A validation study of copy number variant (CNVs) detection to replace constitutional microarray from low resolution whole genome sequencing data

C. Alexander Valencia[1], Zeqiang Ma[1], Gerard Irzyk[1], Edward Szekeres Jr[1], Zdenek Markovic[1], Yang Wang[1], Alice Tanner[1], Christin Collins[1], Alka Chaubey[2], Madhuri Hegde[1,3]

[1]PerkinElmer Genomics, [2]Greenwood Genetic Center, [3]Emory University

ABSTRACT

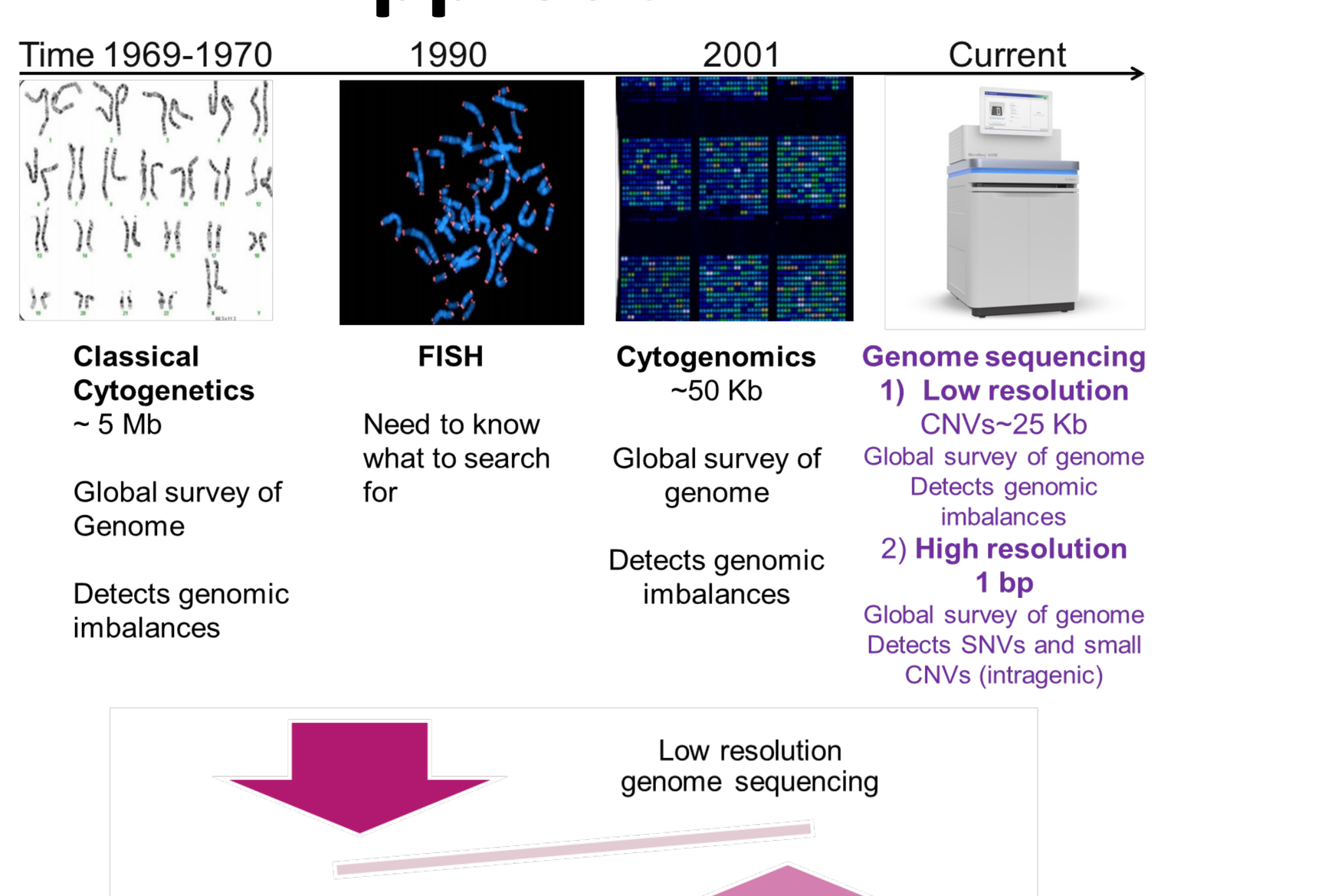
DNA copy-number variants (CNVs) account for up to 300 Mb of sequence variation in a normal human individual. These represent a major class in genome diversity between two different individuals; some of these CNVs are known to be associated with the pathogenicity of a variety of human disorders. The detection of pathogenic CNVs by chromosomal microarray analysis (CMA), including array–comparative genomic hybridization (array-CGH) and single-nucleotide polymorphism array, has been widely used as a gold standard. Compared with CMA, next-generation sequencing (NGS) is an alternative state-of-the-art technology promising improved detection of genetic abnormalities with unprecedented resolution. A recent publication has demonstrated the use of low resolution genome sequencing at an average of depth of 0.25X to detect genomic CNVs. The objective of this validation study was to assess the diagnostic effectiveness, by calculating performance parameters (accuracy, sensitivity, sensitivity and specificity), of using a low-resolution (coverage, 5-10X) whole genome sequencing to detect chromosomal numerical and structural abnormalities in a diagnostic laboratory using four CNV calling tools, namely, BGI CNV, Edico CNV, Variantyx and Golden-Helix CNV. Experimentally, four runs were performed to obtain the global metrics and performance parameters as follows: run 1 had three well characterized Coriell samples (NA12878, NA12891, NA12892) that were sequenced, run 2 was a repeat of run 1, run 3 included 40 positive control samples with a diverse type and size of clinically relevant CNVs such as trisomies (4), interstitial (4) and terminal deletions (4) and duplications, known microdeletions, mosaic deletions, absence of heterozygosity and balanced and unbalanced translocations (4), and run 4 included 20 negative control samples without clinical significant CNVs and 10 positive controls from run 3. The average global genome metrics for the 63 samples were 11.25X average coverage and genome coverage distribution of 9.35% at 0-1X, 0% at 1-2X, 7.71% at 2-5X, 38.45 at 5-10X, 39.04% at 10-20X , 5.11% at 20-30X, 0.19% at 30-40X and 0.125 at >40X. In contrast, samples with an average coverage of 32.64X had a genome coverage distribution of 8.73% at 0-1X, 0% at 1-2X, 0.24% at 2-5X, 0.50 at 5-10X, 6.36% at 10-20X , 38.25% at 20-30X, 37.10% at 30-40X and 8.78 at >40X. At 11X average coverage, the average heterozygous and homozygous variants were 2,569,012 and 1,755,930, respectively, compared to 3,268,943 and 1,801,323 at 32.64X. The accuracy was calculated to be 97% from 30 samples (positive controls in runs 3 and 4) using all CNVs (clinical significant and benign; >250 Kb) detected by from CMA compared to those obtained through WGS CNV tools. However, the accuracy was 100% for clinically relevant CNVs. Importantly, this approach allowed the detection of expected deletions and duplications of varying sizes. For example, the 500,031 bp clinical significant duplication was detected on 3q in a male, arr [hg18] 3q13.33(121,058,076-121,558,107)x3, with autism. Similarly, a small 113,056 bp deletion was detected in male on 22q, arr [hg18] 22q11.22q11.23 (21,390,449-21,978,719)x1, with suspected 22q11.2 deletion syndrome. Similarly, the precision was calculated to be 100% from 26 samples (run 1 versus 2 and positive controls of run 3 and run 4) by comparing all CNVs (clinical significant and benign; >250 Kb) detected in the duplicates. The sensitivity and specificity and were 100% and 99%, respectively, for all CNVs. Significantly, this method permitted the detection of new CNVs missed by CMA of unknown clinical significance. In conclusion, the performance parameters of low resolution WGS were equal or better than those by CMA. Specifically, low resolution WGS is an effective method for the diagnosis of chromosomal diseases or microdeletion/microduplication syndromes. Due to the lower cost, higher resolution and sensitivity, low resolution WGS has the potential to replace CMA.

INTRODUCTION

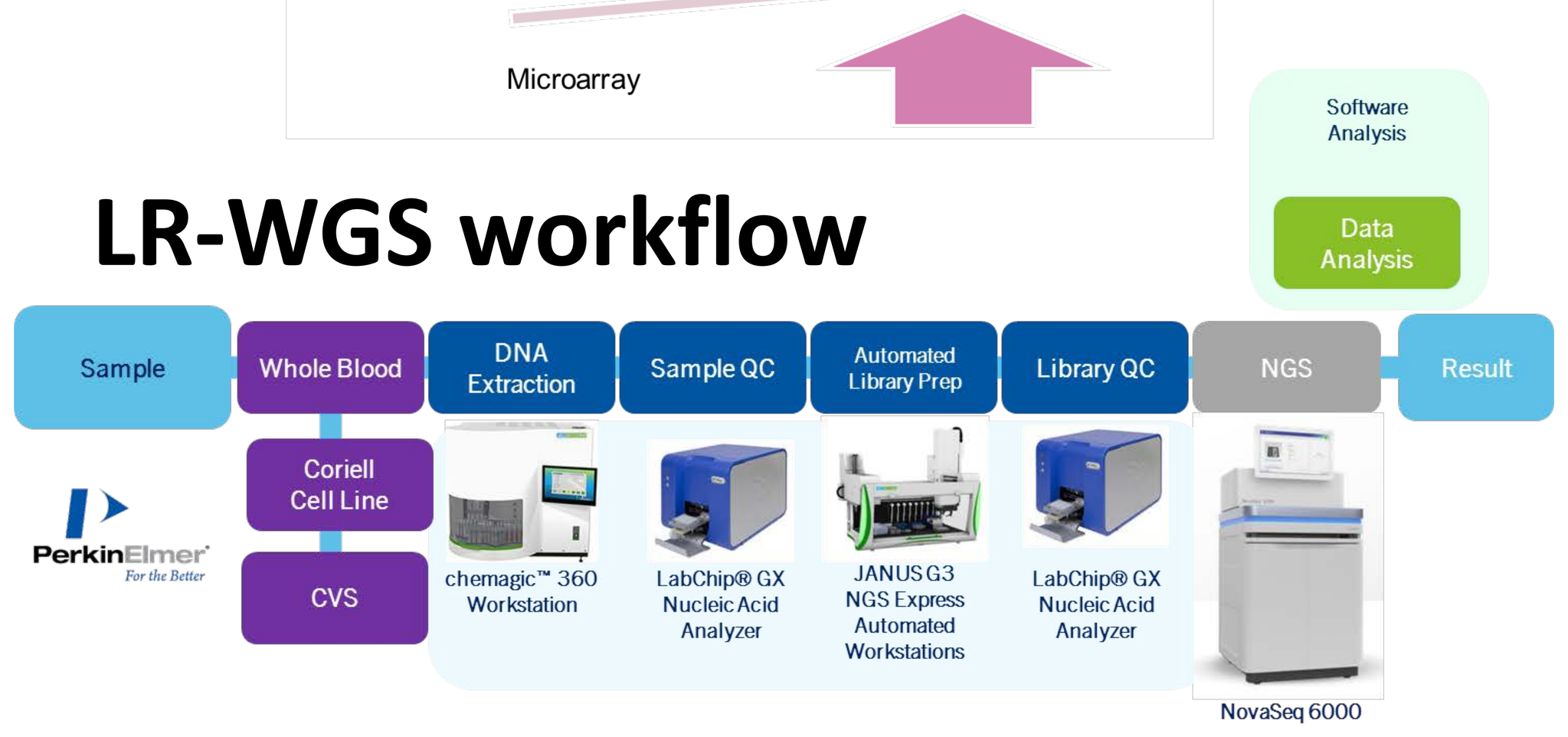
- DNA copy-number variants (CNVs) represent a major class in genome diversity between two different individuals
  - CNVs are known to be associated with a number of human disorders
- The gold standard for the detection of pathogenic CNVs has been chromosomal microarray analysis (CMA), by array-CGH and SNP-array
- **CHALLENGE:** Compared with CMA, next-generation sequencing (NGS) offers a potentially improved detection of CNVs with higher resolution
- **APPROACH:** WGS at a mean 5X coverage (low resolution WGS, LR-WGS) was performed by using the KAPA HyperPlus PCR-free library construction kit on DNAs obtained from a number of samples types. The libraries were subsequently sequenced on the Illumina NovaSeq™ 6000 operating in the 2 x 150 bp mode

METHODS

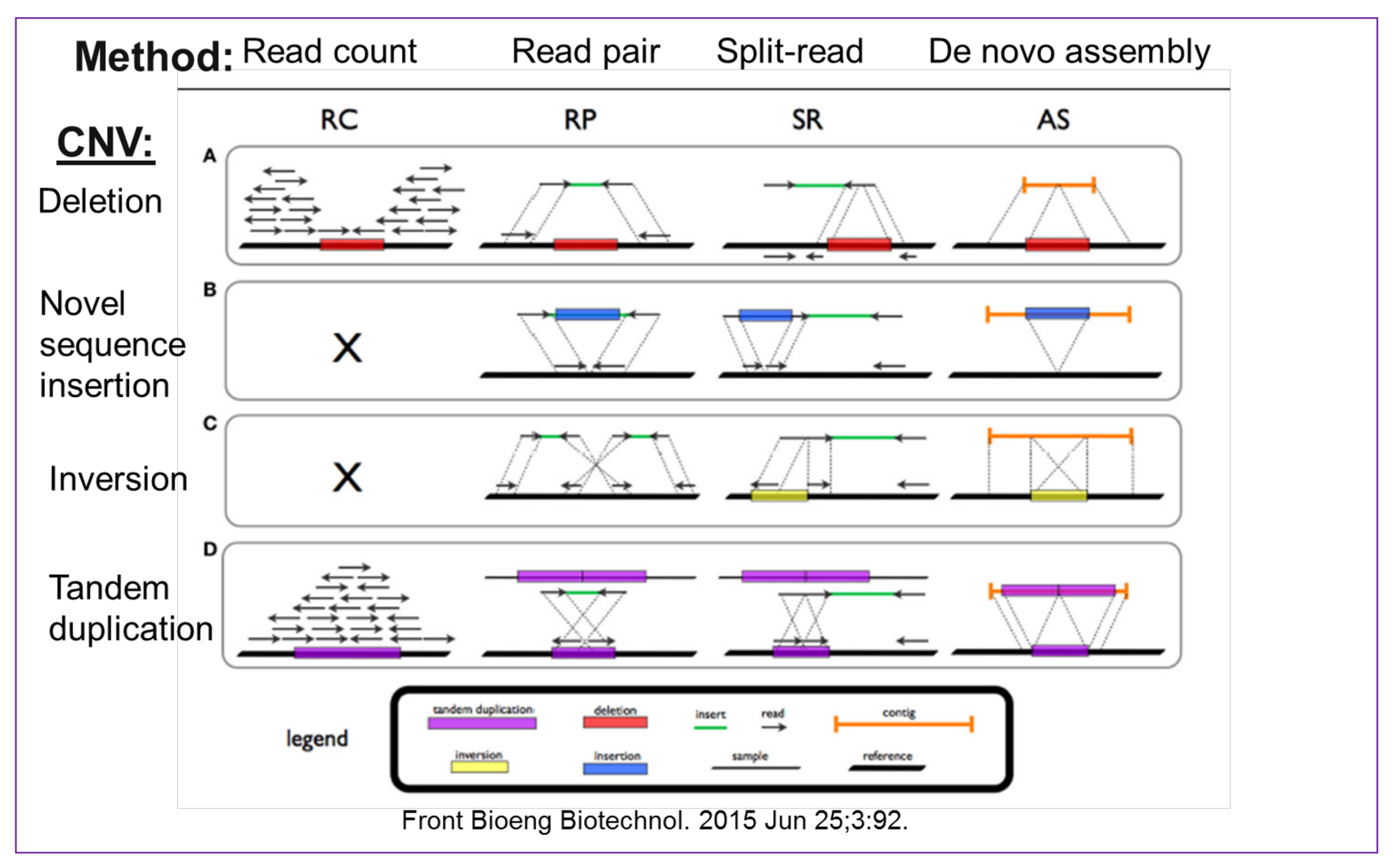
Approach



LR-WGS workflow



Signatures and patterns for structural variants



Front Bioeng Biotechnol. 2015 Jun 25;3:92.

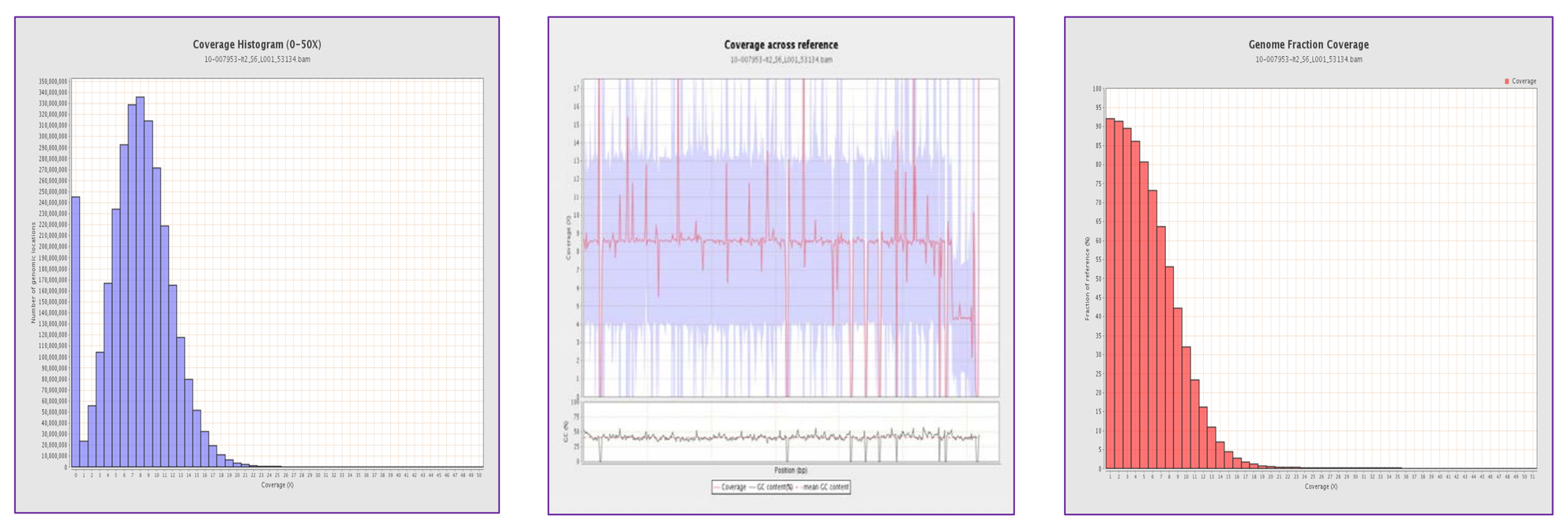
LR-WGS validation plan

- Total number of samples: 77
- CNV negative controls: 35
- CNV (clinically relevant) positive controls: 42
- 4 sequencing runs
- Use Coriell cell line DNA, whole blood, CVS
- Calculated performance parameters: Accuracy, precision, sensitivity and specificity

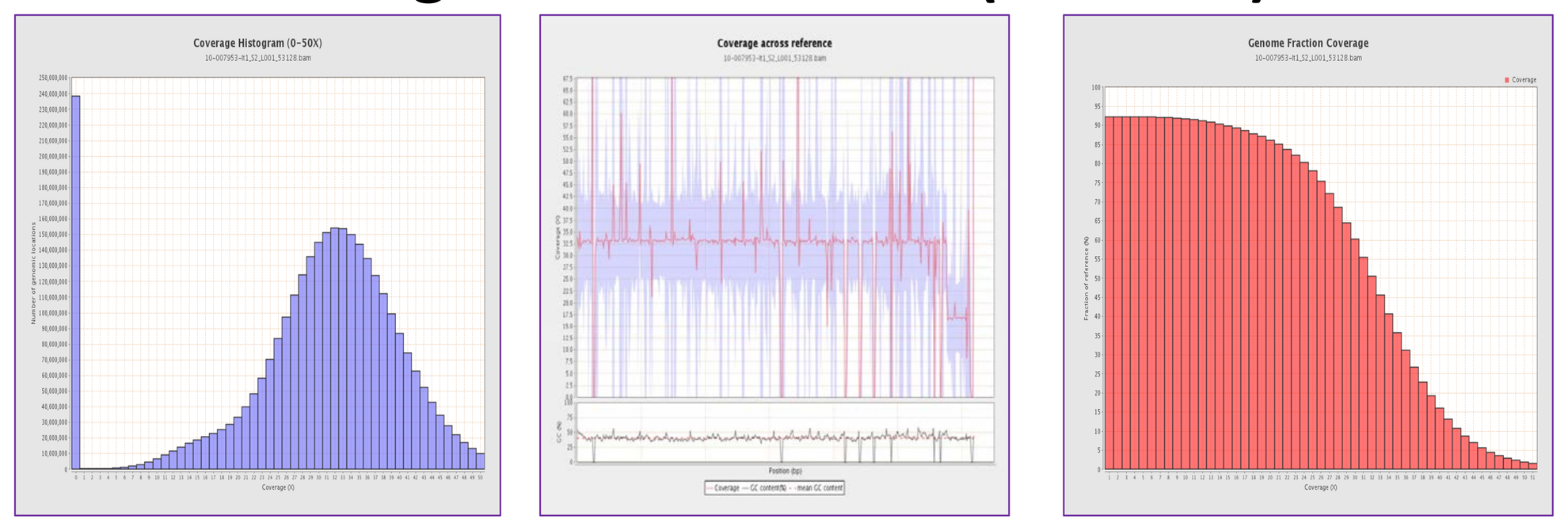
RESULTS

Global Statistics: High quality global key measurements

Coverage statistics at 5X (LR-WGS)

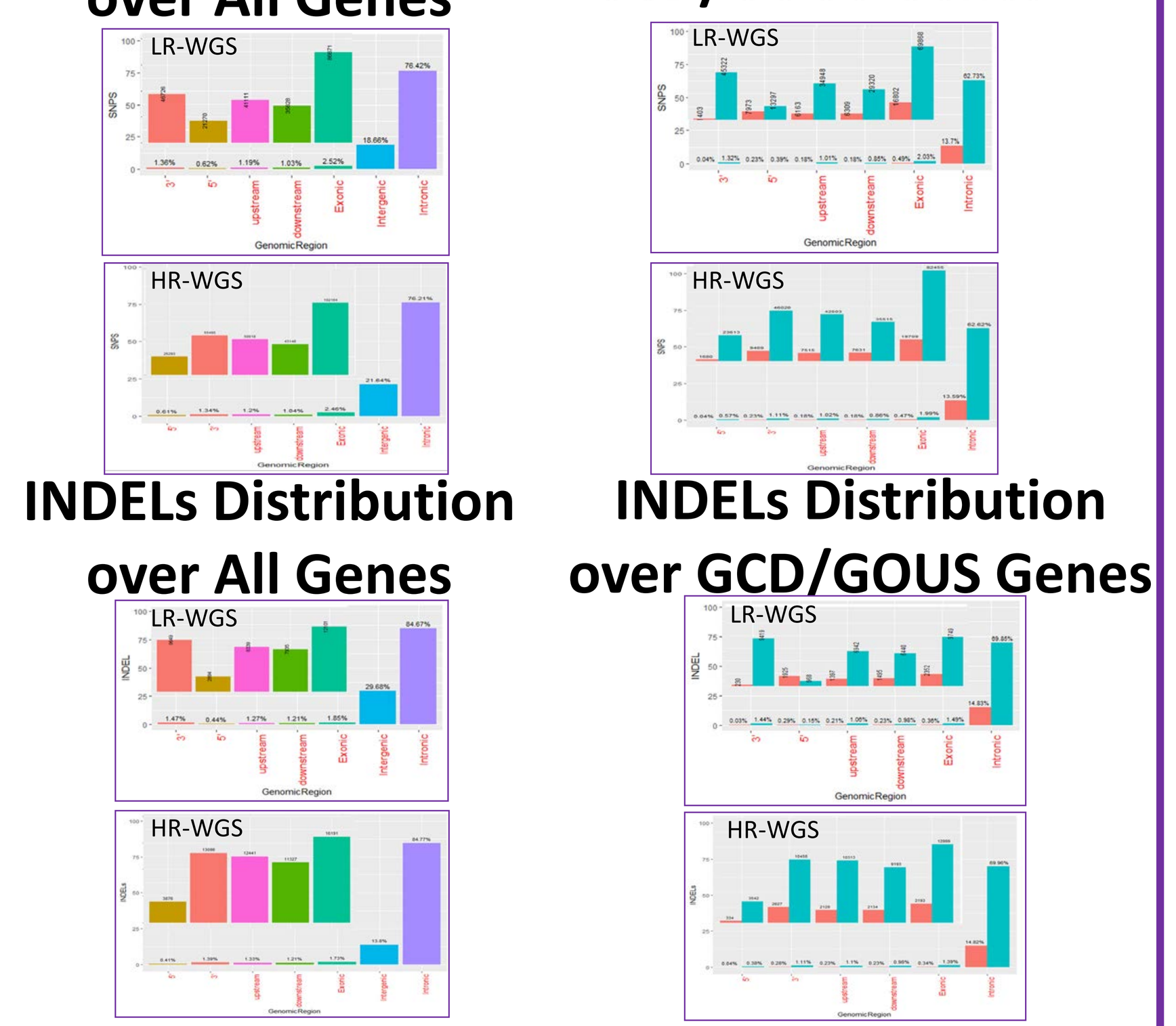


Coverage statistics at 30X (HR-WGS)



- Compared to 30X WGS, LR-WGS clearly has a lower coverage distribution and lower coverage across the reference

SNPs Distribution over All Genes — SNPs Distribution over GCD/GOUS Genes — INDELs Distribution over All Genes — INDELs Distribution over GCD/GOUS Genes



- Comprehensive detection of SNPs and INDELS across entire genome and genes causing disease (GCD) genes in low resolution WGS

Excellent performance and comprehensive detection of CNVs

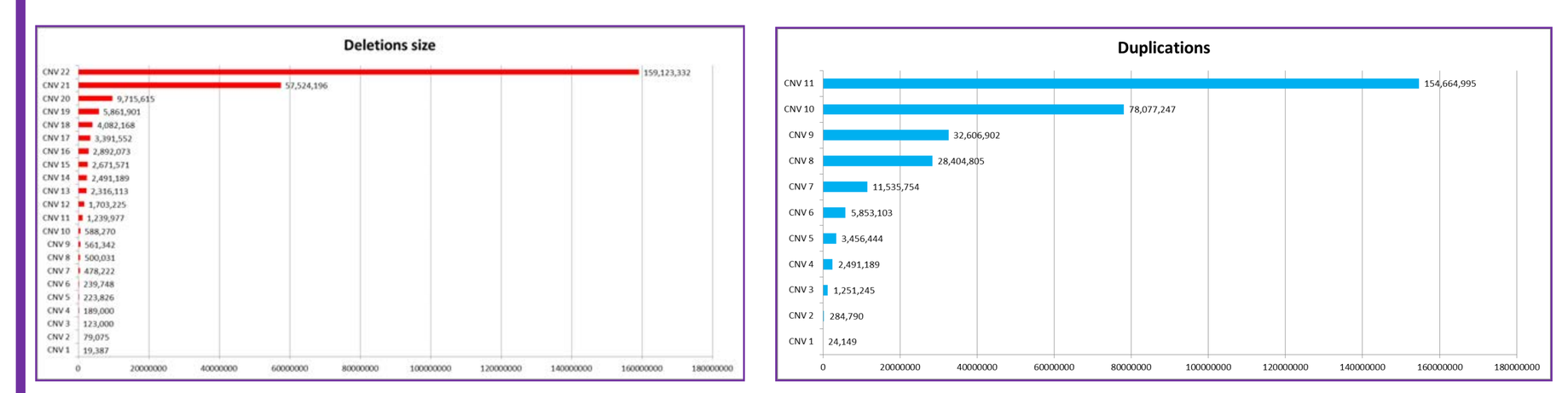
Accuracy is 100%
TP+TN/(TP+TN+FP+FN) = 39+35/(39+35+0+0)

Precision is 100%
TP/(TP+FP) = 39/(39+0)

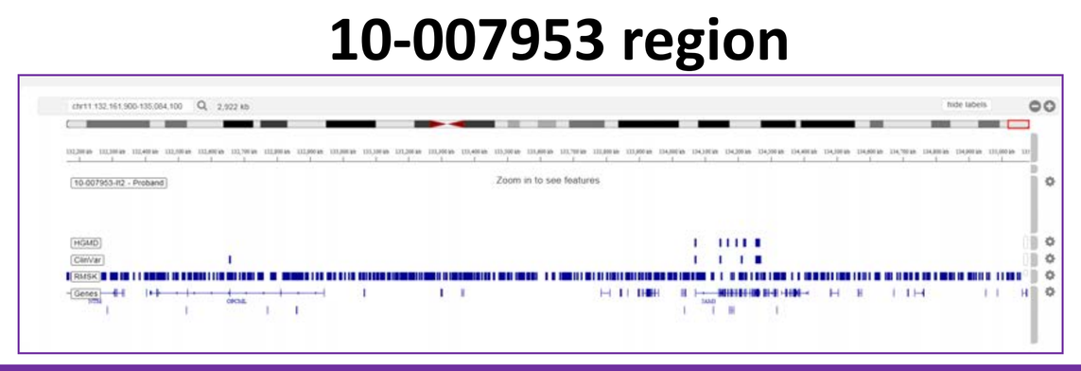
Specificity is 100%
=TN/(TN+FP) = 35/(35/0)

Sensitivity is 100%
=TP/(TP+FN) = 39/(39/0)

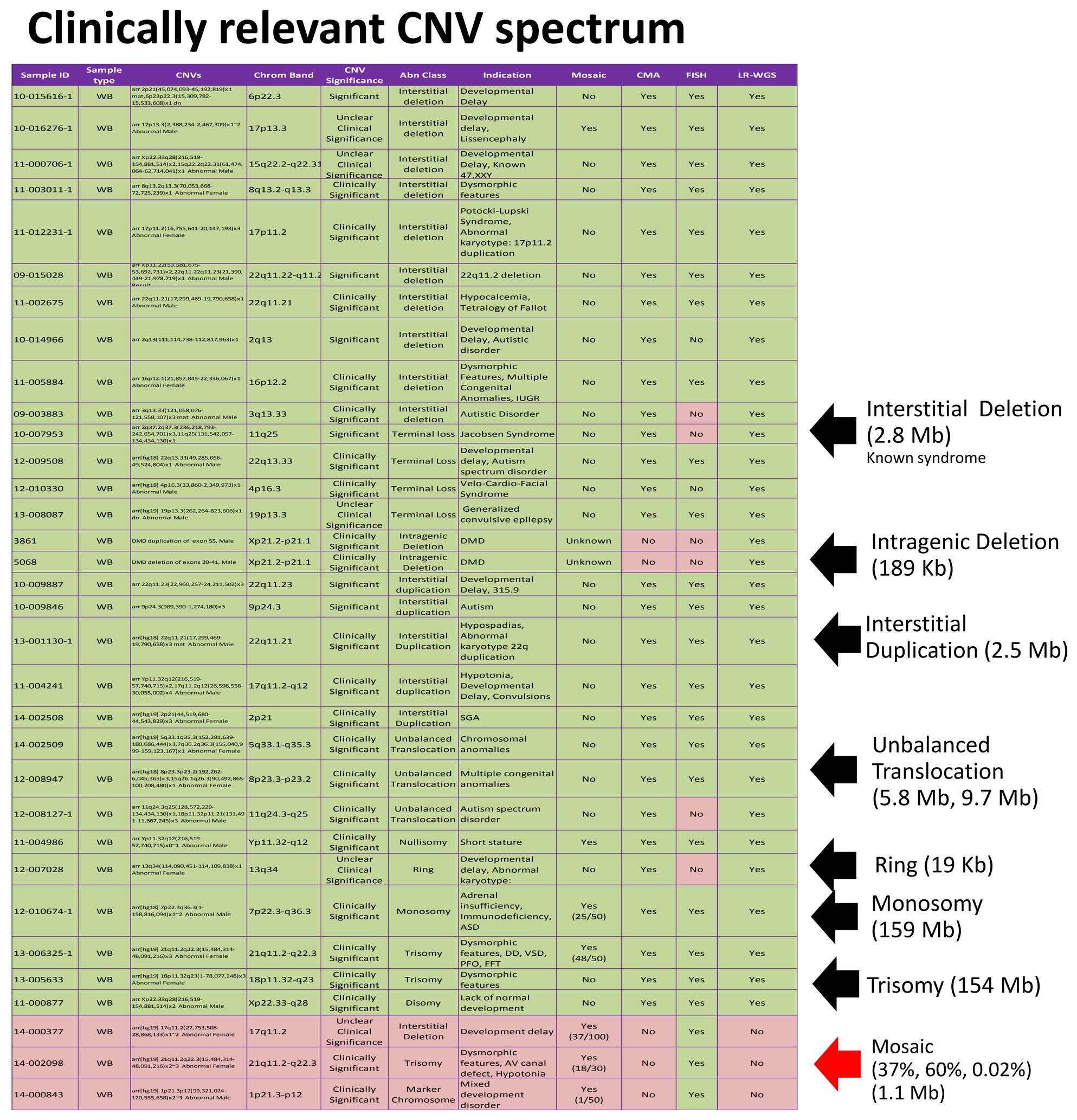
Size of CNVs detected



10-007953 region



Clinically relevant CNV spectrum



DISCUSSION/CONCLUSIONS

- **Significance:** CNV detection explains many genetic disorders
- LR-WGS performance parameters were all >99% for clinically relevant CNVs
  - LR-WGS allowed the detection of expected deletions (19 Kb to 159 Mb) and duplications (24 Kb to 154 Mb) of varying sizes.
  - Detected CNVs of clinical significance
    - Suspected 22q11.2 deletion syndrome: A small 588,270 bp deletion was detected on chr 22q in a male (arr [hg18] 22q11.22q11.23 (21,390,449-21,978,719)x1).
- **Better:** LR-WGS permitted the detection of CNVs missed by CMA of clinical significance
  - *DMD* Exon 55 duplication (123 Kb), *DMD* Exon 20-41 deletion (189 Kb)
- The performance parameters of LR-WGS were equal or better than those by CMA.
- **Cost:** NovaSeq reduced cost significantly and can compete with microarrays
- **Speed:** Workflow is simple and fast
- LR-WGS is an effective method for the diagnosis of chromosomal diseases or microdeletion/microduplication syndromes
- **Marketplace:** Due to the lower cost and higher resolution, LR-WGS has the potential to replace CMA