**A whole genome sequencing-based algorithm for copy number detection at clinical grade level**

Wan-Ping Lee1,2,3,4\*, Qihui Zhu2\*, Xiaofei Yang2,5,6, Silvia Liu2, Eliza Cerveira2, Mallory Ryan2, Adam Mil-Homens2, Lauren Bellfy2, Kai Ye1,6, Chengsheng Zhang1,2,8, and Charles Lee1,2,7,8

1. Precision Medicine Center, The First Affiliated Hospital of Xi’an Jiaotong University, Xi'an 710061, China
2. The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA
3. School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China
4. Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
5. School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
6. MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
7. Department of Life Sciences, Ewha Womans University, Seoul 03760, South Korea
8. Corresponding authors

\*: These authors contributed equally to this work.

Content

[K-mer selection 3](#_Toc64385908)

[DBSCAN 4](#_Toc64385909)

[Droplet Digital PCR (ddPCR) Validation 6](#_Toc64385910)

[Scripts 7](#_Toc64385911)

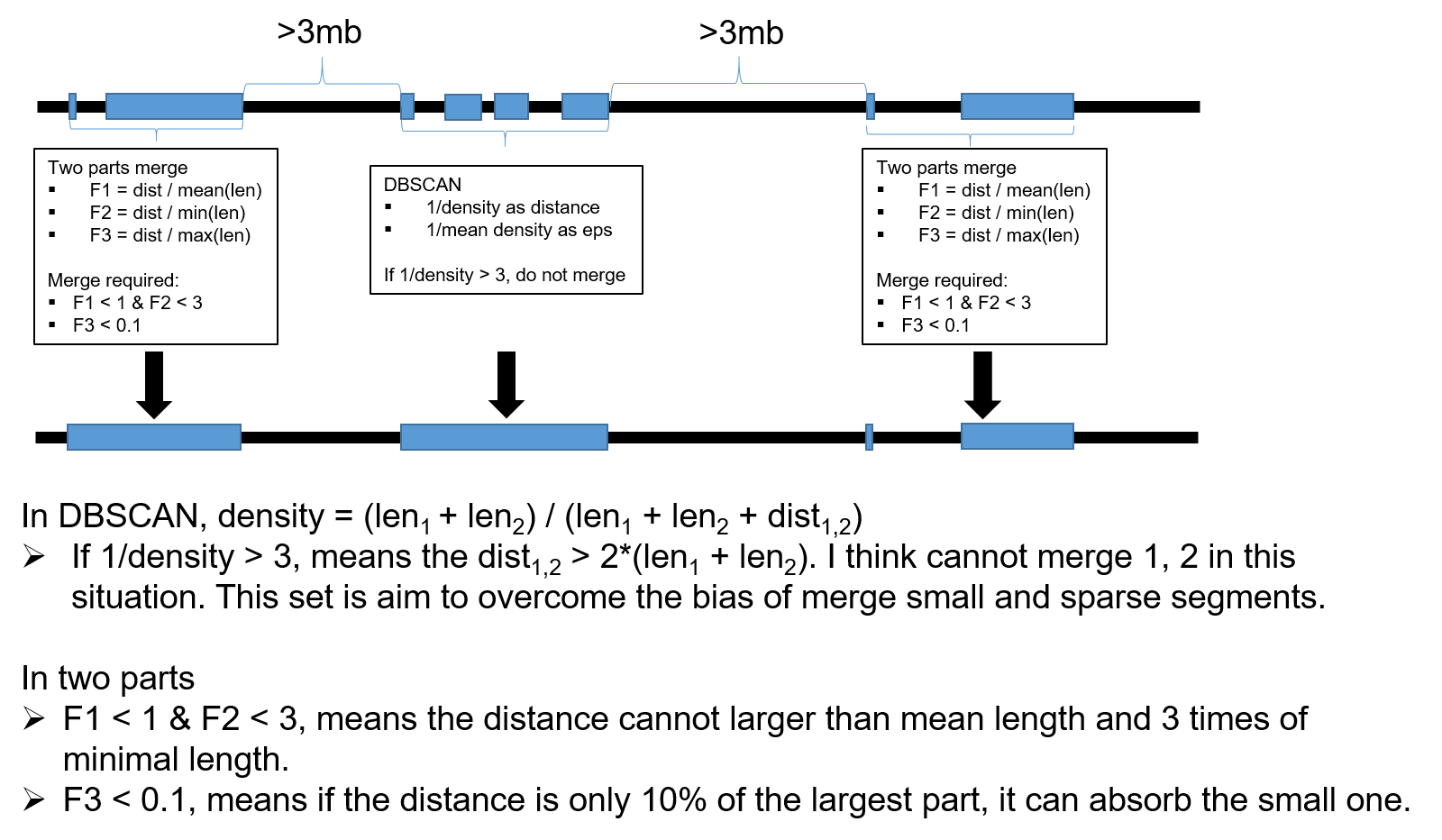
# K-mer selection

Using larger k-mer will increase uniqueness (higher sensitivity) of each k-mer. However, the improvement will saturate and processing larger k-mer requests more computational resource. For example, Jellyfish may use more than 60GB memory for 30-mer.

We choose 25-mer since it maintains good sensitivity and consumes reasonable computational resource. Below is the figure to show the percentage of unique k-mer (gray colored) by applying different k value. Using 5-mer or 10-mer, there is no unique kmer against the human reference genome. From 15-mer to 20-mer, we see a significant increase of unique k-mers that saturates from 25-mer to 30-mer. Since the sensitivity of JAX-CNV does not reply on single factor of uniqueness of a k-mer, the sensitivity was not affected using k-mer larger than 25-mer.

# DBSCAN

Specifically, we first sort the candidate CNV fragments based on their positions. Then, we separate the fragments into different raw clusters by two conditions: a. the distances of any two continuous fragments are < 3,000,000bp; b. the type (e.g. loss, gain) of all fragments located in the raw cluster region are the same. Next, for each raw cluster, we calculate the distance of every continuous fragment pair fi and fi+1 as di,i+1=(ei+1-si)/(li+li+1), where ei+1is the end position of fi+1, siis the start position of fi, and li and li+1 are the length of fi and fi+1. We also calculate the mean distance of the raw cluster as dmean=(E-S)/i=1Nli, where E is the end position of the raw cluster, S is the start position of the raw cluster, and N is the number of fragments in the raw cluster. To overcome the cluster bias on the raw clusters with small and sparse fragments, we set the distance of a continuous fragment pair with d>3 and discontinuous fragment pair as dmean+1. Finally, we apply dbscan function (in dbscan R package) on the distance matrix of each raw cluster with parameters eps = dmean and minPts = 2 to obtain clusters. Afterwards, we update distance matrix and dmean and iteratively apply dbscan function until the cluster results are not change. For the raw clusters with only two fragments (denoted as f1 and f2, and the position of f1 is smaller than that of f2), which can not be clustered by dbscan, we calculate three variables: y1=(s2-e1) / mean(l1,l2), y2=(s2-e1) / min(l1,l2), and y3=(s2-e1) / max(l1,l2). We cluster f1 and f2 when one of the following two conditions is satisfied: a. y1<1 and y2<3; b. y3<0.1. Each final cluster is a CNV and its type is set as the type of fragments in the corresponding raw cluster.



# Droplet Digital PCR (ddPCR) Validation

The ddPCR reactions were created following the Bio-Rad QX200™ system manufacturer protocol. A total of 10ng DNA template was mixed with a 2X ddPCR SuperMix for Probes (no dUTP), HindIII-HF enzyme (2U/reaction) (New England BioLabs, MA, USA), 20X primer/probe, (both FAM and HEX-labeled probes) and water to a final volume of 22 μL. Droplets were generated on the AutoDG™ using the manufacturer’s instructions in a 96-well plate, then heat-sealed with a foil seal. PCR amplification was performed using a C1000 Touch thermal cycler with the following conditions for CNV detection: enzyme activation at 95°C for 10 minutes, denaturation and extension at 94°C for 30 seconds and 60°C for 1 minute for a total of 40 cycles, enzyme deactivation at 98°C for 10 minutes, finished with a 4°C hold. Once completed, the 96-well PCR plate was loaded on the QX200™ Droplet Reader. All experiments had at least two normal controls, and a no-template control (NTC) with water. All samples and controls were run in duplicate, and data from any well with less than 8,000 droplets was treated as failed QC and excluded for downstream analysis. Analysis of the ddPCR data utilized QuantaSoft™ software.

# Scripts

Manta v.1.5.0

manta-1.5.0.centos6\_x86\_64/bin/configManta.py \

--config manta-1.5.0.centos6\_x86\_64/bin/configManta.py.ini \

--bam $BAM \

--referenceFasta $FA \

--runDir $WORK\_DIR

python $WORK\_DIR/runWorkflow.py -m local -j 8

Lumpy v. 0.2.13

lumpy-sv/bin/lumpyexpress -B $BAM -o $OUT\_DIR/$SAMPLE.vcf

Delly v. 0.7.9

delly call \

-g $FA \

-x $DELLY/excludeTemplates/human.hg38.excl.tsv \

-o $WORK\_DIR/$SAMPLE.bcf \

$BAM

CNVnator v.0.3.3

CNVNATOR=CNVnator\_v0.3.3/src/cnvnator

PARAM="-chrom chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX chrY -d refs/hg38/chromosomes"

$CNVNATOR -root $SAMPLE.root -tree $BAM -unique $PARAM

$CNVNATOR -root $SAMPLE.root -his 1000 $PARAM

$CNVNATOR -root $SAMPLE.root -stat 1000 $PARAM

$CNVNATOR -root $SAMPLE.root -partition 1000 $PARAM

$CNVNATOR -root $SAMPLE.root -call 1000 $PARAM > $OUT\_DIR/$SAMPLE.cnvnator

cn.MOPS

We were using SVE (<https://github.com/TheJacksonLaboratory/SVE>) to run cn.MOPS.

SVE/bin/sve call -a cnmops -g hg38 -r $FA -o $OUT\_DIR/ $BAM

JAX-CNV v. 0.1.0

JAX-CNV GetCnvSignal \

-b $BAM -f $FA -k $KMER \

> $OUT\_DIR/$SAMPLE.bed \

2> $OUT\_DIR/$SAMPLE.err

Rscript --vanilla JAX-CNV/JaxCNVMerge.R -i $OUT\_DIR/$SAMPLE.bed