A Copy Number Variation Map of the Korean Population

Jinhwa Kong, Jaemoon Shin, Jeehee Yoon*
Department of Computer Engineering
Hallym University
Chuncheon, Korea
{kjh1112, shinjm, jhyoon*}@hallym.ac.kr

Keonbae Lee
Department of Electronic Engineering
Kyonggi University
Suwon, Korea
kblee@kyonggi.ac.kr

***Abstract*—Copy number variations (CNVs) are structural variations associated with human diseases. We are constructing the first high-resolution CNV map for Korean population, using next-generation sequencing data. High quality CNVs are detected in 400 unrelated healthy Korean individuals by running four calling algorithms, and integrated to identify high confidence CNV regions. We present a pipeline system which is composed of various methodologies for construction of CNV maps. This Korean CNV map will contribute to a more accurate clinical interpretation of CNVs in Korean patients and serve as a starting point for the implementation of personalized health care.**

Keywords—Copy number variation map; Next-generation sequencing; Korean population; Pipeline system

I. Introduction

A copy number variation (CNV) is defined as a DNA segment of 50 base pair or larger and present at a variable copy number in comparison with a reference genome. CNV is an important structural variation contributing to genetic diversity and human evolution. It is also known that CNVs are associated with human diseases such as autism, intellectual disability, epilepsy, schizophrenia, obesity, and cancer [1, 2]. However, the CNV association studies have been impeded due to the incomplete population-specific CNV resources and the lack of standardization in terms of CNV detection and statistical analysis methods [3, 4]. Recent studies focusing on CNVs of specific ethnicities report that there are significant amount of population-specific CNVs [5, 6]. However, most of these assays were based on oligonucleotide or SNP arrays having limitations of noisy signal and low resolution.

We are constructing, to our knowledge, a first CNV map in Korean population using next-generation sequencing data. The aim of the CNV map is to catalogue benign CNVs among healthy individuals of the Korean population. We used whole-genome sequencing data of 400 healthy, unrelated Korean people provided by the Korea National Institute of Health (KNIH). To identify the most accurate set of CNVs of each individual genome, CNVs were assessed using carefully selected four different algorithms and retained only if observed by more than one algorithm. We then separately merged overlapping deletions and duplications across the population to generate population-specific CNV regions (CNVR).

978-1-4673-8988-4/17/$31.00 ©2017 IEEE

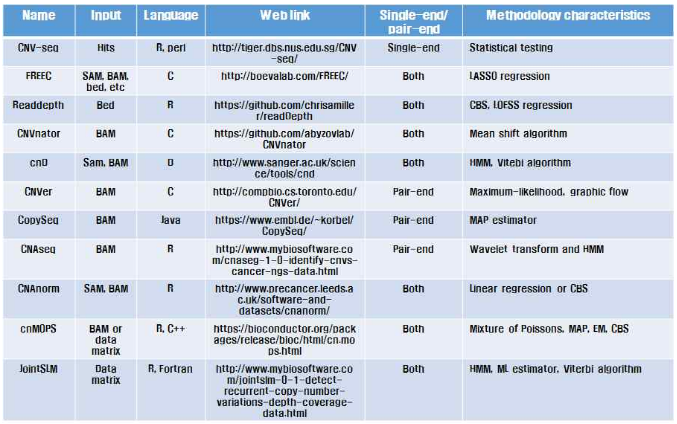
II. Methods

A. Resources

We used whole-genome sequencing data of 400 Korean people provided by the Korea National Institute of Health (KNIH). Sequencing data were generated on an Illumina Hiseq 2000 sequencing platform with an average coverage depth of 30× and alignment data were stored in BAM format. For each sample data we performed a quality control using Picard (http://picard.sourceforge.net/), SAMtools (http://samtools.sourceforge.net/), and GATK (https://software.broadinstitute.org/gatk/) which include the application of filters and the calculation of quality statistics. The post-processed data which passed the strict quality control threshold were only used for all downstream analysis.

Detecting and characterizing CNV from next generation sequencing data is still challenging due to the lack of effective statistical approaches. Currently, no single calling algorithm can detect all types of CNVs in the genome. Table 1 shows a list of selected CNV calling methods. To increase the reliability of CNV detection, we evaluated these publically available CNV calling algorithms using a simulated data set generated by SInC [7] and compared their performances in terms of breakpoint and copy number estimation. We then selected the most efficient four algorithms and analyzed the whole-genome sequencing data of each Korean individual.

TABLE I. SUMMARY OF BIOINFORMATICS TOOLS FOR CNV DETECTION USING WHOLE-GENOME SEQUENCING DATA [8]



Name	Input	Language	Web link	Single-end/ pair-end	Methodology characteristics
CNV-seq	Hits	R, perl	http://tiger.dbs.nus.edu.sg/CNV-seq/	Single-end	Statistical testing
FREEC	SAM, BAM, bed, etc	C	http://boevalab.com/FREEC/	Both	LASSO regression
Readdepth	Bed	R	https://github.com/chrisamiller/readDepth	Both	CBS, LOESS regression
CNVnator	BAM	C	https://github.com/abyzovlab/CNVnator	Both	Mean shift algorithm
cnD	Sam, BAM	D	http://www.sanger.ac.uk/science/tools/cnd	Both	HMM, Vitebi algorithm
CNVer	BAM	C	http://compbio.cs.toronto.edu/CNVer/	Pair-end	Maximum-likelihood, graphic flow
CopySeq	BAM	Java	https://www.embl.de/~korbel/CopySeq/	Pair-end	MAP estimator
CNAseq	BAM	R	http://www.mybiosoftware.com/cnaseg-1-0-identify-cnvs-cancer-ngs-data.html	Pair-end	Wavelet transform and HMM
CNAnorm	SAM, BAM	R	http://www.precancer.leeds.ac.uk/software-and-datasets/cnanorm/	Both	Linear regression or CBS
cnMOPS	BAM or data matrix	R, C++	https://bioconductor.org/packages/release/bioc/html/cn.mops.html	Both	Mixture of Poissons, MAP, EM, CBS
JointSLM	Data matrix	R, Fortran	http://www.mybiosoftware.com/jointslm-0-1-detect-recurrent-copy-number-variations-depth-coverage-data.html	Both	HMM, ML estimator, Viterbi algorithm

B. *A pipeline for generating a copy number variation map*

Our pipeline proceeds in two stages: up and down stages. Fig. 1 shows the overall process of our method. The up stage includes sample selection, quality control (post processing), algorithm-specific CNV detection and sample-specific CNV list generation. The down stage includes CNVR clustering, scoring, filtering and CNV map generation.

First, in the up stage, the BAM file of each sample was processed, sorted and filtered with SAMtools. After removing PCR duplicate reads with MarkDuplicates of Picard, local realignment around indel was performed using the RealignerTargetCreator and IndelRealigner of GATK. We then identified raw CNVs in each sample using four different calling algorithms. Next, within each individual, raw CNVs from all four different algorithms were checked for overlap with each other by at least 50% to detect high quality CNVs that were detected by at least two algorithms. However, whenever two CNVs were observed to overlap reciprocally by 50% or more, the narrower breakpoints were chosen, yielding a shorter and conservative CNV interval. CNVs that passed the comparison were only kept for each individual. We also assigned a score for each CNV (high quality CNV) which reflects the number of algorithms commonly detected the CNV. We then finally generated a high quality CNV list for each sample that contains non-overlapping CNVs that were observed at least twice by the four different algorithms.

In the down stage, CNVs from individual samples were clustered following a 50% reciprocal overlap. As CNVs called from different individuals may estimate partially different CNV intervals, we adopted a CNVR clustering algorithm to identify sets of CNVs in which every possible CNV pair overlaps reciprocally by 50% or more. The CNVs in each cluster were then merged into a CNVR with the outermost coordinates. Two or more neighboring clusters were merged if they overlapped by at least 1 base pair, and the resulting CNVR included the interval of all underlying clusters. In order to score the cluster, we used the score of each CNV component within the cluster. The final score of each cluster which represents the stringency level was then calculated by adding the scores for all CNVs in each cluster.

III. Conclusion

We combined four independently detected raw CNV list from 400 individuals to generate a CNV map at high-resolution in a Korean population. We presented a pipeline system which performs a complete analysis starting from quality control of selected samples to copy number variation map generation. Our population-specific CNV map will serve as a valuable addition to the existing resources for the clinical interpretation of new CNV findings in Korean people.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and future Planning(NRF-2014R1A2A1A11052141, NRF-2017R1A2B4007831) and by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF- 2017R1D1A1B03029186)

References

[1] R. W. Park, T. Kim, S. Kasif, P. J. Park, "Identification of rare germline copy number variations over-represented in five human cancer types," Molecular Cancer, Vol. 14, p. 25, 2015.

[2] H. Stefansson et al., "CNVs conferring risk of autism or schizophrenia affect cognition in controls," Nature, Vol. 505, No. 7483, pp. 361-366, 2014.

[3] M. Zarrei, J. R. MacDonald, D. Merico, S. W. Scherer, "A copy number variation map of the human genome," Nature Reviews Genetics, Vol 16. No. 3, pp. 172-183, 2015.

[4] K. A. Fakhro, N. A. Yousri, et al., "Copy number variations in the genome of the Qatari population," BMC genomics, Vol. 16, No. 1, 2015.

[5] B. Suktitipat, C. Naktang et al., "Copy number variation in Thai population," PloS one, Vol. 9, e104355, 2014.

[6] S. Moon, K. Jung, et al., "KGVDB: a population-based genomic map of CNVs tagged by SNPs in Koreans," Bioinformatics, Vol. 29, No. 11, pp. 1481-1483, 2013.

[7] S. Pattnaik, S. Gupta, A. Rao, B. Panda, "SInC: an accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data," BMC Bioinformatics, Vol. 15, No. 1, 2014.

[8] M. Zhao, Q. Wang, Q. Wang, P. Jia, Z. Zhao, "Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives," BMC Bioinformatics, Vol. 14, (Suppl 11):S1, 2013.

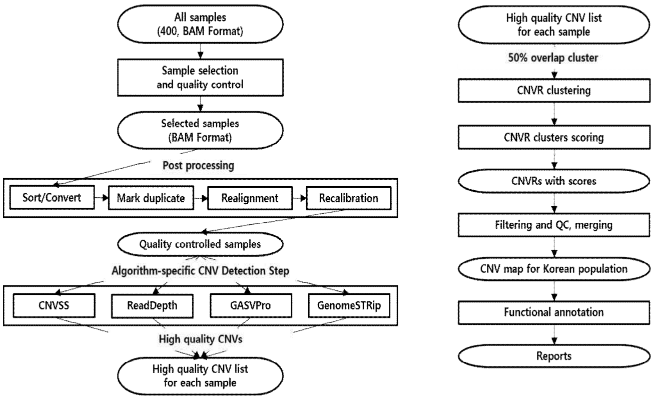


Fig. 1. Overall process of our pipeline system. In the up stage, we discovered high quality CNVs in 400 unrelated healthy Korean individuals using four different calling algorithms. In the following down stage, we integrated these CNVs across the population to generate high confidence population-specific CNV regions.