**Supplementary Material**

Integrative DNA copy number detection and genotyping from sequencing and array-based platforms

Zilu Zhou, Weixin Wang, Li-San Wang, and Nancy Ruonan Zhang

**Supplementary Method**

*SNP log R ratio (LRR) and (ii) B allele frequency (BAF) from array*

In this study, all SNP arrays tested are from Illumina Human660W-Quad, Illumina OmniExpress-12v1 or Illumina Human610-Quad, no Affymetrix array is involved. We used software GenomeStudio 2.0 to extract LRR and BAF signals from raw data in idat format.

*Sequencing data processing and B allele frequency (BAF) from sequencing data*

We used an in-house pipeline form The Alzheimer's Disease Sequencing Project (ADSP) to deal with whole genome sequencing and whole exome sequencing data in this study. First, 100bp paired reads from Illumina HiSeq were mapped to the hg19 reference genome using the BWA algorithm. Read alignments were saved as BAM files and were further processed using PICARD (http://picard.sourceforge.net/) to mark duplicate reads; data from multiple sequencing experiments for the same individual were merged by SAMtools. GATK was used to further improve BAM data quality by local read realignment near known insertions or deletions (indels) sites and base quality score recalibration (BQSR).

To increase the quality of call set, consensus VCFs were generated by applying a QC protocol to the VCFs received from the Human Genome Sequencing Center at Baylor College of Medicine (hereinafter, “Baylor”), which performed genotype calling using Atlas V2 software and to the VCFs from the Broad Institute (hereinafter, “Broad”), which performed genotype calling using GATK-HaplotypeCaller. Both Baylor and Broad derived their VCFs from the same set of BAM files which comprised data from sequencing performed at Baylor, Broad, and Washington University-St. Louis (hereinafter, “WashU”). Both the Baylor and Broad VCFs underwent pipeline-specific quality control to remove low-quality variants and genotypes, and to identify problematic samples. After pipeline-specific QC implementation, the remaining variants and genotypes within the two QCed VCFs were integrated into a single “consensus” set of genotype calls using a novel consensus calling approach. The BAFs provided here are derived using data from genotype-level fields taken from above consensus VCFs generated by the ADSP Quality Control (QC) Working Group.  Genotype components used in these BAF estimations are ‘CS’ (concordance code), ‘AD’ (read depths for REF and ALT, delimited with a comma), and ‘DG’ (overall read depth from GATK).For each individual, the BAF for GATK calls is estimated using only heterozygous samples with CS values of '91' (heterozygous in GATK; uncalled in Atlas) or '11' (heterozygous in both GATK and Atlas).  The formula can be described in this way:

where AD [2] is the count of reads with alternative allele for each qualified het (CS=91 or 11).

In above two circumstances, BAFs are only estimated for bi-allelic SNVs in diploid organisms where all samples are either called as heterozygous or homozygous. They are not calculated for insertion-deletion polymorphisms (indels).

For users without above sophisticated analysis pipelines, we suggest users either can following GATK best practice; or at least do duplication removal and read re-alignment around known indels to reduce the noise, then calling SNVs from quality controlled BAM files by mpile-up module in samtools, and calculate BAF on heterogeneous loci only by dividing DV (Number of high-quality non-reference bases, FORMAT) from DP (Number of high-quality bases, FORMAT).

***Pipeline modification for single platform design***

We modify iCNV pipeline based on the experimental design. In NGS only study, this method detects CNV with both normalized intensity score and BAF called by SAMTOOLs (Supplementary Fig. 1). In array only case, it is very similar to PennCNV with more adaptive emission probability and constrained transition probability (Supplementary Fig. 2).

***Integrated hidden Markov model (HMM)***

In the following section I will introduce the five basis of this integrated HMM: hidden states, observations, transition probability, observation probability and initial state.

Hidden states: We assume there are only 3 possible hidden state at position: deletion (), diploid (), and duplication ().

Observations: Since multiple normalized intensities and BAFs can be observed at same position , our observations for loci are

Transition probability: We implemented a very similar transition probability matrix as XHMM:

|  |  |  |  |
| --- | --- | --- | --- |
| From … to … | Deletion | Diploid | Duplication |
| Deletion |  |  |  |
| Diploid |  |  |  |
| Duplication |  |  |  |

where and D is the mean distance between targets in a CNV; is the CNV rate; and is the mean number of targets in a CNV.

Observation probability consists of two parts: intensity observation probability and BAF observation probability. For easier calculation, we assume these two probability are i.i.d.

Intensity observation probability follows a normal distribution.

where equal to for deletion, diploid and duplication. BAF distribution is modeled by mixture of truncated normal distribution within 0 to 1.

where and . We set in order to give a larger skewness for the deletion BAF distribution and greater penalty for BAF near 0.5.

Initial state: The initial probability is as follow:

|  |  |  |
| --- | --- | --- |
| Deletion | Diploid | Duplication |
|  |  |  |

***Integer copy number inference***

After we identified the CNVs, we further predict the exact copy number in each CNV region by a maximum likelihood approach. As mentioned in the main text, . The likelihood function given intensity and BAF is

where

The BAF probability distribution is similar designed as the emission probability. The intensity distribution, however, has altered. Because there are still deviations of the means between normalized PLR and normalized LRR for each copy number state, we applied platform specific parameter () for intensity distribution estimated by K-means algorithm. In double deletion BAF likelihood, we also add a uniform distribution, since the BAF will have all kinds of value when both alleles are deleted.

***Union and intersection analysis***

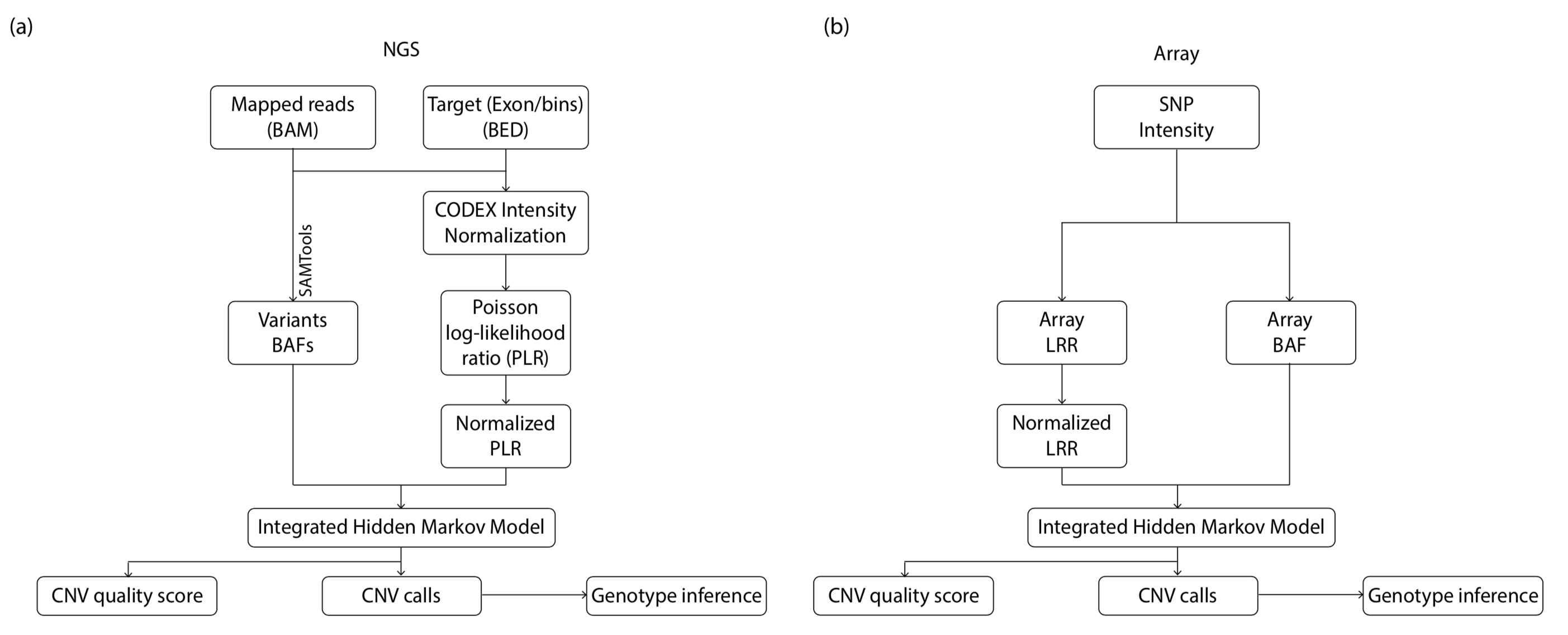
In 3.1 and 3.2, we compared joint platform results with simply union or intersection of single platform calls by iCNV. We mapped our results to the same coordinate followed Fig.2. For any joint locus, we will call a CNV in intersection if all the array SNPs and exons within this region are CNVs; we will call a CNV in Union if any of the array SNPs and exons within this region is a CNV. We further counted the number of joint loci for comparison with joint platform calls showed in Table 1 and Table 2.

***Familial sharing enrichment score***

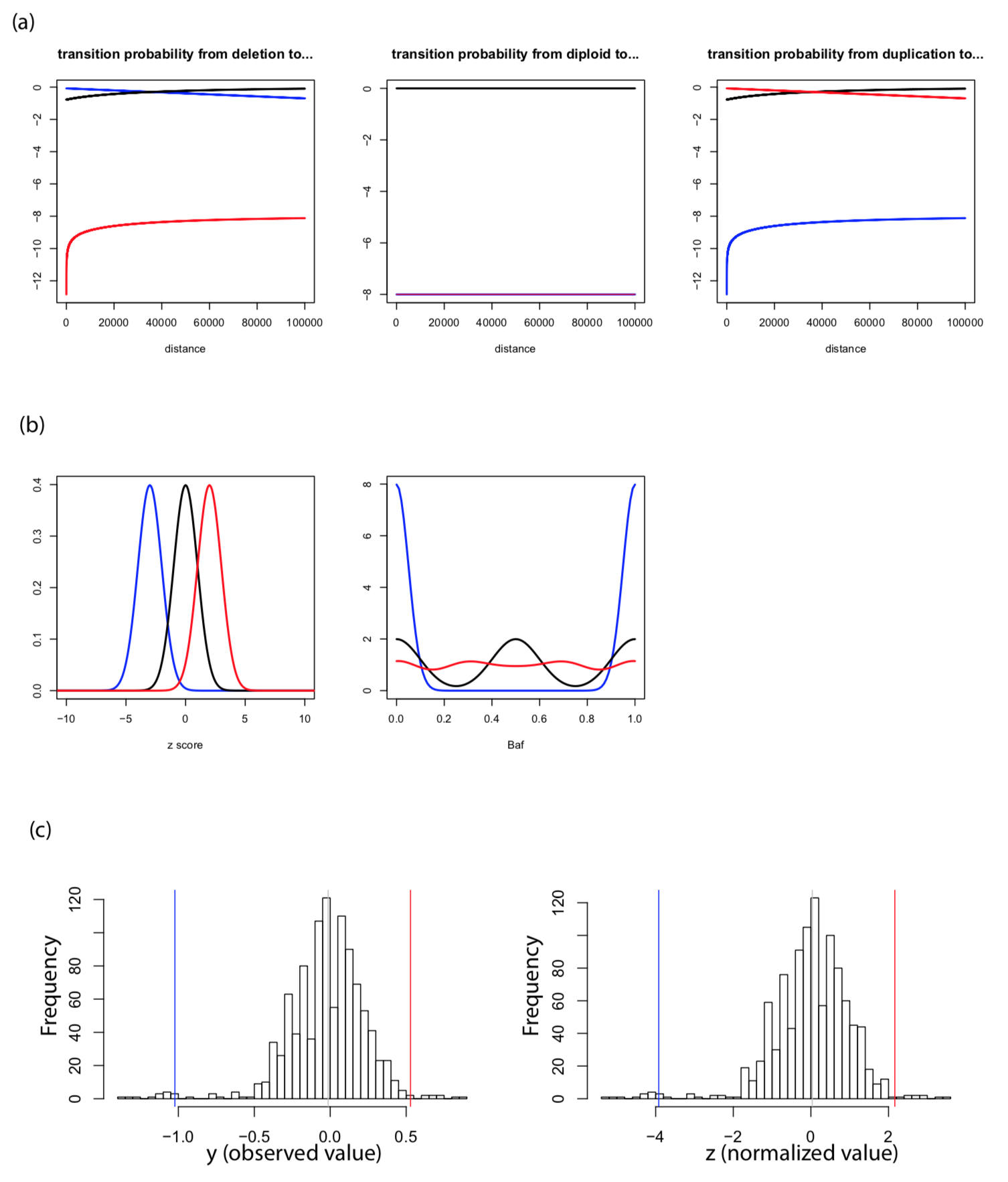
In order for us to compare CNV algorithm, we utilize a dataset with family information. Since individual within the same family should have more similar CNV pattern than random, we calculate a familial sharing enrichment score to capture this enrichment in three steps. First, we calculated the observed familial sharing frequency regarding each locus defined as:

Because the chance of familial sharing CNV increases with CNV population frequency in any locus, we second construct a null sharing frequency curve using permutation. At last, by subtracting from the observed familial sharing frequency its permutation mean and then dividing the difference by its permutation standard deviation, we compute this sharing enrichment score for each locus.

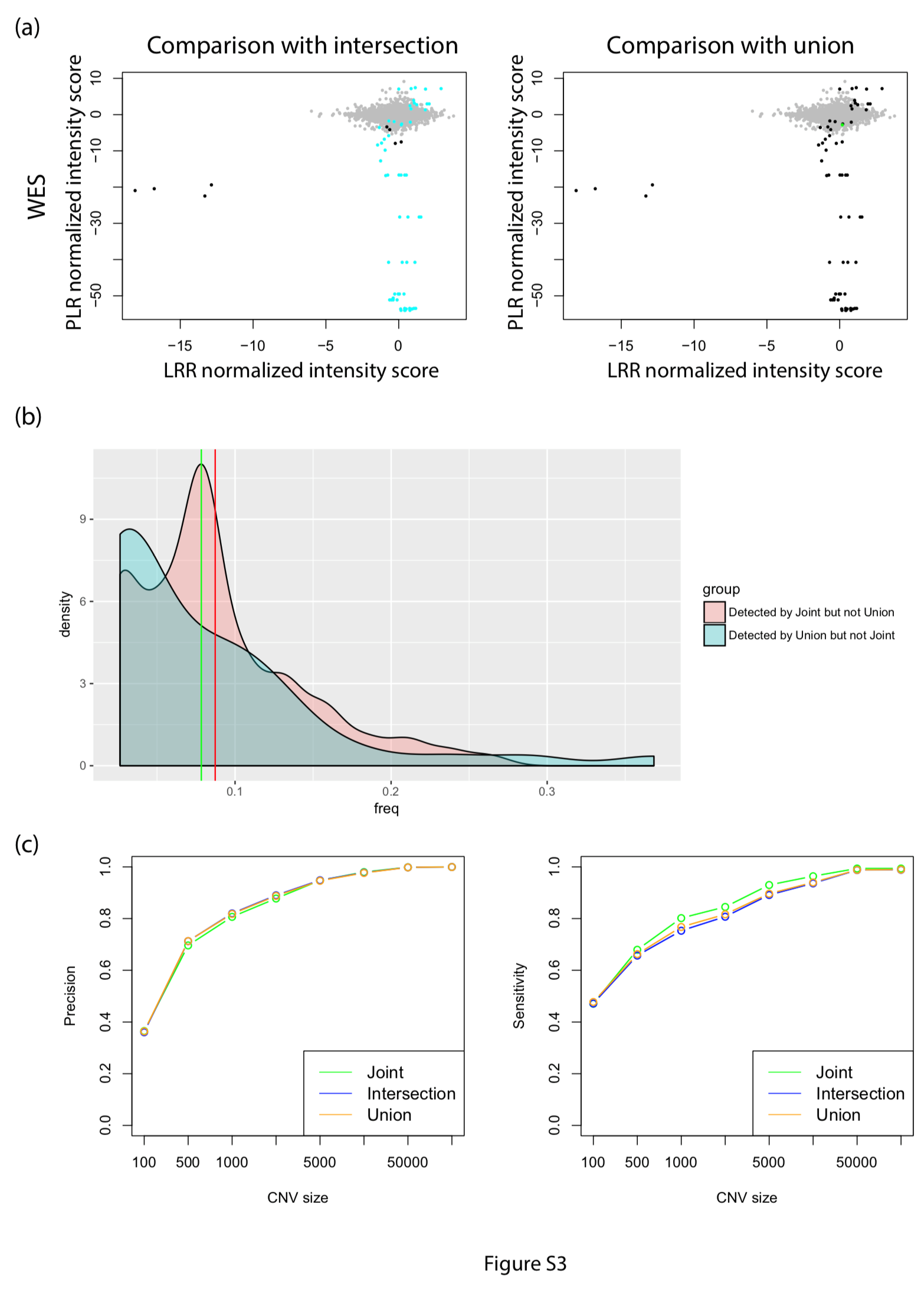
**Figure S1. A flowchart outing the procedures of iCNV in data normalization, CNV calling and genotyping from NGS or array data**. (a) For NGS data, the first step is to normalize intensity data and to calculate a Poisson log-likelihood ratio (PLR) using CODEX. The variants BAF are then calculated with target position using SAMTools. We further convert PLR to a normalized intensity score. The integrated Hidden Markov Model takes the PLR and variants BAF from NGS generating CNV calls and quality score. We further infer their genotypes. (b) . For array data, GenomeStudio calculate a log R ratio - further normalized to a normalized intensity score - and BAF from raw SNP intensity data. The integrated Hidden Markov Model takes the array intensity and BAF from array generating CNV calls and quality score. We further infer their genotypes.



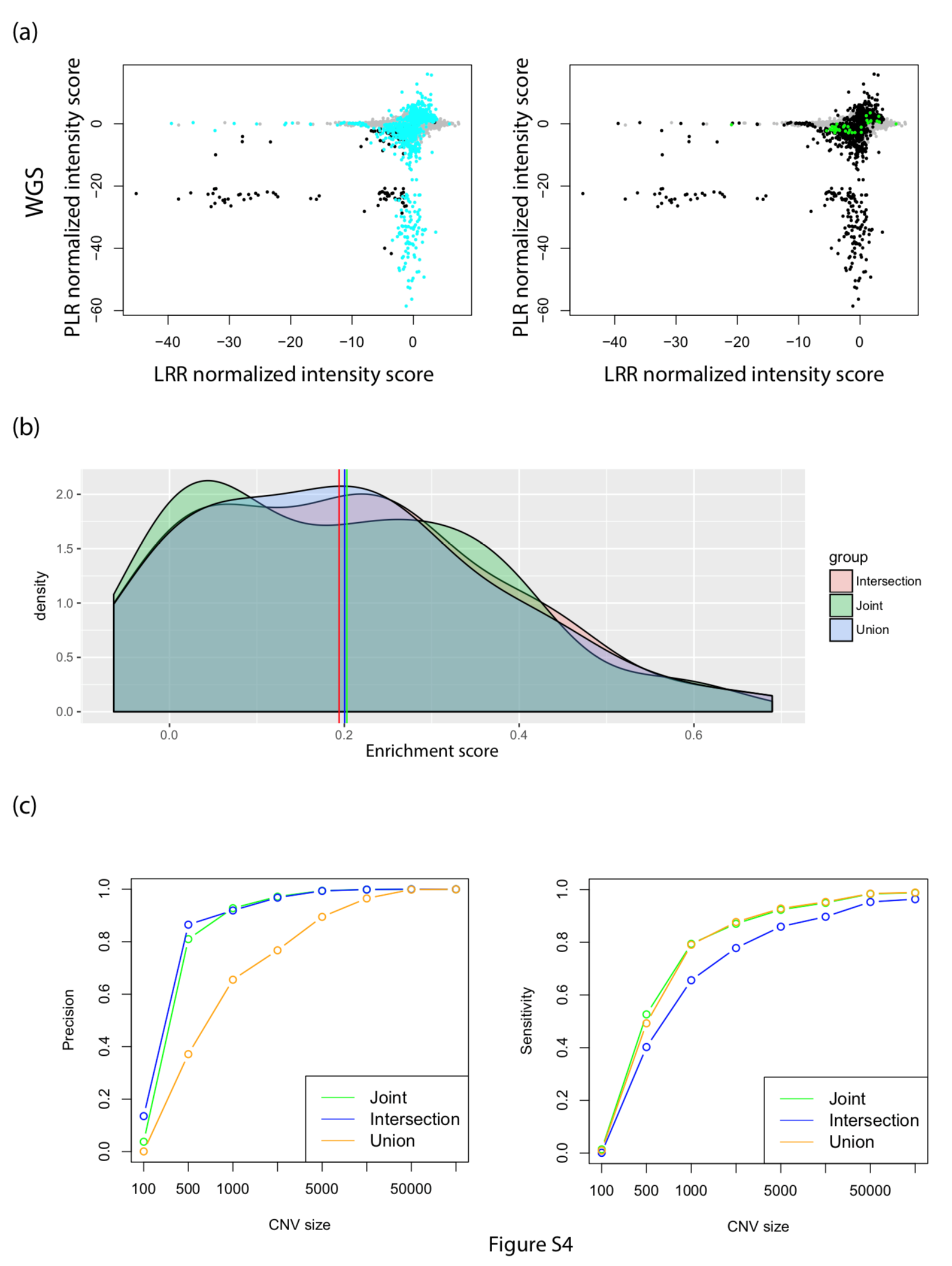
**Figure S2. iCNV transition probability, emission probability and simulation.** (a) Relationship between distance and transition probabilities from each state to deletion (blue line), diploid (black line) and duplication (red line). (b) Emission probability distribution for normalized intensity and BAF in deletion (blue line), diploid (black line) and duplication (red line). (c) In order to study relationship of means between deletion and duplication, we sample observe value from Poisson process with same transformation. The deletion mean (blue line) is more deviated from zero than duplication mean (red line) before and after normalization.



**Figure S3. Results comparison between intersection or union and iCNV.** For (a), we compared the normalized intensity between intersection or union and iCNV joint (WES+array) call set. Grey dots represent all the SNP and NGS overlapped region’s normalized intensity score; Black dots indicate both type of method identify that intensity represent region as CNV; Cyan dots indicate region identified by iCNV but not intersection; and Green dots indicate region identified by iCNV but not union. There is only one green dots in (a) due to very few overlap between exons and SNPs. In (b), we compared the samples shared frequency distribution for CNVs detected by joint (WES+array) but not union (red density, red line indicate mean) and CNVs detected by union but not joint (WES+array, green density, green line indicate mean). CNVs detected by joint but not union are more likely to share between samples (Wilcox-test, p-value=0.038). Subfigure (c) shows precision and sensitivity analysis by in silico spike-in, comparing joint and intersection or union of two individual call set. Results show not much difference in precision; but joint analysis has highest sensitivity.



**Figure S4. Results comparison between intersection or union and iCNV.** For (a), we compared the normalized intensity between intersection or union and iCNV joint (WGS+array) call set. Grey dots represent all the SNP and NGS overlapped region’s normalized intensity score; Black dots indicate both type of method identify that intensity represent region as CNV; Cyan dots indicate region identified by iCNV but not intersection; and Green dots indicate region identified by iCNV but not union. In (b), we compared the z-scores calculated by within family shared frequency distribution for joint (WGS+array, green density, green line indicate mean), union (blue density, blue line indicate mean) and intersection (red density, red line indicate mean CNVs detected by joint are slightly more likely to share between samples (but not statistically significant). Subfigure (c) shows precision and sensitivity analysis by in silico spike-in, comparing joint and intersection or union of two individual call set. Results show that joint calling has precision close to intersection and sensitivity close to union.



**Table S1. CNV case overlap analysis for WES and Array**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CNV case overlap summary (WES, Array) | | | | | |
| Percentage of joint overlapped by/overlapped by joint | WES | Array | WES ∩ SNP | WES ∪ SNP |
| Del | 20.64/96.55 | 74.38/99.52 | 87.19/90.88 | 91.46/90.76 |
| Dup | 42.86/81.82 | 42.86/100 | 85.71/75.41 | 85.71/72.22 |
| Combined | 22.19/94.20 | 72.18/99.54 | 87.09/88.23 | 91.06/87.96 |

**Table S2. CNV case overlap analysis for WGS and Array**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CNV case overlap summary (WGS, Array) | | | | |
| Percentage of joint overlapped by/overlapped by joint | WGS | Array | WGS ∩ SNP | WGS ∪ SNP |
| Del | 90.80/96.91 | 13.14/70.89 | 93.12/87.38 | 97.67/71.54 |
| Dup | 97.11/93.43 | 0.96/100 | 97.11/88.70 | 97.11/84.07 |