UNIVERSITY OF TARTU

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Institute of Computer Science

Computer Science Curriculum

Fanny-Dhelia Pajuste

A novel method for detecting SNV genotypes from personal genome sequencing data

Master's Thesis (30 ECTS)

Supervisor: Maido Remm, PhD

Tartu 2015

A novel method for detecting SNV genotypes from personal genome sequencing data

Abstract:

The genome variation studies are important for many areas like personal medicine, evolutionary analysis or bacterial strain identification. The single nucleotide variants (SNVs) are the most thoroughly studied variations in the genome, associated with different traits and diseases. Genomic studies depend greatly on the ability of detecting the allele variants of these variations present in personal genome. However, the methods used for calling SNV genotypes from personal sequencing data are not very fast nor reliable. The aim of this master's thesis was to develop a novel method for detecting SNV genotypes fast and reliably with a new approach that allows omitting the often error-prone step of read mapping used in the general variant calling pipelines.

A k-mer based approach was introduced in this study for detecting SNV genotypes. A method was developed for using the unique k-mers covering the SNV locations for different allele variants to identify the genotypes of these SNVs. A program was created for compiling a list of unique k-mers for the allele variants of given SNVs and the method was tested using a program for detecting the genotype of these SNVs from the personal genome sequencing data.

The method introduced in this study was tested on both simulated and real sequencing data and the memory and time usage was measured. Some recommendations were made for future work to reduce the time usage of the program as well as improving the detection of SNV genotypes.

Keywords: bioinformatics, personal sequencing data, genome variations, SNV, k-mer

Uudne meetod SNV genotüüpide määramiseks personaalse genoomi sekveneerimisandmetest

Lühikokkuvõte:

Genoomi variatsioonide uuringud on olulised mitme erineva valdkonna jaoks nagu näiteks personaalne meditsiin, evolutsiooniline analüüs või bakteritüvede tuvastamine. SNV-d, üksiku nukleotiidi variandid, on kõige põhjalikumalt uuritud variatsioonid genoomis ning seostatud mitmete tunnuste ja haigustega. Genoomiuuringud sõltuvad olulisel määral genoomist antud variatsioonide alleeli variantide määramise võimekusest, olemasolevad SNV genotüüpide määramise meetodid on aga võrdlemisi aeglased ja ebausaldusväärsed. Käesoleva magistritöö eesmärk on arendada välja uudne meetod SNV genotüüpide määramiseks kiiresti ning usaldusväärselt, jättes vahele kõige vigaderohkema etapi tavalisest SNV määramise töövoost.

Selles töös tutvustati uut, k-meeridel põhinevat lähenemist SNV genotüüpide määramiseks. Arendati välja meetod SNV asukohti katvate unikaalsete k-meeride kasutamiseks antud SNV-de alleeli variantide leidmiseks. Töö käigus loodi programmid etteantud SNV-de jaoks unikaalsete k-meeride leidmiseks ning personaalse genoomi sekveneerimisandmetest genotüübi määramise metoodika testimiseks.

Tutvustatud meetodit testiti nii simuleeritud kui reaalsete sekveneerimisandmetega, ühtlasi mõõdeti programmi aja- ja mälukasutust. Tulevaseks tööks toodi välja ka mõned soovitused programmi ajakulu vähendamiseks ning sekveneerimisandmetest määratud genotüüpide arvu suurendamiseks.

Võtmesõnad: bioinformaatika, personaalsed sekveneerimisandmed, genoomi variatsioonid, SNV, k-meer

Contents

Introduction

The many different types of variations appearing in the DNA are the reason behind the uniqueness of the genome of every individual. Some variations located in the genes may have a great effect on the phenotype, causing different traits, even some major diseases. The genome variation studies have a great importance in many areas such as personal medicine, evolutionary analysis, genetic diversity studies or bacterial strain identification. Many DNA variant detection pipelines have been developed over the past years for these applications, however, there is still a shortage of fast and efficient methods for detecting genome variations reliably.

The aim of this master's thesis is to develop a novel method for variation identification from raw sequencing data without relying on the existing software generally used in variant calling pipelines. The purpose is to use a completely novel approach to detect the variations faster as well as more reliably. This is achieved by skipping the part of the general pipeline that is the most error prone and dependent on the program arguments.

The first section of this work gives a brief overview about the human genome variations, DNA sequencing data and variant calling pipelines. The aims of this study are brought out in the following section. The novel method developed for variant calling is introduced in the third chapter together with a description of the implementation and used data. The fourth section gives an overview of the results of this study. The time and memory usage as well as some other technical aspects are discussed in the fifth section. The conclusion is given in section six.

1 Review of literature

1.1 Human genome variations

Human genome variations are the differences in the human DNA within or between populations. No two human genomes are exactly identical and it is believed that the genomes of two different persons may differ up to 0.5%. The genomes of people from different geographical regions tend to differ more while the genomes of closely related individuals are more similar. DNA variations are the changes that make a person's genome unique and different from the genome of any other person in the world.

There are many types of variations in human genome based on their length and their cause as well as their impact on the person. The most studied variations are the SNVs - single nucleotide variants, insertions - the addition of one or more nucleotides to the genome, deletions - the removal of one or more nucleotides from the genome, Alu elements - repetitive elements in the genome with about 300 base pairs (the pairs of complementary nucleotides) in length. This work focuses on the former type of variants, SNVs. These and other changes in the genome may appear as a result of different mutations occurring either during the division of a cell when DNA is copied or during the meiosis (a type of cell division) in the gametes (sex cells) when parts of the chromosomes are exchanged. The mutations from the latter can be passed down to generations of people. Most of the known SNVs in human genome are very old and have appeared already thousand of years ago. Although some of these variations are very rare, many of the SNVs can be quite common, sometimes found even in about half of the population.

Most of the mutations occur outside of the genes, in the regions of the DNA that do not code proteins. These mutations are usually neutral and do not have any impact on the individual. However, the mutations in different parts of the genes may have a great effect on the functionality of the gene and can thus be harmful. The harmful mutations are less likely to pass on to next generations as they can cause diseases, infertility and other disadvantages that may prevent these individuals from producing offspring. On the other hand, the mutations that give some kind of advantage in the population are more likely to pass on to next generations due to the natural selection.

Human genome variation studies have many different applications in evolutionary analysis, linkage and association studies and personal medicine, to name a few. For

example, the SNV rs1154155 is known to increase the risk for narcolepsy about 1.7 times [Hallmayer et al., 2009], the odds for alcohol dependence are about 1.35 times higher for the G variant allele in rs7590720 SNV [Treutlein et al., 2015] and the risk for the Alzheimer disease can be evaluated based on the genotypes of the SNVs rs429358 and rs7412[Farrer et al., 1997]. Therefore genomic studies are greatly dependent on the ability to detect these mutations in an individual as well as finding new links between the diseases and different variations in human genome. This ability can help finding new causes for different diseases as well as helping people that have a disease or a disposition for it due to some genetic mutations.

1.2 DNA sequencing

A DNA molecule consists of two complementary strands containing four different types of nucleotides - Thymine (T), Adenine (A), Cytosine (C), and Guanine (G). DNA sequencing is the process of determining the order of the nucleotides in a DNA molecule, the DNA sequence. The first breakthrough in DNA sequencing came with Sanger sequencing technique in about 40 years ago. Nowadays, second generation sequencing is the most common sequencing method due to its speed and low cost. This method divides the DNA molecule into small fragments which are then sequenced in millions of parallel functions. The resulting data consists of millions of sequencing reads - short sequenced regions of the genome. The length of the reads is usually about 100 nucleotides. The genome is sequenced with a high sequencing depth, reaching thirty to hundred-fold representation of each nucleotide to decrease the effect of the sequencing errors to the genome. The depth of coverage or the sequencing depth of the data is the average number of reads covering a nucleotide in the genome.

The standard format for storing sequencing reads is the FASTQ format. FASTQ format normally uses four lines per read containing the sequencing id and description, the sequence of the read and quality values for the sequenced base pairs.

1.3 SNV calling pipelines

Most of the SNV calling pipelines map the sequencing reads to the reference genome, a standard reference sequence used as a representative example of human genome. Human genome is a diploid genome: there are two copies of each chromosome, one from the

mother and one from the father. The human reference genome is a haploid compilation of DNA sequences of different persons, which means that each chromosome is represented only once and the sequence does not correspond to any actual individual.

The mapped sequencing reads are used to detect the differences between the reference genome and the genome of the given individual to identify SNVs. If there were no sequencing errors, then for a high depth of coverage, it would be rather easy to detect the SNV genotype. If the nucleotide in the SNV location would be the same for all the reads covering this location, lets say A for example, then the person would have this allele variant in both of the chromosomes, i.e. the person would be a homozygote. If half of the reads would have one and half another nucleotide in this position, for example A and C, then the person would be a heterozygote which means that the allele variant would be different in the two chromosomes. Either way it would be easy to say which genotype the individual has: AA, CC or AC. If the genotype cannot be detected, it is marked as NN.

However, the sequencing reads contain errors and bias, and sequencing with a high depth of coverage is expensive, which makes the task a lot harder in reality. Usually some probabilistic methods are used to estimate the probability of each genotype. Bayes' theorem is often used to find the probability of a genotype being the true genotype given the observed data. SNV calling methods differ based on the algorithm for calculating the prior probabilities of the genotypes and modelling the distribution of the observed data. Some of the most used probabilistic methods for detecting SNV genotypes are implemented in the SAMtools[Li, 2011] package and the Genome Analysis Toolkit[Depristo et al., 2011] (GATK). In cases where the data does not correspond to the assumptions of the probabilistic models, heuristic methods are preferred for detecting the SNV genotype. These methods use different heuristic factors like minimum allele counts or read quality cut-offs to determine the right genotype.

1.4 Overview of the read mappers

Although there is a great number of different mappers[Fonseca et al., 2012], there are some basic methods used for most of these tools. For example, some programs hash the read sequences and scan through the reference sequence, which is efficient in memory, but can take a lot of time as the whole genome might have to be scanned. Another group

of programs hash the genome instead of the reads and can be easily parallelized to run with less time, but need a lot of memory to index the whole genome. Apart from these and some other approaches, there is a newer group of aligners using Burrows-Wheeler transform[Healy et al., 2003]. This approach is often preferred for read mapping as it is quite fast as well as efficient in memory usage.

Two main methods used for mapping the reads to the genome are both based on the Burrows-Wheeler transform: BWA[Li, 2013] and Bowtie2[Langmead and Salzberg, 2015]. BWA, the Burrows-Wheeler Alignment method, uses the backward search with Burrows-Wheeler Transform to align the reads against the reference genome. The backward search uses the suffix array of the prefix trie and is equivalent to the top-down traversal on the prefix trie itself, but without holding the whole trie in memory. BWA also allows mismatches and gaps, which is implemented using a bounded traversal and backtracking. The memory usage can be reduced by only using two bits per nucleotide and not holding the whole suffix array in memory as it can be reconstructed from only a part of it.

Bowtie indexes the reference genome based on the Burrows-Wheeler transform and FM index. In addition to the usual exact-matching algorithm to search in a FM index, this method uses backtracking to allow mismatches. Excessive backtracking is avoided by another extension called double indexing.

Although both of these methods are quite fast and efficient in memory compared to the other mappers, they are still unable to map all the reads and have a considerable percentage of erroneous alignments. Also, the result is greatly dependent on the parameters used when running the tools, which makes it hard to reproduce the same results. The mapping process is quite error prone and time consuming, so are the pipelines used for SNV calling. In this work a novel method is introduced and its applicability for calling SNVs from raw personal sequencing data without the read alignment process is tested.

2 The aims of the study

The main goal of this master's thesis was to develop a novel method for detecting SNV genotypes from raw personal sequencing data fast and reliably. The method should be independent from the read mapping tools that can give unreliable results which vary greatly depending on the argument values.

The first aim of the study was to create a program for finding unique k-mers, the short DNA sequences with the length of k, to describe the SNV variations in the human genome. A list of these k-mers and the corresponding SNVs could be formed using this program.

The second aim of this work was to develop a tool for detecting SNV genotypes from personal sequencing data of a given individual using the list of the unique k-mers. The program should use a statistical framework to determine the right genotype of the SNV based on the frequencies of these k-mers.

3 Method and implementation

3.1 SNV data

Three different datasets of SNVs were used in this study. First, 719 666 SNVs from HumanOmniExpress chip were used for testing the program for finding unique k-mers and for determining the appropriate k-mer length. This set contains a great amount of common single nucleotide variants that are often used in genome-wide association studies.

The second dataset used in this study contained the SNVs from the Homo sapiens Short Variation set (GRCh38.p2) from Ensembl[Cunningham et al., 2015] Variation 79 database. About 55 million single nucleotide variants were drawn from the dataset containing SNVs and indels. These SNVs were further filtered, removing those that did not contain the variant from reference genome, that were assigned to multiple locations or that would have had more than one other SNV in a 25-mer. The latter were removed to avoid creating k-mers with all possible combinations of the allele variants of different SNVs. Some locations were merged when they occurred in the dataset multiple times with different SNV id-s, but same allele variants. Those locations which occurred with different SNV id-s and also different allele variants were removed. The final dataset had about 40 million SNVs with two allele variants, these SNVs were used for compiling a list of unique k-mers and testing the program for detecting SNV genotypes.

The third dataset contained the Estonian specific SNVs drawn from the variations of 57 individuals sequenced by Estonian Genome Centre. This data was used in addition to the 40 million SNVs for creating a list of unique k-mers to reduce the effect of the coexistence of different SNVs to the results of genotyping. This dataset contained rare Estonian specific SNVs that might not be present in the previous set of 40 million SNVs, but could be relevant when studying the genomes of Estonian individuals.

3.2 Creating a list of SNVs with unique k-mers

For every SNV in the genome, exactly k sequential k-mers cover its location. For a SNV with two possible allele variants, there are k pairs of k-mers so that both k-mers in the pair start from the same position, but have a different nucleotide at the location of the SNV. As some SNVs have more than two allele variants marked, then for these variations the sets of three or even four k-mers can be used. An example of a SNV with two possible

allele variants (C and A) and the corresponding 8 8-mer pairs can be seen in Figure 1.

If a k-mer pair for a SNV is unique in the genome, i.e. neither of these k-mers occur in more than one place in the genome, then it can be used to identify the genotype of this SNV for an individual. More precisely, if a k-mer that is seen only in the location of a SNV in the genome for one SNV allele variant, is found in the sequencing data in an expected amount, it can be concluded that the person has this allele variant. Therefore, the unique k-mers can be used to determine which allele variant the person has and if the person is homozygote (the same variant in both chromosomes) or heterozygote (has different variants in chromosomes).

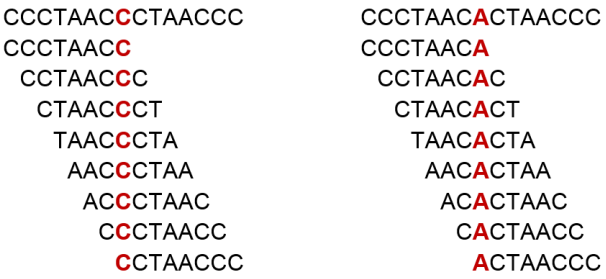

Figure 1: 8-mers covering a location of SNV with two possible variants C and A

The pipeline for finding the unique k-mer pairs for SNVs contained the tools from the GenomeTester4[Kaplinski et al., 2015] toolkit. First, the k-mer counting tool GListMaker was used to build a list file of all the k-mers present in the reference genome. The human reference genome version GRCh38 was used in this study. GListMaker takes either the sequencing reads in FASTQ format or longer sequences in the FASTA format, parses the files with a sliding window of a given length and counts all the k-mers with this length present in the given files by sorting the array of found k-mers and walking through the array. The tool outputs a binary list file with an array of k-mers stored as 64-bit unsigned integers and their frequencies as 32-bit unsigned integers. As every nucleotide is represented by two bits, the longest k-mer length allowed by this tool is 32. The list file contains only the smaller integer from this of a k-mer and its reverse complement (the complementary sequence in the other strand of the DNA), therefore both the k-mer and its reverse complement are present in the list file only once, with the sum of their frequencies.

In the next step, the k-mers covering the SNVs were found for every allele variant. The

location of the SNV was used to get the surrounding region (the location of SNV itself with k-1 bp from both sides) from the reference genome. The regions for other known SNV variants were found by exchanging the nucleotide in the SNV location. Then the GListQuery tool was used to divide the sequences into k-mers and find their frequencies from the list file created earlier using the GListMaker tool. Every k-mer (or its reverse complement) is searched from the list file using a binary search. From these results, only those k-mers were drawn that did not have a frequency bigger than 1 in the list file. The unique k-mer pairs (i.e. the k-mers starting from the same location, but with different allele variants) were then found for the given SNVs. The pipeline for finding these unique k-mer pairs is also described in the flowchart in Figure 2.

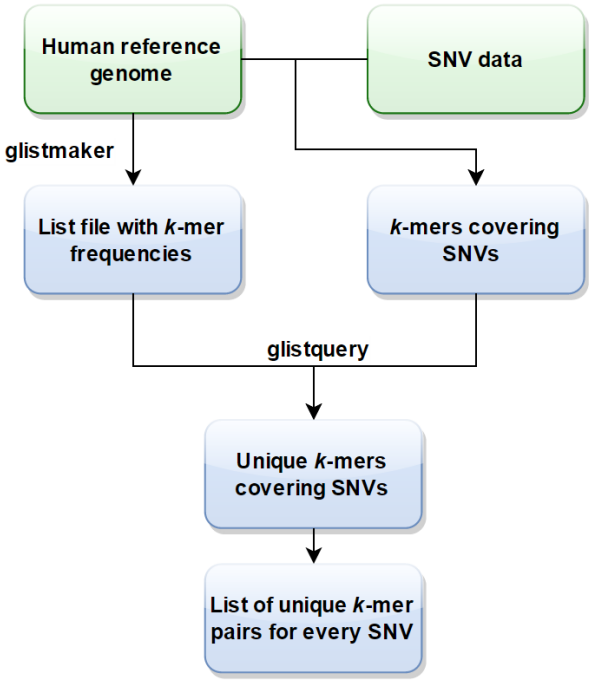

Figure 2: Creating list of unique k-mers

The SNVs that had unique k-mer pairs were written to a file containing the id of the SNV, the location in the genome, known allele variants and a list of unique k-mer pairs. For every SNV the file also contained an integer that is a representation of how these k-mers are located in relation to the SNV: if this number is converted to binary, then for every k-mer pair that covers this SNV, 1 represents its presence (uniqueness), 0 its absence (not unique in the genome). This information can be used for detecting the SNV genotype of an individual.

The pipeline can be further improved by changing the GListQuery argument of the number of allowed mismatches to 1 when finding the unique k-mers from the list file. This changes the conditions so that to be considered unique, the sum of the frequencies of the given k-mer and all these k-mers that have 1 mismatch compared to it, can be at most 1. Using the k-mers found this way would make the results less affected by sequencing errors and other SNVs.

3.3 Detecting SNV genotypes from personal sequencing data

3.3.1 Counting unique k-mers

The first step for detecting the SNV genotype from personal sequencing data is to get the frequencies of the k-mers in the list of unique k-mer pairs found for this SNV. GListMaker tool is used to create a binary list file from the files in FASTQ format given by the user with the reads of a sequenced individual. GListCompare tool from the GenomeTester4 toolkit may be used to take an intersection of these k-mers and the unique k-mers found for SNVs, to create a smaller list file for faster k-mer searching. The frequencies are drawn using the GListQuery tool and later used for detecting the genotype of the given SNVs. A flowchart of the pipeline can be seen in Figure 3.

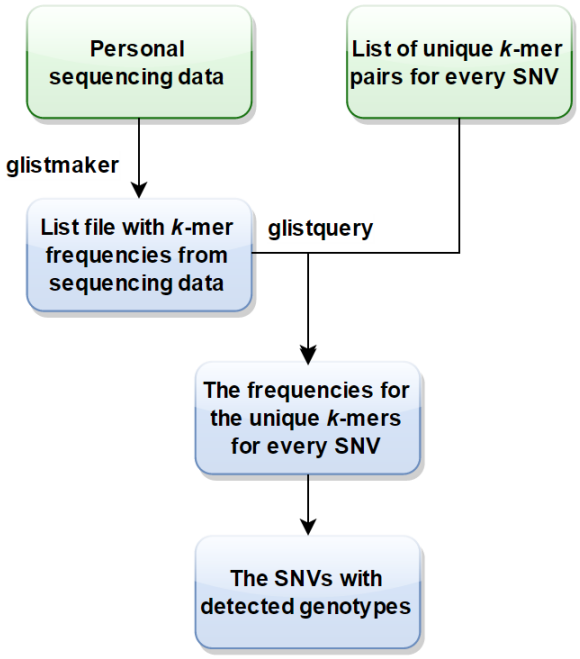

Figure 3: Using found unique k-mers for SNVs to detect the genotype in an individual

3.3.2 Statistical framework

The genotype is the combination of the allele variants of the two chromosomes. For instance for a SNV with two possible allele variants C and A, the genotype in an individual can be either CC, CA or AA. The method introduced in this work uses the frequencies of the unique k-mer pairs and the information about how these k-mers are located for detecting the SNV genotype. To identify the particular genotype for which the observed frequencies could be seen, the expected value of the frequencies for every possible genotype must be found. This section gives an overview of the notations and derived formulas for calculating the expected value and variance of the frequencies of the k-mers for a certain genotype.

Let S_i, $i \in 1, .., L$ where L is the read length in base pairs, be the number of reads starting from a certain position so that it would cover i base pairs beginning from the SNV location. To simplify, an example can be seen in Figure 4. Assuming that reads are distributed randomly across the genome and they do not cluster, the number of times a base is sequenced, i.e. the coverage, follows Poisson distribution. Therefore, $S_i \sim Poi(\lambda)$ where $\lambda = \frac{C}{L}$ and C represents the coverage. As $C = \frac{N \cdot L}{G}$ where N is the number of reads sequenced and G is the length of the target genome in base pairs, then $\lambda = \frac{N \cdot L/G}{L} = \frac{N}{G}$. As the target genome used in this study is the human genome, then $G \approx 3 \cdot 10^9$.

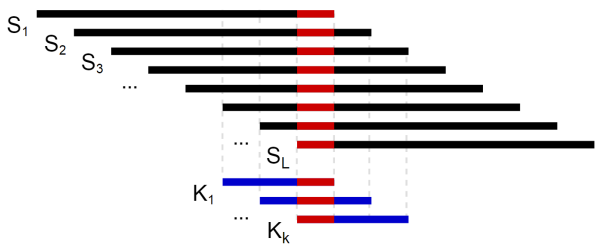

Figure 4: Reads (black) and k-mers (blue) covering a SNV location (red)

Let K_j, $j \in 1, .., k$ be the value that shows if a k-mer pair is described as unique in the genome:

$$K_j = \begin{cases} 1, & \text{if } j^{th} \text{ } k\text{-mer pair is unique} \\ 0, & \text{otherwise.} \end{cases}$$

Let F_j be the frequency of the j^{th} k-mer, i.e. the number of times this k-mer was seen in the sequencing data. The frequency of a k-mer can be found as a sum of the

number of reads containing this k-mer: $F_j = S_j + S_{j+1} + ... + S_{L-(k-j)}$. So, the sum of the frequencies of all the k-mers covering a SNV for one allele variant would be:

$$F_1+...+F_k = S_1+2S_2+...+(k-1)S_{k-1}+\overbrace{kS_k + ... + kS_{L-(k-1)}}^{L-2(k-1)}+(k-1)S_{L-(k-2)}+...+2S_{L-1}+S_L.$$

Let T be the sum of the frequencies of these k-mers that are marked as unique. It can be seen that

$$T = K_1F_1 + ... + K_kF_k = K_1S_1 + (K_1 + K_2)S_2 + ... + (K_1 + ...K_{k-1})S_{k-1}+$$
$$+\overbrace{(K_1 + ... + K_k)S_k + ... + (K_1 + ... + K_k)S_{L-(k-1)}}^{L-2(k-1)}+(K_2 + ... + K_k)S_{L-(k-2)} + ... + K_kS_L.$$

The expected value of T is then

$$E(T) = E(K_1S_1) + E[(K_1 + K_2)S_2] + ... + E[(K_1 + ...K_{k-1})S_{k-1}]+$$
$$+\overbrace{E[(K_1 + ... + K_k)S_k] + ... + E[(K_1 + ... + K_k)S_{L-(k-1)}]}^{L-2(k-1)}+$$
$$+ E[(K_2 + ... + K_k)S_{L-(k-2)}] + ... + E[K_kS_L].$$

From here we get

$$E(T) = K_1E(S_1) + (K_1 + K_2)E(S_2) + ... + (K_1 + ...K_{k-1})E(S_{k-1})+$$
$$+\overbrace{(K_1 + ... + K_k)E(S_k) + ... + (K_1 + ... + K_k)E(S_{L-(k-1)})}^{L-2(k-1)}+$$
$$+ (K_2 + ... + K_k)E(S_{L-(k-2)}) + ... + K_kE(S_L).$$

Since $S_i \sim Poi(\lambda)$, then $E(S_i) = \lambda$, therefore

$$E(T) = K_1\lambda + (K_1 + K_2)\lambda + ... + (K_1 + ...K_{k-1})\lambda+$$
$$+\overbrace{(K_1 + ... + K_k)\lambda + ... + (K_1 + ... + K_k)\lambda}^{L-2(k-1)}+(K_2 + ... + K_k)\lambda] + ... + K_k\lambda.$$

After simplifying, the expected value can be found as following:

$$E(T) = \lambda(L - k + 1)\sum_{i=1}^{k} K_i.$$

Since we assumed that S_i, $i \in 1, .., L$ are independent, then D(T) can be found similarly to E(T):

$$D(T) = D(K_1S_1) + D[(K_1 + K_2)S_2] + ... + D[(K_1 + ...K_{k-1})S_{k-1}]+$$
$$+\overbrace{D[(K_1 + ... + K_k)S_k] + ... + D[(K_1 + ... + K_k)S_{L-(k-1)}]}^{L-2(k-1)}+$$
$$+ D[(K_2 + ... + K_k)S_{L-(k-2)}] + ... + D(K_kS_L),$$

$$D(T) = K_1^2 D(S_1) + (K_1 + K_2)^2 D(S_2) + ... + (K_1 + ...K_{k-1})^2 D(S_{k-1}) +$$
$$+ \overbrace{(K_1 + ... + K_k)^2 D(S_k) + ... + (K_1 + ... + K_k)^2 D(S_{L-(k-1)})}^{L-2(k-1)} +$$
$$+ (K_2 + ... + K_k)^2 D(S_{L-(k-2)}) + ... + K_k^2 D(S_L).$$

Since $S_i \sim Poi(\lambda)$, then also $D(S_i) = \lambda$, so

$$D(T) = K_1^2\lambda + (K_1 + K_2)^2\lambda + ... + (K_1 + ...K_{k-1})^2\lambda +$$
$$+ \overbrace{(K_1 + ... + K_k)^2\lambda + ... + (K_1 + ... + K_k)^2\lambda}^{L-2(k-1)} + (K_2 + ... + K_k)^2\lambda] + ... + K_k^2\lambda.$$

After simplifying, the variance can be found as

$$D(T) = \lambda[\sum_{i=1}^{k-1}(\sum_{j=1}^{i} K_j)^2 + [L - 2(k-1)](\sum_{j=1}^{i} K_j)^2 + \sum_{i=2}^{k}(\sum_{j=i}^{k} K_j)^2].$$

3.3.3 Detecting SNV genotype

Lets assume a SNV has two allele variants denoted by V_1 and V_2. A person can then have either only one of these variants (i.e. is homozygote) or both of them (is heterozygote) depending on if the variants are the same for both chromosomes. So, the genotype of this SNV in a person can be either V_1V_1, V_1V_2 or V_2V_2. The method developed in this study uses three competitive null hypothesis to test for each genotype if the observed frequencies could be seen if this was the true genotype.

For the first case, the genotype V_1V_1, the sum of the frequencies of the k-mers for allele variant V_1 would be approximately from a normal distribution with the mean $\mu = E(T_{V_1})$ and the variance $\sigma^2 = D(T_{V_1})$. In the second case the number of reads containing these k-mers should be two times smaller for the allele variant V_1, thus we should use 0.5λ instead of λ, so the sum of the frequencies would be from the following distribution: $N(0.5\mu, 0.5\sigma^2)$. As there should be no reads containing k-mer for variant V_1 in the third case, then in this case $T_{V_1} \sim N(0, 0)$. So, for these three cases, the sum of the frequencies of the k-mers for variant V_1 would be from the following distributions:

$$V_1V_1\text{: } T_{V_1} \sim N(\mu, \sigma^2)$$
$$V_1V_2\text{: } T_{V_1} \sim N(0.5\mu, 0.5\sigma^2)$$
$$V_2V_2\text{: } T_{V_1} \sim N(0, 0)$$

As the λ value is the same for both variants and also the same k-mers are used (because uniqueness was determined for pairs of k-mers), then $E(T_{V_2}) = E(T_{V_1}) = \mu$ and

$D(T_{V_2}) = D(T_{V_1}) = \sigma^2$. The sum of the frequencies for the k-mers for variant V_2 would be thus from these distributions:

$$V_1V_1\text{: } T_{V_2} \sim N(0, 0)$$
$$V_1V_2\text{: } T_{V_2} \sim N(0.5\mu, 0.5\sigma^2)$$
$$V_2V_2\text{: } T_{V_2} \sim N(\mu, \sigma^2).$$

Now, the difference $T_{V_1} - T_{V_2}$ would be for these three cases from the distributions $N(\mu - 0, \sigma^2 + 0)$, $N(0.5\mu - 0.5\mu, 0.5\sigma + 0.5\sigma)$ and $N(0 - \mu, 0 + \sigma^2)$ respectively, therefore:

$$V_1V_1\text{: } T_{V_1} - T_{V_2} \sim N(\mu, \sigma^2)$$
$$V_1V_2\text{: } T_{V_1} - T_{V_2} \sim N(0, \sigma^2)$$
$$V_2V_2\text{: } T_{V_1} - T_{V_2} \sim N(-\mu, \sigma^2).$$

These differences can be calculated using the frequencies of the unique k-mers for both allele variants. Then, for every genotype, Z-test can be used to determine if the difference of the frequencies could be from this particular distribution. The value of the test statistic can be computed for the three cases in the following way:

$$V_1V_1\text{: } Z = \frac{T_{V_1} - T_{V_2} - \mu}{\sigma} \sim N(0, 1)$$
$$V_1V_2\text{: } Z = \frac{T_{V_1} - T_{V_2}}{\sigma} \sim N(0, 1)$$
$$V_2V_2\text{: } Z = \frac{T_{V_1} - T_{V_2} + \mu}{\sigma} \sim N(0, 1).$$

The corresponding p-values can be found using a z-table. As the null hypothesis is that the seen observation (found difference) is from this certain normal distribution, then if a p-value is bigger than the given significance level for one of these three cases, it can be assumed that the person has the variants of this particular case - either V_1V_1, V_1V_2 or V_2V_2. In cases where none or multiple p-values are significant, the genotype cannot be detected using the given significance level.

3.3.4 Implementation and source code

The programs for finding unique k-mer pairs and for detecting the SNV genotypes from sequencing data were implemented in Python (version 3.4). The source code is available in the following github repository: https://github.com/fannydhelia/SNV-finder.

4 Results

4.1 Determination of optimal k-mer length

The first task in this study was to choose an appropriate k-mer length for detecting SNV genotypes from personal genome sequencing data. For this, a small amount of the common SNVs used in genome-wide association studies from HumanOmniExpress chip, was used. For every SNV, all the unique k-mer pairs (or sets) were found for k-mer lengths 16, 20, 24, 28 and 32 to evaluate the fraction of SNVs that could be detected when using different k-mer lengths. The maximum length value that could be used was 32 as the GListMaker, GListQuery and GListCompare tools from the GenomeTester4 toolkit used in this pipeline do not allow using longer k-mers. The unique k-mer pairs were found based on the human reference genome.

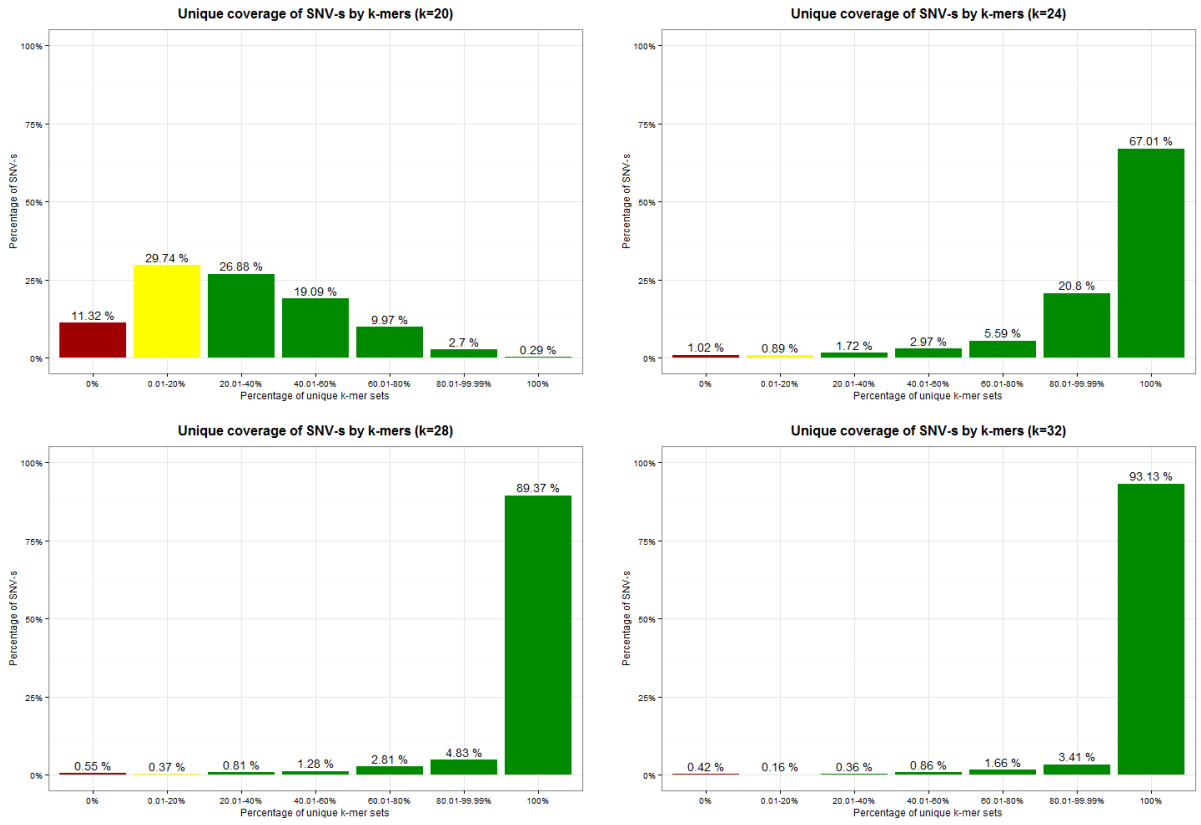

Figure 5: Unique k-mer coverage for different k-mer lengths (found with one mismatch). The plots show the fraction of the SNVs with the given percentage of unique k-mer pairs. Red indicates the SNVs that could not be detected, SNVs with unique k-mers are in the green or yellow (if there were only a few unique k-mers) parts.

These results already showed that values smaller or equal to 16 could not be used as the k-mer length since 36% of the SNV-s had no unique 16-mer sets. For the other lengths (20, 24, 28 and 32), the unique k-mers were also found using one mismatch, which means that the k-mer was unique if the sum of the frequencies of the given k-mer itself and all these k-mers that differed in only one nucleotide, would not be bigger than 1 (only one of them could be present in the genome). This was important for considering the possible impact of other SNVs or sequencing errors to the uniqueness of the k-mers. Figure 5 shows the results on the fraction of SNVs with or without unique k-mer pairs for the given lengths using one mismatch.

It can be seen that when using the k-mer length 20, about 11% of the SNVs had no unique k-mers and approximately 30% had only a few. For k=24, most of the SNVs could be detected. As the memory usage increases with the k-mer length and using k-mers longer than 24 base pairs would not give any significant advantage according to these results, the appropriate k-mer length for this method should be between 24 and 28 for human genome. The k-mer length chosen in this study for compiling the list of unique k-mer pairs and testing the detection of SNV genotypes was 25.

4.2 Compilation of the list of unique k-mers for known variants

To create a list of the unique k-mer pairs for all known variants, the unique 25-mers were found for the 40 million SNVs filtered from the database of human shot variations. Finding the unique k-mers based on the reference genome only, the situations where the allele variants of other SNVs would create the same k-mers as an allele variant of the given SNV, would be ignored. For this reason, the list files with k-mer frequencies were not created only from the reference sequence, but also using the sequences of the SNV surrounding regions for the allele variants not present in the reference genome.

First, the k-mers with the allele variants not in the reference genome for the same 40 million SNVs were added to the list file. Next, another list file was created by adding the sequences with the SNVs of the 57 individuals from Estonian Genome Centre to the reference genome to also consider the variations specific to Estonians. Using population specific unique k-mers could improve the results when detecting the genotypes. The fractions of SNVs with different number of unique k-mer pairs based on these lists can be seen in Figure 6 and 7.

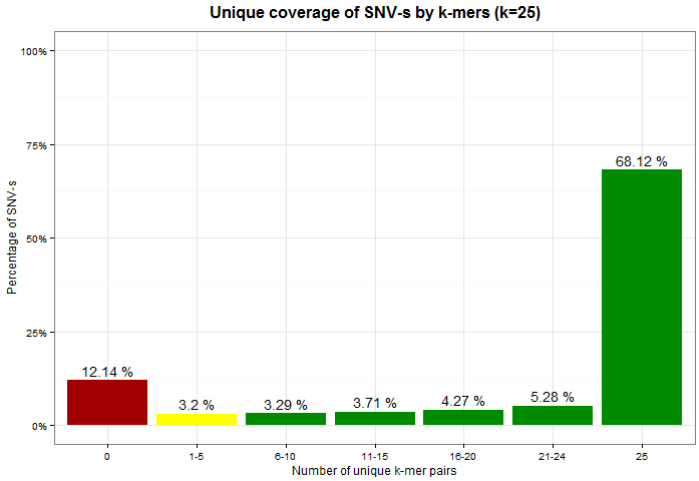

Figure 6: Unique k-mer coverage found based on both the reference genome and the allele variants of the 40 million SNVs. The plots show the fraction of the SNVs with the given number of unique k-mer pairs. Red indicates the SNVs that could not be detected, SNVs with unique k-mers are in the green or yellow parts.

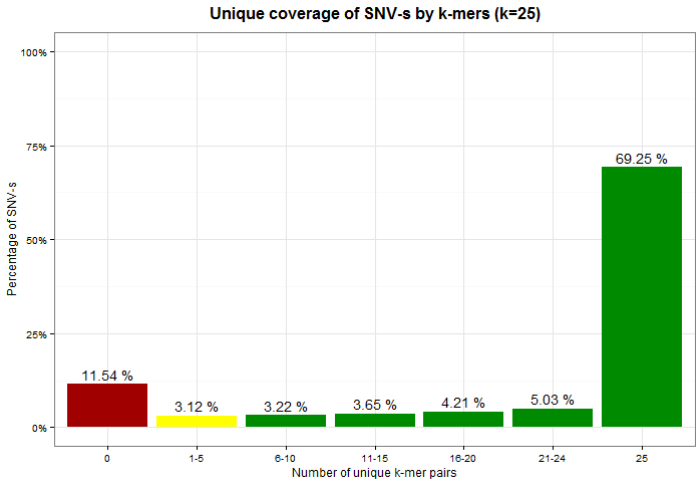

Figure 7: Unique k-mer coverage found based on both the reference genome and the allele variants of the SNVs from 57 Estonian individuals. The plots show the fraction of the SNVs with the given number of unique k-mer pairs. Red indicates the SNVs that could not be detected, SNVs with unique k-mers are in the green or yellow parts.

The intersection of the unique k-mers found for these two datasets was used in this study to compile a list of SNVs with unique k-mer pairs. The number of unique k-mer pairs for the 40 million SNVs based on the intersection of the previous results can be seen in Figure 8.

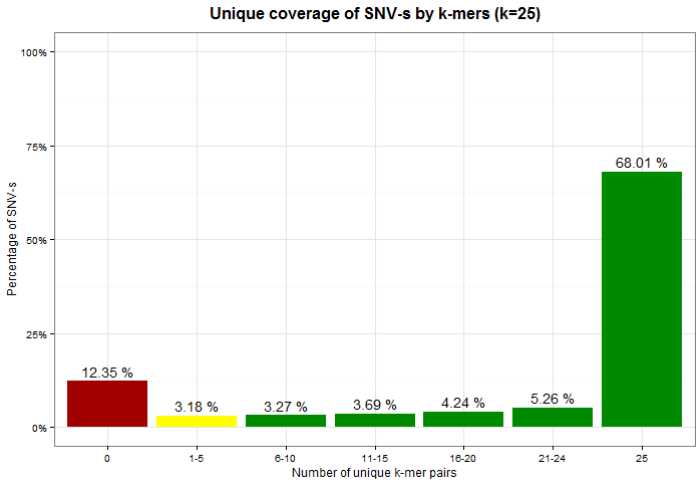

Figure 8: Unique k-mer coverage found based on both the reference genome and other SNV allele variants. The results are found using the allele variants of the 40 million SNVs as well as the allele variants of the SNVs of 57 Estonian individuals. The plots show the fraction of the SNVs with the given number of unique k-mer pairs. Red indicates the SNVs that could not be detected, SNVs with unique k-mers are in the green or yellow parts.

It can be seen that about 88% of these 40 million SNVs had unique k-mer pairs. The lists of unique k-mer pairs were compiled for these SNVs, the produced file had about 35 million SNVs that were used for testing the program for detecting SNV genotypes. Quite often the genotypes have to be detected for these SNVs that are more commonly used in association studies or that are located in coding areas of the genes. For this reason, it was also found that from the dataset of the 40 million SNVs, 95% of these that were located in the coding regions had unique k-mers and could be therefore detected using this method. From the HumanOmniExpress SNVs in this dataset, about 99% had unique k-mer pairs.

4.3 Identifying SNV genotypes from sequencing data

The unique k-mer pairs for the 35 million SNVs were used for testing the detection of SNV genotype from raw sequencing data. Three different datasets were used for testing the program for genotype identification. Two of these datasets consisted of simulated reads created from the reference genome with a depth of coverage of 30. One of these datasets contained the sequences from the reference genome, the simulated reads of the other set contained the allele variants of dbSNP variations that were not in the reference genome. For testing the method on real sequencing reads, the data of a sequenced individual from 1000 Genome Project was used with a coverage of about 10.

The genotypes of about 80.5% of the 35 million SNVs with unique k-mers were detected using the simulated data, for the real sequencing data, 78.2% of the SNVs were identified. For other SNVs the genotype could probably not be detected due to other variations, sequencing errors and low coverage which violated the assumption that the required k-mers are unique and present in the genome if the individual has the corresponding SNV allele variant.

From the SNVs that were in the coding regions of the genome, the genotypes were detected for about 83% from simulated data and 50% from the real sequencing data. The result for HumanOmniExpress chip SNVs that are often used in genome-wide association studies, was approximately 82% for both the simulated and real sequencing data.

The significance level of 0.1 was used in this study. In the cases where the genotype could not be detected, none or multiple p-values were bigger than the given significance level, thus the results vary depending on the significance level chosen by the user. In future, the effect of different significance values to the results on genotype detection should be studied to find the optimal value for these statistical tests, also the detected genotypes should be compared to the results from other SNV calling pipelines.

5 Time and memory usage

5.1 Time and memory usage of the program for finding unique k-mer pairs

The time and peak memory usage was measured for the pipeline for finding unique k-mers for different k-mer lengths to see the impact to the running time of the program as well as to the amount of used memory. These measurements were made while finding the unique k-mer pairs for the HumanOmniExpress SNVs without using mismatches. The results do not contain the time and memory used by the GListMaker tool for creating the list files from the reference genome sequence. No parallelization was used in the program. The measured time and memory usage can be seen from Figure 8.

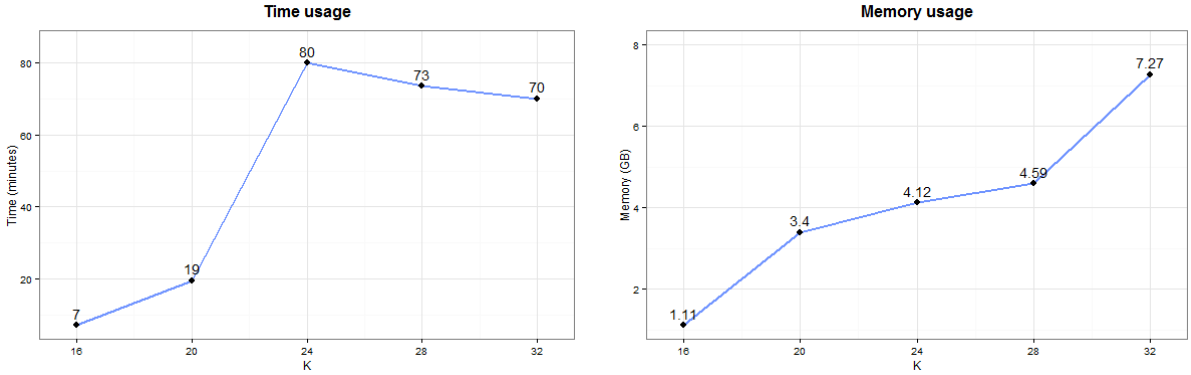

Figure 9: Time and memory usage for different k-mer lengths

5.2 Time and memory usage of the program for detecting SNV genotypes

The time and memory usage of the program for detecting SNV genotypes depends on the coverage depth of the data and the number of sequencing errors. The step with the biggest differences for these measures is creating the list file from the sequencing data with GListMaker. The amount of memory used for creating the list and the size of the resulting list file depend on the quantity of the sequencing errors in the data rather than the coverage depth. Sequencing errors produce new k-mers that have to be stored in the array, but a higher coverage depth only increases the frequencies of the k-mers already present in the array. The time also increases with the amount of sequencing errors as

sorting the array of k-mers takes more time for a bigger array.

For the simulated data with the coverage depth of about 30, the process of creating the list file with GListMaker and reducing its size with GListCompare took about 2.5 hours. This was measured by running the GListMaker tool with the default value of the number of threads used, which was 8. The same process took an hour longer for the real data although the coverage was smaller and there were fewer reads, because the sequencing errors in the real data produced a lot of new k-mers with small frequencies and the list file created from this data was two times bigger than the one created from simulated data. The amount of memory used by GListMaker was 214 GB for the real sequencing data and 185 GB for simulated data which was also the peak memory usage of the whole pipeline.

Finding the k-mer frequencies for unique k-mers with GListQuery took about 1.5 hours for both the simulated and real data. This value does not vary much for different datasets if the GListCompare tool is previously used to create a smaller list file containing only the k-mers that are later searched by GListQuery. However, the time of GListQuery tool could be further reduced by allowing multiple parallel searches, which is not allowed by this tool at this time.

The most time-consuming step was using the frequencies of the unique k-mers for finding the p-values and detecting the genotype of the SNVs. The time of this process does not depend on the sequencing data, but on the number of SNVs to genotype, in this work it was about 35 million. The measured time of this process for each dataset was about 7 hours. The time of this step could be reduced by parallelizing and detecting the genotype for multiple SNVs at the same time.

The total time of detecting the SNV genotypes from raw personal sequencing data is about 11-12 hours. This time could be significantly reduced in future by using multiple threads for detecting the genotype based on the k-mer frequencies and changing the GListQuery tool to search multiple query k-mers in parallel.

6 Conclusion and future work

The aim of this thesis was to develop a novel method for detecting SNV genotypes from raw personal genome sequencing data. A program was created for compiling a list of k-mer pairs that are unique for the target genome for every given SNV. A statistical framework was developed for using the lists of unique k-mer pairs to detect the genotypes of SNVs and a program was created to fulfill this task. Also, the appropriate k-mer length was evaluated for human genome and the list of unique k-mer pairs were found for 35 million SNVs that could be identified from sequencing data using this method. The program for detecting SNV genotypes was tested on both simulated and real data using the 35 million SNVs with unique k-mers. The genotypes of about 80% of these SNVs could be determined.

The time usage of the program for detecting SNV genotypes was measured to be about 11-12 hours. The time could be significantly reduced by parallelizing the process of genotype detection based on the k-mer frequencies, also by using more threads when running GListMaker tool or improving the GListQuery tool to search multiple k-mers in parallel.

The results of the program could be further tested for different significance level values to determine the best value for detecting the genotype from sequencing data. In addition, a comparison with other SNV calling pipelines could be made by measuring their time and memory usage, the amount of SNVs detected and comparing the genotypes found for these SNVs.

References

[Cunningham et al., 2015] Cunningham, F., Amode, M. R., Barrell, D., Beal, K., et al. (2015). Ensembl 2015. *Nucl. Acids Res.*, 43 Database issue:D662–D669.

[Depristo et al., 2011] Depristo, M., Banks, E., Poplin, R., et al. (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat. Genet.*, 43:491–498.

[Farrer et al., 1997] Farrer, L., Cupples, L., Haines, J., Hyman, B., Kukull, W., Mayeux, R., Myers, R., Pericak-Vance, M., Risch, N., and van Duijn, C. (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease. a meta-analysis. apoe and alzheimer disease meta analysis consortium. *JAMA*, 278:1349–1356.

[Fonseca et al., 2012] Fonseca, N., Rung, J., Brazma, A., and Marioni, J. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28:3169–3177.

[Hallmayer et al., 2009] Hallmayer, J., Faraco, J., Lin, L., et al. (2009). Narcolepsy is strongly associated with the t-cell receptor alpha locus. *Nat. Genet.*, 41:708–711.

[Healy et al., 2003] Healy, J., Thomas, E., Schwartz, J., and Wigler, M. (2003). Annotating large genomes with exact word matches. *Genome Res.*, 13:2306–2315.

[Kaplinski et al., 2015] Kaplinski, L., Lepamets, M., and Remm, M. (2015). Genometester4: a toolkit for performing basic set operations with k-mer lists. Manuscript submitted for publication.

[Langmead and Salzberg, 2015] Langmead, B. and Salzberg, S. (2015). Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9:357–359.

[Li, 2011] Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27:2987–2993.

[Li, 2013] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv*, 1303.3997.

[Treutlein et al., 2015] Treutlein, J., Cichon, S., Ridinger, M., et al. (2015). Genome-wide association study of alcohol dependence. *Arch. Gen. Psychiatry*, 66:773–784.

Non-exclusive license to reproduce thesis and make thesis public

I, Fanny-Dhelia Pajuste (date of birth: 17th of May 1991),

1. herewith grant the University of Tartu a free permit (non-exclusive license) to:

1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Finding Human Genome Variations from Personal Sequenced Data

supervised by Maido Remm

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 21.05.2015