Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Detection of microsatellite instability from genomic data obtained by next generation sequencing

Bachelor Thesis

2020
Dária Čárska

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Detection of microsatellite instability from genomic data obtained by next generation sequencing

Bachelor Thesis

Study Programme:	Bioinformatics
Field of Study:	Computer Science and Biology
Department:	Geneton, PRIF UK
Supervisor:	Mgr. Jaroslav Budiš, PhD.

Bratislava, 2020
Dária Čárska

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Dária Čárska
Študijný program: bioinformatika (Medziodborové štúdium, bakalársky I. st., denná forma)
Študijné odbory: informatika
biológia
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Detection of microsatellite instability from genomic data obtained by next generation sequencing
Detekcia mikrosatelitovej instability z genomických údajov získaných sekvenovaním novej generácie

Anotácia: Moderné technológie sekvenovania umožnili cenovo dostupné testovanie na viaceré typy genetických ochorení. Veľký potenciál má ich aplikácia v oblasti onkológie, kde stále chýba spoľahlivý nástroj na detekciu a sledovanie priebehu rakovinového ochorenia. Perspektívnou cestou je identifikácia genomického materiálu z nádorového tkaniva, pre ktoré sú typické formy, ktoré sa nenachádzaju v zdravej populácii.
Práca sa zameria na špecifický typ genomickej variácie, mikrosatelity, ktoré sú známe nestabilným prejavom pri prebiehajúcich onkologických ochoreniach. Študentka vyhľadá na genóme ich pozície a na základe agregovaných genomických údajov určí ich typické formy v slovenskej populácii. Netypické formy detekované u sekvenovaného jedinca by budú následne slúžiť ako indikátor prebiehajúceho ochorenia.

Vedúci: Mgr. Jaroslav Budiš, PhD.
Katedra: FMFI.KI - Katedra informatiky
Vedúci katedry: prof. RNDr. Martin Škoviera, PhD.

Dátum zadania: 29.10.2019

Dátum schválenia: 29.10.2019 doc. Mgr. Bronislava Brejová, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Dária Čárska
Study programme:	Bioinformatics (Joint degree study, bachelor I. deg., full time form)
Field of Study:	Computer Science, Informatics Biology
Type of Thesis:	Bachelor´s thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Detection of microsatellite instability from genomic data obtained by next generation sequencing

Annotation: Modern sequencing technologies have allowed affordable testing for several types of genetic diseases. Although their application in oncology has a great potential, the field still lacks a reliable tool for detecting and monitoring oncological diseases. A promising way is to identify genomic variants that are not found in a healthy population, and thus potentially originated in a tumor tissue.
The work will focus on a specific type of genomic variation, microsatellites, which are known to be unstable in oncological diseases. The student will locate them on the human genome and determine their typical forms based on aggregated genomic data from Slovak population. Atypical forms detected in a tested subject would indicate an ongoing oncological disease.

Supervisor:	Mgr. Jaroslav Budiš, PhD.
Department:	FMFI.KI - Department of Computer Science
Head of department:	prof. RNDr. Martin Škoviera, PhD.
Assigned:	29.10.2019
Approved:	29.10.2019 doc. Mgr. Bronislava Brejová, PhD. Guarantor of Study Programme

.....
Student

.....
Supervisor

Acknowledgment: I would like to express my thanks to my supervisor Mgr. Jaroslav Budiš, PhD. for his guidance, support and positive approach throughout the elaboration of this bachelor thesis.

Abstrakt

Mikrosatelity sú krátke repetitívne DNA sekvencie, ktoré vykazujú vysokú variabilitu v počte opakovaní medzi jedincami v populácii. Niektoré typy rakoviny vykazujú mikrosatelitovú instabilitu, ktorá je charakterizovaná vysokým výskytom mutácií v krátkych repetitívnych oblastiach genómu. V práci sme sa konkrétne zamerali na analýzu mononukleotidových mikrosatelitov, ktoré sú taktiež nazývané homopolyméry. Analyzovaním genotypov slovenskej populácie sa nám podarilo určiť populačné frekvencie tohto typu variability. Následne sme porovnávali dĺžky homopolymérov medzi geneticky nepríbuznými populáciami a taktiež medzi zdravými jedincami a onkologickými pacientmi. V našej analýze poukazujeme na značnú odchýlku homopolymérov identifikovaných zo vzoriek onkologických pacientov od bežne sa vyskytujúcich foriem v populácii. Z toho dôvodu by mohli homopolyméry potenciálne slúžiť ako genetické markery pre diagnostiku a monitorovanie onkologických ochorení.

Kľúčové slová: mikrosatelity, mikrosatelitová instabilita, krátke tandemové repetície, genomická variabilita, ľudský genóm

Abstract

Microsatellites are repetitive stretches of short DNA sequences with a highly variable number of repetitions between individuals in the population. Certain cancer types are associated with the microsatellite instability that refers to the hypermutability of short repetitive sequences. In the bachelor thesis we specifically focused on monomeric microsatellites also called homopolymers. We analysed genotypes of the Slovak population and determined the population-specific frequencies. Then we compared the genetic variants of genetically remote populations as well as healthy individuals with oncological patients. In our analysis, we show that homopolymers identified in samples of oncological patients significantly divert from common forms observed in general populations, and thus may be used as a potential biomarker for detection and monitoring of oncological diseases.

Keywords: microsatellites, microsatellite instability, short tandem repeats, genomic variability, human genome

Contents

List of Figures

List of Tables

Introduction

Nowadays, genomic analyses are gaining in popularity and simultaneously genetic tests are becoming more accessible. Modern sequencing technologies enable generating up to billions of reads per run representing DNA fragments of an examined sample in digital form in a relatively short period of time. Therefore, large amounts of genomic data can be analysed and used in various fields of research. However, there are still substantial costs associated with the sequencing process which can be limiting especially for large-scale genome-wide studies, that are typically based on the aggregation of sequencing data from thousands of individuals. It has been shown that the re-use of low-coverage sequencing data acquired from routine prenatal testing can be used as an affordable method for the detection of small variation [1] and certain types of structural variation [19] in a population. The detection of common microsatellite variation has been however neglected in these studies, even if it represents an invaluable source of genomic information with wide use in forensics, genealogy, clinical diagnostics, even detection and monitoring of certain oncological diseases.

We analysed repetitive stretches of short DNA sequences also called microsatellites that are a rich source of genomic variability in a population. Although this type of genetic variation has been already widely used in many areas of research such as forensic analysis, paternity testing, genetic mapping, and population genetics, the application of microsatellites in the field of oncology still lacks a reliable tool for detecting and monitoring oncological diseases. Therefore, we aimed to study microsatellites in the context of cancer research focused on the Slovak population. Oncological diseases are known to be accompanied by a dysregulation of the repair mechanism called mismatch repair resulting in the accumulation of mutations within the microsatellite loci. This condition of genetic hypermutability present in the short repetitive regions of a genome is called microsatellite instability and can be detected by identification of artificial forms of microsatellites in the examined sample that are not observed in the common population.

Firstly, we explained the used terminology and described the biological background. Next, we illustrated the steps that are typically involved in the genomic analysis. Due to technical limitations of analysed sequenced reads (low coverage and short read lengths), we decided to focus only on monomeric microsatellites called homopolymers.

We identified genotypes of the Slovak population and calculated the population-specific frequencies of homopolymer lengths over more than 7,000,000 genomic loci. We utilized computed frequencies to determine a microsatellite instability status of tested subjects indicating the presence or absence of an oncological disease.

In our work we also compared the genomic variation of homopolymers among genetically distinct populations as well as examined differences of homopolymer lengths between healthy samples and samples of oncological patients.

Chapter 1

Biological background

In this chapter, we are going to introduce the biological background and explain the used terminology.

1.1 DNA

The *deoxyribonucleic acid* (DNA) is a molecule, which carries the specific genetic information of each organism and contains the important instructions needed for the synthesis of proteins and the regulatory processes in the cell. The DNA consists of subunits called nucleotides which are joined by phosphodiester bonds to form a DNA strand. Each nucleotide is composed of a phosphate group, a sugar group deoxyribose, and a nitrogen base. There are four possible types of nitrogen bases in the DNA: adenine (A), thymine (T), guanine (G) and cytosine (C) and their first letters are used as the abbreviation to represent easily the order of nucleotides in the DNA chain as a string. The DNA is present in the cell in form of a double helix which means that there are two separate strands of nucleotides joined together with hydrogen bonds between the nitrogen bases according to the complementary base pairing rule (adenine with thymine and guanine with cytosine). The double-stranded DNA molecule is packaged with the proteins into a compact aggregate called chromosome and a genome is a set of all chromosomes in the nucleus of the cell and the mitochondrial DNA. A human genome consists of 23 pairs of nuclear homologous chromosomes and altogether a human haploid genome is made up of about 3 billion base pairs of DNA [22].

1.2 Genome variability

A genome variability describes the differences in the genome among individuals of the same species. Although most of the DNA sequences are the same within a species, there are virtually no two humans having identical genome, not even monozygotic twins. The

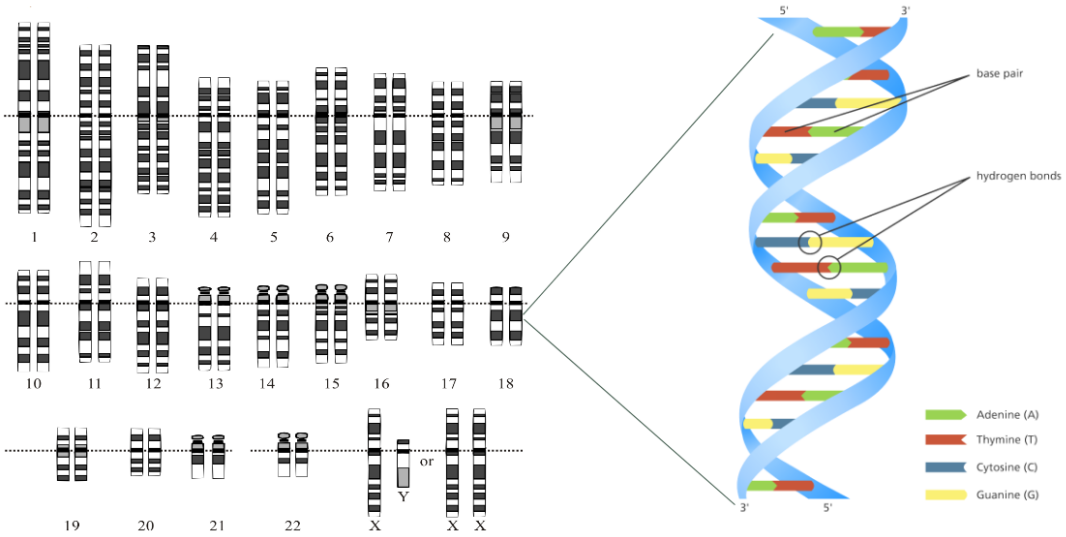


Figure 1.1: Human chromosomes and the double helix structure of DNA. One of the DNA strands within the double helix represents the following sequence of nucleotides read in the 5' to 3' direction: ATGACACTGTGACA. Source: https://www.yourgenome.org/facts/what-is-dna

ability of DNA sequences to vary enables individuals to be unique in appearance and behavior.

Mutation is a change in a DNA sequence and it is one of the main sources of the variability. It can have positive, neutral or negative impact on life of an organism. There are parts of the genome containing strictly conserved sequences that do not change over time and remain the same within and many times also across the species. These sequences usually fulfill some essential vital functions and therefore, the variability does not commonly appear within these genomic regions because a mutation could cause disfunction of mechanisms important for life. However, a genome consists also of nonfunctional DNA sequences that are usually not conserved and thus are rich source of variability across individuals and species.

We distinguish between inherited, germ-line, and somatic variants. The inherited variation appears in the parental germ cells (a sperm or an egg) and such genetic variant can be inherited from one generation to another. Therefore, an inherited variant from a parent occurs in every cell of an offspring. Another source of genetic variability are de novo mutations. In case a de novo mutation arises in a germ-line cell of an organism, the resulting mutant variant will occur in each descendant cell derived from the cell carrying a novel variant. The somatic variation can be acquired during the life and affects the somatic cells. This alteration occurs only in one affected somatic cell and can be replicated during cell division, but can not be passed down the generations. [22]

There are different types of a genetic variation spreading over one to many nu-

cleotides of a DNA sequence. The difference in one nucleotide between genomes is called a *single nucleotide polymorphism* (SNP) and it is the most common variation type in the human population [22]. The variation can also arise from insertion or deletion mutations on a smaller scale and such genetic variants are called indels. The next type of variability is called a structural variation, which usually occurs over larger regions of the DNA sequence and changes the structure of a chromosome. It includes deletion, duplication and insertion events and chromosomal rearrangements like inversions and translocations. Furthermore, the sequence variations include repetitive tracts of DNA (microsatellites, minisatellites and satellites) where the genetic variation appears in the length of a repetitive sequence between genomes among the individuals.

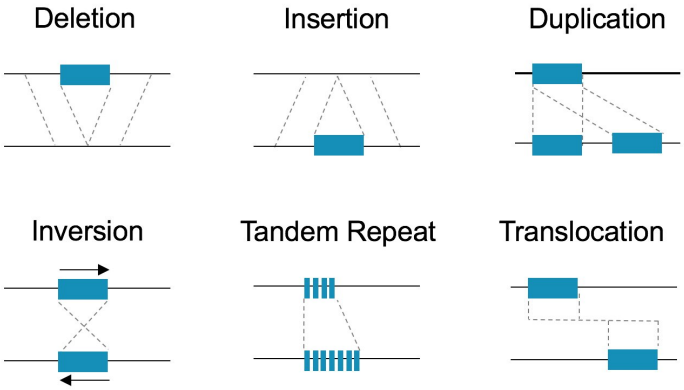


Figure 1.2: Different types of structural variation, source: https://www.pacb.com

1.3 Short tandem repeats

Short tandem repeats (STRs) are also called microsatellites and are composed of a motif sequence, which is repeated consecutively and consists of 1 to 6 base pairs (bp) [5]. Different authors do not always agree on the definition of a microsatellite and therefore there is no consistency in the motif length as well as in the total length of repeat units but a microsatellite locus is typically considered to contain up to 100 nucleotides (nt) [15].

There are different types of STRs depending on the length of a repeated motif. When the STR variant consists of one nucleotide, it is called a homopolymer or a monomeric STR. When the unit of two nucleotides is repeated, these are dimeric STRs. We call the STRs trimeric, when the motif sequence contains 3 nucleotides and STRs with longer repetitive patterns are called tetra-, penta- and hexameric respectively. Usually the longer a repeated unit the less often is this type of STR present in a genome [4]. STRs can be also classified according to their structure as perfect, imperfect,

interrupted or composite. Perfect microsatellite consists of only identical copies of the motif within the repetitive sequence while in an imperfect STR there are one or more mismatches present which do not match the motif sequence. An interrupted STR contains a small sequence different from a motif which has been inserted into the repeated sequence and in the case of a composite STR there can be found more than one type of motif [17].

STRs are ubiquitous, highly polymorphic and cover approximately 3% of the human genome [4]. Their placement across the chromosomes is not uniform and in the human genome STRs are located most densely on the chromosome number 19 [23]. Most of STRs are located within noncoding DNA while approximately 8% occur in the coding regions [3]. Especially motifs consisting of 3 or 6 nt are most frequently found in the exon sequences because of their triplet structure which represents codons encoding amino acids [20].

Majority of STRs do not have any known function, are not conserved DNA sequences and therefore such STRs are typically rich source of variability. We can observe their extensive length polymorphisms across species and populations because mutations can be accumulated within these nonfunctional DNA sequences without any phenotypic alteration causing some disfunction or having negative impact on life. However, there are also STRs that fulfil some functions. It has been already proven that microsatellites have an effect on the regulation of a gene expression and also some molecular phenotypes [5]. They may also encode proteins, be involved in regulating the transcription and affect recombination and maintenance of chromatin spatial organization [4, 27, 9].

Microsatellites exhibit a mutation rate ranging approximately from 10^{-4} to 10^{-3} per locus per generation which is far higher in comparison to single nucleotide polymorphisms having mutation rate 10^{-8} nucleotides per generation in the human genome [24]. The main cause why microsatellites tend to mutate more often than other loci of a genome is the slippage of the DNA polymerase during the replication process of a DNA strand 1.3 [4]. The slippage occurs when replicating DNA strands temporarily dissociate from each other and realign in a different position. As a result, the newly synthesized DNA strand ends up with either more or fewer motif copies of the STR locus. These mistakes in replication during meiosis or mitosis are induced by the repetitive character of tandem repeats and when the failure occurs and is not corrected, the replicated STR variant is fixed and carries permanently a different number of repetitions in comparison to the template. The other possible but not so frequent mechanism for the STR mutation is an unequal crossing over in meiosis where the satellites of two homologous chromosomes are exchanged unequally [4].

Slippage events during DNA replication happen more often than is the actual amount of mutations resulting from the slipped strand mispairing. The reason behind this is the correction mechanism called a *mismatch repair* (MMR), which is capable of

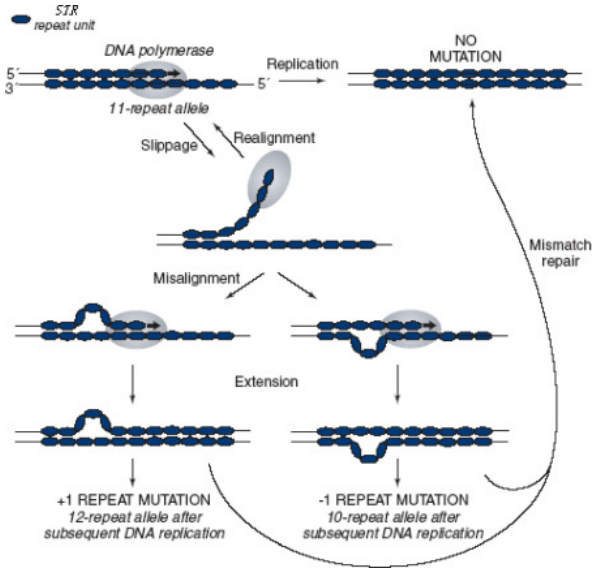


Figure 1.3: Schematic illustration of the strand slippage during DNA replication [4]

finding, removing, and correcting the errors in the DNA sequences and thus eliminates permanent mutations.

The mutation rate of STRs is influenced by many factors including [4]:

- *repeat number* - the longer the STR the higher rate of mutations, usually extension of a STR locus (gain of nucleotides) is present more often in short sequences and reduction of STR (loss of nucleotides) in longer ones,
- *repeat unit* - for instance, it is proven that dimers mutate more frequently than trimers,
- *repeat structure* - some studies imply that the more complex the repeat structure is, the less frequently mutations arise,
- *base composition of repeat unit* - AT-rich sequences (genomic regions containing a higher proportion of nucleobases adenine and thymine) mutate more often than GC-rich sequences because of the lower stability of the template,
- *sex* - sperm cells are replicated more times than eggs and therefore their mutation rate is higher,
- *age* - the mutation rate in the sperms depends on the age of men, higher mutability is present in the sperm cells of men whose age is greater, because they undergo more mitoses,
- *interruptions in STRs* - mutations such as transitions, transversions, single-nucleotide insertions or deletions and others can change the STR sequence and thereby the mutation rate of this locus will be different.

1.3.1 Importance for genetic testing

Many inherited genetic diseases are associated with specific forms of STRs including Mendelian diseases like cystic fibrosis or Gilbert syndrome, complex traits, and cancer [5]. It is also well known that specific trinucleotide repeat expansions cause neurological disorders such as Huntington's disease, fragile X syndrome, myotonic dystrophy, and others [18].

STRs are highly variable in their lengths and alleles in a population and thanks to these attributes microsatellites are extremely useful in identification of people in forensic analysis and also in determining parentage by comparing the number of repeats at the specific multiple STR loci between a parent and a child and expecting the similarities to some extend [6, 26, 4]. Another application of STR data is in research of evolution processes. STRs are informative for determining the relationships between closely related species in an evolution and estimating distances in phylogenetic trees [25]. They are used as well to study a human evolution and a history of migration [4].

1.3.2 Importance for oncology

Several types of oncology diseases are characteristic with the *microsatellite instability* (MSI) that is a state of the genome to be hypermutable which means that also the STR loci are more prone to mutations and spontaneously lose or gain of nucleotides in their sequences. This instability in a genome induces the production of the novel forms of microsatellites that differ in their lengths from germline. MSI is the result of a dysregulated reparation mechanism called MMR. MMR repairs the errors in a DNA right after the replication under normal circumstences but when this mechanism works abnormally, the mutation rate is higher and errors are accumulated in the DNA sequences including microsatellites in a great measure. It is estimated that homopolymers could be suitable markers for detection and analysis of the MSI status, along with the underlying oncology disease [21].

Many cancers are usually caused by an acquired somatic mutations rather than those who are hereditary. The somatic variability can lead to the disfunction of MMR and then the cell growth can get out of control. In a human genome the growth of the cells is strictly controlled by many genes, but mutations present at one or more of these genes can promote the growth of abnormal cells resulting in a cancer. Cancers where the sporadic MSI status has been already detected include colon, colorectal, endometrial, ovarian, and gastric cancers [21].

MSI detection has an important clinical significance and because of the high mutation rate of STRs, they can be useful as genetic markers for a cancer diagnosis, prognosis and also monitoring of the disease.

Chapter 2

Analysis of genomes

In this chapter we will focus on description of steps which are typically involved in an analysis of a genome. Generally, the process of analysing genomes at first involves sequencing where the order of nitrogen bases in the DNA fragments isolated from a sample are converted into digital form. As a next step, the obtained sequences are compared with the *reference* genome, that is a representative genome assembly, and finally the identified differences and genetic variants are further analysed and interpreted.

2.1 Sequencing

Sequencing is the process of determining the order of 4 nucleotide types in a DNA chain. The sequencers are machines that can automatically sequence the DNA strands. Due to the limitations of the underlying laboratory processing, the sequencers are typically not able to analyze a whole DNA molecule. At first, genetic material containing the DNA of a tested organism is extracted from a sample and fragmented (DNA strands are split into shorter sequences where the length of fragments depends on a type of technology that is used). This is the genetic input material for the sequencers which can be further analyzed. The output of sequencing are reads which are the strings of letters A, T, C, G representing the nitrogen bases (adenine, thymine, cytosine and guanine) of the sequenced DNA fragment. Sequencers usually detect also the quality scores for each base indicating the confidence that a nucleotide, that has been read, was determined correctly. The quality values are subsequently reported in an output. Nowadays, there are three generations of sequencers differing in the used technologies and in the guarantees of the output reads.

The first generation sequencers are based on Sanger sequencing method which uses the chain termination technique. It requires the clonal amplification of the DNA fragments to detect the nucleotides and produces reads of the average lengths ranging from

400 to 900 bp [11]. This sequencing method is low-throughput and only one read is produced per run. It is cheap for low numbers of DNA strands but very expensive and extremely time-consuming for sequencing whole genomes. The guaranteed accuracy is approximately 99.7% [11].

The second generation is called *next generation sequencing* (NGS) and differs from the first generation in the number of sequenced fragments. It is also called the massive parallel sequencing because of its ability to process millions of fragments simultaneously per run. The fragments have to be amplified to get a signal which is strong enough to distinguish the bases. There is a limitation in the length of the fragments which can be approximately 35-150 bp long but on the other hand, the accuracy is high (more than 99%) [11].

The third generation sequencing is different from the previous two generations because the amplification of the fragments is not needed anymore. These sequencers can analyze single-molecule templates and produce long reads (tens to hundreds of kilobase pairs) [11]. The disadvantage of this method is its accuracy because the error rate is relatively high in the range from 12% to 15% [11].

The most popular sequencers are from the second generation because of their high accuracy, low price, and high speed of analysis. Their advantage is also the ability to sequence large amounts of fragments at a time and to cover whole genome [11].

NGS enables a genome-wide analysis and determines a primary structure of DNA. The sequencing process involves more stages and the quality of resulting data depends on the precision of each step of the analysis. At first genetic material needs to be processed. DNA is isolated from a sample, fragmented to the desired length, and modified by adding specialized adapters to the ends of DNA fragments which are designed to interact with an NGS platform. As a next step, a DNA library is prepared by attaching the fragments to either a flow-cell (Illumina) or beads (Ion Torrent) [8]. This is followed by an amplification of fragments whereby the clonal DNA colonies called DNA clusters, each arising from a single library fragment, are formed. Finally, the sequencing procedure is carried out where the sequence of each cluster from a library is read. It is realized by repeated cycles of nucleotide incorporation at which each added nucleotide into a newly synthesized complementary strand is monitored by a fluorescence detection or by changes in electrical charge depending on the used technology during the sequencing process [13].

Sequenced reads are typically stored in a FASTQ file. Each read in the FASTQ format is represented by 4 consecutive lines 2.1. The first line starts with a "@" character followed by a sequence identifier that unambiguously specifies sequencer, sequencing run, and the physical location of the sequenced cluster. The nucleotide sequence itself is present on the second line and is typically composed of abbreviations A, T, C, G, N representing adenine, thymine, cytosine, guanine, and uncertain nucleotide observa-

tion, respectively. The third line serves as a separator and contains a "+" character. On the fourth line quality scores corresponding to bases of the sequence from the second line are provided and are represented by ASCII characters. An ASCII code of each character can be converted to the actual quality score which is linked to the probability of an incorrect base call.

There are two options how the DNA fragments can be sequenced. In case a DNA strand is sequenced from only one end, as a result a single read is generated for each fragment. There is also a paired-end sequencing where both ends of each DNA fragment are sequenced and can overlap. Generally, for standard genetic testing the sequencing process typically generates about millions reads per run [11]. However, some sequencers are also capable of generating up to billions reads per run which are required in whole genome analyses with high coverage [11].

```
@NB501192:13:HCM5CBGXY:1:11101:23913:1152 1:N:0:TCACGCGC+CTTCGCCT
CCTGTATCTTCGTGATGCAGTGACCAGTGGTTGGT
+
6AA6AEEE6EEEAE6EEEEEEEEEEEEE/EEEEE/
```

Figure 2.1: An example of a single entry in a FASTQ file

```
NB501192:13:HCM5CBGXY:1:11101:23913:1152 99 chr7 107101273 42 35M = 107101391 152
CCTGTATCTTCGTGATGCAGTGACCAGTGGTTGGT 6AA6AEEE6EEEAE6EEEEEEEEEEEEE/EEEEE/ AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0
```

Figure 2.2: An example of a single entry in a SAM file

2.2 Mapping

A fundamental step in high-throughput sequencing analysis is finding positions of sequencing reads on the reference genome where reads most likely come from. The process of aligning generated reads against a reference genome is called mapping. There are numerous mapping tools including BWA, Bowtie, Bowtie2, MAQ, and others [7]. Mapping programs typically require to build an index of the reference genome which speeds up their alignment algorithms. The resulting aligned reads to a reference genome are usually stored in SAM 2.2 and BAM file. A BAM file is the compressed binary version of a SAM file, thus takes up less storage space, while preserving the same information. A BAM file is usually sorted by genomic coordinates and indexed which allows faster access to reads aligned within a certain genomic region of interest. On the other hand, SAM is a text-based format and therefore is appropriate for viewing a file by operator.

Using the Samtools view command it is possible to easily convert between SAM and BAM formats [14]. Every SAM or BAM file contains an optional header section with the information about a reference which is followed by an alignment section. Each alignment line representing a mapped read consists of 11 mandatory fields providing essential alignment information such as the read identifier, a bitwise set of information describing the alignment (FLAG), the sequence of a read and its quality, mapping position, mapping quality reflecting the alignment confidence, CIGAR string indicating the presence of any SNP's or indels in the read and an information about its mate read (if paired-end sequenced), and other optional fields for additional information.

2.3 Variant calling

Variant calling is the process of identifying genetic variants from mapped reads and consists of two fundamental steps: genotype assignment and variant identification. It is a very specific stage of a genome analysis depending on the type of studied variation. Typically, aligned reads are compared with the sequence of the reference genome and identified variants and detected differences can be further analysed. Various bioinformatics tools have been already developed for a variant calling such as Samtools mpileup, Genome Analysis Toolkit, Freebayes and Ion Proton Variant Caller [10] but custom variant callers designed specifically for the examined type of genetic variation are used as well. There are also many databases of genomic variants which are available and helpful for comparing the variants of examined samples with population frequencies. Nowadays, it is still of great importance in research to collect amounts of genetic data and create population specific frequencies of various DNA variants. The data sets comprising human genotypes are especially essential for different biomedical applications. On the other hand, retrieval of large quantities of genetic information destined for large-scale population studies is still associated with substantial costs. However, population frequencies are extremely helpful and important in clinical studies, provide valuable insights into causes of diseases and their underlying risk factors, and later on can be used for a diagnostic assessment.

Chapter 3

Analysis of homopolymers

In our work we focused specifically on homopolymers. We studied a genotype of each individual from available sample sets and in the analyses we used various bioinformatics tools as well as the custom scripts. As the output we wanted to obtain the genotypes of all fully covered homopolymer loci for each sample. Subsequently, we were able to further analyse and compare identified genetic variants.

3.1 Sequenced DNA

Our data set of reads was sequenced using NGS technology, concretely Illumina sequencing platform. The samples, where DNA molecules were extracted from, were gained from blood plasma of pregnant women from Slovak and Indian populations and were originally dedicated to non-invasive prenatal testing (NIPT). We reused these sample sets [1] for the purpose of our analyses. The data sets that we used were very specific in certain aspects and nonstandard for the type of our study. We had to deal with the following attributes of the data:

- The sequenced reads were 35 bp long and due to their short length we decided to limit our study to analyse only homopolymers. Specifically, we studied homopolymers with the number of repetitions ranging from 2 to 29 nt.
- Samples were sequenced using low-coverage massively parallel whole-genome sequencing method, which means that there were nucleotides and regions of a genome that have not been sequenced, and thus not all homopolymers from each individual sample could be read and detected. Approximately 14.46% of homopolymers were covered in average per sample and that was also predominantly performed by a single read.
- Positions of detected homopolymers were different among samples and thus could not be straightforwardly compared.

- The data set comprised of the large amount of Slovak samples (10,645 individuals) dedicated to calculate the population frequencies. On the other hand, there were also many homopolymer loci across the human genome which could be analysed. However, the usage of the data generated during NIPT testing for population specific frequency determination of small sequence variants such as single nucleotide and insertion-deletion variants has been already presented and published in the study [1]. In our case we had even considerably larger data set in comparison to the amount of samples (1,501) used in the above mentioned study. Therefore, we assumed that the quantity of samples could reasonably balance and overcome the problems associated with the coverage deficiency and the length of reads.

Illumina allows sequencing from both ends of the DNA fragments, which has been used in our case, and therefore we worked with paired-end reads. Paired-end reads are convenient in an aligning process of reads to a reference genome because it is expected that each pair of reads obtained from one DNA fragment should map within a certain distance of each other and in a certain order and thus improves the quality of mapping and reduces the problem of multi-mapping. Thus, there were two FASTQ files generated by a sequencing process from each sample, one file with all reads in forward direction, and the other one with all reads in reverse-complement orientation to its corresponding mate pairs from the first file.

3.2 Data sets

We worked with the following data sets:

- samples of healthy individuals from the Slovak population dedicated to be used for population frequency determination of genetic variants (SVK-POP),
- Slovak samples of healthy individuals used as the independent control data set in comparisons using population frequencies (SVK),
- samples of healthy individuals from the Indian population (IND),
- samples of oncological patients obtained from the Slovak population (ONC).

All sample sets we used consisted of sufficiently large amount of samples which enabled us to sensibly compare samples of different data sets 3.1. We decided to analyse the genetic variation of homopolymers from 2 different aspects. Firstly, we wanted to look at the differences between unrelated populations where we used Indian and Slovak sample sets and secondly, we wanted to examine homopolymer lenghts in context of MSI via samples obtained from oncological patients and healthy individuals of the same

Data set	Number of samples	Proportion of covered homopolymers		
		Minimum value	Maximum value	Median
SVK-POP	10,645	1.1226e-05	0.506	0.1028
SVK	200	0.0082	0.367	0.0998
IND	306	0.0455	0.3706	0.1079
ONC	82	0.07501	0.4027	0.1932

Table 3.1: Properties of data sets

population. The data set of oncological samples was special because contained always two samples per one patient. One set of samples (ONC0) comprised genetic material obtained from plasma on the day of an operation before surgery and the other half of samples (ONC3) contained DNA isolated from blood plasma which was taken from patients three days (in two cases one day) after the surgery.

3.3 Calling homopolymers

Due to the specific nature of our sequenced reads, we designed our own method for genotyping STR loci. At first, we identified locations of homopolymer loci on the reference genome. Then, we extracted sequenced reads that were mapped to them. Finally, we extracted the number of repetitions for each covered locus and summarised them into numeric vectors.

3.3.1 Identification of homopolymer loci in the reference genome

We decided to use the latest major version of human reference genome GRCh38, which is available online for free download for our analysis. Firstly, we created an index over this genome, enabling faster retrieval of genomic content in regions of interest, using the Samtools faidx command. Subsequently, we identified all homopolymers and their positions throughout the human genome using custom Python scripts and statistically evaluated the occurrence of homopolymers in the human genome. For passing through all regions of the genome we used the Pysam Python module which allows reading and writing different formats including e.g. SAM, BAM, BED, FASTA and FASTQ file formats and also supports random access to the genomic data through indexing. The minimal length of searched homopolymers was set to 6 nt and we looked only for perfect monomeric STRs without any mismatches within the repeat. We stored identified homopolymers in the BED file format which is a standard file format for recording coordinates of genomic regions. Generally, a BED file is a tab-delimited text file and consists of 3 required and 9 additional optional fields. In our BED file only

the mandatory fields and one extra nonstandard BED field containing the repeated base within the homopolymer were present. The required BED fields consisted of the name of the chromosome or scaffold, the 0-based starting position, and 1-based end position of the locus where the homopolymer was identified. By analysing the human reference genome we also created a file containing lengths of monomeric microsatellites and its counts. The results confirmed our expectations that with the increasing length of a homopolymer the count of such microsatellite systematically decreases across the human genome 3.1. Altogether, we detected 7,749,621 homopolymer loci with length at least 6 nucleobases in the human genome. We counted homopolymers separately for each nucleotide and length as well. Interesting fact resulting from this data was much higher occurrence (approximately one to two orders of magnitude higher) of the nucleobases adenine and thymine in comparison to cytosine and guanine within the same length of a homopolymer.

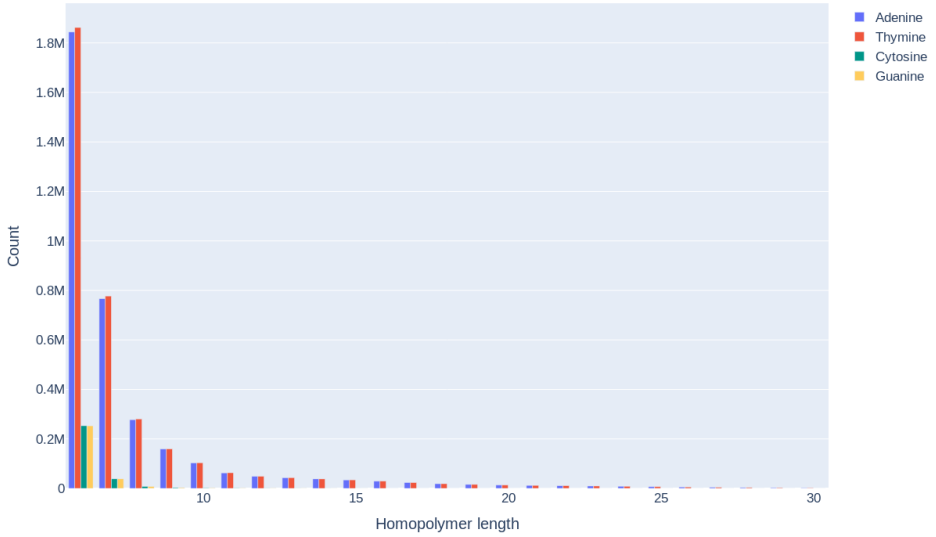


Figure 3.1: Counts of homopolymers in the human genome

3.3.2 Genotyping of sequenced samples

The first step of analysing samples was to map all reads against the reference genome. We created the mapping index over the human genome reference with the Bowtie2 tool. We also used Bowtie2 tool for mapping reads from both FASTQ files coming from one sample to the reference genome and as a result, a SAM file was created for each individual sample. Each SAM file was subsequently converted to a BAM format using Samtools view and each BAM file was sorted using Samtools sort and indexed using Samtools index tool. Finally, each generated sorted BAM file was processed by

our own Python script. By running this script we excluded all reads which were not mapped to any homopolymer from the reference genome from further analyses. Also, the script identified and stored the lengths of all homopolymers which were properly aligned to the reference genome based on the following criteria:

- *W*e set the condition that each analysed aligned read needed to have a mapping quality higher than or equal to 2 to eliminate reads with uncertain mapping location.

- *W*e specified the minimal length of a homopolymer to 2 nt.

- *E*ach homopolymer in the read, we took into consideration, had to be surrounded from its both ends by at least 3 flanking bases. The maximal usable homopolymer length, given 35 bp long reads in our data set, could be therefore 29 nt. Via this filtering step we especially aimed to exclude incompletely covered homopolymers from further analyses.

We used a compressed, one-dimensional NumPy [16] array to store detected genotypes of each sample from Slovak, Indian and oncological data sets. Positions in a vector representing homopolymers that were not identified from any read and had an unknown length contained a default value 0. We opted to use the NumPy objects because NumPy library in Python provides among other things efficient numerical computation and contains sophisticated broadcasting functions allowing to effectively operate on the NumPy arrays. Another advantage of using NumPy objects is that they can be used as input parameters of functions when creating graphs or performing statistical tests on data.

We worked with a large amount of FASTQ files which needed to be processed in the same way. Therefore, the process of analysing all reads of Slovak, Indian and oncological samples was performed using very useful tool called Snakemake [12] which enables distributed running of scripts and shell commands on many files. We created a Snakefile consisting of several rules defining the workflow which automatically processed each sample and took two FASTQ files containing reads as an input and generated a NumPy vector, which was determined as a target output file, for each sample.

3.3.3 Calculation of population frequencies

The next step of analysing monomeric microsatellites was the aggregation of 10,645 randomly chosen vectors out of all 10,845 one-dimensional NumPy arrays obtained from the sample set of the Slovak population. We kept 200 samples, which we did not involve in aggregation we performed on the vectors, as the control data set for comparisons and statistical evaluations. We wrote a Snakefile with workflow description

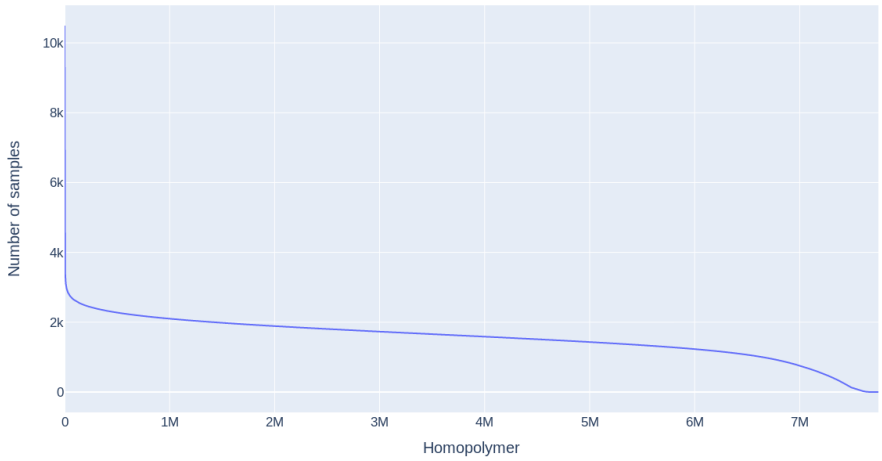


Figure 3.2: Coverage of homopolymer positions from Slovak sample set

to process automatically the above mentioned arrays and we also wrote a Python script to manipulate with arrays and store the result of aggregated numpy vectors in a two-dimensional NumPy array. We decided to represent genotypes of detected homopolymer lengths from the Slovak population in a matrix G. The i-th row, G_i, corresponded to a single homopolymer locus and the j-th column, G_j, represented genotypes of the j-th individual. The homopolymer length itself of the i-th locus and j-th sample or the value 0 in case the locus was not covered was stored in the matrix at position $G_{i,j}$. Subsequently, we calculated occurrences of homopolymer lengths separately for each position from the matrix G. The counts of genetic variants identified in the Slovak population were stored in the NumPy matrix C in a such way that the i-th row, C_i, stood for the i-th locus out of all homopolymers and the order of homopolymer loci was the same as stored positions of homopolymers in the BED format. Therefore, rows of the two-dimensional array C corresponded to lines in the BED file. Each column index represented the number of repetitions and in the array itself the counts of homopolymers, which were identified in the Slovak sample set, were stored at positions of the matrix determined by a row and a column.

Relative frequency distribution is defined as the percentage or proportion of data elements in each class. In our case to get the population frequencies, we computed the relative frequencies from obtained homopolymer counts observed in our sample set of the Slovak population using the following formula:

$$F_{i,j} = \frac{C_{i,j}}{\sum_{k=2}^{29} C_{i,k}}$$

where F is the matrix containing resulting relative frequencies and C is the matrix of observed counts of homopolymers in the Slovak individual samples. The determination

of population specific frequencies has a great potential to be further widely used in various clinical applications and in research focused on monomeric STRs.

Using different metrics we wanted to examine how significantly monomeric STRs differ among populations. We purposely studied Slovak and Indian individuals which are considered to be genetically highly remote because it was expected that there would be observable variability in their homopolymer lenghts. However, the comparison of these two populations also particularly served as the baseline for the evaluation if oncological patients dispose of higher deviation from healthy variants in comparison to determined interpopulation differences. We processed all NumPy vectors containing counts of detected monomeric microsatellites from Indian, oncological, and control sample sets in such way that we took only positions where nonzero values were present and thus the lengths of homopolymers were detected at these loci. Subsequently, we stored the corresponding population frequencies of observed homopolymer lengths in a new NumPy vector which was used and processed in further analyses. We decided to look at the resulting data from various views and compare Slovak versus Indian population, Slovak versus oncological samples and pairs of oncological samples using the resulting relative frequencies of monomeric STRs calculated from the Slovak population according to several different metrics. We calculated average and median values of the vector elements for the estimation of genetic distance between the mentioned populations. We also compared the occurrences of such homopolymers that appeared rarely, based on the computed frequency values, in the Slovak population and thus had lower relative frequency values. This was crucial in our study because we wanted to focus especially on those homopolymers which showed atypical lengths for the Slovak population and thus assess the variability. We decided to look separately at 3 categories and set 3 different thresholds for the relative frequencies which were as follows:

- *observed homopolymers with frequencies equal to 0* - in this case the lenghts of these homopolymers were not identified and present in samples from the Slovak population,

- *homopolymers with frequencies lower than 0.1 and lower than 0.05* - lengths of these homopolymers were not common in the Slovak population.

As the metric for rarely occurring lengths, which we subsequently compared between our sample sets, we took a count of those detected homopolymers within the vector which satisfied the given condition and divided it by number of nonzero vector elements.

Finally, we created graphs using Python graphing library called Plotly to display the distributions of computed metrics within each sample set.

Chapter 4

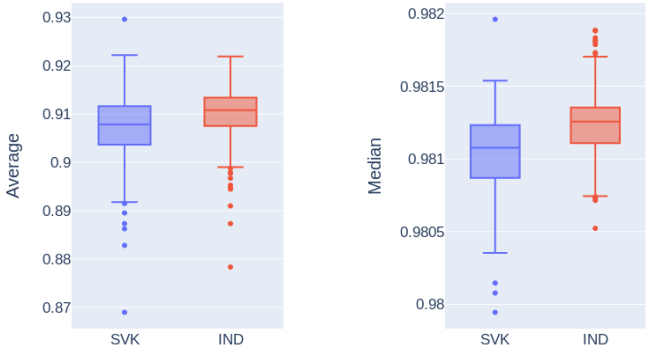
Results

In this chapter we will interpret the obtained results of our analyses.

We used non-parametric statistical tests to assess the difference of measured metrics between data sets. To confirm that compared distributions are not normally distributed, we performed the normality test using Scipy.stats package [2]. The corresponding function tests the null hypothesis that a sample comes from a normal distribution. Using this statistical function we got the calculated p-value much lower than 0.05 for each distribution which confirmed that the null hypothesis could be rejected and thus our data did not follow the normal distribution.

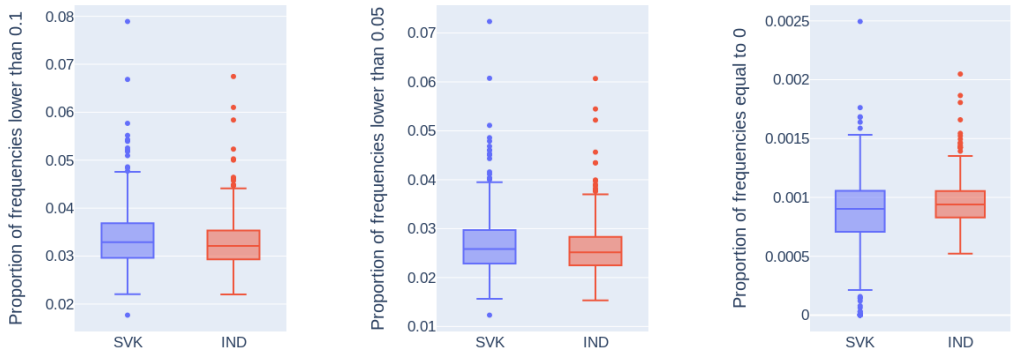
4.1 Comparison of Slovak and Indian samples

We looked at the frequency distributions of Slovak samples from a control data set and Indian samples. These two populations are not considered to be closely related and therefore we expected that relative frequencies of homopolymers from Indian samples should take in general lower values than samples from the control data set. From the graphs we created, it was not obvious if there is the significant difference and variability in homopolymer lengths between these two populations. To prove if lengths of homopolymer loci varied between Slovak and Indian populations only by chance or if there was the trend that the distributions in the two groups differed significantly we performed a statistical test. Concretely, we used Mann–Whitney U test from Scipy.stats package which is designed for 2 groups of independent data with not normal distribution. P-values in almost each analysed distribution of specific metric showed that there is a significant difference between Indian and Slovak populations.



(a) Average of frequency values (Mann–Whitney U = 22159, n_1 = 200, n_2 = 306, P = 7.64507e-08, two-tailed)

(b) Median of frequency values (Mann–Whitney U = 17537, n_1 = 200, n_2 = 306, P = 2.26578e-16, two-tailed)



(c) Frequencies lower than 0.1 (Mann–Whitney U = 27894, n_1 = 200, n_2 = 306, P = 0.04623, two-tailed)

(d) Frequencies lower than 0.05 (Mann–Whitney U = 28033, n_1 = 200, n_2 = 306, P = 0.05523, two-tailed)

(e) Frequencies equal to 0 (Mann–Whitney U = 26250, n_1 = 200, n_2 = 306, P = 0.00341, two-tailed)

Figure 4.1: Distributions of Slovak and Indian sample sets using different metrics applied to relative frequency values

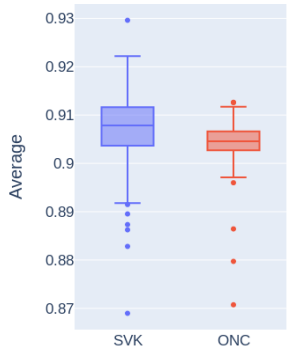
4.2 Comparison of Slovak and oncological samples

When comparing Slovak and oncological sample sets visually via box plots it had been already noticeable that there were differences between the homopolymer length distributions. We proved again our hypothesis by Mann–Whitney U statistical test. For

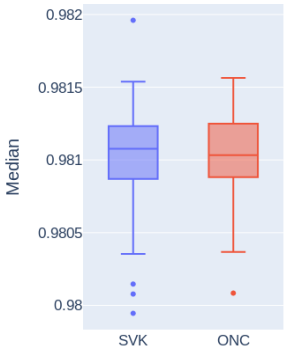
each observed metric except the median of relative frequencies we got a p-value much lower than 0.05, which confirmed our expectations, and thus we could declare that the differences in homopolymer lengths between healthy and oncological samples are not caused by chance. Therefore, we assume that the MSI status should be detectable in the oncological samples using monomeric microsatellites as the genetic markers.

4.3 Analysis of oncological sample pairs

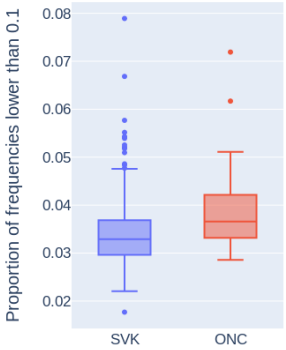
As we have already mentioned we had always a pair of samples coming from one patient in our data set of oncological samples. One sample was obtained before surgery and the other one 1 or 3 days after. We expected that the MSI status would decrease after the surgery because the tumour was removed and thus cancer cells should not be present in the body of a patient at higher levels in comparison to the condition before surgery. However, the data did not show clearly what we have expected and therefore did not confirm our assumption. For statistical evaluation of oncological sample pairs, we applied the two-sided Wilcoxon signed-rank test for the paired groups of data which are not independent of each other and do not follow the normal distribution. Almost all p-values resulting from this test were higher than threshold 0.05 and thus we could conclude that there was no significant difference in homopolymer lengths between the groups. On the other hand, these results could be negatively affected by many factors, and thus our hypothesis that the MSI status should be lower after the surgery could be further analysed in future work.



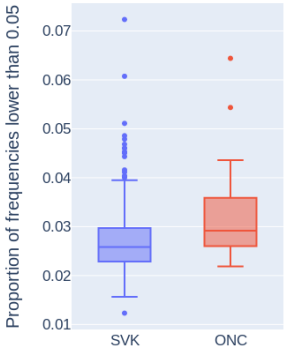
(a) Average of frequency values (Mann–Whitney U = 5344, n_1 = 200, n_2 = 82, P = 2.20002e-06, two-tailed)



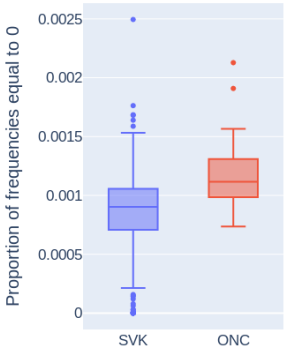
(b) Median of frequency values (Mann–Whitney U = 8178.5, n_1 = 200, n_2 = 82, P = 0.48653, two-tailed)



(c) Relative frequencies lower than 0.1 (Mann–Whitney U = 5263, n_1 = 200, n_2 = 82, P = 1.16896e-06, two-tailed)

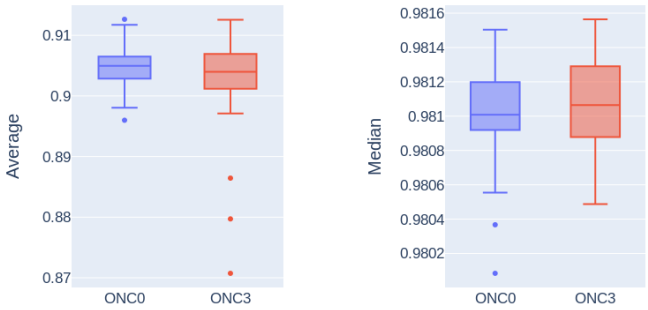


(d) Relative frequencies lower than 0.05 (Mann–Whitney U = 5214, n_1 = 200, n_2 = 82, P = 7.91084e-07, two-tailed)



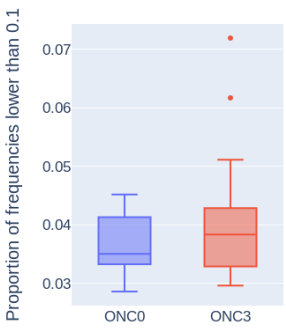
(e) Relative frequencies equal to 0 (Mann–Whitney U = 3739, n_1 = 200, n_2 = 82, P = 3.68497e-13, two-tailed)

Figure 4.2: Distributions of Slovak healthy and oncological samples using different metrics applied to relative frequency values

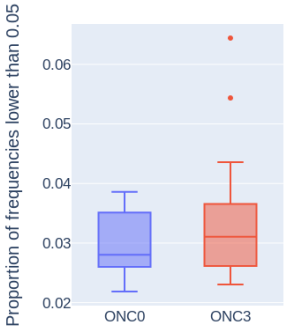


(a) Average of frequency values (Wilcoxon signed-rank test Z = 359, n = 41, P = 0.3542)

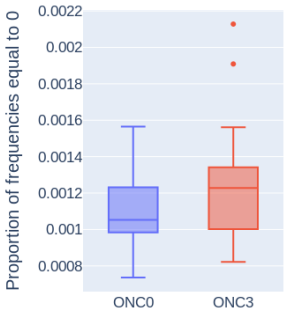
(b) Median of frequency values (Wilcoxon signed-rank test Z = 306, n = 41, P = 0.1621)



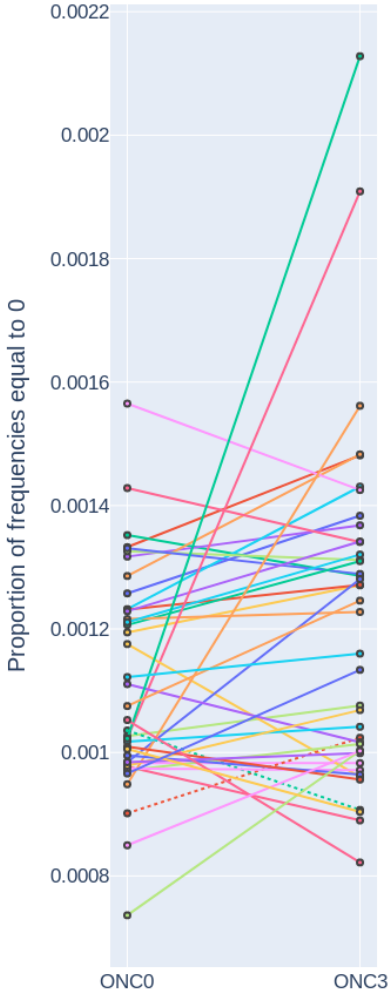
(c) Frequencies lower than 0.1 (Wilcoxon signed-rank test Z = 330, n = 41, P = 0.1928)



(d) Frequencies lower than 0.05 (Wilcoxon signed-rank test Z = 312, n = 41, P = 0.1246)



(e) Frequencies equal to 0 (Wilcoxon signed-rank test Z = 254, n = 41, P = 0.0222)



(f) Frequencies equal to 0

Figure 4.3: Distributions of Slovak and Indian sample sets using different metrics applied to relative frequency values

Conclusion

Microsatellites have been widely studied in genealogy and forensics since their highly variable nature is well-suitable for comparison of individuals. On the other hand, their potential as oncology biomarkers have emerged recently and should be more explored to fully assess their usability in clinical diagnostics. We analysed large cohorts of sequenced genomic data to determine common forms of homopolymers in the Slovak population. We studied the interpopulation differences of homopolymer lengths with a spatially distant Indian population and compared genomic variants of a healthy population with variants of oncological patients.

At first, we located all homopolymer positions across the human genome. Next, we mapped the reads of all analysed samples to the reference genome and identified the genetic variants. We aggregated the obtained genomic data and determined the typical forms and relative frequencies of homopolymer lengths separately for each of 7,749,621 loci in the Slovak population. The computed population-specific frequencies are of great importance for clinicians who can use this data for different biomedical applications. Our custom scripts can be also further used to calculate the frequencies of homopolymer lengths for any population or can be applied to any studied data set of samples.

Using different metrics we were able to compare the frequency distributions between various data sets. Via applying the statistical test to each performed comparison (Slovak vs. Indian, Slovak vs. oncological and oncological before vs. oncological after surgery) we showed that there was a significant difference between the Slovak control and Indian data sets as well as Slovak control and oncological samples whereby this trend was much more notable in the oncological samples. Therefore, we imply that the determined population frequencies have great potential usage, especially in oncology to detect and monitor the ongoing oncological disease. However, the determination of a microsatellite instability status of a tested subject itself have the potential to provide relevant prognostic information and guide therapeutic choices. The comparison of oncological paired samples did not reflect our expectations. There are different possible options that could explain the unexpected results. The tumour could be incompletely removed, cancer biomarkers could be at that time still present in blood and thus samples should be taken from patients later than 3 days after surgery or the statistical

modelling should be improved.

Although there was a significant trend that oncological samples divert from the common forms of homopolymers in the population, the identified differences were not sufficient for an explicit classification of healthy and oncological samples. Therefore, more advanced statistical modelling could be used in a future work that could possibly involve the development of a classifier that would be capable to recognise and differentiate between pathogenic and healthy samples based on microsatellite variants. It could be also interesting to focus separately on different types of cancers and examine a microsatellite instability status within each class because certain cancers are more commonly associated with high instability status.

Bibliography

[1] Jaroslav Budis, Juraj Gazdarica, Jan Radvanszky, Maria Harsanyova, Iveta Gazdaricova, Lucia Strieskova, Richard Frno, Frantisek Duris, Gabriel Minarik, Martina Sekelska, et al. Non-invasive prenatal testing as a valuable source of population specific allelic frequencies. *Journal of biotechnology*, 299:72–78, 2019.

[2] Ralph B d'Agostino. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–348, 1971.

[3] Hans Ellegren. Heterogeneous mutation processes in human microsatellite dna sequences. *Nature genetics*, 24(4):400–402, 2000.

[4] Hao Fan and Jia-You Chu. A brief review of short tandem repeat mutation. *Genomics, Proteomics & Bioinformatics*, 5(1):7–14, 2007.

[5] Melissa Gymrek. A genomic view of short tandem repeats. *Current opinion in genetics & development*, 44:9–16, 2017.

[6] Holly A Hammond, Li Jin, Y Zhong, C Thomas Caskey, and Ranajit Chakraborty. Evaluation of 13 short tandem repeat loci for use in personal identification applications. *American journal of human genetics*, 55(1):175, 1994.

[7] Ayat Hatem, Doruk Bozdağ, Amanda E Toland, and Ümit V Çatalyürek. Benchmarking short sequence mapping tools. *BMC bioinformatics*, 14(1):184, 2013.

[8] Steven R Head, H Kiyomi Komori, Sarah A LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R Salomon, and Phillip Ordoukhanian. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, 56(2):61–77, 2014.

[9] Shaun M Heale and Thomas D Petes. The stabilization of repetitive tracts of dna by variant repeats requires a functional dna mismatch repair system. *Cell*, 83(4):539–545, 1995.

[10] Sohyun Hwang, Eiru Kim, Insuk Lee, and Edward M Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific reports*, 5:17875, 2015.

[11] Mehdi Kchouk, Jean-François Gibrat, and Mourad Elloumi. Generations of sequencing technologies: from first to next generation. *Biology and Medicine*, 9(3), 2017.

[12] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[13] Jerzy K Kulski. Next-generation sequencing—an overview of the history, tools, and "omic" applications. *Next Generation Sequencing–Advances, Applications and Challenges*, pages 3–60, 2016.

[14] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[15] Eyal Nadir, Hanah Margalit, Tamar Gallily, and Shmuel A Ben-Sasson. Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. *Proceedings of the National Academy of Sciences*, 93(13):6470–6475, 1996.

[16] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[17] Eder Jorge Oliveira, Juliano Gomes Pádua, Maria Imaculada Zucchi, Roland Vencovsky, and Maria Lúcia Carneiro Vieira. Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, 29(2):294–307, 2006.

[18] Harry T Orr and Huda Y Zoghbi. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.*, 30:575–621, 2007.

[19] Ondrej Pös, Jaroslav Budis, Zuzana Kubiritova, Marcel Kucharik, Frantisek Duris, Jan Radvanszky, and Tomas Szemes. Identification of structural variation from ngs-based non-invasive prenatal testing. *International journal of molecular sciences*, 20(18):4403, 2019.

[20] Guy-Franck Richard, Alix Kerrest, and Bernard Dujon. Comparative genomics and molecular dynamics of dna repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.*, 72(4):686–727, 2008.

[21] Stephen J Salipante, Sheena M Scroggins, Heather L Hampel, Emily H Turner, and Colin C Pritchard. Microsatellite instability detection by next generation sequencing. *Clinical chemistry*, 60(9):1192–1199, 2014.

[22] D Peter Snustad and Michael J Simmons. *Principles of genetics*. John Wiley & Sons, 2015.

[23] Subbaya Subramanian, Rakesh K Mishra, and Lalji Singh. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome biology*, 4(2):R13, 2003.

[24] James X Sun, Agnar Helgason, Gisli Masson, Sigríður Sunna Ebenesersdóttir, Heng Li, Swapan Mallick, Sante Gnerre, Nick Patterson, Augustine Kong, David Reich, et al. A direct characterization of human mutation based on microsatellites. *Nature genetics*, 44(10):1161, 2012.

[25] Naoko Takezaki and Masatoshi Nei. Empirical tests of the reliability of phylogenetic trees constructed with microsatellite dna. *Genetics*, 178(1):385–392, 2008.

[26] A Urquhart, CP Kimpton, TJ Downes, and P Gill. Variation in short tandem repeat sequences—a survey of twelve microsatellite loci for use as forensic identification markers. *International journal of legal medicine*, 107(1):13–20, 1994.

[27] WAYNE P Wahls, LJ Wallace, and PETER D Moore. The z-dna motif d (tg) 30 promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture. *Molecular and cellular biology*, 10(2):785–793, 1990.

Appendix: content of electronic appendix

In the electronic appendix of this work are included the following files:

- analysis.py — script analysing sample and creating an NumPy array
- bed_format.py — script locating homopolymer positions and storing them in the file ref.bed
- count_homopolymers.py — script that counts homopolymers of each length
- counts.npz — NumPy matrix with homopolymer counts in the Slovak population
- create_matrix.py — script aggregating NumPy vectors into a NumPy matrix counts.npz
- freqs.npz — NumPy matrix containing relative frequencies
- homopolymer_counts, homopolymer_counts_input — files containing counts of homopolymers in the reference genome
- ref.bed — file containing homopolymer positions
- ref_names
- Snakefile1 — file for distributed running of the script analysis.py
- Snakefile2 — file for running the script create_matrix.py
- SVK — samples from the Slovak control data set
- SVK-POP — samples from the Slovak data set
- analysis.ipynb — jupyter notebook (creating graphs, calculation of the population frequencies)