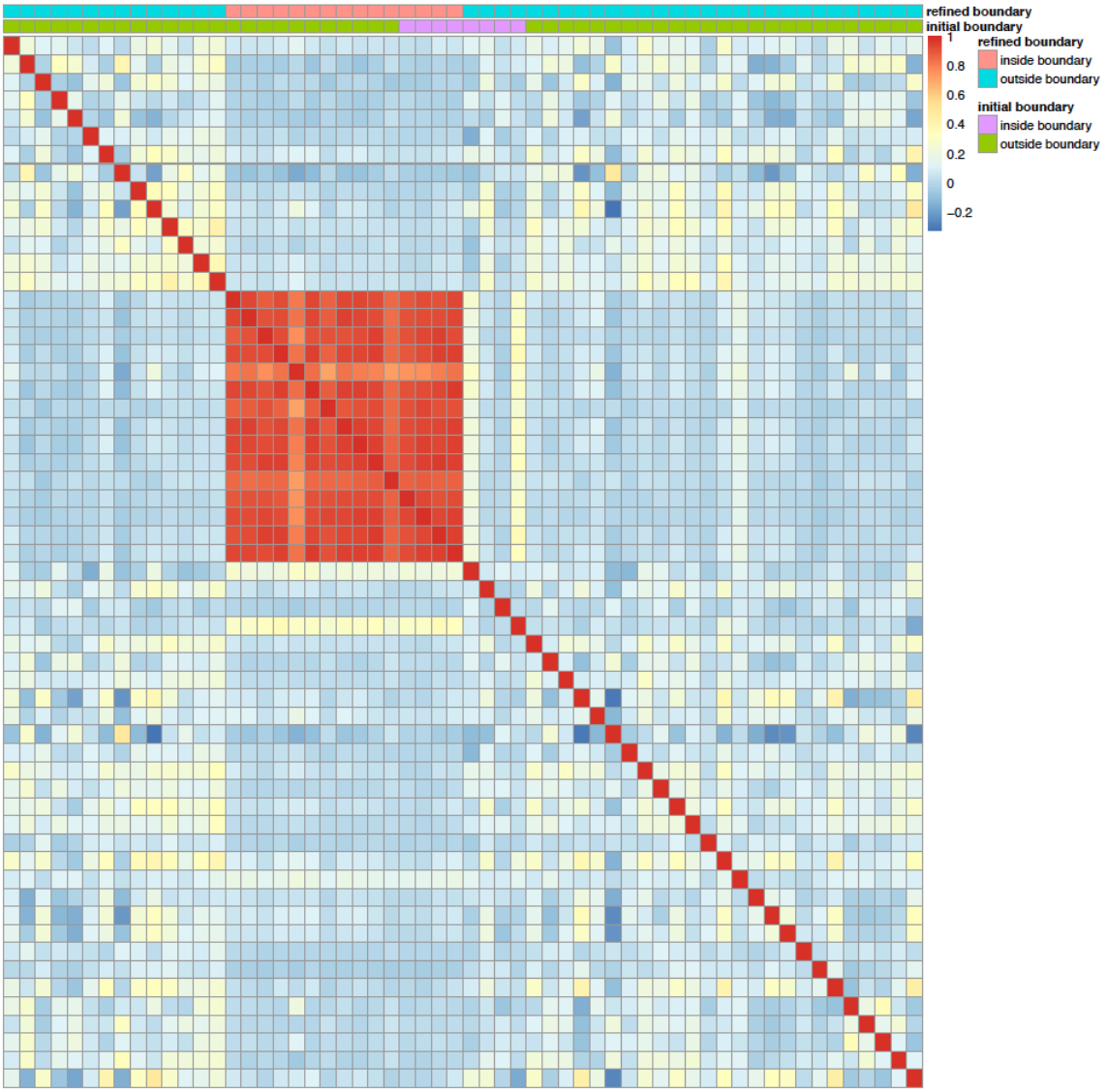

Figure S7: Example of CNVR boundary refinement

The pair-wise correlation matrix of probes within the initial CNVR boundaries and nearby regions is visualized by heatmap. The highly correlated block highlights the actual boundaries of the CNVR. The bars above the heatmap indicate the initial boundaries and the refined boundaries.

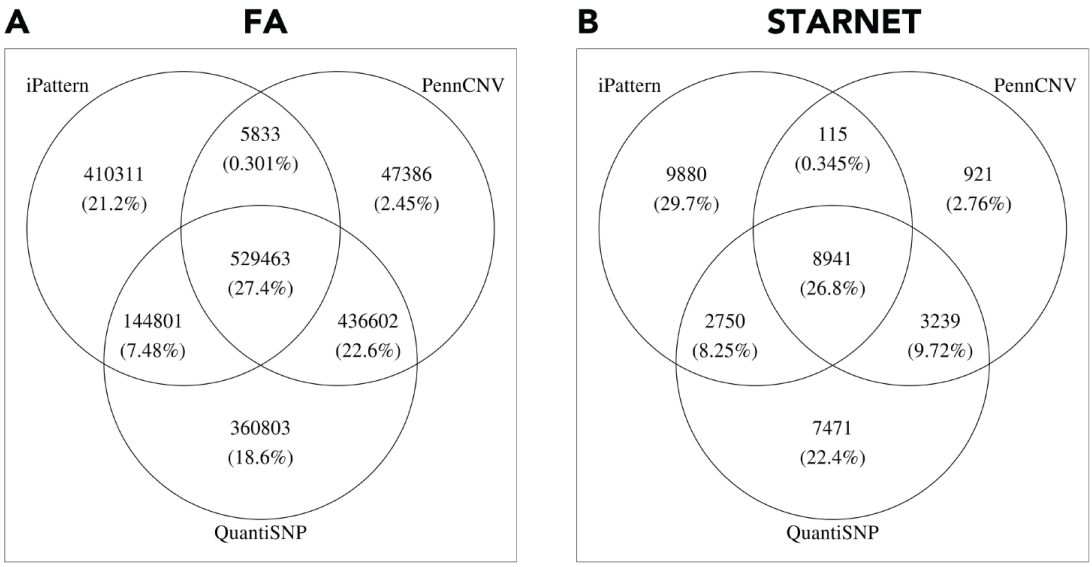

Figure S8: Venn diagram of the number of CNVs detected by iPattern, PennCNV, QuantiSNP as well as the "intersection" and "union" integration methods

(**A**) FA dataset; (**B**) STARNET dataset. For the "intersection" method, the number of CNVs detected is the sum of numbers shown in the regions covered by at least two of the three individual methods; for the "union" method, the number of CNVs detected is the sum of numbers in all the regions covered by any of the three individual methods.

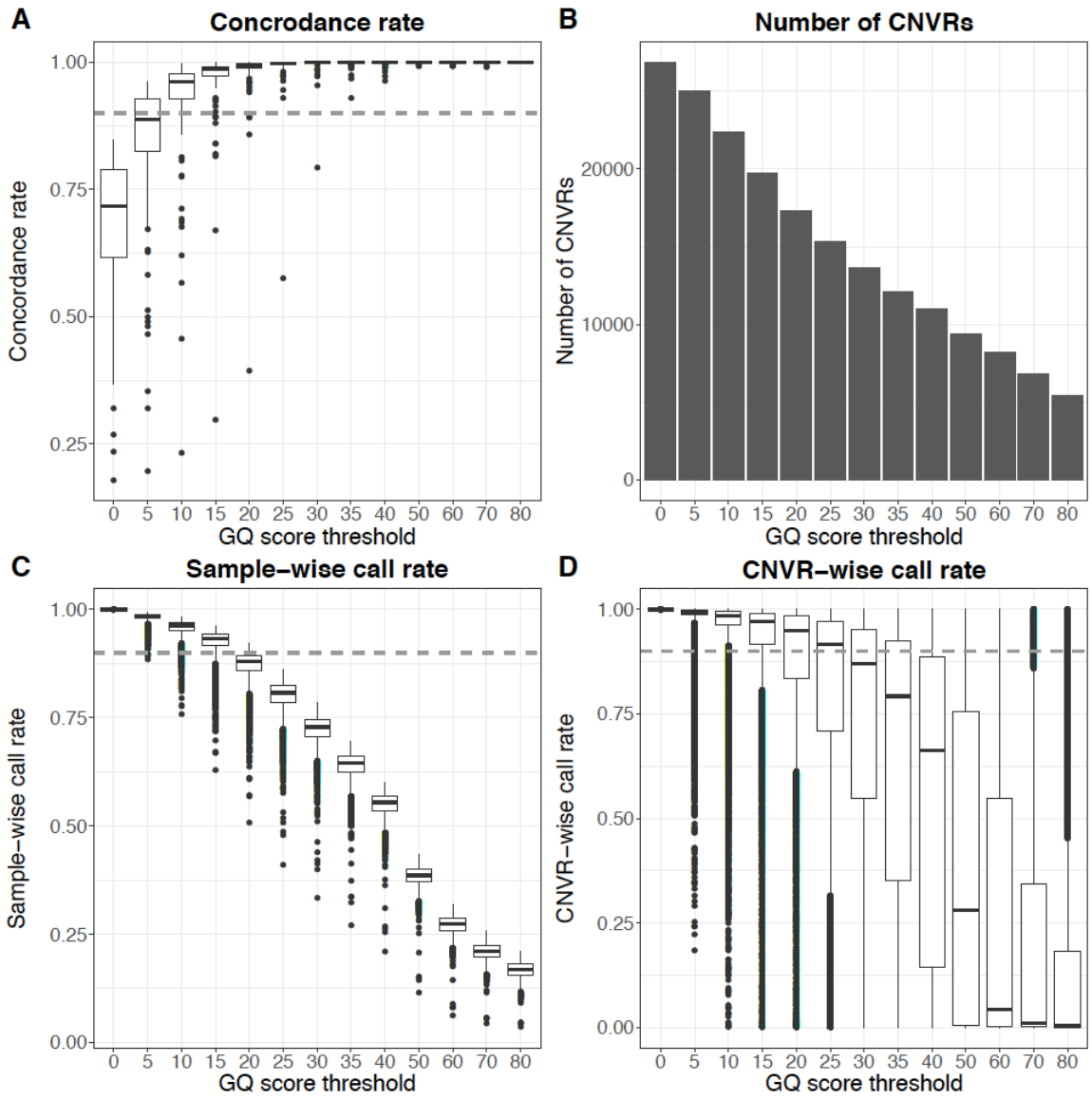

Figure S9: Results from ensembleCNV at different GQ score thresholds in the FA dataset

As the GQ score threshold increases, shown is the trend of change in (**A**) the concordance rate between duplicated pairs, (**B**) the total number of CNVRs, (**C**) the sample-wise call rate and (**D**) the CNVR-wise call rate.

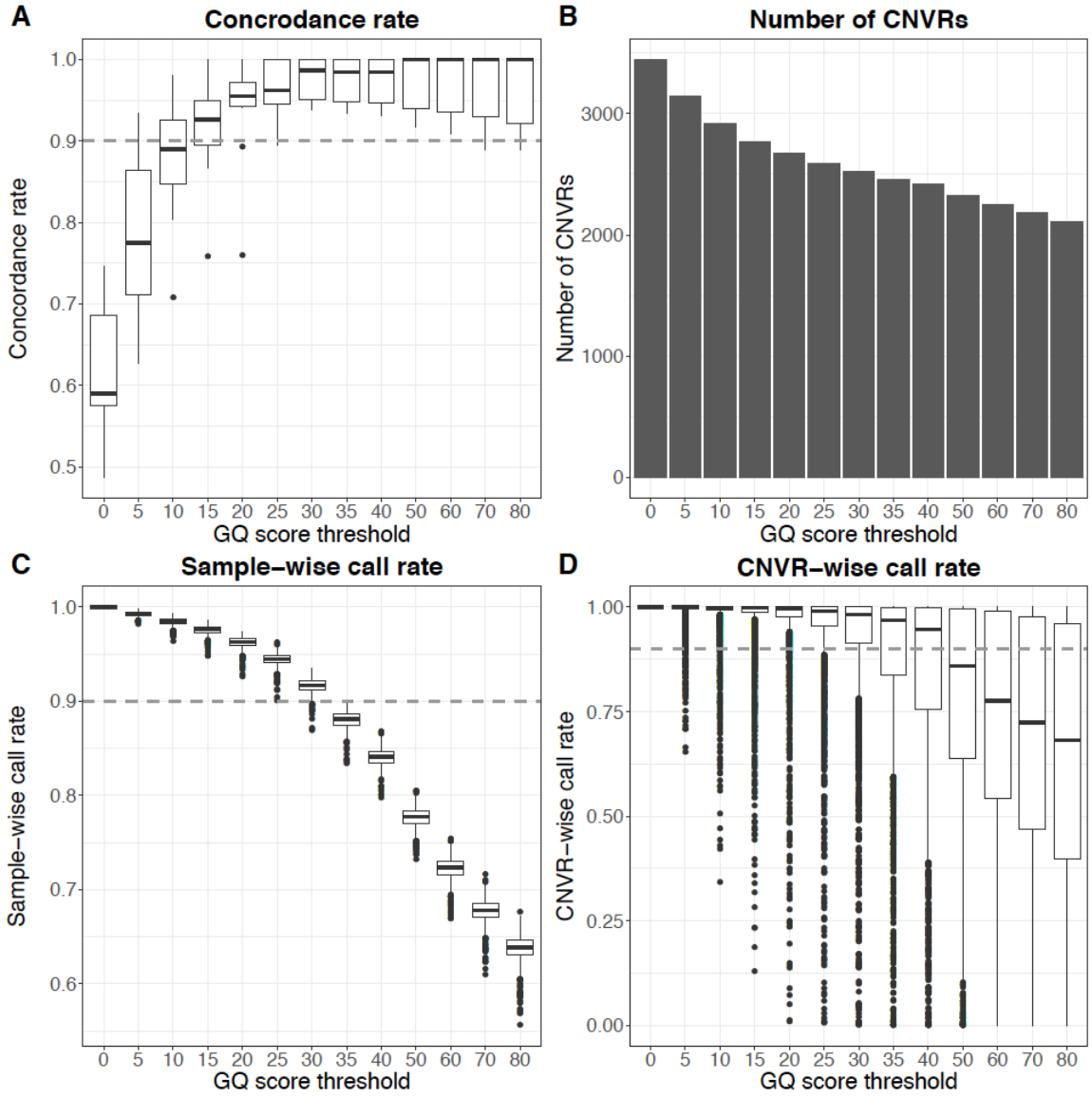

Figure S10: Results from ensembleCNV at different GQ score thresholds in the STARNET dataset

As the GQ score threshold increases, shown is the trend of change in (**A**) the concordance rate between duplicated pairs, (**B**) the total number of CNVRs, (**C**) the sample-wise call rate and (**D**) the CNVR-wise call rate.

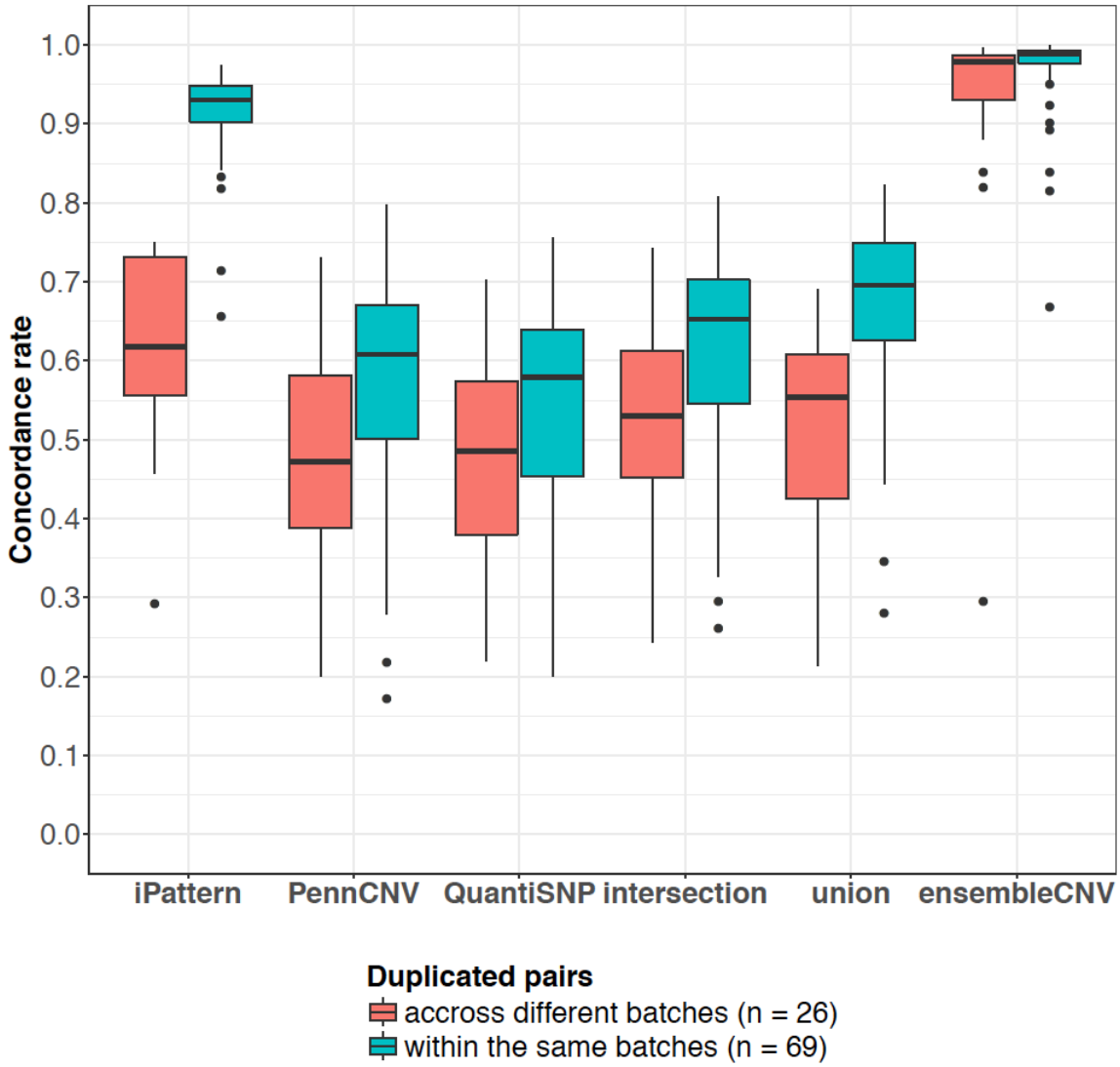

Figure S11: Concordance rate in technical duplicates within the same batches vs. across different batches of the FA dataset

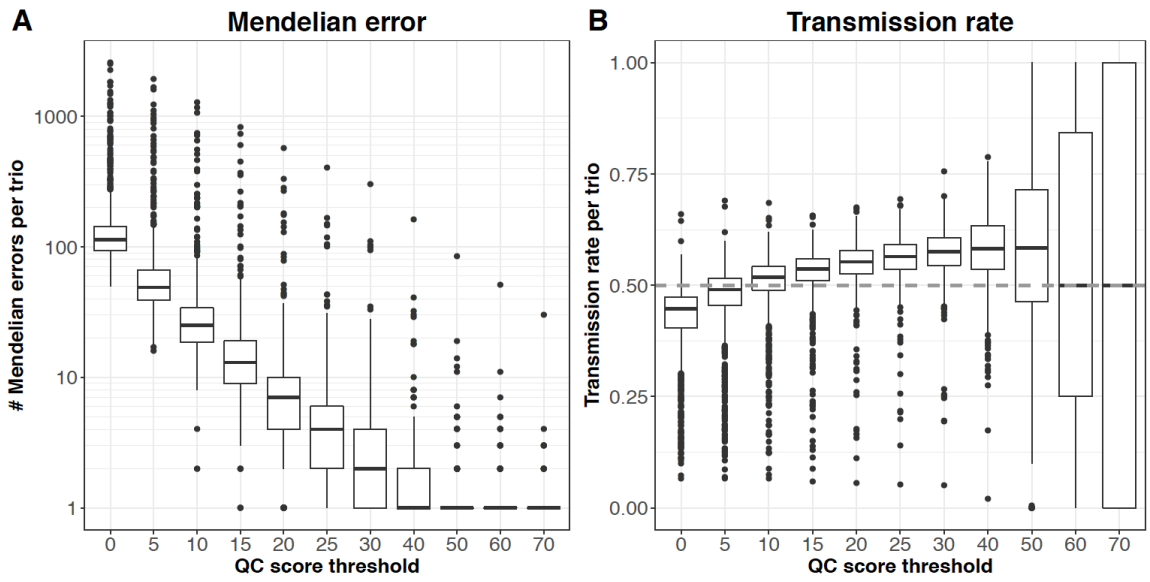

Figure S12: Mendelian error and transmission rate at different GQ score thresholds in FA trios

As the GQ score threshold increases, shown is the trend of change in (**A**) the number of Mendelian errors per trio and (**B**) the estimated transmission rate in each trio.

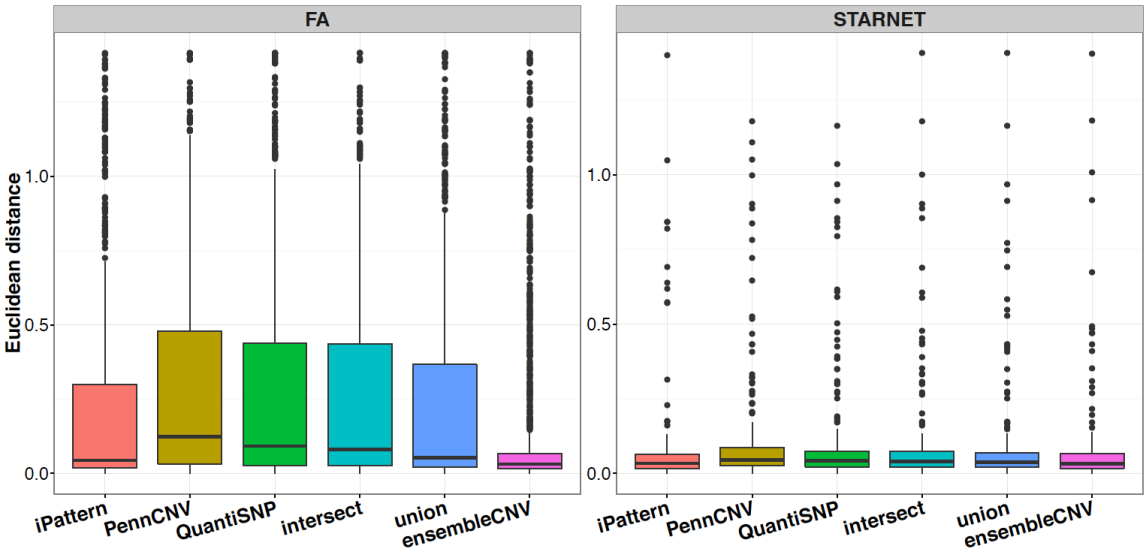

Figure S13: Comparison of CN genotype frequencies between detected CNVs and KGP CNVs

For each of the common CNVs (frequency > 1%) documented in the KGP dataset and also detected in the FA or STARNET datasets, the frequencies of CN = 0, 1, 2, 3 genotypes were compared in terms of their Euclidean distance. For the CNVs in the KGP dataset, the CN > 3 genotypes were collapsed with CN = 3, and the data from European populations were used for calculation. The CN genotype frequencies in the FA and STARNET datasets were calculated based on the unrelated individuals – the data from parents and all individuals were used for the FA and STARNET studies, respectively; from each duplicated pair, only one sample was used for calculation.

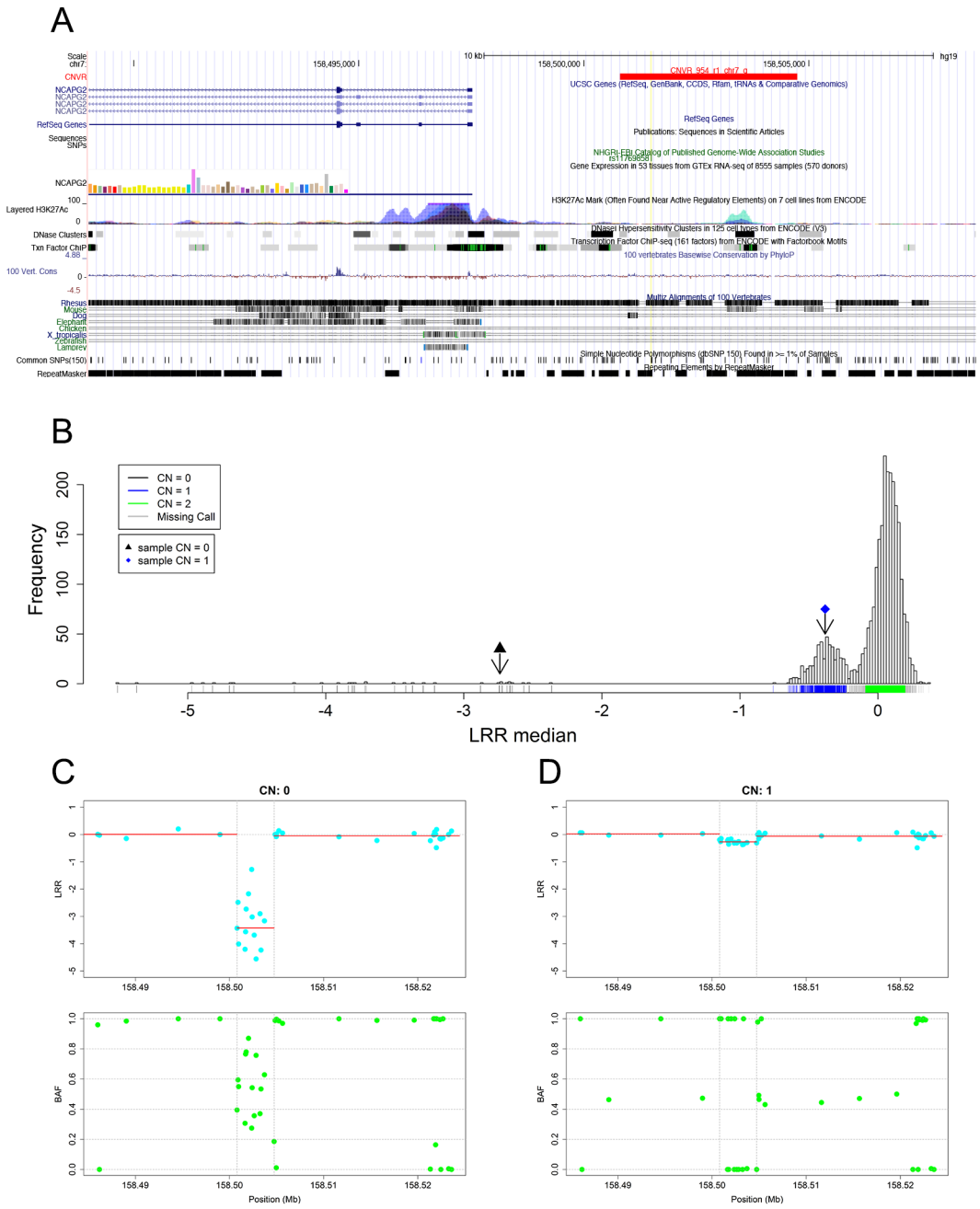

Figure S14: A CNVR detected in the FA dataset at 7q36.3 locus

(**A**) The CNVR_954_r1_chr7_q detected in the FA dataset is located upstream of the gene encoding NCAPG2. The CNVR is indicated as the red bar in the UCSC genome browser along with several epigenetic and genomic annotations. (**B**) Histogram of median LRR values of the 15 probes within the CNVR. Examples of LRR and BAF signals around the CNVR are displayed for (**C**) CN = 0 and (**D**) CN = 1.

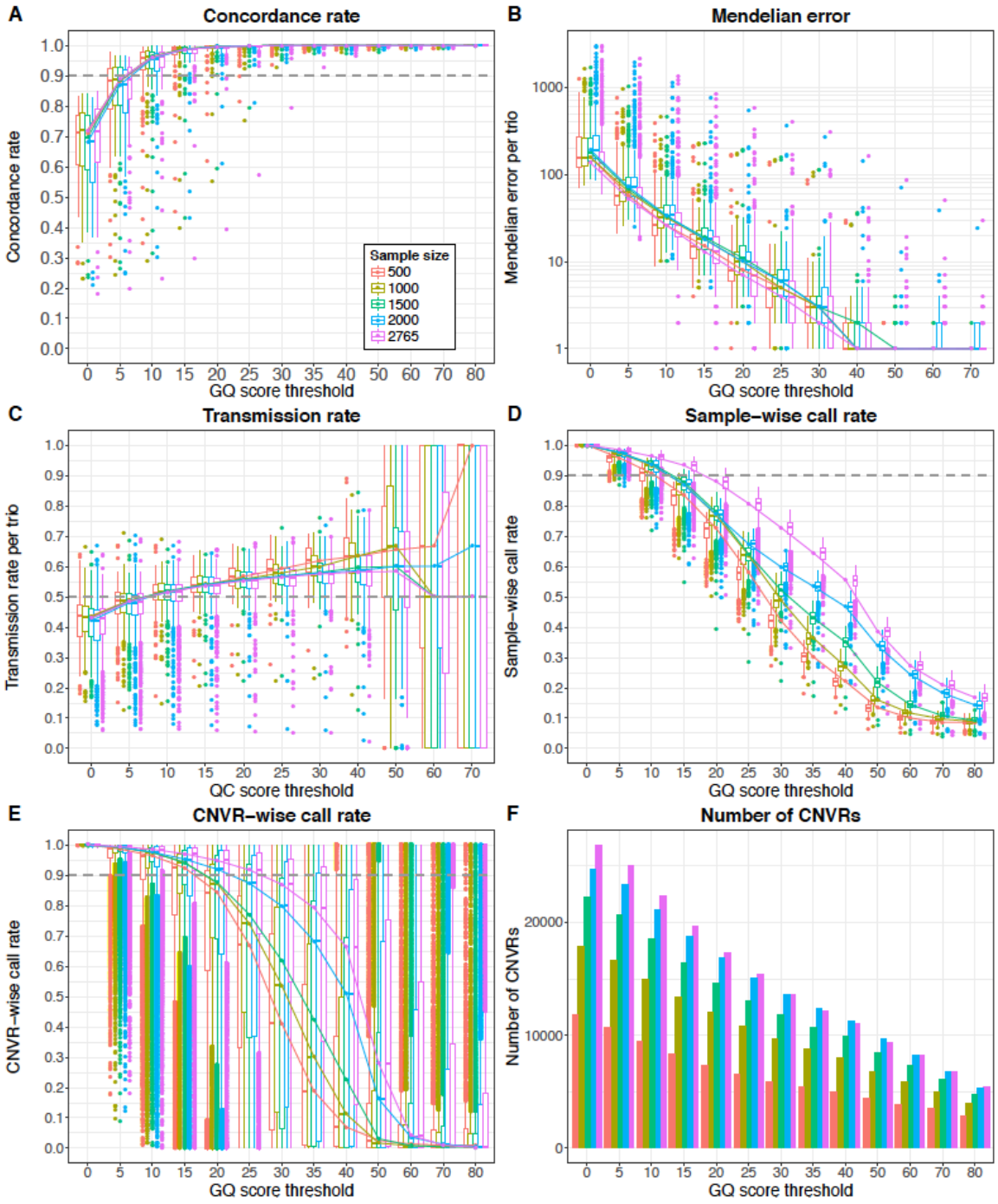

Figure S15. EnsembleCNV results at different sample sizes drawn from the FA dataset

Subsets of 500, 1000, 1500 and 2000 samples from the full dateset (n = 2765) were generated by keeping the duplicated pairs and randomly drawing complete trios. As the GQ score threshold increases, shown is the trend of change in (**A**) the concordance rate between duplicated pairs, (**B**) the Mendelian error in trios, (**C**) the transmission rate in trios; (**D**) the sample-wise call rate, (**E**) the CNVR-wise call rate, and (**F**) the total number of CNVRs. The y-axis in (B) is on log10 scale.

Supplementary Tables

Table S1: Summary statistics of the FA and STARNET datasets

	FA	STARNET
Platform	HumanOmni1-Quad-v1-0_B	HumanOmniExpressExome-8-v1-0-B
# probes	1140419	951117
# SNP probes	1048713	951117
# CNV probes	91706	0
Mean distance b/w probes (kb)	2.63	3.23
Median distance b/w probes (kb)	1	1
# samples (before/after QC)	2790/2765	874/834
European	2386 (85.5%)/2365 (85.5%)	874 (100%)/834 (100%)
African American	46 (1.6%)/46 (1.7%)	-
East Asian	58 (2.1%)/58 (2.1%)	-
Other	300 (10.8%)/296 (10.7%)	-
# duplicated pairs (before/after QC)	100/95	12/12
# families (before/after QC)	839/835	N/A

Table S2: Summary statistics of the 1000 Genomes Project (KGP) CNV dataset

		KGP[a]	KGP detectable in FA[b]	KGP detectable in STARNET[c]
# CNVRs		49929	6456	3571
# deletions		40975	3870	1537
# duplications		6025	1331	1226
# complex CNVRs (del/dup)		2929	1255	808
# CNVs/sample (mean ± s.d.)		1682 ± 171	634 ± 28	50 ± 12
	AFR[d]	1838 ± 134	626 ± 22	46 ± 8
	AMR	1630 ± 147	636 ± 29	50 ± 10
	EAS	1620 ± 145	608 ± 18	45 ± 8
	EUR	1614 ± 154	655 ± 24	54 ± 13
	SAS	1641 ± 140	649 ± 22	58 ± 14
# CNVs/sample (median ± MAD)		1711 ± 224	633 ± 30	48 ± 10
	AFR[d]	1811 ± 178	625 ± 21	44 ± 6
	AMR	1593 ± 185	638 ± 27	49 ± 10
	EAS	1559 ± 172	608 ± 16	43 ± 7
	EUR	1555 ± 169	654 ± 22	52 ± 7
	SAS	1684 ± 196	648 ± 25	58 ± 18
# singleton CNVs/sample (mean ± s.d.)		8.2 ± 5.0	0.5 ± 2.3	0.5 ± 2.2
	AFR[d]	8.3 ± 4.7	0.0 ± 0.0	0.0 ± 0.0
	AMR	7.2 ± 4.4	0.0 ± 0.0	0.0 ± 0.0
	EAS	9.5 ± 4.5	0.0 ± 0.0	0.0 ± 0.0
	EUR	7.2 ± 5.8	2.7 ± 4.5	2.3 ± 4.5
	SAS	8.4 ± 4.7	0.0 ± 0.0	0.0 ± 0.0
# singleton CNVs/sample (median ± MAD)		8.0 ± 4.4	0.0 ± 0.0	0.0 ± 0.0
	AFR[d]	8.0 ± 4.4	0.0 ± 0.0	0.0 ± 0.0
	AMR	7.0 ± 4.4	0.0 ± 0.0	0.0 ± 0.0
	EAS	9.0 ± 4.4	0.0 ± 0.0	0.0 ± 0.0
	EUR	6.0 ± 4.4	2.0 ± 1.5	2.0 ± 1.5
	SAS	8.0 ± 4.4	0.0 ± 0.0	0.0 ± 0.0
CNVR size (kb) (mean ± s.d.)		17.19 ± 40.86	39.68 ± 58.52	63.75 ± 68.43
CNVR size (kb) (median ± MAD)		3.88 ± 5.18	17.21 ± 22.14	40.87 ± 38.49
# rare CNVR (freq < 1%) / # common CNVR (freq >= 1%)		43227 / 6702	4263 / 2193	3278 / 293
	AFR[d]	16535 / 8140	1571 / 2017	999 / 262
	AMR	11085 / 5576	1442 / 2072	800 / 300
	EAS	12019 / 4615	1229 / 1501	764 / 221
	EUR	11704 / 5042	4508 / 1948	3301 / 270
	SAS	11789 / 5209	1466 / 1823	894 / 262

[a] A subset of the 1000 Genomes Project (KGP) structural variant dataset including deletions, duplications and complex CNVs with multiple variant alleles

[b] A subset of the KGP CNV dataset that is detectable in the FA dataset with CNVs spanning at least 5 probes of the SNP array used

[c] A subset of the KGP CNV dataset that is detectable in the STARNET dataset with CNVs spanning at least 5 probes of the SNP array used

[d] AFR: African; AMR: American; EAS: East Asian; EUR: European; SAS: South Asian

Table S3: Summary statistics of CNVs detected by different methods in the FA and STARNET datasets

	iPattern	PennCNV	QuantiSNP	intersect	union	ensembleCNV
(a) FA						
# CNVRs	10200	14486	25085	13204	28636	19695
# deletions	6917	8470	11475	7676	13248	12614
# duplications	2212	2759	6419	2425	7062	3397
# complex CNVRs (del/dup)	1071	3257	7191	3103	8326	3684
# CNVs/sample (mean ± s.d.)	395 ± 67	362 ± 145	476 ± 182	404 ± 129	700 ± 231	635 ± 138
# CNVs/sample (median ± MAD)	380 ± 31	326 ± 82	426 ± 98	372 ± 77	630 ± 125	620 ± 46
# singleton CNVs/sample (mean ± s.d.)	1.0 ± 6.6	1.6 ± 7.5	3.5 ± 17.6	1.4 ± 7.4	3.7 ± 17.2	1.5 ± 8.7
# singleton CNVs/sample (median ± MAD)	0.0 ± 0.0	1.0 ± 1.5	2.0 ± 1.5	1.0 ± 1.5	2.0 ± 1.5	1.0 ± 1.5
CNVR size (kb) (mean ± s.d.)	43.04 ± 148.44	38.51 ± 176.29	29.43 ± 115.11	37.52 ± 168.18	32.45 ± 136.52	36.65 ± 153.86
CNVR size (kb) (median ± MAD)	13.39 ± 18.33	9.38 ± 12.14	9.63 ± 11.79	8.34 ± 10.89	11 ± 13.64	11.67 ± 14.50
# rare CNVR (freq < 1%)	7886	11304	20994	10116	23706	16588
# common CNVR (freq >= 1%)	2314	3182	4091	3088	4930	3107
(b) STARNET						
# CNVRs	1495	2293	3595	2106	4006	2670
# deletions	817	1167	1420	1063	1570	1298
# duplications	615	901	1653	846	1864	1113
# complex CNVRs (del/dup)	63	225	522	197	572	259
# CNVs/sample (mean ± s.d.)	26 ± 5	16 ± 6	26 ± 11	18 ± 5	40 ± 14	38 ± 8
# CNVs/sample (median ± MAD)	26 ± 4	15 ± 4	24 ± 6	18 ± 4	37 ± 7	38 ± 7
# singleton CNVs/sample (mean ± s.d.)	1.0 ± 1.5	1.7 ± 2.2	2.5 ± 3.6	1.5 ± 1.6	2.9 ± 4.4	1.7 ± 2.0
# singleton CNVs/sample (median ± MAD)	1.0 ± 1.5	1.0 ± 1.5	2.0 ± 1.5	1.0 ± 1.5	2.0 ± 1.5	1.0 ± 1.5
CNVR size (kb) (mean ± s.d.)	98.08 ± 180.52	67.97 ± 314.24	54.38 ± 134.87	76.34 ± 167.70	57.57 ± 254.06	79.19 ± 316.87
CNVR size (kb) (median ± MAD)	41.97 ± 47.38	21.59 ± 26.5	16.11 ± 20.85	25.23 ± 31.12	16.23 ± 20.65	24.85 ± 29.86
# rare CNVR (freq < 1%)	1255	2048	3199	1860	3535	2226
# common CNVR (freq >= 1%)	240	245	396	246	471	444

Table S4. CNVs in 1000 Genomes Project (KGP) data detectable by commercial Illumina SNP arrays

	MEGA	MEGAEX	Omni Express Exome	Omni Express2.5 Exome	Omni Express5 Exome	GSA	Multi Ethnic
# probes	1705969	2036060	960943	2612381	4559489	642848	1748274
# all CNVs	14921	16631	10618	19590	25186	9144	15018
# common CNVs	959	1031	288	728	1267	242	472
# rare CNVs	13962	15600	10330	18862	23919	8902	14546

* 1000 Genomes Project CNV dataset: 48152 CNVs, including 39565 deletions, 5789 duplications, 2798 multi-allelic CNVs; 5839 common CNVs (frequency >= 0.01), 42313 rare CNVs (frequency < 0.01).

** A CNV is defined as detectable with respect to a SNP array if it encompasses at least 5 probes in the array.