A Copy Number Variation Map of the Korean Population

Jinhwa Kong[1], Jaemoon Shin[1], Jeehee Yoon[1] and Keonbae Lee[2]
[1]Department of Computer Engineering, Hallym University, Korea
[2]Department of Electronic Engineering, Kyonggi University, Korea

- **Copy number variations (CNVs)**
  - 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.
  - is an important structural variation contributing to genetic diversity and human evolution.
  - are associated with human diseases such as autism, intellectual, disability, epilepsy, schizophrenia, obesity, and cancer.

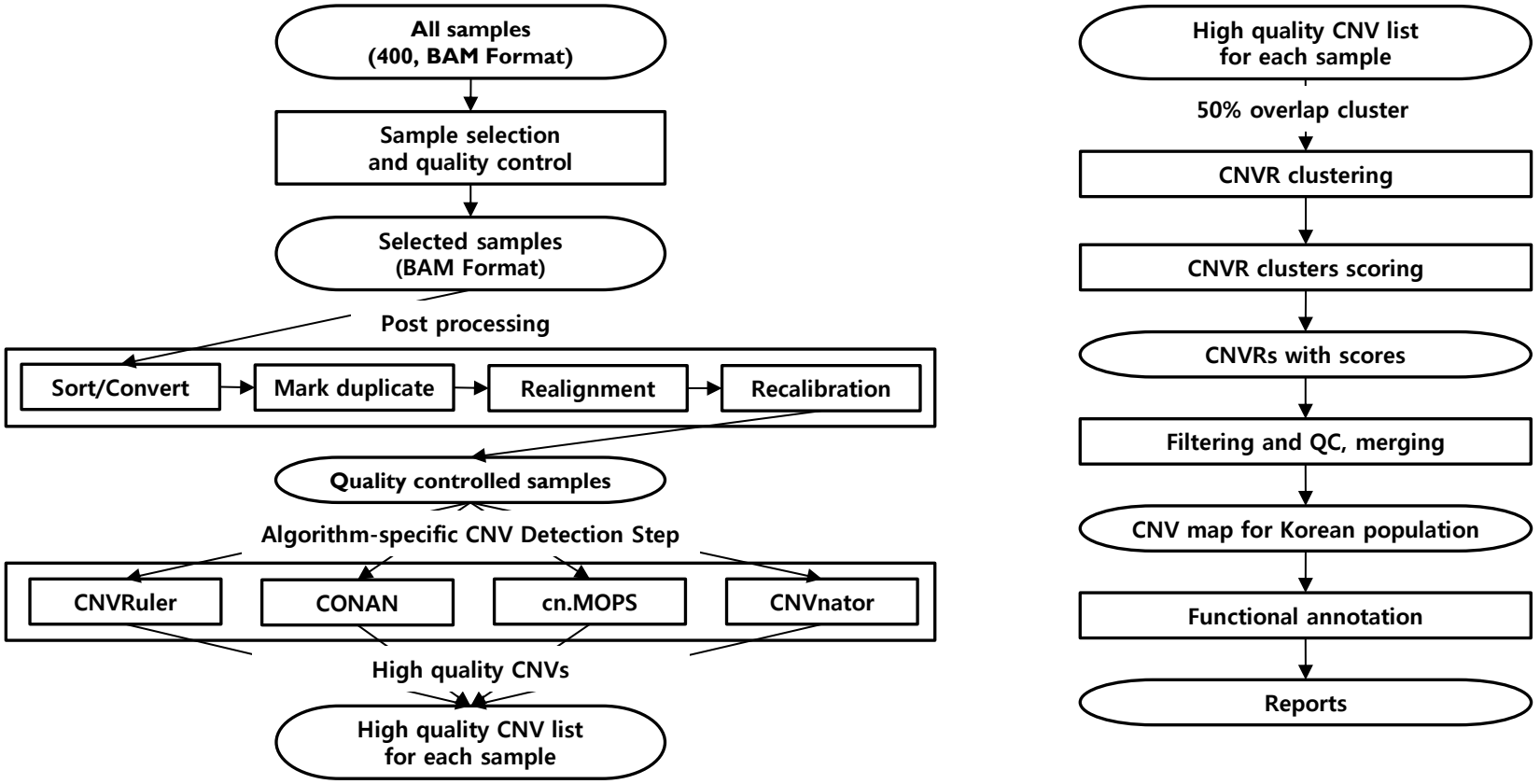
- Our Work
  - We are constructing, to our knowledge, a first CNV map in Korean population using next-generation sequencing data.
  - To identify the most accurate set of CNVs of each individual genome, CNVs were assessed using carefully selected four different algorithms and retained only if observed by more than one algorithm.

Method

- Resources
  - used whole-genome sequencing data of 400 Korean people
    - provided by the Korea National Institute of Health(KNIH).
    - Illumina Hiseq 2000 sequencing platform
    - Average coverage depth of 30x and alignment data were stored in BAM format(48Terabyte)

- A pipeline for generating a copy number variation map
  - Our pipeline proceeds in two stages: up and down stages.
    - Up stage : sample selection, quality control, algorithm-specific CNV detection and sample-specific CNV list generation.
    - Down stage : CNVR clustering, scoring, filtering and CNV map generation.

Overall process of our pipeline system.



- 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.