A Novel Method for Detecting SNV Genotypes from Sequencing Data

Fanny-Dhelia Pajuste
Institute of Computer Science, University of Tartu, Tartu, Estonia
Curriculum: Computer Science, Master's 2013/2014
fanny@ut.ee
Supervisor: Maido Remm
Repository: https://github.com/fannydhelia/SNV-finder

SNVs

SNVs (single nucleotide variants) are the most studied genome variations and linked to multiple traits and diseases. Most of the SNVs have two possible allele variants - nucleotides that may appear in this location of the genome. For example the SNV in human genome shown in Figure 1 has two possible allele variants - some people have nucleotide A in this position in the genome as other people have nucleotide G. Because of the diploid genome, humans have two copies of each chromosome and thus there are three possible genotypes for every bi-allelic SNV. If the person is a homozygote, he or she has the same allele variant in both chromosomes, the nucleotides in the SNV location are different if the person is a heterozygote. Thus, for example for an SNV with two allele variants A and G, the SNV genotype for an individual may be either AA, GG or AG.

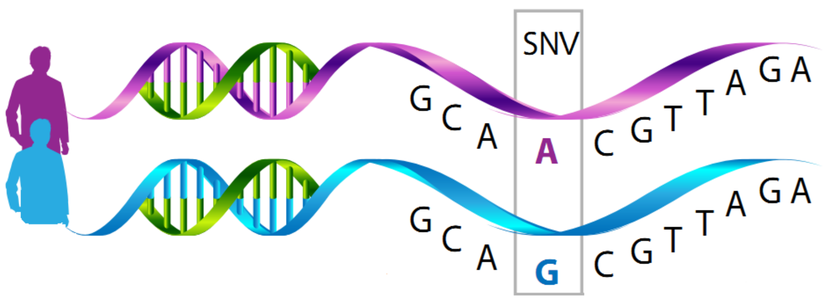

Figure 1: A bi-allelic SNV in human genome with allele variants A and G

Motivation

SNV genotyping is used in several areas such as association studies, evolutional analysis, personal medicine and bacterial strain identification. Therefore, genomic studies depend greatly on the ability of detecting SNV genotypes. With the growing amount of next generation sequencing data that contains millions of short DNA sequences called reads, a great number of different methods have been developed for detecting SNV genotypes from this kind of data. However, most of the methods used for this purpose map the reads to the reference genome. This is a time consuming process and the results are not reliable for certain regions of the genome. The mapped reads are further used for SNV genotyping using either some probabilistic or heuristic methods. Therefore, the pipelines used for SNV genotype detection are slow and cannot identify reliably the genotypes of SNVs located in some regions of the genome.

The Aims of the Project

The aim of this project was to develop a novel method for fast detection of SNV genotypes from these genomic regions that allow reliable genotyping. The second goal of this work was to implement the method and test it on sequencing data for a set of given SNVs.

Method

The method is based on *k*-mers, short DNA sequences with the length of *k*. Exactly *k* sequential *k*-mers cover one nucleotide in the genome. From figure 2 an example can be seen with an SNV with two allele variants. For both of these variants, 8 8-mers cover the nucleotide in the SNV location. If some of these *k*-mers are unique in the genome and occur only in this location and for the given allele variant, then the frequencies of these *k*-mers from the sequencing data can be used to identify the SNV genotype of the individual.

```
...ATAGGCAACGTTAGA...      ...ATAGGCAGCGTTAGA...
  ATAGGCAA                   ATAGGCAG
   TAGGCAAC                   TAGGCAGC
    AGGCAACG                   AGGCAGCG
     GGCAACGT                   GGCAGCGT
      GCAACGTT                   GCAGCGTT
       CAACGTTA                   CAGCGTTA
        AACGTTAG                   AGCGTTAG
         ACGTTAGA                   GCGTTAGA
```

Figure 2: 8 8-mers covering the nucleotide in the SNV location for allele variants A and G

Implementation

This method is implemented in Python (version 3.4) and uses *k*-mer counting tools from GenomeTester4 toolkit. The implementation has two parts.

1. Creating the list of unique *k*-mer pairs for given SNVs.
   The unique *k*-mers can be detected based on the reference genome.

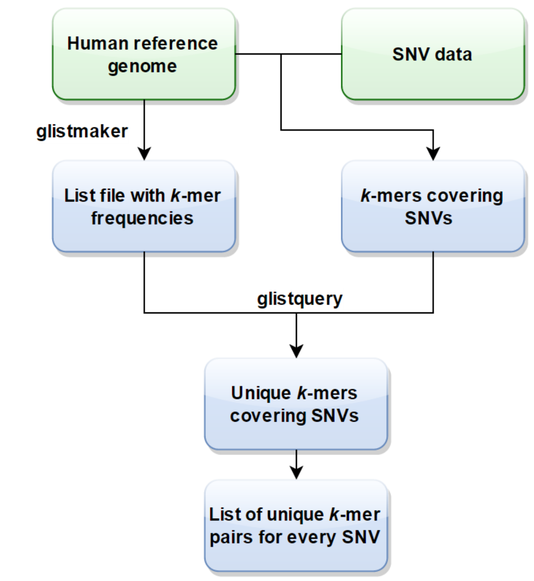

Figure 3: Pipeline for creating unique *k*-mer lists

2. Detecting SNV genotypes from the sequencing data.
   The frequencies of the previously detected unique *k*-mers are found from sequencing data. For every possible genotype, a statistical test is used to determine if the found *k*-mer frequencies could be seen for this particular genotype.

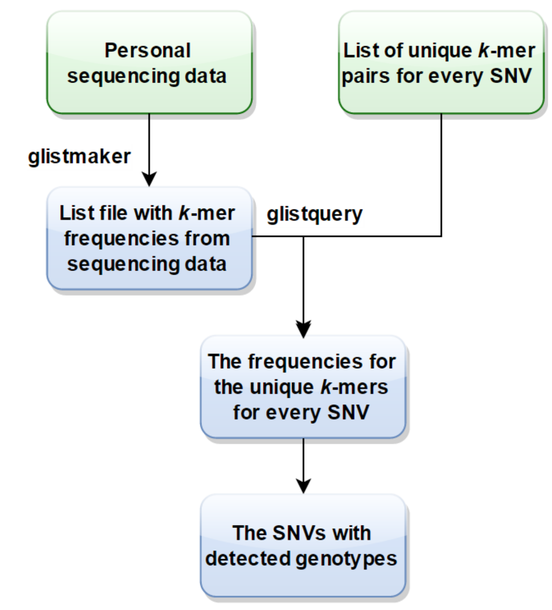

Figure 4: Pipeline for detecting SNV genotypes

Testing

The method was tested on simulated data (30x coverage, read length 102) and real data (10x coverage, read length 76, from 1000 Genome Project). The unique *k*-mers were found based on human reference genome and some additional sequences of other possible SNV allele variants. From 40 million SNVs filtered from Ensembl Human Short Variations database, about 35 million had unique *k*-mer pairs and could therefore be identified. The genotypes were detected for these 35 million SNVs from both simulated and real data.

Time and Memory Usage

The process of detecting SNV genotypes from sequencing data given the unique *k*-mer lists took about 7 hours using a maximum of 8 parallel processes. The maximum memory usage of this pipeline was about 200 GB. The memory usage can be reduced by improving the *k*-mer counting tool or using temporary files, in addition the time usage could be further improved by parallelizing the process of finding the frequencies of unique *k*-mers.

This work was supported by the EU ERDF through the Estonian Centre of Excellence in Genomics and by Estonian Ministry of Education and Research through institutional grant IUT34-11.