# COMPUTATIONAL METHODS FOR ANALYZING NGS DATA TO DISCOVER CLINICALLY RELEVANT MUTATIONS

by
BEKİR ERGÜNER

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Sabancı University
July 2017

COMPUTATIONAL METHODS FOR ANALYZING NGS DATA TO DISCOVER CLINICALLY RELEVANT MUTATIONS

APPROVED BY:

Prof. Dr. İsmail Çakmak
(*Thesis Supervisor*)

Prof. Dr. Osman Uğur Sezerman

Prof. Dr. Yücel Saygın

Assoc. Prof. Dr. Devrim Gözüaçık

Assoc. Prof. Dr. Muhammed Oğuzhan Külekci

DATE OF APPROVAL: 24/07/2017

ABSTRACT

COMPUTATIONAL METHODS FOR ANALYZING NGS DATA TO DISCOVER CLINICALLY RELEVANT MUTATIONS

BEKİR ERGÜNER

Ph.D. Dissertation, July 2017

Supervisor: Prof. Dr. İsmail Çakmak

**Keywords:** Genome, next generation sequencing, structural variation, mutation, Mendelian disorders

The advent of Next Generation Sequencing platforms started a new era of genomics where affordable genome wide sequencing is available for everyone. These technologies are capable of generating huge amounts of raw sequence data creating an urgent demand for new computational analysis tools and methods. Even the simplest NGS study requires many analysis steps and each step has unique challenges and ambiguities. Efficiently processing raw NGS data and eliminating false-positive signals have become the most challenging issue in genomics. It has been shown that NGS is very effective identifying disease-causing mutations if the data is processed and interpreted properly. In this dissertation, we presented an effective whole genome/exome analysis strategy which has successfully identified novel disease-causing mutations for Cerebrofaciothoracic Dysplasia, Klippel-Feil Syndrome, Spastic Paraplegia and Northern Epilepsy. We also presented a *k*-mer based method for finely mapping genomic structural variations by utilizing *de novo* assembly and local alignment. Compared to the mapping based read extraction method, the *k*-mer based method improved detection of all types of structural variations, in particular detection rate of insertions increased 21%. Moreover, our method is capable of resolving complete structures of complex rearrangements which had not been accomplished before.

# ÖZET

## KLİNİKLE İLİŞKİLİ MUTASYONLARIN KEŞFİNDE ETKİN YENİ NESİL DİZİLEME VERİSİ ANALİZ METOTLARI

BEKİR ERGÜNER

Doktora Tezi, Temmuz 2017

Tez Danışmanı: Prof. Dr. İsmail Çakmak

**Anahtar Kelimeler:** Genom, yeni nesil dizileme, yapısal varyasyonlar, mutasyon, Mendel hastalıkları

Yeni nesil dizileme (YND) teknolojileri sayesinde, genom çapında dizileme yapmanın herkes tarafından erişilebilir olduğu bir devir başladı. Bu teknolojiler aracılığıyla devasa boyutlarda veri üretilmesi yeni analiz metotlarının ve yazılımlarının geliştirilmesi için acil ihtiyaçlar doğurdu. En basit YND çalışması bile birçok analiz basamağı gerektirmektedir. Bununla birlikte her bir analiz basamağı da kendine özgü zorluklara ve yanılsamalara sahiptir. Günümüzde, ham YND verisini verimli bir şekilde analiz ederken yanlış pozitif sinyallerin de düşük miktarda tutulması genomik sahasının en önemli sorunu haline gelmiştir. YND verisinin doğru analiz edilmesi ve yorumlanması sayesinde kalıtsal hastalıklara yol açan mutasyonların keşfinde çok etkili olduğu birçok araştırma tarafından gösterilmiştir. Bu çalışmada, özgün mutasyonların bulunmasında çok etkin bir tüm genom ve tüm ekzom verisi analiz yöntemi sunulmuştur. Geliştirdiğimiz bu yöntemle Serebrofasiotorasik Displazi, Klippel-Feil Sendromu, Spastik Paraparezi and Kuzey Epilepsi hastalıklarına sebep olan özgün mutasyonları keşfetmeyi başardık. Bunun yanı sıra, yapısal varyasyonların hassas haritalanması için kullanılan, *de novo* birleştirme ve lokal hizalamadan faydalanan *k*-mer bazlı bir metot geliştirdik. Haritalama verisine bağlı metoda kıyasla *k*-mer bazlı metot her çeşit yapısal varyasyonun tespitinde daha iyi sonuç verdi. Ayrıca geliştirdiğimiz bu metot daha önce başarılamamış olan kompleks yapıdaki yapısal varyasyonların çözümünü de yapabilmektedir.

*To her...*

## ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to Prof. Dr. Uğur Sezerman for being a wonderful supervisor and a great mentor to me. His guidance helped me to complete this work and shape my career as a researcher. I am grateful for being his student for the past 12 years. I also wish to thank my dissertation supervisor Prof. Dr. Ismail Çakmak and jury members Prof. Dr. Yücel Saygın, Assoc. Prof. Dr. Devrim Gözüaçık, and Assoc. Prof. Dr. Muhammed Oğuzhan Külekci for their valuable review and comments.

Special thanks goes to my collegues at IGBAM, Buğra Özer, Ahmet Çakmak, Zeliha Görmez, Betül Yücetürk, Ömer Faruk Gerdan, Mete Akgün, Pınar Kavak, Yağmur Gök, Aydan Saraç, Bayram Yüksel, Hüseyin Demirci and Mahmut Şamil Sağıroğlu for their support and friendship. It had been a wonderful experience working at BILGEM thanks to their dexterity and companionship. I owe special thanks to Prof. Dr. Nurten Akarsu, Assoc. Prof. Dr. Fatih Bayraklı, Dr. Yavuz Şahin and Dr. Arda Çetinkaya for their invaluable collaborative work in clinical studies.

I would like to thank The Turkish Ministry of Development and TUBITAK-BILGEM for funding The Advanced Genomics and Bioinformatics Research Center (IGBAM). Without IGBAM, I would not be able to continue my career as a bioinformatician and create this work.

Last but not least, I want to express my most sincere gratitude to my dear wife Şerife Rabia. It would not be possible to complete this dissertation without her help and support. I appreciate everything she has done for me and I owe her for a lifetime.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## ABBREVIATIONS

| | |
|---|---|
| DNA | Deoxyribonucleic acid |
| NGS | Next generation sequencing |
| WGS | Whole genome sequencing |
| WES | Whole exome sequencing |
| SNV | Single Nucleotide variation |
| Indel | Insertion and deletion |
| SNP | Single nucleotide polymorphism |
| SV | Structural variation |
| CNV | Copy number variation |
| bp | Base pairs |
| Kbp | Kilo base pairs |
| GB | Gigabytes |
| PCR | Polymerase chain reaction |
| STR | Short Tandem Repeat |
| LCR | Low complexity region |
| SR | Short read |
| ELMO2 | Engulfment and Cell Motility Protein 2 |
| ROH | Runs of homozygosity |
| CFT | Cerebro Facio Thoracic dysplasia |
| TMCO1 | Transmembrane and Coiled-Coil Domain-Containing Protein 1 |
| KFS | Klippel-Feil syndrome |
| MEOX1 | Mesenchyme Homeobox 1 |
| SP | Spastic paraplegia |
| KLC4 | Kinesin Light Chain 4 |
| NCL | Neuronal ceroid lipofuscinosis |
| CLN8 | Ceroid-Lipofuscinosis, Neuronal 8 |
| OCL | Overlap consensus layout |
| DBG | de Bruijn graph |

# 1 INTRODUCTION

Genome is the complete genetic information of an organism which defines its biological traits. Most of the structural, functional and regulatory information such as sequences of RNA and proteins is encoded and stored in the genome. Although more than 99% of the genomic sequence is common among the individuals of the same species, it is the small percentage of the genome which gives them their identities and many phenotypic properties. Most of today's genetics studies are oriented towards the discovery of these small differences among the genomes of individuals in order to associate discrete genomic regions (i.e. genes) with specific biological traits such as hereditary disorders. In the past decade, the number of such discoveries increased dramatically with the help of latest DNA sequencing technologies. The accumulation of data from genetic studies enabled clinicians to suggest optimal treatment strategies to the patients based on their genomic structure. Today, the advent of Next Generation Sequencing (NGS) technologies is paving the way for personalized medicine and pharmacogenomics [1, 2].

The latest developments in the Next Generation Sequencing technologies greatly increased the sequencing throughput while decreasing the costs. It took 15 years and cost 3 billion US dollars for The Human Genome Project to completely sequence the first draft sequence of the human genome [3]. In contrast, it is now possible to sequence whole genome sequence of a person for as low as one thousand US dollars and it is projected the prices will keep falling in the near future [4]. These figures show that genome sequencing will be more and more commonly used in the near future. Whole exome sequencing (WES), a method where the DNA from coding regions is captured for sequencing, has been the major method used by researchers during the last decade. It is mainly used for detecting deleterious single nucleotide variations (SNV) and small insertions and deletions (indels) in the translated part of the genes which are important for their clinical implications. Whole genome sequencing (WGS) is used for sequencing almost every region in the genome to conduct more comprehensive analysis. It is possible to detect SNVs and small indels as well as large genomic structural variations in the non-coding regions of the genome using WGS.

The most widely used NGS platforms can produce millions or even billions of short read sequences with sizes ranging between 100 and 300 base pairs (bp). These short reads are then aligned back to a reference genome in order to construct the sample's genome sequence. After the alignment process, genomic variations specific to the sample can be identified under two categories, first category is the SNVs and small indels and the second one is the large structural variations. Detection, functional analysis and interpretation of these two variation categories have different kinds of difficulties and considerations. SNVs and indels can be detected with relatively high accuracy and sensitivity by short read sequencing [5,6]. However, because SNVs and small indels appear in vast numbers it is challenging to confirm which ones have deleterious effect on the genes and whether they are clinically relevant or not. On the other hand, structural variations would usually cause a deleterious effect if they appear inside a gene but it is more complicated to detect and validate them using short read sequences [7].

The aim of this dissertation is to develop methods for accurately identifying clinically relevant mutations from the genome data generated by using next generation sequencing platforms. For this purpose I present a computational framework for analyzing WGS and WES data which has been successful to identify the causative mutations of 4 rare Mendelian disorders. I also present a novel method which can be used for fine mapping complex genomic rearrangements using WGS or targeted sequencing data. A *k*-mer based read extraction strategy is used in this method which increased the detection rate of both large deletions, insertions and inversions.

# 2 BACKGROUND AND RELATED WORK

## 2.1 Next Generation Sequencing

"*Next generation sequencing"* is the term used for various high throughput sequencing platforms that are capable of sequencing millions of short DNA fragments at each run. They generate short DNA sequences, called short reads, with sizes ranging between 50 bp and 600 bp. Recently, a new generation of sequencing platforms have been developed which can sequence much longer DNA sequences with lower throughput and lower accuracy. In order to differentiate these long read generating platforms some scholars started using the term "*second generation sequencing*" for short read generating high throughput sequencing platforms. In this dissertation, we continue to use the term *NGS* to address the second generation sequencing platforms for simplicity and easy understanding.

As in any DNA sequencing process the NGS methods start with extraction of genomic DNA (Figure 2.1A). Extraction can be done manually by using conventional methods or using commercially available extraction kits. The main objective for the extraction step is to extract ample amount of DNA without causing too much fragmentation. The extracted DNA is then carefully sheared into smaller fragments in order to obtain fragments with sizes closer to the optimal value required by the sequencing platform. The sheared fragments goes into a size selection process for selecting the fragments with desired size and discarding shorter and longer fragments. Sequencing adapters which are specific to the sequencing platform are ligated at both ends of the selected DNA fragments. Indexed adapters can be used for multiplexing to sequence multiple samples together. Depending on the amount of DNA fragments, polymerase chain reaction (PCR) can be used to amplify and adjust the final concentration of the DNA library.

The prepared DNA library can be sequenced using one of the available NGS platforms. Currently the most widely used platforms are Illumina's HiSeq and MiSeq sequencers, and Ion Proton-Torrent sequencer from Life Technologies. In general Illumina's platforms can generate better quality reads [8–10]. Illumina also offers pair-

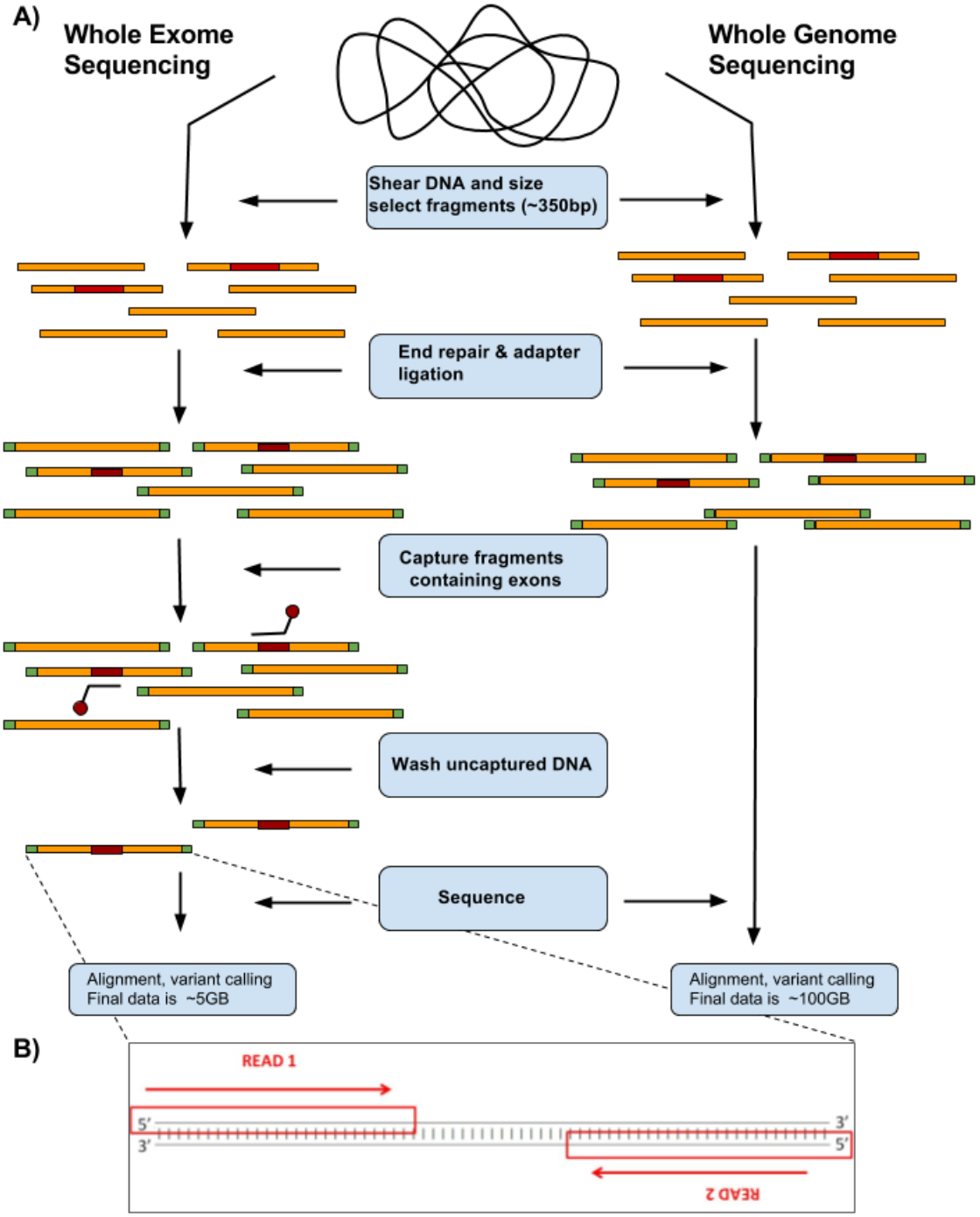


**Figure 2.1:** NGS sequencing procedure. A) Whole exome sequencing versus whole genome sequencing processes. B) Pair-end sequencing of the DNA fragments in the library.

end sequencing method on its platforms by which pairs of short reads can be generated by sequencing both ends of the DNA fragments. Pair-end sequencing is widely adopted in genomics community for being useful in detection of large structural variations and scaffolding contigs in *de novo* genome assemblies. The sequence data used in this dissertation was also generated by using pair-end sequencing on Illumina's platforms (Figure 2.1B).

### 2.1.1 Whole Genome Sequencing

Whole genome sequencing is performed for sequencing every possible region of the genome. It is possible to detect SNVs, small indels and large structural variations which appear in intergenic, intronic and exonic regions using WGS. It is also possible to detect copy number variations (CNV) of large (more than 1 Kbp) sections of the genome by read count analysis [11, 12]. In order to produce good results with WGS, the data should have uniform coverage throughout the whole genome. It is also important to have tight distribution of fragment sizes (insert sizes) in pair-end sequencing data for detection of structural variations with high sensitivity [7, 13]. The major downside of WGS has been its high sequencing costs however the costs have been constantly dropping and it will be much more affordable to use WGS in the near future [4]. Another concern is the difficulty of storing, transferring and analyzing large amounts of data generated by WGS. Modern cloud based genome analysis services become available for such computational demands. However, researchers and clinicians should still carefully plan for whole genome data analysis if they want to sequence many samples. WGS should be conducted only if the sufficient computing infrastructure would be available for the analysis. A typical WGS experiment with 40x mean depth of coverage yields more than 100 GB of compressed sequence data per sample. Even the variant data has formidable size, more than 3 million of SNVs and small indels are generated per sample. Production of such large volumes of data made development of efficient and effective genome sequence analysis software more important than ever.


**Figure 2.2:** Visualization of aligned short reads. NGS data from WGS (upper half) and WES (lower half). The blue blocks at the bottom indicates some of the exonic regions of the KLC4 gene. Note that highly variable sequence depth (size of the peaks in the middle band) among different exons in WES data. Image created by IGV [14].

### 2.1.2 Whole Exome Sequencing

In whole exome sequencing only the captured DNA fragments coming from exonic sites are sequenced (Figure 2.1). An exome capture array is used for enrichment of exonic DNA fragments during the sample preparation process. Since only about 1% of the genome is translated into proteins, targeting these regions dramatically decreases the required sequencing throughput for generating high depth of coverage at exonic sites. WES is commonly used in clinical researches and diagnostics for its cost effectiveness and for being lightweight to analyze. WES is preferred instead of WGS if it is highly likely that the suspected mutations are in the coding region. Numerous studies have shown that WES was very effective for detection of rare germline mutations causing Mendelian disorders [15] as well as somatic mutations from cancer samples [16–18].

In contrast to WGS, the ability to sequence protein coding regions with high depth of coverage is the main advantage of WES (Figure 2.2). It is most effective for detection of SNVs and small indels occurred at the exonic sites. Because intronic and intergenic

regions are not sequenced and variations in those regions cannot be detected. This may not be a concern because such variations are unlikely to affect the structure of the genes. However it is also possible to miss structural variations, especially inversions and translocations, because of the lack of sequence information outside the exonic regions. This is also valid for the SVs that span across exonic sites if their breakpoints are in intronic or intergenic regions. Copy number variation analysis conducted with WES data is also less sensitive [19] because of the read depth bias introduced by the exome enrichment process during sample preparation (Figure 2.2). Therefore WGS is the better option for detection of structural variations and copy number variations.

## 2.2 NGS Data & Analysis

### 2.2.1 Raw NGS Data

Short reads generated by NGS are stored in a text-based format which is called FastQ. In a FastQ file there are four lines for every short read; the unique read label, the read's nucleotide sequence, a separator and the string for base quality scores (Figure 2.3). The base quality scores show how much confidence does the sequencing platform has for calling the particular base. The characters in the quality string are ASCII encodings of base quality scores in Phred scale:

$$Q = -10 \log_{10} P$$

where $Q$ is the base quality score and $P$ is the probability of the base being an incorrect base call. For current NGS platforms these scores range between 2 and 40. The base quality of 40 means 1 in 10000 of the bases would be incorrect. Although base quality scores are good indications of the sequencing accuracy, the machine generated scores might not represent the reality accurately [9, 20]. Most notably, every NGS platform loses accuracy sequencing short tandem repeat regions (STR) and low complexity regions (LCR). Ion Proton and Roche FLX platforms struggle guessing the length of homopolymer regions which introduces false indels at those sites. In addition to such biases platforms can also introduce pattern specific sequencing errors [21, 22].

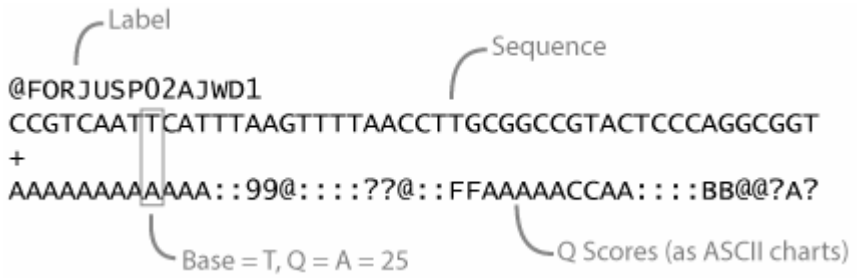


**Figure 2.3:** Demonstration of a short read stored in FastQ format. Here the offset for the base scores is 40 because ASCII equivalent of *A* is 65.

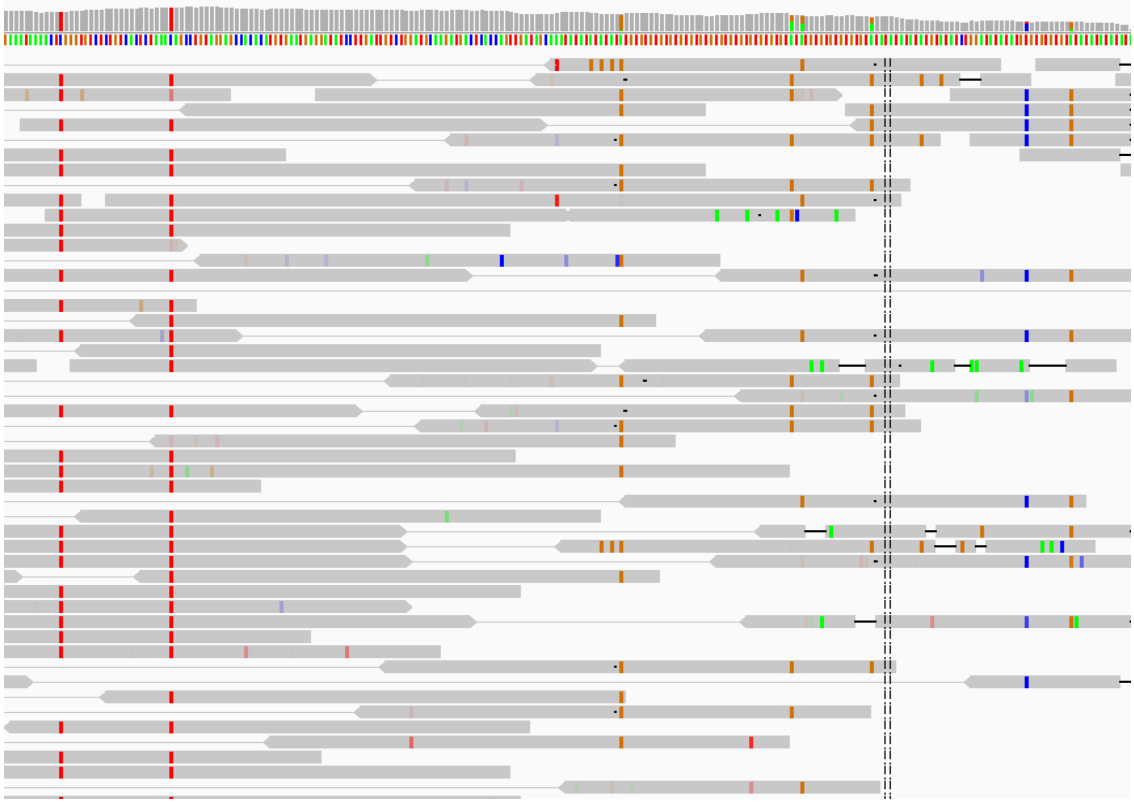

**Figure 2.4:** Visual overview of STR and LCR regions. The two homozygous SNVs on the right side of the image are true, all the mismatches on the reads coincide with each other. Many false variants appear towards the right side caused by sequencing and alignment errors caused by STR and LCR regions.

The sequencing accuracy of NGS platforms is 99% on average [10]. For most of the sequencing projects this is not an acceptable rate, especially when it is compared to 99.9% accuracy rate of capillary/Sanger sequencing. In order to increase their accuracy and sensitivity for variant calling, NGS platforms use their massively parallel sequencing ability to sequence genomes with high depth of coverage (read depth). Although higher sequencing depths can significantly increase the variant calling accuracy, it should be kept in mind that these improvements are less effective in regions such as STR and homopolymer regions, where system specific errors are abundant (Figure 2.4).

### 2.2.2 *de novo* Genome Assembly

*de novo* assembly is the process where short pıeces of sequences (reads) are joined together in order to recreate the original sequence. In genomics, the aim is to generate complete sequences of chromosomes of organisms in order to reveal their whole genome. The actual order of the short sequences is predicted based on alignment of start/end regions between them. This means every short sequence needs to be aligned to every other sequence resulting in a computational complexity of $O(n^2)$ which makes genome assembly a computationally intensive task. Moreover, large amounts of memory is required because most of the processes have to be kept in the memory for increased speed. Therefore assembly of short reads generated by NGS technologies can be an overwhelming task even with the most advanced computing infrastructure because the number of reads generated by NGS can easily reach billions. In particular, *de novo* genome assembly for organisms with large genomes, such as plants and vertebrates, is known to be greatly challenging due to the excessive number of reads required for covering their genomes entirely. In order to overcome these challenges many different algorithms and software have been developed specializing in assembly of short reads generated by NGS platforms.

Most of the available genome assembly software are optimized in order to achieve best performance with minimum memory footprint. In general there are two types of approaches for the genome assembly, the string based approaches and graph based approaches. Many of the string based short read assembly tools implement the Greedy-extension algorithm [23]. Such tools [24–27] are mainly used for highly accurate assembly of small genomes. Compared to the string based methods, graph based methods

are more efficient and can be used for handling the assembly of large and complex genomes [28–30]. The overlap-consensus-layout (OCL) graphs and *de Bruijn* graphs (DBG) are the most common algorithms used by the graph based genome assemblers. In OCL algorithm the assembly graph is created based on the overlaps between reads longer than a certain threshold value (Figure 2.5). DBG are created by first chopping the reads into much smaller pieces (*k*-mers) and edges are formed between adjacent *k*-mers. OCL based tools [31–33] are more suited for assembly of longer reads with low depth of coverage whereas DBG based tools [30,34,35] are better for assembling shorter reads with high coverage data [36]. In contrast to OCL based tools, DBG based tools can be configured by changing the *k*-mer size in order to use less memory. Therefore they have been the primary choice when assembling gigabase-long genomes [28, 29].

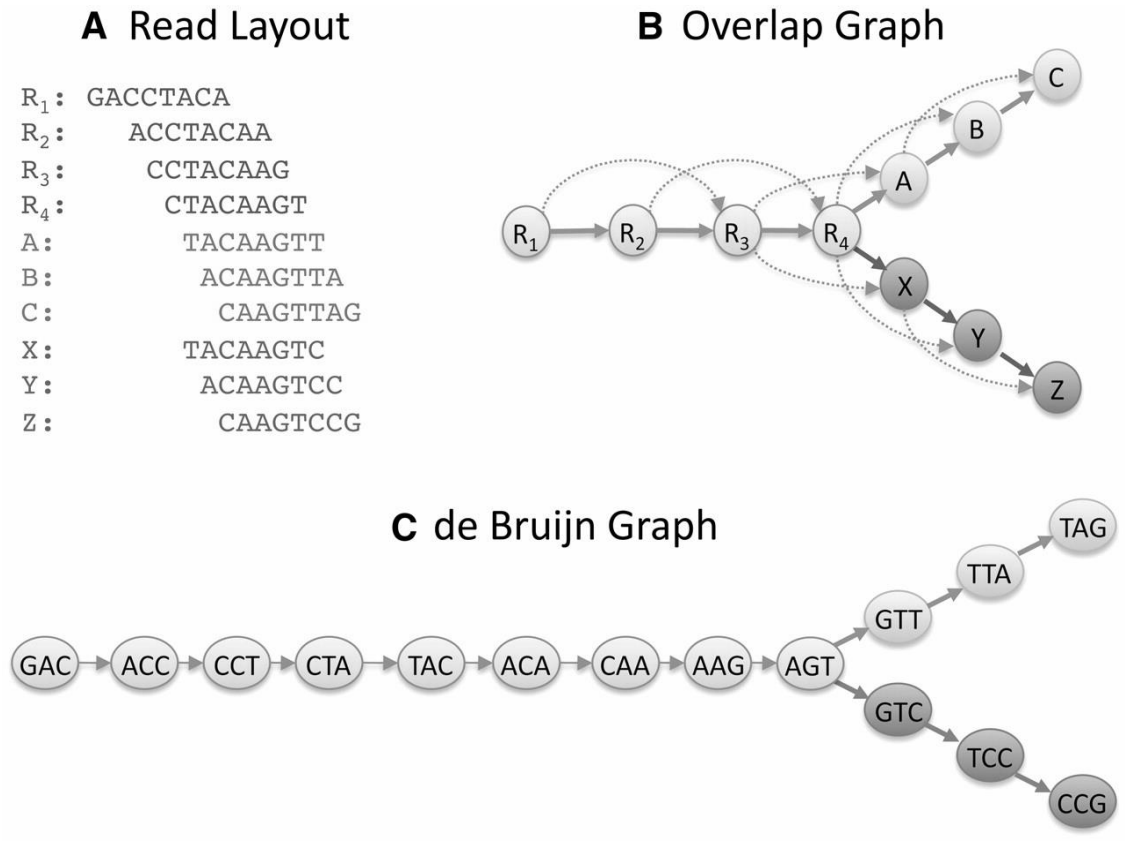


**Figure 2.5:** Comparison of OCL and *de Bruijn* graphs. A) Overlapping reads. B) Overlap graph of the reads. C) DBG generated by using 3-mers in the reads.

For regenerating the original sequence, assemblers search the graphs in order to find a path that visits every node only once, which is called the Euler path. Ideally there should be a single Eulerian path visiting all the nodes in the graph. However in many

cases finding such a path is not possible due to the bubbles, spurs and cycles [37]. Bubbles are formed when an alternate path diverging from the main path and then converging back arises (Figure 2.6A). Such paths are usually caused by polymorphisms or platform specific sequencing errors. Assembly tools can either discard the alternate paths or save it as an alternative assembly. Spurs are short dead-end divergent branches usually caused by sequencing errors (Figure 2.6C). Most assemblers prune and discard these branches and give priority to other paths. Cycles are formed when the main path converges back to a previous point on itself (Figure 2.6B). Such cases appear when there are repeat regions which are longer than the reads or the *k*-mers. Assemblers have to separate the assembly at cyclic regions in the absence of long reads spanning across the repeat region. These separate contiguous assemblies are called contigs.

Most of the large genome assembly projects started with creating relatively short (~10 - 100 kbp) contigs by assembling short (~100 - 300 bp) reads. These contigs were then connected in the correct order using pair-end or mate-pair libraries in order to create longer assemblies in a process called scaffolding [38]. It is very important to have high depth of coverage when using short reads for assembly because increasing the number of reads at each loci would increase the chance of having longer overlapping regions between the reads. Fewer ambiguities like cycles appear by having longer overlapping regions between the reads during creation of the assembly graphs. BAC or fosmid clones can be used in order to solve the assembly of the repeat regions longer than the available short reads. For example, it would be impossible to determine the length and the sequence of a tandem repeat region of length 500bp by using 100 bp short reads. Recently, this approach has been changed by the introduction of the long read sequencing platforms, namely the third generation sequencing platforms. These platforms can provide up to 100 kbp long sequences but with very high error rate, about 10%. In the latest assembly projects, long read data with low coverage is used for creating long contigs with lengths more than several megabases. Then high coverage and high quality short read data is used for correcting sequencing errors on the contigs [39, 40].

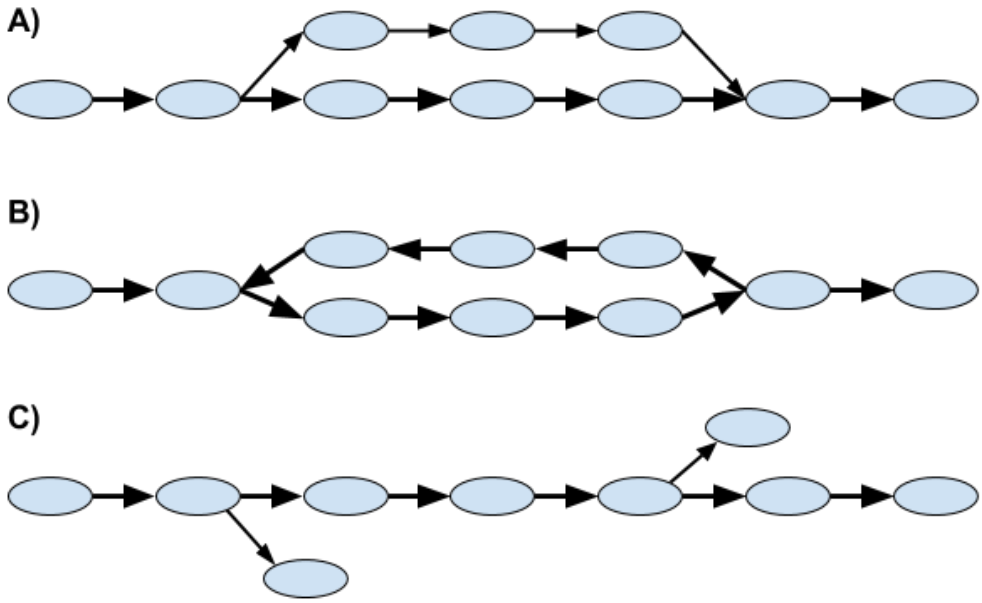


**Figure 2.6:** Depiction of bubbles, cycles and spurs seen in assembly graphs. A) Bubbles are alternative paths diverging from the main path and converging back again. B) Cycles are formed when the main path converges back on itself. C) Spurs are short dead-end divergent branches.

### 2.2.3 Short Read Alignment

The raw NGS sequence data does not have the information of genomic locations where the short reads originates from. In order to build the genomic sequence of the sample, short reads must be either assembled or aligned to a reference genome. Because there is a high quality reference genome available for humans the alignment process is preferred for efficiency and speed in most applications. The most challenging part of the alignment process is to provide high speed, high sensitivity and high precision all at the same time. In practice, however, providing high speed alignment has become the dominant factor because of the huge number of short sequences generated by NGS platforms. For this purpose specialized alignment software, called short read mappers, have been developed. These tools can map thousands of reads per second to the human reference genome [41] making alignment of human WGS data, which can have more than one billion reads, a feasible process.

Every short read mapper has its own advantages and disadvantages depending on the algorithm and the implementation method that it's using. The most commonly used mappers use some form of Burrows-Wheeler transform to generate an FM-index from reference genome and search the indexed genome for matching substrings from the reads. BWA [42], Bowtie [43], Bowtie2 [44], SOAP2 [45] are some of the well-known short read mappers that use FM-index. The other commonly used method is called *seed & extend method* where the indexed reference genome is searched for exact matching seeds (*k*-mers) from the reads for finding candidate locations. Then the reads are fully aligned to candidate locations with Smith-Waterman algorithm. MAQ [46], mrFast [47], mrsFast [48], SHRiMP [49], BFAST [50, 51], SSAHA2 [52] are several of the commonly used implementations of seed-extend method. In general, FM-index based tools are faster, especially for mapping exactly matching reads, and require less memory compared to the seed-extend based tools. In theory, seed-extend based methods should be more sensitive at increased mismatch and indel rates. However, in practice FM-index methods can also provide adequate sensitivity when high quality sequence data is provided [41, 49]. This is because newer implementations of FM-index aligners can use subsequences from the reads as seeds and use the Smith-Waterman algorithm to complete the alignment. As a result, FM-index based mappers are more commonly used than seed-extend based mappers in large scale human genome sequencing projects such as 1000 genomes project [53].

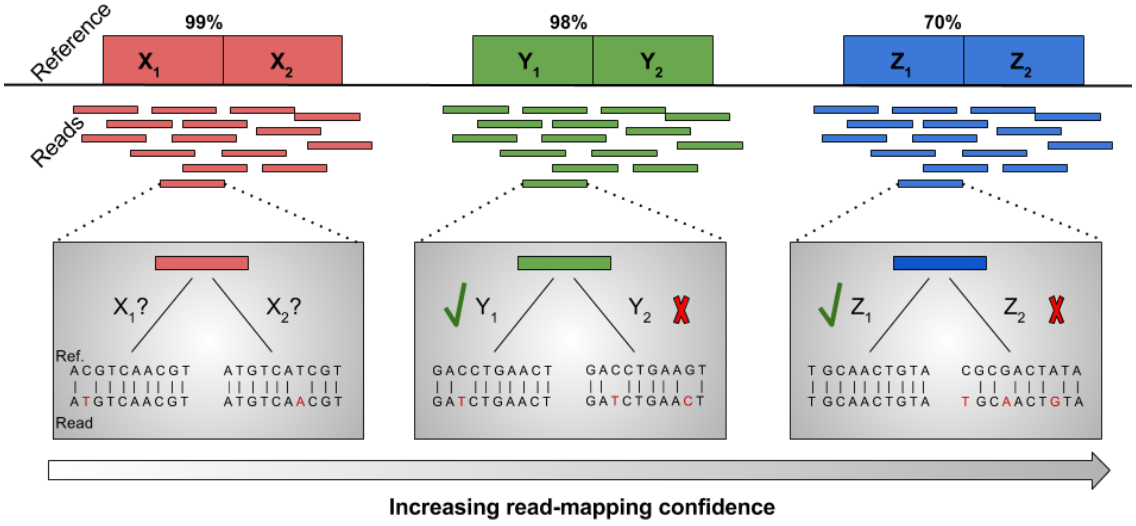


**Figure 2.7:** The effect of repeat regions on short read mapping. Mapping confidence increases as the sequence identity between repeats decreases.

The most difficult aspect of short read alignment process is generating accurate results regardless of the repeat regions existing in the genome. Most organisms have large portions of duplicated regions within their genomes. It was reported that about 50% of the human genome is comprised of repeats of various sizes [54]. On the other hand, only 5% of the human exome fall within the repetitive regions [55]. This increases the chance to accurately detecting deleterious mutations in the exome. However, the variants detected outside the unique regions have to be validated with another sequencing technology before making any decision. The repeat regions cause ambiguities in short read mapping process primarily due to the length of the repeat elements can be longer than the short reads (Figure 2.7). The reads originating from repeat regions can be mapped to multiple locations without any distinction between them. The problem is exacerbated when the mismatches caused by SNPs and sequencing errors are taken into account. The only solution for overcoming repetitiveness of the genome is by increasing read sizes without decreasing sequencing accuracy. By using the current NGS technologies, it is inevitable to have uncertainties in read mapping even with the most accurate read mappers. Therefore the discrepancies caused by the repeat regions should be taken into consideration during downstream analyses.

During the alignment process short reads from repeat regions align to multiple locations on the genome. Because of this property these reads are commonly called as multi-reads. They can either match exactly to multiple locations or there can be several mismatched bases between the alternative locations. Depending on the goal of the sequencing project there are three mapping strategies for handling the multi-reads; single-mapping, multi-mapping or all-mapping and best-mapping. In the single-mapping strategy, multi-reads are randomly assigned to one of the alignment positions which have same identity with the read. Their mapping quality is set as 0 or 255 in order to indicate that they are multi-reads so that the variant callers may discard these reads or they give very low quality scores to the variants detected on them. Single mapping is the most commonly adopted strategy in WGS/WES projects because it is fast especially if the FM-index based aligners were used, such as BWA, Bowtie and SOAP. Most of the aligners can also output a given number of alternative alignment locations for multi reads up to a certain mismatch ratio. Seed-extend based methods like MrFast and BFAST can output all of the possible alignment positions allowing a given mismatch rate. This strategy is applied only in special occasions because it can take thousands of CPU hours mapping a

WGS sample. For example Variationhunter [50] uses all-mapping information in order to increase the detection sensitivity of structural variations. In case of best-mapping, the alignment position with the highest mapping score is reported. This process requires first performing an all-mapping alignment and then selecting the highest scoring alignment. Therefore it requires even longer CPU times than all-mapping. Some of the few mappers such as MAQ and SHRiMP can utilize base quality scores of mismatching bases in order to calculate mapping scores as accurate as possible. This functionality also costs higher computation times therefore best-mapping is only used when the utmost accuracy is required.

### 2.2.4 Calling SNVs and Small Indels

The sole purpose of sequencing the genome of an individual is revealing genetic variations carried by the individual. Therefore detection of variants is the most crucial part of NGS analysis. There are many software tools developed for detecting SNVs and small indels such as GATK [56], VarScan [57], Freebayes [58], Samtools [59] and Strelka [60]. These tools specialize in variant calling based on the coinciding mismatched bases from the reads that were aligned to the same region. The confidence of calling a variant depends on the base qualities of the mismatching bases, mapping confidences of the reads, allele balance and strand bias of the variant. Allele balance is the ratio of reads carrying the variant to the total number of reads overlapping the variant. For a high confidence heterozygous variant call the allele balance should be close to 0.5. Strand bias represents the bias between the number of reads mapped to the forward or the reverse strand of the reference. By using these criteria a confidence score is calculated for every variant which represents the chance of the variant being a false-positive. In addition to the variant score, a genotype score which reflects the confidence for genotyping a particular sample is given for each sample. These scores are important measures to differentiate the true variants from the sequencing and alignment artifacts.

Availability of high quality sequence data together with high depth of coverage is vital for generating high confidence variant calls [61]. Depth of coverage is particularly important for deciding whether the genotype of the called variant is heterozygous or homozygous [61, 62]. The required depth of coverage varies considerably depending on the aim of the sequencing. While 35x mean coverage depth is considered adequate for

calling germline mutations [63], at least 80x coverage is recommended for detection of somatic mutations from tumor samples [64]. The uniformity of the read depth also plays an important role. An average of 80x coverage is required for WES to cover 85%-95% of the target bases due to the differences between capturing efficiencies of the probes [65]. It was shown that SNP data from the population databases can be used in order to increase calling sensitivity at low coverage regions [66]. Also, variant calling from father-mother-child trio data has become a common practice in clinical applications for increasing sensitivity [67].

### 2.2.5 Detecting Structural Variations

Genomic alterations which are larger than 20 bp in size are typically considered as structural variations (SV) [68]. Size of a structural variation can be millions of base pairs. There are cases where entire arms of chromosomes can be deleted, duplicated or inversed. SNVs and small indels are more numerous than the SVs, however, the total size of the genome affected by structural variations is larger [69][70]. In general, SVs can be categorized as insertions, deletions, inversions, duplications and translocations (Figure 2.8). Most of the short read mappers cannot perform complete alignment of reads affected by the SVs due to the alignment limitations forced by performance concerns [71]. Therefore standard variant callers that depend on the alignment information of individual reads cannot identify SVs. Special methods have been developed to detect SVs using additional information such as insert size distribution, mate/pair orientations of pair-end/mate-pair reads and split-read alignment to identify large structural variations.

The most commonly used information by SV detectors is the insert sizes of pair-end reads. Insert size is the distance between the first base of the downstream read and last base of the upstream read. Indels would change the mean, median and variance values of the insert sizes at the affected sites (Figure 2.8A-B). This metric is especially effective for detecting the large deletions because the insert sizes increase considerably at the affected regions. Therefore many tools such as BreakDancer [72], CLEVER [68], GASV [73], DELLY [74], HYDRA [75], MoDIL [76], PEMer [77], VariationHunter [50] were developed that use statistics based on the insert sizes for detecting large indels. The methods which solely rely on pair/mate information cannot give the exact positions and the sizes of the structural variants, instead they estimate border positions which contain

the structural variations. Moreover, as the variance of the insert size distribution of DNA libraries increases such methods suffer from loss of sensitivity for smaller indels [68]. In order to achieve higher precision several SV detectors, such as NovelSeq [78], PINDEL [79], SOAPindel [80], Splitread [81], BreakSeq [82] and BreaKmer [83] have been proposed which utilize split-reads and/or one-end-anchored reads. Split-reads contain breakpoints caused by the SVs and they can only be partially aligned by the short read mappers. They are usually output as soft-clipped reads. Unmapped reads with mapped mates are called one-end-anchored reads. Various algorithms are used to extract the exact locations of breakpoints of SVs from the soft-clipped reads and one-end-anchored reads. For example, PINDEL uses a pattern growth algorithm on one-end anchored reads to find indels. BreakMer assembles the novel *k*-mers in order to find and validate breakpoints. NovelSeq focuses on revealing the sequences of novel insertions by assembling the unmapped reads. Another important feature of the pair-end reads is the relative alignment strands of the mates with respect to each other. For normally mapped (concordant) pair-end reads the reads should be on the opposite strands. However inversions causes both mates to be aligned on the same reference strand (Figure 2.8E). Such discordant mates are considered as evidence for existence of an inversion at the affected region. HYDRA, PEMer, DELLY and many other SV detectors can utilize this feature to detect inversions [7].

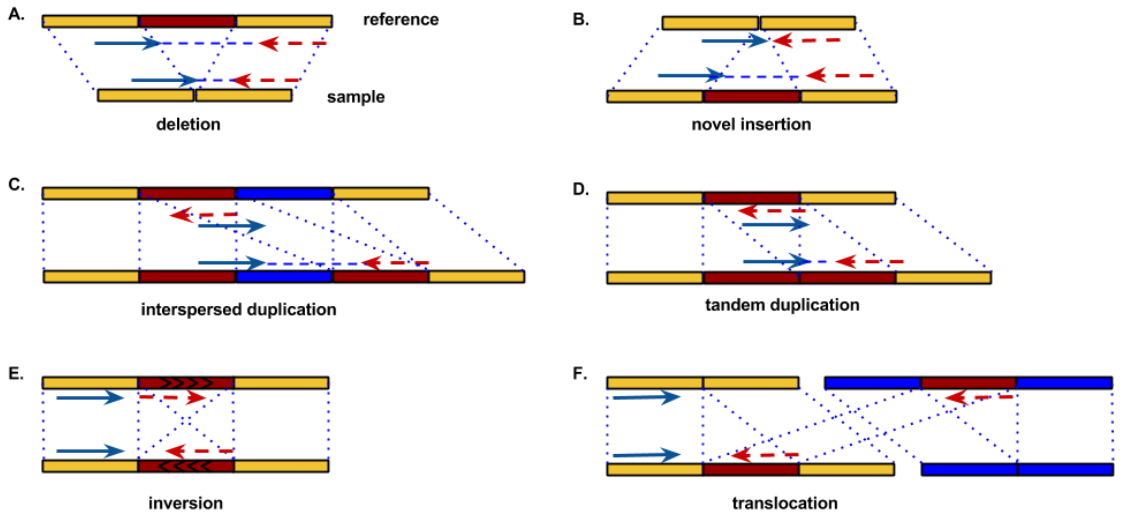


**Figure 2.8:** Demonstration of structural variations. It can be seen how the pair-end read mapping is affected by each SV type.

### 2.2.6 Variant Annotation

Development of computational tools for annotating large sets of variants generated by NGS has become extremely important. Such tools can give insight about the functional impacts of variations on genes. The information generated by the annotation software is crucial for finding the disease-causing mutations. There are many kinds of information which can be inserted into the variant information. The most commonly added information is database IDs and allele frequencies of variations from population SNP databases, such as dbSNP [84] and ExAC [85], which can be used to select rare mutations specific to individuals or cohorts. Another very important information is the effect of exonic variations on the translated protein sequence. SnpEff [86] and ANNOVAR [87] can be used to annotate missense, nonsense, frameshift and splice site mutations. Missense mutations can be further annotated with predictive values reflecting their impact on the phenotype. These values are generated by variant effect predictors such as SIFT [88], PolyPhen 2 [89], MutationTaster [90]. Essential information for the assessment of clinically relevant mutations can be retrieved from HGMD [91], OMIM [92] and ClinVar [93] databases. Any details with genomic location information in BED [94] or VCF [95] format can be added to the variant data by using SnpSift [96], VCFtools [95] and GATK Variant Annotator [56]. The availability of rich and versatile annotation information makes NGS data more powerful for clinical research and diagnostics.

## 2.3 NGS and Mendelian Disorders

Mendelian disorders are mostly monogenic and rare diseases which have inheritance pattern fitting the Mendel's inheritance model. Although each rare disorder individually affects small number of people, it was reported that more than 4% of the newborns have been affected with some type of Mendelian disorder [97]. More than 3000 disorders in OMIM catalog [92] have reports related to their molecular basis. However there are still more than 3500 disorders without any information on their genetic origin. During the last decades great efforts have been made to discover associated genes and mutations with Mendelian disorders. Before NGS, conventional methods used to rely on linkage analysis to narrow down the set of candidate genes into a small number (<300) [98]. These candidate genes were then sequenced one by one using Sanger sequencing in

order to find the disease-causing mutation. This type of genetic study was feasible only if there were enough number of affected samples available [99] and it was not possible to discover *de novo* mutations.

The introduction of NGS methods into rare disease studies has revolutionized the way candidate mutations and genes are found. For autosomal recessive diseases, the number of candidate mutations can be lowered down to 10 or fewer if there are 4 or 5 samples available from the same family. It is even possible to pinpoint the causative mutation from a single patient's whole exome data from non-consanguineous families [100,101]. WGS and WES methods generate large sets of variants which broadly cover the genetic landscape of samples. On the other hand, the deluge of variants detected by NGS methods also contains large number of trivial polymorphisms together with false-positive variants arising from the sequencing and alignment errors. Complex filtering methods should be devised for discarding unrelated and false variants while prioritizing a small set of candidates.

# 3 DISCOVERING CLINICALLY RELEVANT MUTATIONS

## 3.1 Motivation

Today's NGS platforms were engineered to generate the most amount of sequence data for the best price. As a result, there has been an exponential growth in the amount of sequence data [102] in the last decade. Such rapid growth of data volume has created a demand for efficient analysis pipelines each specific to the type of NGS application used. Many analysis steps such as assembly, alignment, variant calling, annotation etc. are necessary for reaching the desired results for all types of NGS applications. Because of the aforementioned sequencing artifacts and technical limitations each analysis step introduces some noise to the resulting data. It is crucial to choose the correct tool at each step in order to minimize the false-positive results while maximizing the recall rate. The correct choice of tools greatly depends on the purpose of the NGS experiment. Here we present an efficient WGS and WES analysis pipeline compiled for detecting germline SNVs and small indels by using the best practices and tools reported in the literature. We utilized parallel processing and distributed computing

methods for increasing the pipeline's throughput to the full capacity of the available computing cluster. We also describe filtering and prioritization strategies devised for revealing the causative mutations of rare Mendelian disorders in the WGS/WES variant data.

## 3.2 Methods

### 3.2.1 Preprocessing Raw Sequence Data

Our pipeline starts with the conversion of base calling files generated by the NGS platforms to FastQ files by using vendor provided software. The platforms used in our project were Illumina HiSeq 2000 and 2500 (Illumina Inc., San Diego, CA, USA). Therefore we used the *bcl2fastq* conversion tool from the Casava software package (version 1.8.2 or later). The conversion tool also performs demultiplexing of the data from multiplexed sequencing runs. We allowed 1 mismatch in the indexes for demultiplexing process in order to avoid single base sequencing errors and collect as many reads as possible. This is the maximum number of mismatches allowed without causing index collision for the KAPA single indexed adapters (Kapa Biosystems, MA, USA) which were used in our WES projects. We separated the FastQ data into multiple files containing 4 million reads to enable distributed parallel processing during the sequence alignment step. The converted and demultiplexed FastQ files were first adapter trimmed using an in-house trimming script. We discarded reads that are shorter than 35 bp to reduce ambiguously aligned reads [103]. We also removed any low quality reads with average base quality score less than 20 during the adapter trimming process in order to lessen false variants caused by sequencing artifacts [104].

### 3.2.2 Sequence Alignment

Short read alignment is the most crucial step for any genome resequencing project. Results of all subsequent analysis steps depend on the sensitivity and the precision of the alignment tool used. Especially for large scale sequencing projects performance has high priority for selecting the right alignment software. We used the BWA's [42] MEM tool which had been extensively benchmarked [41,105] showing

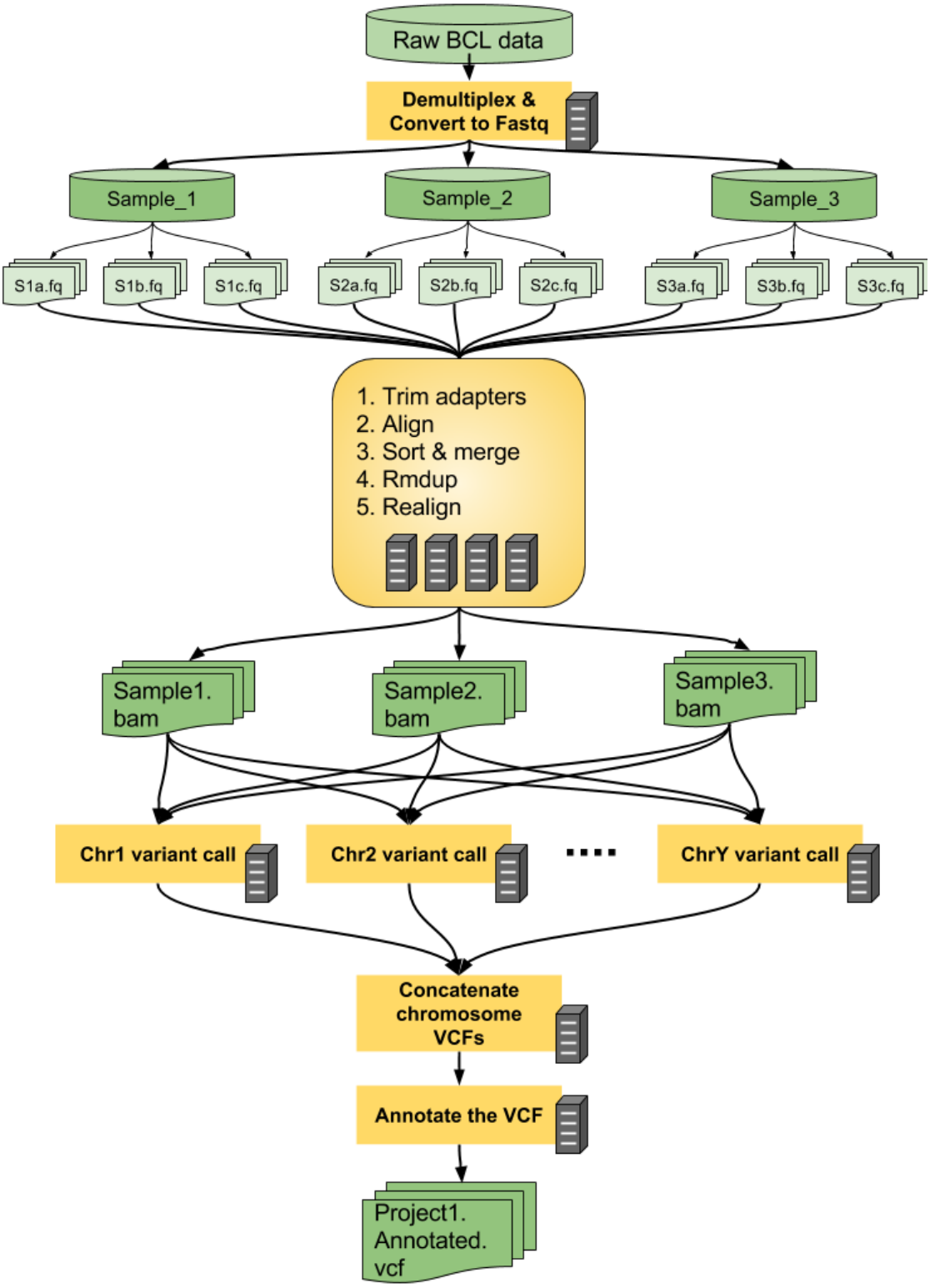


**Figure 3.1:** Workflow diagram of the WGS/WES data analysis pipeline.

that it has very good performance and accuracy. In our experience we have seen that it is one of the most reliable and well maintained SR mapping software available for academic use. We used the default parameters of the BWA which were already optimized for mapping Illumina short reads onto human genome. Pair-end short reads were mapped onto the reference human genome GRCh37 released under the GATK resource bundle (v2.5)[106]. Appropriate read group names were added to aligned sequence data by setting *-r* parameter of the BWA. For parallelization, individual FastQ files were aligned independently by multiple BWA instances running on multiple machines (Figure 3.1). Each instance of the BWA was also run in multithreaded mode for maximum resource utilization. Aligned reads were stored in binary SAM (.bam) files and bam files were position-sorted by using Samtools sort tool [59]. Sorted bam files belonging to the same sample were merged with Samtools merge tool. PCR duplicates were removed to prevent biases in allele balances of variants [107] by processing the merged bam files with Samtools rmdup (Figure 3.2). Finally, GATK IndelRealigner was used to perform multiple sequence alignment around indel sites for correcting the alignment errors caused by the indels (Figure 3.3) [108].

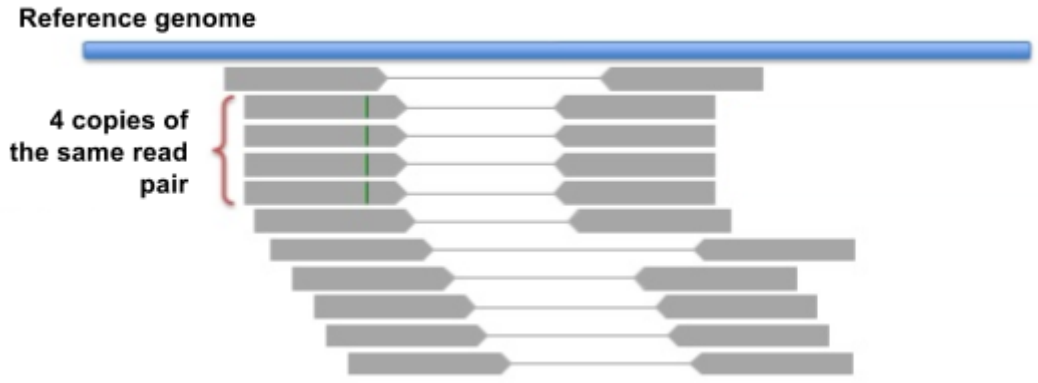


**Figure 3.2:** Bias created by PCR duplicates. PCR duplicates show up as pairs of reads aligning at the same positions. The mismatches on the duplicated reads are probably PCR or alignment artifacts. If the duplicate removal is not performed they might be considered as evidence for a variant.

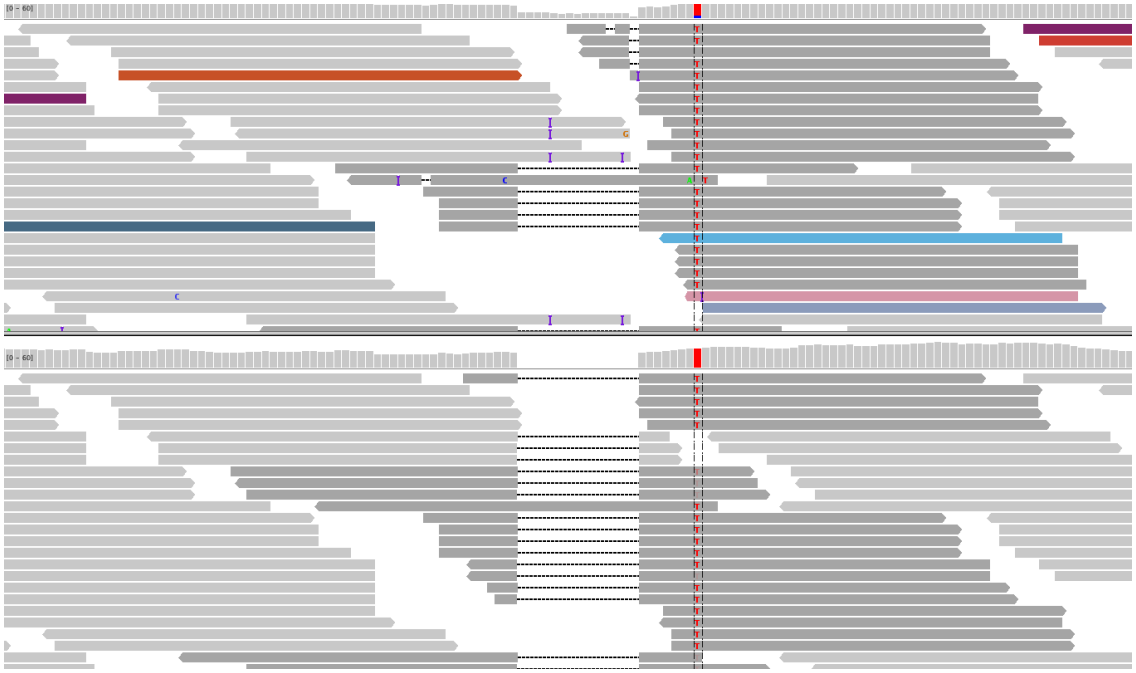

**Figure 3.3:** Effect of indel realignment. Depiction of short read alignments at an indel site before (top) and after (bottom) the multiple sequence alignment. Multiple false indels were eliminated to reveal one large true deletion.

### 3.2.3 Variant Calling

Variant calling is another key step in NGS analysis workflows where the results are highly dependent on the algorithms and statistical methods used for detection of the variants. We have chosen GATK UnifiedGenotyper tool for calling variants because of its high sensitivity for both SNVs and small indels [5]. UnifiedGenotyper allows pooling multiple samples together for increasing sensitivity at low read depths. It is especially important to pool the samples from the same cohort during variant calling because the genotype based filtering is more effective when the variant set is common for all the samples included in the filtering. In addition to pooling we supplied the dbSNP (version 132 or later) variants for increasing the sensitivity. Variant calling is a very computationally intensive task especially for WGS data. In order to distribute the workload and increase speed of the process we performed variant calling for each chromosome separately on multiple machines and then concatenated the output variant files in the same order as the reference genome file. Finally, the variants were inserted into an in-house SQL database for storing population wide variant data. The database

was updated after each batch of samples were analyzed and the population wide allele frequencies of variants were calculated to annotate variants from the next batch.

### 3.2.4 Annotation of the Variants

Availability of rich annotation information increases the efficacy of variant filtering and prioritization for discovering disease causing genes. We used SnpEff [86] to predict the effects of missense mutations and indels. We also used SnpSift [96] to add custom information gathered from various databases to the variants. The detailed information about the annotation fields can be seen on Table 3.1. The final annotated variant data is stored in VCF format [95] for downstream analysis.

### 3.2.5 Discovering Disease Associated Mutations

In this study we have devised an effective variant filtering strategy for rare Mendelian diseases to eliminate unrelated variants and prioritize the potentially related mutations. Our strategy depends on both the segregation analysis of the affected families and prioritization of the variants based on various annotation information. For an efficient analysis the variant data of all the samples from the affected families were collected in a single VCF file. Before starting the filtering process we eliminated low quality variant calls by filtering out the ones with coverage less than 4x and genotype score lower than 15. We used VarSifter [109] tool for applying our filtrations criteria on the annotated VCF files. We began with filtering out the variants with more than 1% allele frequency in the public databases. We also discarded the variants existing in the in-house database (IGBAM, TUBITAK-MAM, Turkey) if the disease had not been studied before. Otherwise an appropriate filtering value, such as 4%, was applied on the in-house allele frequency depending on the scarcity of the disease. We then select the variants which segregated in the family according to the heredity pattern of the disease. For autosomal recessive disease, variants that were heterozygous in parents and homozygous in the affected children were selected in consanguineous families. If the family was non-consanguineous compound heterozygosity was also considered. For autosomal dominant disorders; the variants heterozygous in the affected samples and homozygous-reference in the control samples were selected. Furthermore, if there is information about linkage regions from previous studies then the variants in the linkage regions are selected. For

recessive diseases, we used HomSI [110] to identify runs-of-homozygosity (ROH) regions on the affected samples. ROH regions were used for evaluating variants in the absence of linkage information. The selected variants were further evaluated based on their potential deleterious effects on the proteins. The loss of function (LOF) mutations such as start/stop codon gain/loss variants and frameshift indels were given the highest priority. Missense (nonsynonymous) variants were prioritized based on the collective information gathered from mutation effect predictors. Moreover, variants within the conserved regions (PhastCons score) were also given higher priority. Finally, the short read alignments around the candidate variants were visually scrutinized using IGV [14]. Variants from the regions where sequencing, alignment and variant calling errors were abundant, such as repeat regions (segmental duplications), STR regions and low complexity regions, were discarded or given low priority.

## 3.3 Results

### 3.3.1 Nonsense TMCO1 Mutation Causes CFT

Whole exomes of two family trios (Figure 3.4A) were sequenced and annotated variant data was created with our standard WGS/WES pipeline. The children from both families were phenotypically diagnosed with cerebrofaciothoraic dysplasia (CFT) [122]. Based on the preliminary studies about the families it was devised that the disease was inherited via autosomal recessive model. Therefore we applied our variant filtering strategy for recessive disorders on both families' variant data (Table 3.2). Initially, we filtered out common variants in population with more than 1% minor allele frequency and eliminated the variants existing in our in-house control variant database. The variants with low genotype quality and inadequate coverage were also discarded. From the remaining variant set we selected the variants that are heterozygous in parents and homozygous in affected children. The stop gain mutation (p.Arg87Ter, c.259C>T) on the TMCO1 gene was the only mutation residing in the target haplotype region (2.28 Mbp ROH region in Figure 3.4B) which fitted the segregation pattern of the disease in both families. Sanger sequencing was used to validate the mutation was not an artifact caused by sequencing or alignment errors (Figure 3.4C). Therefore we identified the mutation as the sole candidate causing CFT.

**Table 3.1:** Annotation fields and their explanation.

| Annotation Field | Type | Tool | Description | Source | Reference |
|---|---|---|---|---|---|
| Gene name | String | SnpEff | RefSeq Gene name | RefSeq (SnpEff) | [111] |
| Type of the variant | String | SnpEff | Effect on the translated product: missense, stop_gain. Etc. | SnpEff | [86] |
| dbID | String | SnpEff | dbSNP variant ID | dbSNP | [84] |
| 1000G MAF | Float | SnpSift | 1k Genomes Project allele frequency | 1K Genomes Project | [53] |
| ESP6500 MAF | Float | SnpSift | ESP6500 exome allele frequency | ESP | [112] |
| ExAC MAF | Float | SnpSift | Europe exome allele frequency | The Exome Aggregation Consortium | [113] |
| Icelanders AF | Float | SnpSift | Iceland WGS allele frequency | deCODE | [113, 114] |
| TRdb Het MAF | Float | SnpSift | Ratio of heterozygous carriers | In-house | |
| TRdb Hom MAF | Float | SnpSift | Ratio of homozygous carriers | In-house | |
| Mutation Disease | String | SnpEff | Previous disease associations of the variant | HGMD,ClinVar, OrphaNet | [91], [93], [115] |
| Gene Disease | String | SnpEff | Previous disease associations of the gene | HGMD,ClinVar, OrphaNet | |
| PhastCons score | Float | SnpEff | Conservation score among vertebrates | PHAST | [116] |
| Segmental Duplication | Float | SnpEff | Identity of the repeat regions | SnpEff | |
| Gene Process | String | SnpEff | Gene ontology | MsigDB | [117] |
| PolyPhen2 score | Float | SnpSift | PolyPhen 2 effect score for missense mutations | DBNSFP | [118], [89] |
| PolyPhen2 pred | String | SnpSift | PolyPhen 2 effect prediction | DBNSFP | |
| SIFT pred | String | SnpSift | SIFT effect prediction | DBNSFP | [119] |
| SIFT score | Float | SnpSift | SIFT effect score for missense mutations | DBNSFP | |
| MutationTaster pred | String | SnpSift | MutationTaster effect prediction | DBNSFP | |
| MutationTaster score | Float | SnpSift | MutationTaster effect score for missense mutations | DBNSFP | |
| CADD pred | String | SnpSift | CADD effect prediction | DBNSFP | [120] |
| CADD score | Float | SnpSift | CADD effect score for missense mutations | DBNSFP | |
| HGVS | String | SnpEff | Human genome variant server entry | HGVS | [121] |
| Transcript ID | String | SnpEff | NCBI transcript ID(s) | SnpEff | |
| Amino Acid Change | String | SnpEff | Amino acid change caused by missense mutations | SnpEff | |

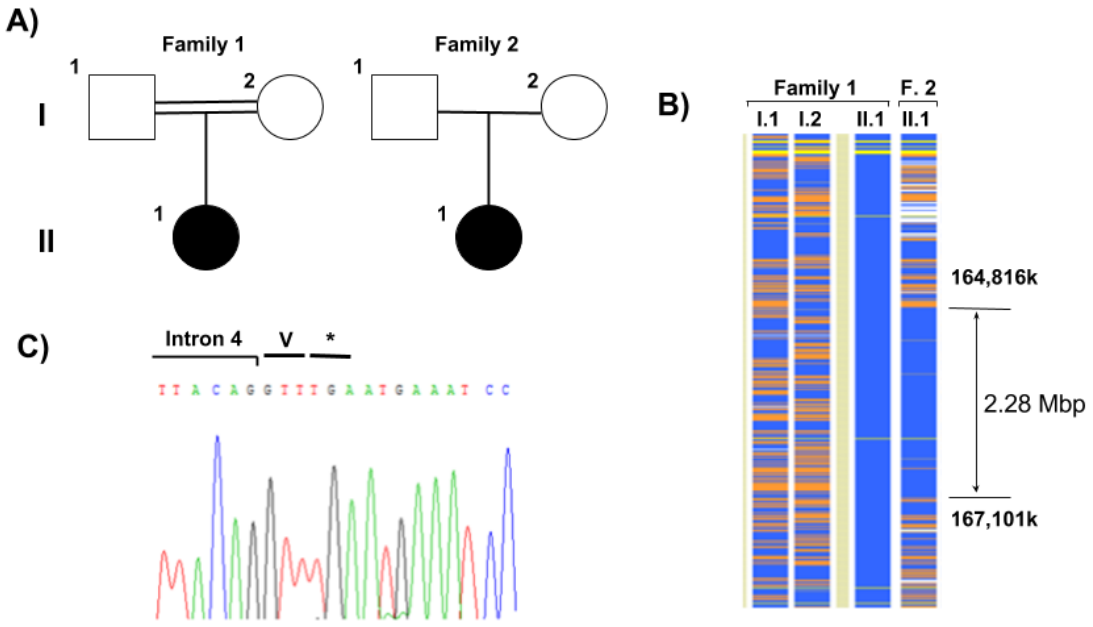


**Figure 3.4:** CFT family. A) Pedigrees of the two families diagnosed with cerebrofaciothoraic dysplasia. B) HomSI output showing the 2.28 Mbp homozygous region in blue color. C) Sanger sequence validating the TMCO1 p.Arg87Ter (c.259C>T) mutation.

**Table 3.2:** Variant counts after each filtering step for CFT families.

| | Family 1 | | | Family 2 | | |
|---|---|---|---|---|---|---|
| | I.1 | I.2 | II.1 | I.1 | I.2 | II.1 |
| Total # of variants | 181,946 | 235,603 | 234,340 | 164,958 | 212,319 | 209,188 |
| >4 coverage and >15 Genotype score | 92,113 | 148,641 | 132,673 | 90,858 | 114,109 | 123,587 |
| Not found in dbSNP 135 or GMAF <0.01 | 23,474 | 39,981 | 35,054 | 24,734 | 30,195 | 32,933 |
| Not found in in-house DB (n=136) | 1,968 | 2,651 | 2,624 | 2,954 | 3,590 | 3,590 |
| Heterozygous in parents, homozygous in children | 65 | | | 3 | | |
| In target haplotype region (chr1:164816956-167101983) | 1 | | | 1 | | |
| Common across families | 1 | | | | | |

Further investigations were performed in order to confirm that the candidate mutation in homozygous state was causing CFT. Expression analysis confirmed that patients with homozygous c.259C>T mutation were TMCO1 deficient while heterozygous carriers and homozygous reference individuals were able to produce TMCO1 [122]. Several studies prior to our findings had also found results supporting our discovery. A previous study reported that other Turkish families which had the same

phenotype were carrying the same p.Arg87Ter mutation [123]. Moreover results from another study [124] had shown that TMCO1 frameshift mutation in an Amish family had caused similar phenotypic properties. These results strongly indicate that the variation discovered by our method was indeed the causative mutation of CFT.

### 3.3.2 Nonsense MEOX1 Mutation Causes Klippel-Feil Syndrome

Whole exome sequencing was performed for 8 individuals from a consanguineous family with 5 children affected by the Klippel-Feil syndrome (KFS) (Figure 3.5). The inheritance pattern of the disease was found to be autosomal recessive [125]. We have analyzed the exome data with our standard pipeline and generated the annotated variant data. We selected the variants that have less than 1% minor allele frequency in the dbSNP database and discarded the variants with less than 50 variant quality score generated by GATK UnifiedGenotyper. For segregation based filtering we selected the variants that are homozygous in affected children and heterozygous in mother. We discarded the variants that are homozygous in the unaffected sibling. A preliminary genome-wide linkage analysis had identified a linkage region between 17:36410559-52907886 with 4.2 LOD score. After applying our filtration strategy we identified 6 mutations in the linkage region (Table 3.3).

Comprehensive literature investigation was performed in order to reveal the most prominent candidate out of the 6 mutations. AOC3 gene is related to leukocyte trafficking and it is expressed on the surface of endothelial cells. AOC3 deficient mice were shown to be healthy and fertile [126]. Mitochondrial ribosomal L27 protein is encoded from MRPL27 gene. Defects in these family of proteins were observed to cause deficiencies in oxidative phosphorylation [127]. The mutation observed in the KRTAP4-11 gene was not located in the conserved region. ORMDL3 gene had been associated with asthma [128]. No functional significance had been reported for GHDC gene. As a result, we prioritized the nonsense mutation p.Q84X occurred on the MEOX1 gene which truncated the ⅔ of the 254 amino acid long protein. It was confirmed with Sanger sequencing that the mutation was homozygous in affected children and heterozygous in parents and in an unaffected child (Figure 3.5). Another study concurrently identified a frameshift deletion in MEOX1 causing KFS [129]. Moreover a previous study had shown that deficiency of MEOX1 and MEOX2 genes caused KFS like phenotype in mice embryos [130]. In the

light of such strong supporting information we have concluded that the nonsense mutation in MEOX1 gene had caused the KFS in this family.

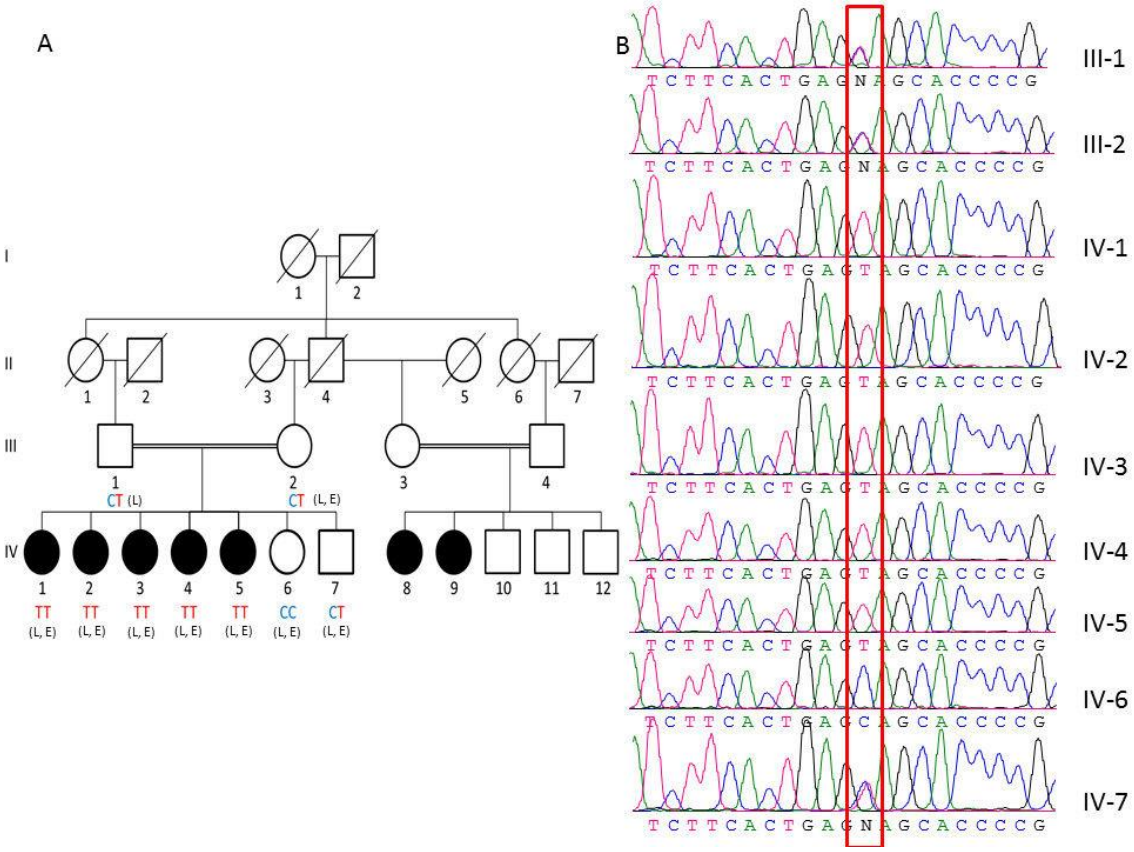


**Figure 3.5:** KFS family. **A)** Pedigree of the extended family. Exome sequencing was performed for II-2, IV-1, IV-2, IV-3, IV-4, IV-5, IV-6, and IV-7. **B)** Sanger sequence validating the nonsense mutation in 9 family members. (Copyright Bayrakli et al. 2013)

**Table 3.4:** The remaining 6 candidate variants after filtering for KFS.

| Chromosome | Genomic locations and variants | Amino acid change | Gene Name |
|---|---|---|---|
| 17 | 38078188_38078205 delCCCATCTTTCCCCAAC | (UTR3) | ORMDL3 |
| 17 | 39274194C>T | p.C125Y | KRTAP4 |
| 17 | 40344303_40344305delCAC | p.W281del | GHDC |
| 17 | 41003401T>C | p.L14P | AOC3 |
| 17 | 41738653G>A | p.Q84X | MEOX1 |
| 17 | 48447413G>A | p.R74S | MRPL27 |

### 3.3.3 Deletion on KLC4 Causes Hereditary Spastic Paraplegia

3 children from a consanguineous family (Figure 3.6A) were clinically phenotyped with progressive complicated spastic paraplegia (SP). We sequenced whole exomes of 3 affected children and their parents. Based on the inheritance pattern observed in the family we filtered the variant data for selecting autosomal recessive mutations (Table 3.5). First, we discarded low quality (<50) variants. Then we selected the variants which were homozygous in affected children and heterozygous in parents. Finally we removed the variants seen in our in-house database and other population wide polymorphism databases. As a result of our filtration process 2 novel mutations had remained. We gave priority to the frameshift deletion (c.853_871del19) on the KLC4 gene because it was inside a 6.5 Mbp ROH region (Figure 3.6B). The existence of the mutations and its segregation pattern was validated with Sanger sequencing [131].

Further investigation confirmed that the 19 bp frameshift deletion caused a termination signal at the 277th codon of the transcript truncating more than half of the 619 amino acid long protein [131]. KLC4 is one of the four isoforms of the KLC proteins which are from the kinesin family. Kinesins are known to be involved in intra cell transportation and microtubule regulation. The truncation caused by the c.853_871del19 deletion removes 4 of the 5 tetratricopeptide regions which are necessary for cargo binding and transportation. Therefore the deletion renders the KLC4 protein completely dysfunctional in the homozygous patients. It was known that defects in microtubule-based transportation mechanism hindered neuronal activities [132]. Moreover the clinical phenotype of the affected children showed strong correlation between KLC1 gene knockout models of drosophila and mice [133,134]. Such findings are strong evidences supporting the association of the KLC4 gene with SP phenotype.

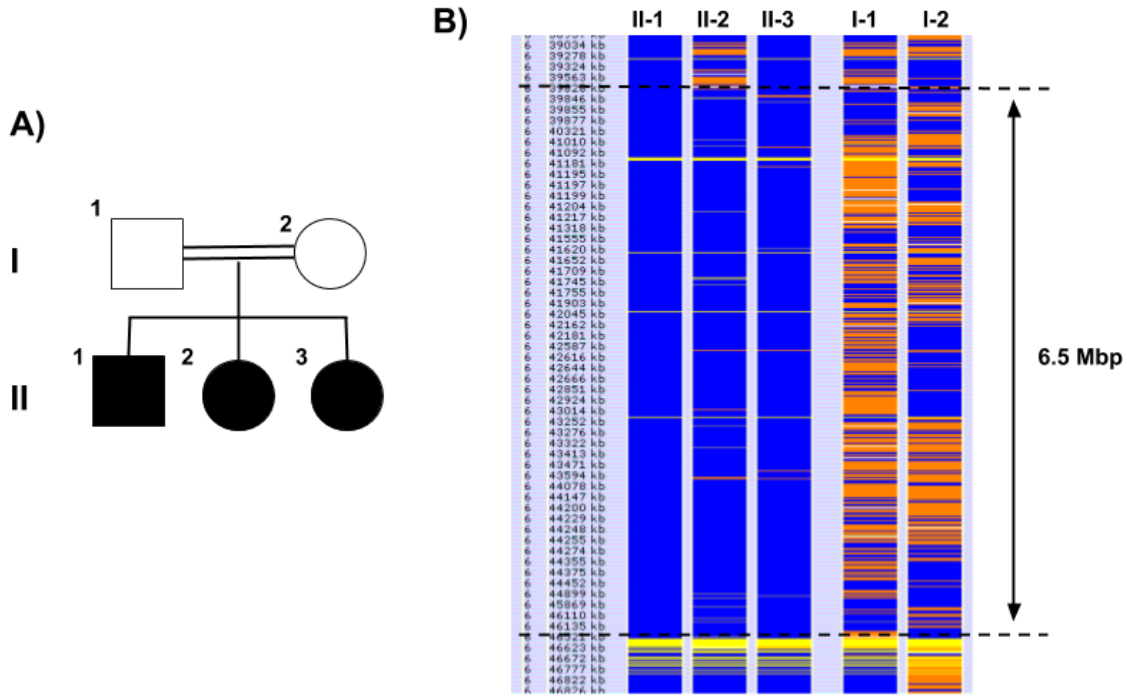


**Figure 3.6:** SP family. **A)** Pedigree of the affected family by SP. **B)** 65 mbp ROH region observed in the SP patients colored in blue.

**Table 3.5:** Number of variants matching the filtering criteria in the SP family.

| | **II-1** | **II-2** | **II-3** | **I-1** | **I-2** |
|---|---|---|---|---|---|
| Covered at Least 5x | 97.82% | 95.68% | 98.09% | 97.22% | 97.83% |
| Number of All Variants | 277,500 | 263,729 | 281,438 | 275,019 | 282,422 |
| Number of All Variants QS > 50 | 266,584 | 251,676 | 269,585 | 263,938 | 270,598 |
| Total Number of Variants | 390,006 | | | | |
| Total Number of Variants QS > 50 | 358,607 | | | | |
| ... homozygous non-reference variants (patients) | 107,368 | 107,124 | 106,776 | | |
| ... heterozygous non-reference variants (father and mother) | | | | 164,412 | 167,537 |
| ... Father and mother are heterozygous & patients are homozygous | 738 | | | | |
| ... exonic or splice site | 95 | | | | |
| ... novel (does not exist in in-house DB and dbSNP 138) | 2 | | | | |
| ... inside linkage region (chr6:40Mb-46Mb) | 1 | | | | |

### 3.3.4 Missense CLN8 Mutation Causes Northern Epilepsy

Whole exome sequence data of 7 individuals from a consanguineous Turkish family (Figure 3.7A) was analyzed with our standard WGS/WGS pipeline. The annotated variant data was filtered based on the autosomal recessive model (Table 3.6). Variants with inadequate coverage (<4) and genotype quality scores (<15) were neglected. The variants that are homozygous in affected children and heterozygous in parents were selected. 12 variants remained after discarding the polymorphisms in the public variant databases with more than 1% allele frequency. Only one loss of function mutation, c.677T>C (p.Leu226Pro) on the CLN8 gene, remained after the filtration. After observing that the mutation resides in a 6.5 Mbp ROH region (Figure 3.7B) we decided that this mutation is the most prominent candidate for the family.

The family was diagnosed with a subtype of neuronal ceroid lipofuscinosis (NCL) [135]. Clinical data indicated that the phenotype of the Turkish family matched with a subtype of NCL which was first described in Finland as Northern Epilepsy (NE) [136]. Numerous studies were done confirming that CLN8 mutations caused NE [137–141]. Hence, we concluded that c.677T>C (p.Leu226Pro) on the CLN8 is the causative mutation of the disease observed in this family. Finally, Sanger sequencing was used to validate that the mutation was not an NGS artifact (Figure 3.7C).

**Table 3.6:** Variation filtering results from the NCL family WES data.

| | Patients | | | | Parents | | |
|---|---|---|---|---|---|---|---|
| Filtering conditions / individuals | V-1 | V-2 | V-4 | V-5 | IV-1 | IV-2 | IV-3 |
| Average coverage | 72 | 52 | 46 | 58 | 48 | 59 | 43 |
| Percentage of >4 coverage | 98.00% | 97.00% | 97.00% | 98.00% | 97.00% | 98.00% | 97.00% |
| Total number of variants | 493,535 | 444,975 | 425,165 | 450,862 | 429,634 | 465,550 | 420,307 |
| Genotype Quality ≥15 and Coverage ≥4 in all individuals | 163,996 | 164,083 | 163,559 | 162,129 | 169,574 | 168,968 | 167,829 |
| Homozygous in patients and heterozygous in parents | 134 | | | | | | |
| GMAF <0.01 in dbSNP, ESP6500, in house Turkish exome database n=978) | 12 | | | | | | |
| Loss of function mutations in exonic region | 1 | | | | | | |
| In shared homozygous region (chr8:0-6.5 Mbp) | 1 | | | | | | |
| Selected mutation | CLN8:p.Leu226Pro/c.677T>C | | | | | | |

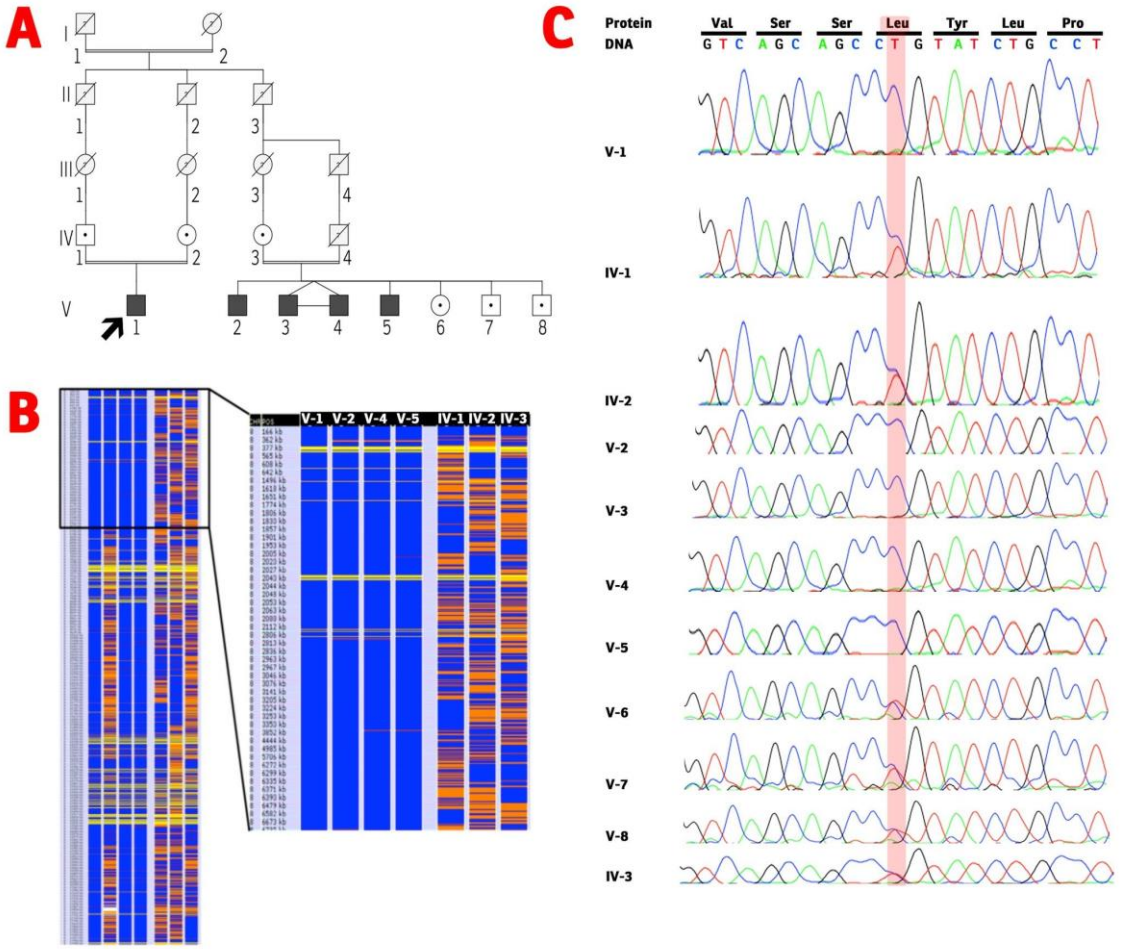


**Figure 3.7:** NCL family. **A)** Pedigree of the NCL family. Exomes of IV-1, IV-2, IV-3, V-1, V-2, V-4 and V-5 were sequenced. **B)** HomSI analysis showing the homozygous region in the affected patients. **C)** Sanger sequence data confirming the c.677T>C (p.Leu226Pro) mutation. (Copyright Sahin et al. 2016)

## 3.4 Discussion

NGS is a powerful technology providing researchers and clinicians with plenty amount of genomic data. The large volume of the sequence data combined with the systematic artifacts embedded in it create special challenges regarding the performance of the analysis as well as accuracy of the results. For any NGS experiment many steps must be taken in order to achieve the desired results. Every tool in every step has its own advantages and disadvantages. None of the software tools has hundred percent accuracy and sensitivity. In this study we presented an efficient analysis pipeline for WGS and

WES data which utilizes the best practices reported in the literature. We also chose well maintained and stable software in order to increase versatility of the pipeline under heavy loads. Choosing well maintained software is especially important for reproducibility of the results by other researchers in the medical genetics community. Reproducibility is necessary both for further validation of the results and for being beneficial to the future researches about the related clinical cases. An important aspect to reproducibility is compatibility of the data format used to store the output values. In the history of bioinformatics, many number of data formats have been used to store biological data. Incompatibility between the data formats always hindered sharing the information between different research projects [142]. For avoiding this problem, we have adopted file formats such as FastQ, SAM and VCF, which are the most commonly used formats by the genetic community [143]. By doing this we were also able to use the variants detected in the previously sequenced samples as control data. Moreover, we applied simple yet powerful parallelization methods in order to distribute the workload between multiple computing nodes and increased the overall speed of the standard analysis. Although speed is not the primary concern for most of the research projects, it has vital importance for clinical diagnostics. Whole exome sequencing of mother-father-child trios has become a common practice for diagnosis of rare diseases in clinical setting [144]. Quickly generating the diagnostic reports increases chance of successful treatment and patient's satisfaction. Our analysis pipeline can be taken as an example model for high speed NGS data analysis for promptly generating diagnostic reports. Finally, the results showed that our pipeline successfully generated annotated variant data which is very effective for detecting disease-causing mutations in the clinical setting.

We have also shown that WES data can be very useful for discovery of novel rare disease-causing mutations if it's analyzed correctly. The capacity and affordability of sequencing large portions of the genome increases our chance to discover novel mutations causing the diseases. However, the vastness of the numbers of variants detected from WGS/WES data creates a "needle in a haystack" problem. It is difficult to differentiate the real disease associated variants from the trivial polymorphisms without the help of relevant annotation information. In the pre-NGS era only a small number of genes or regions were sequenced based on the preliminary linkage studies. If a variant from the target region had missense or loss of function effect then it was considered a strong indication for disease association [15]. However, the effect of the variant solves

only a small part of the problem for WES/WGS experiments because every individual carry hundreds of missense/loss of function variants [114]. Thus, careful experiment design as well as plausible filtering and prioritization strategies are vital in order to single out the correct causative mutation. If the candidate genes or loci related to the phenotype are unknown prior to WES/WGS, it is necessary to sequence multiple affected and control samples preferably from the same family. Variant data of the negative controls that are closely related to the affected individuals is highly effective for filtering out the rare but trivial variants inherited in the family. As seen in the Klippel-Feil syndrome and the Northern Epilepsy cases, segregation based filtering dramatically decreased the number of suspected variants. Segregation based filtering was much more effective when an additional family with the same phenotype had been sequenced in case of CFT dysplasia. This indicates that positive controls from different families can be more effective probably because distant families have fewer shared polymorphisms and there is a higher chance that one of the shared variants are actually related to the common phenotype. Another effective way of discarding irrelevant variants is to use variant database of negative controls from the same population. This is especially essential in cases where small number of samples were sequenced. For example, there were nearly one hundred candidate variants left after the segregation based filtering in the Spastic Paraplegia case. By using the in-house Turkish variant database (TUBITAK-MAM) in conjunction with dbSNP, we were able to reduce the number of candidates to only 2 variants. Hence, it has utmost importance to create and update their own in-house variant databases for genome centers so that they can use the unaffected population data as negative controls.

In many cases segregation and population based filtering may not be enough to single out the disease associated gene or mutation. For example, there were 6 potential candidates for KFS even though 8 samples from the same family had been sequenced. In such cases it is necessary to carefully investigate functional information of each candidate gene in order to figure out which one of them is more relevant to the phenotype. The first thing to look for is whether the candidate genes were associated with any disease or phenotypes in the previous studies. This was the decisive information which concluded the Northern Epilepsy case because CLN8 was associated with the disease by multiple studies in the past. In the case of KFS, however, further investigation regarding expression profiles of the genes, biological functions of the proteins, interspecies

conservation rates of the mutation sites and phenotypic effects of knock-out experiments on model organisms for all the candidates was necessary. In addition to validating the candidate mutation with Sanger sequencing it is also necessary to show the effects of the mutation on transcription and translation. It is expected that loss of function mutations would cause the defective mRNAs to be quickly digested by the nonsense-mediated mRNA decay mechanism [145,146]. Quantitative real-time PCR and northern blot methods were used to confirm that the suspected mutations are causing the mRNA degradation in CFT and SP cases.

Determining the candidate variations for disease causing mutations is the most important step in clinical researches. NGS technologies have made the discovery of candidate mutations much more easy and affordable. However, multiple validation steps are necessary for confirming that the selected variant is not a false positive. First, it must be validated with Sanger sequencing and/or PCR that the mutation is not an NGS artifact. Furthermore, molecular and functional studies about the phenotype, and concordant results from multiple families are necessary to conclude that the selected mutation is actually causing the particular phenotype. Caution is necessary even if there is only “1 in a million” chance of false association, because NGS can easily discover millions of variants. Therefore it was crucial to validate our results with as many samples as possible. Our results were also confirmed with multiple families, functional and molecular experiments.

# 4 FINE MAPPING STRUCTURAL VARIATIONS

## 4.1 Motivation

Genomic rearrangements and indels larger than 20 bp are classified as structural variations. Special methods which are different from those used for calling SNVs and small indels are used for detecting structural variations. For finding structural variations, most of the SV detectors search for signals such as soft clipped reads and discordantly mapped read pairs clustered together. Even the most advanced SV detectors rely on the mapping information generated by the short read mappers and do not utilize unmapped reads. The main reason why SV callers depend on short read mappers is because their goal is scanning through the whole genome in order to discover as many SVs as possible in the shortest amount of time. However, it is known that sensitivity and specificity of short read mappers can be greatly reduced by deletions and insertions [41]. Hence, most of the SV callers cannot use the discordant reads which are mapped to erroneous locations due to discrepancies created by SVs. Relying on short read mappers may provide SV callers with high performance but it may also cause losing potentially valuable reads for revealing exact structure of individual SV regions. The existing SV callers allow this compromise because scrutinization of individual SV regions is not a priority for them but defining approximate regions of all the SVs is the main priority.

The majority of the SV detectors were designed with only the purpose of listing the candidate regions which may contain SV breakpoints. Some of the SV callers have used *k*-mers for detecting SVs at single base pair resolution. BreaKmer [83] scans the reads mapped to the region of interest for finding novel k-mers not found in the reference genome and assembles them to reveal the breakpoints. Both BreaKmer is memory and CPU intensive and can only be used for targeted sequencing experiments. novoBreak [147] finds novel *k*-mers by comparing *k*-mers in raw read data of tumor-blood pairs in order to discover somatic SVs. novoBreak does not require short read mapping but it can only be used with case-control or tumor-blood paired data. Local reassembly of split-reads is another method for determining SVs at single base pair resolution. TIGRA [148] and HYDRA [75] assembles discordantly mapped and one-end-anchored reads that are in close proximity to the SV sites. These tools require less memory and they are fast

enough to scan WGS data however they heavily rely on the mapping information coming from the SR mappers. In addition to SV detectors there are several tools which focus on graphical visualisation and manual validation of the suspected SV sites. PyBamView [149] is a web based SAM/BAM alignment viewer which focuses on visualisation of insertions. Bambino [150] detects SNVs and small indels using the alignment information in SAM files and visualizes them. Svviz [151] takes a suspected SV as an input and visualizes the supporting reads in comparison to normal reads in order to help validating the SV. Although some of these tools perform realignment or reassembly at the local level, they still require mapped read data and ultimately rely on the short read mappers.

Here we present a method for fine mapping of structural variations based on their *k*-mer content. The goal of our method is to reveal exact structure and sequence of already detected or suspected structural variant regions rather than discovering novel variant regions. In our method, the reads that are possibly associated with the SVs are extracted based on the shared *k*-mer content between the reads and the SV regions. This enables us to retrieve every possibly relevant without depending on mapping information and even the unmapped reads can be utilized. The extracted reads are then assembled by using *de novo* assembly and the assembled contigs are aligned onto the reference genome by using Blast [152,153]. Because *de novo* assembly is a key step in our method we compared results from three different assembly software; SPAdes [34], Velvet [35] and ABySS [30] in order to assess performance of our method regardless of the assembler. Finally, the local hits, together with junction information, are plotted against the SV regions for revealing the final structures of the variations. Using *de novo* assembly alongside with local alignment enables us to report multiple genomic rearrangements in a single region. To the best of our knowledge [7, 13], our method is the only method which can accomplish such a feat. Hence, our method can be crucial for understanding highly variable DNA shuffling regions. Moreover, extracting reads based on their *k*-mer content rather than relying on the mapping information improved detection rate of SVs, especially for insertions. SVMap, the Python implementation of our method, is available at https://github.com/berguner/svmap.

## 4.2 Methods

### 4.2.1 *k*-mer based read extraction

The key improvement provided by our method is that short reads are extracted from raw sequence data based on their k-mer content rather than location information generated by the mapping software. Basically, any read sharing at least one *k*-mer with the SV regions that are of interest for detailed scrutinization is selected for *de novo* assembly. Any SV region detected by a SV detection method can be given as input. We use the term "SV region" referring to a subsequence from the reference genome including the SV event together with flanking sequences on both sides. Inclusion of anchoring/flanking sequences is vital because the *k*-mers collected from these regions will enable capturing reads and/or pairs of reads crossing the breakpoints of SVs (Figure 4.1). For single-end sequencing the size of these anchoring/flanking sites depends on the read size whereas for pair-end sequencing it depends on the insert size of the library. For example, three times the mean insert size from both flanking regions would be a safe value for most of the pair-end genome sequence data. It is best to adjust the size of flanking regions large enough to cover majority of the sequence data based on the DNA library properties and features of the sequencing platform.

To start the procedure every *k*-mer existing in the given SV region(s) that are of interest is extracted from the reference genome sequence (Figure 4.2). By definition, *k*-mer refers to every possible subsequence of length *k* existing in a given sequence. For a given region of length *L*, *L-k+1 k*-mers will be extracted. We collect every subsequence of length *L* starting from the first base of the SV region while shifting 1 base pair in every iteration until the last *k*-mer is reached. In addition, reverse-complements of the extracted *k*-mers are also collected because reference genome files have only forward strand but reads can be either on the forward or the reverse strand. Hence the total number of extracted *k*-mers will be *2(L-k+1).* The extracted *k*-mers and the information of which SV region do they belong should be stored in a data structure which can be efficiently searched. We used a hash table based dictionary structure for our implementation where every key (*k*-mer) search takes constant, O(1), amount of time regardless of the number of *k*-mers stored in the dictionary. This allows looking up every *k*-mer existing on every

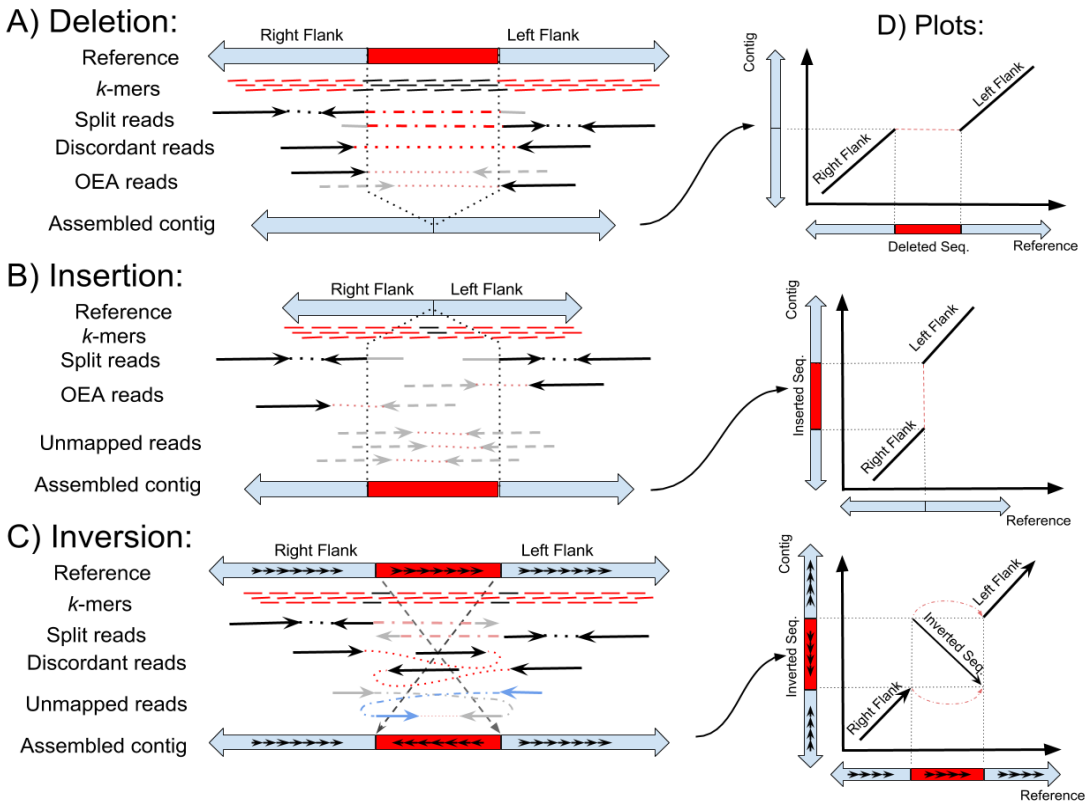


**Figure 4.1:** Visual demonstration of SV events, *k*-mers and extracted reads that are used for assembly of variant region. The *k*-mers that are indicative for supporting reads for SV events are shown in red color. Our method is able to extract split reads, discordant reads, one-end-anchored (OEA) reads as well as unmapped reads. The line plot representation for each type of SV is shown on the right.

read on the dictionary to be completed in a practical amount of time. It can be argued that looking for exact matches might cause missing relevant reads because of the single nucleotide polymorphisms or sequencing errors. However, such artifacts would be avoided mostly because we search for every *k*-mer with 1bp shifts on the read.

A crucial consideration for our approach is balancing specificity versus sensitivity of *k*-mer based read extraction. The key measure affecting this balance is the size of *k*-mers; longer k-mers would be more specific while shorter *k*-mers would be more sensitive. Therefore selected size must be a value allowing to capture as many relevant reads as possible without gathering too many irrelevant reads. In our tests we took the *k*-mer size as 26 because 80% of the *26*-mers in human genome are unique and increasing the size does not increase specificity significantly whereas decreasing it would dramatically reduce specificity [103]. Besides determining the size of *k*-mers, avoiding the systematic repeats in the genome is an important step. Because of the repetitive nature of human genome there are many *k*-mers with exceptionally high frequency regardless of their size. Eliminating such *k*-mers is especially crucial when working with whole genome data because thousands of irrelevant reads from telomeric, centromeric and various other repeat regions may be selected which would adversely affect the assembly process. We used "aln" and "samse" functions of BWA [42] to detect and eliminate highly repetitive *26*-mers in the human genome.

The *k*-mer dictionary becomes ready for searching and extracting relevant reads after the cleanup process. For the extraction process, every *k*-mer in each read from the raw sequence file(s) is searched in the target *k*-mer dictionary. If a read has at least 1 common *k*-mer with a particular SV region, the read is selected and stored for *de novo* assembly of that region. The mate of the read is also selected for the pair-end or mate-pair sequence data.

### 4.2.2 Assembly

There are many *de novo* genome assembly software using specialized algorithms. Each one these software has their own advantages and disadvantages depending on read size, read quality, library properties and sequence content of the genome. We tried three of the most well-known assemblers for assembling the SV regions; SPAdes [34], Velvet

[35] and ABySS [30]. For our test set, the assemblers were run with recommended settings for human genome with a minimum coverage cut off value of 3. *K*-mer size is set to 31 and 64 for velvet and ABySS respectively. We used the default setting for Spades where various *k*-mer sizes are used iteratively to improve the assembly. All of the assemblers are run with pair-end read option in order to maximize contig/scaffold sizes.

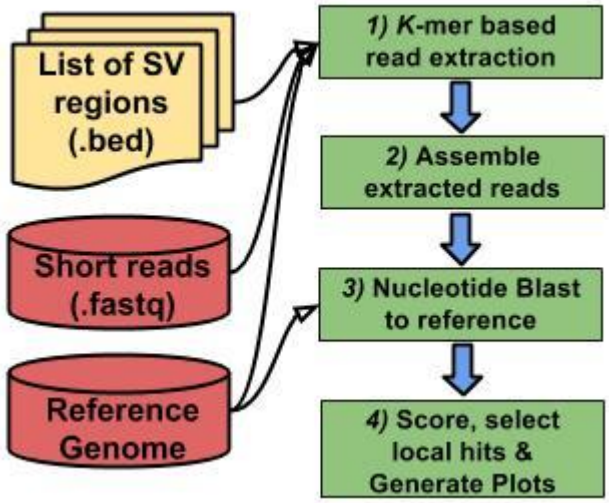


**Figure 4.2:** The chart showing the workflow in of SVMap.

### 4.2.3 Basic Local Alignment

In order to assess the final structure of the variant with respect to reference genome we used NCBI's Blastn [152,153] software for aligning the assembled contigs back to the reference genome database. Local alignment allows us to find the most significant locally matched regions of contigs to reference chromosomes. We created the Blastn search database from the reference human genome assembly (GRCh37.p13) using the word size as 11. We set the maximum number of hits per target sequence (chromosome) and the maximum number of target sequences to 10. The default parameters of Blastn were changed for searching interspecies genomic sequence matches with high sensitivity. Since we are aligning human samples to human reference, we lowered the reward and the penalty scores to the lowest possible values (-reward=1, -penalty=-5) in order to prioritize highly specific shorter hits. Even so, it is inevitable that irrelevant hits will occur frequently because 51% of the human genome is made up of repetitive DNA [54] and Blast favors longer hits because of the scoring algorithm it uses to estimate the e-value. We have managed to prevent irrelevant long hits from completely

hiding relevant short hits by adjusting the reward and penalty parameters. Finally, the results are stored in XML file format for easy and flexible parsing options.

### 4.2.4 Fine mapping SVs

The basic assumption of our mapping approach is that flanking regions (Figure 4.1) of SVs in the assembled contigs would align back to the SV region on the reference genome. If a contig has been assembled from reads that support an SV event there will be multiple Blast hits for that contig rather than having a single continuous alignment. Alignment positions and orders of hits relative to each other would therefore help visualizing and revealing the type, size and location of the genomic rearrangement(s) (Figure 1D). For example, if a contig is carrying a deletion of size 150 bp there will be two Blast hits for that contig which are 150bp apart from each other. The sequence of the contigs carrying genomic rearrangement(s) would reveal the structure of the variant in single base level because we also include split reads from breakpoints in the assembly process. Furthermore, the combination of *de novo* assembly with local alignment allows us to explain multiple genomic rearrangements happened in the same region since we are taking into account multiple Blast hits from the contigs which are assembled independently from the reference genome.

Blastn result files are lists of similar subsequences between the query sequences and the chromosomes in the reference genome sorted by their alignment *e*-values. In our implementation, the 10 most significant hits for the 10 most relevant chromosomes are listed for every contig from each region's assembly. This gives 100 hits per each contig in the assembly file. Evidently, it is highly unlikely for a correctly assembled contig to have 100 real hits because it would mean that there are 100 different rearrangement events occurred in the same region. Hence we need to select the most significant hits for each contig which could explain the actual rearrangement events. The quick solution would be to choose the hits with the lowest e-values -highest significance- however longer hits would dominate our selection. This is a common problem in long interspersed nuclear elements (LINE) in the genome because e-value of a hit decreases exponentially as the length increases and long hits from LINE regions can eliminate the shorter, more relevant hits. Because of such consequences we used bit scores for scoring the local hits. Bit scores are significance scores of local hits normalized by their lengths. We

implemented an iterative method to maximize the total bit score of local hits selected for explanation of a given contig. To make a decision between overlapping hits we selected the hits which have higher per-base-bit score for the overlap regions. The final score for the contig is calculated by multiplying the total bit score of selected hits by the average read depth of the contig. Scoring the contigs this way would help prioritizing multiple contigs associated with the same region based on the number of supporting reads and the sequence identity between the contig and the SV region.

Finally, the selected local hits are sorted based on their position on the contigs and the likely events such as, insertion, deletion and inversion, are deduced based on the relative positions and strands of consecutive hits on the reference genome. To put it simply, if there is a gap between two consecutive hits on the reference side then it is considered as a deletion, but if the gap is on the contigs' side then it is called an insertion. Also, it is noted as an interchromosomal translocation if the consecutive hits are on different chromosomes. The event is called an inversion when the consecutive hits are on different strands. We also create a graphical representation of likely rearrangement events by plotting the selected hits against the SV region in the reference genome using line plots. The pairwise line plot representation that we used can show both positions of the SV breakpoints and the relative strandedness of local hits from the contigs. Therefore it is possible to visually interpret exactly what kind of genomic rearrangements have occurred in the region of interest (Figure 4.1).

### 4.2.5 Dataset

We used 2 different datasets in order to test effectiveness our method. The first dataset contains a complex rearrangement event discovered in a rare disease study [154]. Raw data consists of high coverage pair-end sequence data targeting a 3.27 Mb region in the chromosome 20 where the rearrangement has occurred. The second dataset contains high confidence structural variations with sizes of up to 10kb which are confirmed by multiple SV detection methods [155] from the genome of the HapMap [156] individual NA12878. Pair-end whole genome sequence data of NA12878 archived under study SRX485062 is downloaded from the SRA database [157]. To prepare the sequence data for comparing *k*-mer based and mapper based methods, we mapped the sequences to

GRCh37 reference human genome by using BWA MEM [42]. The aligned sequences were sorted and the PCR duplicates were removed by using the samtools [42].

**Table 4.1:** Detailed information about the sequence data used in our tests.

| Data | Read Size | # of Reads | Size of Sequenced Region | Average Coverage |
|---|---|---|---|---|
| NA12878 | 100x2 (Pair-end) | 1,482,602,390 | Whole genome | 49x |
| 2341, targeted sequence | 75x2 (Pair-end) | 13,192,022 | 3.27 Mb (chr20:43,655k-46,924k) | 302x |

## 4.3 Results

### 4.3.1 Targeted Sequencing Data

We applied our method to targeted pair-end sequence data of a patient in which the bases between 43,655,000-46,924,000 of the chromosome 20 were sequenced. The targeted region covered a complex genomic rearrangement causing ELMO2 gene to lose its function. The approximate location of the rearrangement event was first identified by BreakDancer [72], it reported 5 inversions and 2 deletions with high scores between 45,021,000-45,040,000 (Supplementary Table A1). After visually inspecting the region with IGV [14] it was evident that a complex rearrangement has occurred affecting the first three exons of the ELMO2 gene (Supplementary Figure A1). Although BreakDancer was able to identify some of the SV events it was insufficient to understand the rearrangement completely and it could not report the breakpoints accurately. We ran BreaKmer [83] for detecting breakpoints in the affected region more accurately. BreaKmer was able to discover all of the breakpoints, however it classified them as inversions and/or translocations while neglecting the deletion and the insertion events (Supplementary Table A2).

We utilized our method in order to understand the complete scope of rearrangements happened in the region. We extracted the *k*-mers from the bases between

45,022,000 and 45,037,129 for creating the dictionary which would be used for read extraction. The extracted reads were then assembled with all three of the assemblers. Finally, the assembled contigs were locally aligned to the reference genome. Our scoring algorithm successfully prioritized a 3257 bp contig assembled by SPAdes which contained the rearranged genome sequence. The rearrangement event was composed of one large deletion, one novel insertion, one inverted duplication and one inversion events all happened in a frame approximately 6kb in size (Figure 4.3). Capillary (Sanger) sequencing was used for validating that the sequence of the selected contig matches exactly with sequence of the actual rearranged genome [154] (Supplementary Figure A3). Furthermore, our scoring algorithm was also able select and prioritize the correct local hits exactly explaining the rearrangement events (Supplementary Figure A4).

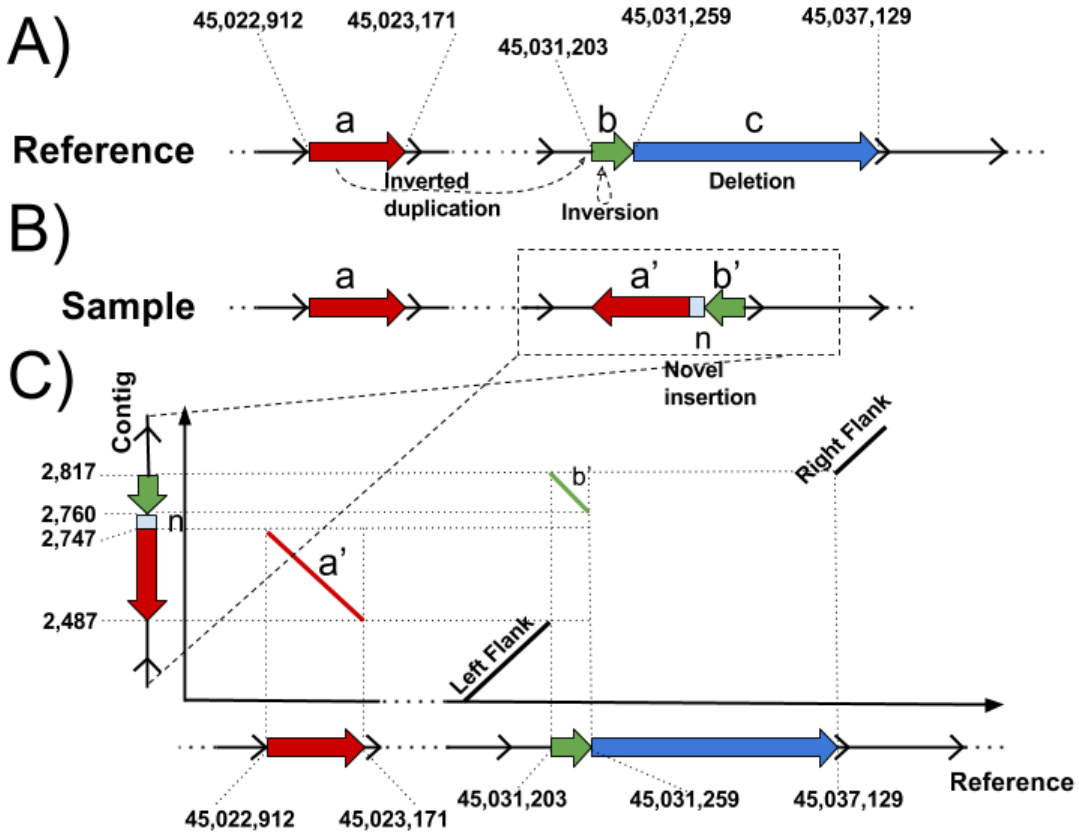


**Figure 4.3:** Visual demonstration of complex rearrangement happened on ELMO2 gene. A) Shows the sequence of chromosome 20 from hg19 reference genome. B) Shows the sample's sequence after the rearrangement. C) Shows the explanation of the plot generated by our method. Segment a (260bp) was inversely-duplicated and inserted upstream of the segment b together with a novel 13 bp insertion. Segment b (57bp) was inverted and segment c (5870bp) was deleted.

Our scoring and selection algorithm played a crucial role revealing the complex nature of this rearrangement event. This is especially important for this case because ELMO2 has an isoform on chromosome 15 with 92% sequence identity. Because of such high similarity the targeted sequence data was also covering the bases between 22,770,816 and 22,802,215 on chromosome 15. Many reads, which actually belong to the isoform region, were selected for assembly because they contained *k*-mers extracted from the original region. As a result, the assembly output contained contigs belonging to chromosome 15 also. Our scoring and selection algorithm performed successfully on two levels for elimination of the noise caused by the isoform. First, it gave the highest score to the contigs assembled from the reads coming from the actual region of interest rather than those coming from the isoform. Second, out of 100 hits it selected the 4 relevant local hits explaining the individual rearrangement events on the contig, even though some of them were small and have relatively high e-values. We observed that adoption of per-base bit scores for scoring has played a key role for the successful outcome.

### 4.3.2 Whole Genome Dataset

For the purpose of assessing effectiveness of our method we applied our method to detect and explain structural variations found in the whole genome of HapMap individual NA12878. We used high confidence SVs reported in a study for benchmarking SV discovery tools [155]. Although there were more than 4000 SVs in the dataset we used the deletions and insertions found on chromosomes 1, 2, and 3. Such a selection was necessary because considerable amount of time was needed for manual scrutiny of each SV. The number of SVs found on the first three chromosomes should be enough to demonstrate how effective our method for general use is. On the other hand, we kept all of the inversions because there were already a few of them. The final test set consisted of 998 SVs including 544 deletions, 404 insertions and 50 inversions (See supplementary spreadsheet for details).

The main objective of our test was to demonstrate how effective is *k*-mer based read extraction compared to the mapping location based read extraction for detecting structural variations. Our *k*-mer based read extraction process was used to extract reads belonging to each of the SV regions in the test set. The reads which were mapped to the SV regions were extracted using Samtools and converted to fastq format using a custom

bam-to-fastq conversion script. In an attempt to eliminate assembler induced biases we used 3 different *de novo* assemblers to assemble reads extracted by both methods. Finally, the high scoring local hits from the local alignment results were scrutinized for inclusion of the actual SV events. We also included joint detection metrics where results from both extraction methods and/or 3 assembly methods were joined together.

The test results showed that our *k*-mer based method aided detection of additional structural variations for all 3 types of structural variations (Table 4.2). Moreover, *k*-mer based method performed better compared to mapper based method for all types of tested structural variations. Compared to mapping based extraction the most distinguished advantage was seen in detecting homozygous insertions. 21% more homozygous insertions were detected using the *k*-mer method with SPAdes only or with joint assembler output. Homozygous deletions were the easiest SV type to detect while heterozygous insertions were the most difficult. A similar ranking between the SV types is observed in almost all of the current SV discovery tools. Although the recall rates are not perfect, they on par or better than most of the SV discovery tools [68,158]. Looking at the test results, it is evident that detection rate heavily depends on the assembler. For all types of SVs SPAdes performed notably better than Velvet and ABySS. It can be argued that such difference is due to the read correction and iterative *k*-mer size selection capabilities built into the SPAdes. There was also an option in SPAdes for consideration of large rearrangements for scaffolding diploid genome assemblies which was not available in Velvet or ABySS. In contrast to other cases *k*-mer based extraction performed worse only for detecting deletions while using ABySS. Perhaps the reason behind this result was the difficulties in handling reads with shared *k*-mer content coming from unrelated genomic regions which can be inferred from the fact that there is less difference in detection rates for heterozygous deletions.

One of the most powerful features of our method is that it can reveal exact sequence of the genomic region after SV events. It is not restricted to categorize a genomic rearrangement under any one of the recognized SV types. Hence, it can report multiple types of rearrangement events in a given region which has been shown in the ELMO2 case. After inspecting 998 SVs in NA12878 dataset we have encountered 18 occasions where multiple SV events happened adjacently. We have also detected that 13 of the reported insertions were actually translocation or trans-duplication events

happened within a 2k bp region. Such findings show that our method is useful when a complex rearrangement has been found in a region of interest but could not be explained using currently available SV discovery tools.

**Table 4.2:** Recall rates of SVMap run with k-mer based method vs. mapping based method. "Joint" columns represents the resulting recall rate when any one of the methods was able to detect a given SV. The "joint" row at the bottom represents joint recall rates of all of the assemblers.

| | Deletion (544) | | | | | | Insertion (404) | | | | | | Inversion (50) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Homozygous (229) | | | Heterozygous (315) | | | Homozygous (272) | | | Heterozygous (132) | | | Homozygous & Heterozygous | | |
| | kmer | map | joint | kmer | map | joint | kmer | map | joint | kmer | map | joint | kmer | map | joint |
| SPAdes | **94.8** | 93.9 | 97.8 | **54.9** | 40.0 | 59.7 | **67.3** | 48.2 | 71.7 | **56.8** | 51.5 | 64.4 | **57.1** | 53.1 | 63.3 |
| Velvet | **62.9** | 59.0 | 71.6 | **6.7** | 3.5 | 8.9 | **10.3** | 5.5 | 12.9 | **5.3** | 2.3 | 7.6 | **16.3** | 14.3 | 18.4 |
| AbySS | 46.7 | **71.2** | 80.3 | 35.9 | **41.0** | 47.9 | **9.6** | 8.1 | 14.0 | **13.6** | 7.6 | 17.4 | 36.7 | 36.7 | 42.9 |
| joint | **96.9** | 95.2 | 97.8 | **64.4** | 56.5 | 69.5 | **70.6** | 49.6 | 75.7 | **61.4** | 54.5 | 65.9 | **67.3** | 65.3 | 71.4 |

## 4.4 Discussion

It is certain that whole genome sequencing will be much more accessible and widely used for different purposes such as personal medicine and cancer genomics. These studies require discovery and explanation of every genomic variation existing in the individual's genome. 10 years have passed since the first individual human genome had been sequenced [159] and discovery of structural variations still remains a difficult challenge after many developments in sequencing technologies. There are 40 different tools and methods [7] developed for discovering structural variations which aim detection of structural anomalies throughout the genome. However, little has been done for understanding individual genomic rearrangement regions. We proposed SVMap to address such needs and showed that *k*-mer based read extraction would be beneficial for revealing the underlying structure of all types of structural variations. novoBreak also uses *k*-mers for extracting SV related read pairs and it is independent from short read mapper but it requires tumor-blood paired data for selecting the aberrant *k*-mers in tumor samples. SVMap can be used without a control sample however it analyses only the target regions. BreaKmer is also confined to the target regions but it is also dependent on the

mapped short read data. In contrast to BreaKmer and novoBreak, SVMap utilizes all of the reference k-mers in the SV region for the purpose of collecting all possibly relevant read pairs and assembles them to reveal the complete structures of the genomic rearrangements. BreaKmer and novoBreak focus on detecting only the breakpoints rather than solving the structures of the rearrangements.

Our method is the first method to use *de novo* assembly in conjunction with basic local alignment in order to explain genomic rearrangement events. HYDRA and TIGRA als uses local reassembly but they only utilise discordant or split reads. Discordance or concordance of the read pairs is determined by the relative alignment positions of the mate and this decision strongly depends on the short read mapper and can be biased based on the DNA library properties. Therefore, it is highly likely that they miss relevant reads because of such complications. HYDRA, TIGRA and novoBreak aim detection of structural variants on the entire genome therefore they make some compromises in order to make the analysis more practical in terms of computational resources. SVMap, on the other hand, focuses on analysis of selected important regions and does not make such sacrifices. Hence, it excels at the local level compared to general SV detectors. A common case for using SVMap would be scrutinizing intragenic SV candidates detected by one of the SV detectors. PyBamView, Bambino and svviz also focus on scrutinizing selected candidate SV regions but they are more focused on the graphical representation of the read alignments individually. In addition to the its immunity to insensitivity of SR mapping tools, to the best of our knowledge, SVMap is the only method which can report multiple types of rearrangements in a single region. These features makes SVMap a powerful new tool useful for better understanding structural variations with high importance.

By investigating our test results we saw that in only one test scenario, deletion detection by ABySS, *k*-mer based read extraction underperformed compared to mapping based read extraction. We can argue that the loss of sensitivity for ABySS is cause by the inclusion of more reads that are not supporting the deletion events into the assembly with the *k*-mer method. It is common for assemblers to prune some paths in the assembly graph for avoiding false assemblies. Such behaviour might be the reason for losing the contigs including the deletion events. The fact that this adverse effect is less pronounced in the heterozygous deletions supports our claim. Even in ABySS's case *k*-mer based read

extraction was able to detect additional deletions which were not detected by mapping based read extraction. It was also observed that using both extraction methods improved detection rate for all three types of SVs. The results also showed that outcome of our method is highly dependent on the underlying assembly method. In order to achieve the best possible results it is crucial to use multiple assemblers. However, every one of the tested assemblers performed poorly for heterozygous SVs. This indicates that a special study is necessary for optimizing assembly of subregions from large diploid genomes. Even if every step of the process is optimized to the maximum capability the limitations of short read sequencing cannot be overcome. Most notably assembly and mapping of large novel insertions is not possible where the size of reads and/or insert size of the mates/pairs are not long enough to completely cover the inserted sequence. Because of this limitation the reads captured by using the anchoring (flanking) regions cannot be connected to span the whole extent of the insertion resulting in two separate contigs from each side. This limitation similarly affects assembly of inversion sites, especially the heterozygous inversions. Another limitation is the need for high depth of coverage and necessity of high quality reads for successful assemblies. Because such limitations arise from the sequencing technologies, they can be overcome only by the availability of long and high quality reads in abundance.

# 5 CONCLUSION

In our work we have investigated the computational methods used for analyzing the NGS genome data in order to compile best practices for discovering disease causing mutations. We studied the mutations under two categories, first is the SNVs and small indels, and second is the complex genomic structural variations. In contrast to structural variations the methods and software tools for calling SNVs and small indels are well established and have high sensitivity and accuracy rates. The challenging aspect of small variant analysis is the large number of clinically irrelevant and/or false variant calls. Accurately selecting the disease-causing mutation among irrelevant variants is prone to false discoveries because several hundred thousand small variants are called from a typical human WES sample. We have shown that our standard analysis pipeline and mutation-disease association strategy is a good implementation of effective WES analysis methods. Structural variants, on the other hand, are not as numerous as small variants and they are most probably deleterious on the genes. But it is challenging to define the exact structure of the rearrangement occurred by using currently available tools. In this context, we developed a fine mapping method which was capable of solving even the more complex genomic rearrangements.

It is evident that the introduction of NGS technologies has completely changed the landscape of genomic area. The new sequencing platforms are generating ever more data with lower costs than the previous platforms. The accessibility and affordability of genomic data have presented unprecedented research opportunities to clinical researchers. However, the plethora of data generated by these technologies is presenting great challenges for current computational analysis tools and methods. It is crucial to develop new methods and tools which can keep up with the pace of genomic data accumulation in order to achieve groundbreaking discoveries. It is also vital to increase recall rate while decreasing the error rate because the there is greater risk of making false decisions due to the increased amount of data. It is forecasted that new and better sequencing technologies coupled with advanced analysis methods will minimize the false discovery rates in the future.

## BIBLIOGRAPHY


1. Rabbani B, Nakaoka H, Akhondzadeh S, Tekin M, Mahdieh N. Next generation sequencing: implications in personalized medicine and pharmacogenomics. Mol Biosyst. 2016;12: 1818–1830. doi:10.1039/c6mb00115g

2. Hong H, Zhang W, Su Z, Shen J, Ge W, Ning B, et al. Next-Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine. Omics for Personalized Medicine. 2013. pp. 39–61. doi:10.1007/978-81-322-1184-6_3

3. Toriello J. The Human Genome Project [Internet]. The Rosen Publishing Group; 2002. Available: https://books.google.com/books/about/The_Human_Genome_Project.html?hl=&id=l5iliQ3yHaoC

4. The Cost of Sequencing a Human Genome. In: National Human Genome Research Institute (NHGRI) [Internet]. [cited 5 Jun 2017]. Available: https://www.genome.gov/27565109/The-Cost-of-Sequencing-a-Human-Genome

5. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5: 17875. doi:10.1038/srep17875

6. Cornish A, Guda C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res Int. 2015;2015: 1–11. doi:10.1155/2015/456479

7. Guan P, Sung W-K. Structural variation detection using next-generation sequencing data: A comparative technical review. Methods. 2016;102: 36–49. doi:10.1016/j.ymeth.2016.01.020

8. Ergüner B, Üstek D, Sağıroğlu MŞ. Performance comparison of Next Generation sequencing platforms. Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. IEEE; 2015. pp. 6453–6456. Available: http://ieeexplore.ieee.org/abstract/document/7319870/

9. Na HS. Performance Comparison of Benchtop Next-generation Sequencing Systems. J Bacteriol Virol. 2014;44: 208. doi:10.4167/jbv.2014.44.2.208

10. Fox EJ, Reid-Bayliss KS. Accuracy of Next Generation Sequencing Platforms. Journal of Next Generation Sequencing & Applications. 2014;01. doi:10.4172/2469-9853.1000106

11. Zhang Z, Hao K. SAAS-CNV: A Joint Segmentation Approach on Aggregated and Allele Specific Signals for the Identification of Somatic Copy Number Alterations with Next-Generation Sequencing Data. PLoS Comput Biol. 2015;11: e1004618. doi:10.1371/journal.pcbi.1004618

12. Brynildsrud O, Snipen L-G, Bohlin J. CNOGpro: detection and quantification of CNVs in prokaryotic whole-genome sequencing data. Bioinformatics. 2015;31: 1708–1715. doi:10.1093/bioinformatics/btv070

13. Abel HJ, Duncavage EJ. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genet. 2013;206: 432–440. doi:10.1016/j.cancergen.2013.11.002

14. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29: 24–26. doi:10.1038/nbt.1754

15. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12: 745–755. doi:10.1038/nrg3031

16. Shimizu T, Marusawa H, Chiba T. Recurrent Somatic Mutations in Human Gastric Cancers Identified by Whole Exome Sequencing. Gastroenterology. 2012;143: 1385–1387. doi:10.1053/j.gastro.2012.09.027

17. Chang VY, Basso G, Sakamoto KM, Nelson SF. Identification of somatic and germline mutations using whole exome sequencing of congenital acute lymphoblastic leukemia. BMC Cancer. 2013;13. doi:10.1186/1471-2407-13-55

18. Hirotsu Y, Zheng T-H, Amemiya K, Mochizuki H, Guleng B, Omata M. Targeted and exome sequencing identified somatic mutations in hepatocellular carcinoma. Hepatol Res. 2016;46: 1145–1151. doi:10.1111/hepr.12663

19. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14 Suppl 11: S1. doi:10.1186/1471-2105-14-S11-S1

20. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics. 2012;13: 341. doi:10.1186/1471-2164-13-341

21. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39: e90–e90. doi:10.1093/nar/gkr344

22. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, et al. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl Environ Microbiol. 2014;80: 7583–7591. doi:10.1128/AEM.02206-14

23. Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS One. 2011;6: e17915. doi:10.1371/journal.pone.0017915

24. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res. 2007;17: 1697–1706. doi:10.1101/gr.6435207

25. Bryant DW Jr, Wong W-K, Mockler TC. QSRA: a quality-value guided de novo

short read assembler. BMC Bioinformatics. 2009;10: 69. doi:10.1186/1471-2105-10-69

26. Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, et al. Extending assembly of short DNA sequences to handle error. Bioinformatics. 2007;23: 2942–2944. doi:10.1093/bioinformatics/btm451

27. Warren RL, Sutton GG, Jones SJM, Holt RA. Assembling millions of short DNA sequences using SSAKE. Bioinformatics. 2007;23: 500–501. doi:10.1093/bioinformatics/btl629

28. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of the giant panda genome. Nature. 2010;463: 311–317. doi:10.1038/nature08696

29. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20: 265–272. doi:10.1101/gr.097261.109

30. Simpson JT, Wong K, Jackman SD, Schein JE, J.M. Jones S, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Res. 2009;19: 1117–1123. doi:10.1101/gr.089532.108

31. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 2008;18: 802–809. doi:10.1101/gr.072033.107

32. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. Science. 2000;287: 2196–2204. Available: https://www.ncbi.nlm.nih.gov/pubmed/10731133

33. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380. doi:10.1038/nature03959

34. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19: 455–477. doi:10.1089/cmb.2012.0021

35. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18: 821–829. doi:10.1101/gr.074492.107

36. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, et al. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics. 2012;11: 25–37. doi:10.1093/bfgp/elr035

37. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95: 315–327. doi:10.1016/j.ygeno.2010.03.001

38. Pop M. Genome assembly reborn: recent computational challenges. Brief Bioinform. 2009;10: 354–366. doi:10.1093/bib/bbp026

39. Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, et al. A hybrid

approach for de novo human genome sequence assembly and phasing. Nat Methods. 2016;13: 587–590. doi:10.1038/nmeth.3865

40. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res. 2015;25: 1750–1756. doi:10.1101/gr.191395.115

41. Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV. Benchmarking short sequence mapping tools. BMC Bioinformatics. 2013;14: 184. doi:10.1186/1471-2105-14-184

42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

43. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10: R25. doi:10.1186/gb-2009-10-3-r25

44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

45. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25: 1966–1967. doi:10.1093/bioinformatics/btp336

46. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18: 1851–1858. doi:10.1101/gr.078212.108

47. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41: 1061–1067. doi:10.1038/ng.437

48. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, et al. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Methods. 2010;7: 576–577. doi:10.1038/nmeth0810-576

49. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol. 2009;5: e1000386. doi:10.1371/journal.pcbi.1000386

50. Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics. 2010;26: i350–7. doi:10.1093/bioinformatics/btq216

51. Homer N, Merriman B, Nelson SF. BFAST: An Alignment Tool for Large Scale Genome Resequencing. PLoS One. 2009;4: e7767. doi:10.1371/journal.pone.0007767

52. Ning Z. SSAHA: A Fast Search Method for Large DNA Databases. Genome Res. 2001;11: 1725–1729. doi:10.1101/gr.194201

53. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

54. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012; doi:10.1038/nrg3164

55. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011;7: e1002384. doi:10.1371/journal.pgen.1002384

56. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

57. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25: 2283–2285. doi:10.1093/bioinformatics/btp373

58. Website [Internet].

59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

60. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Keira Cheetham R. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics. 2012;28: 1811–1817. doi:10.1093/bioinformatics/bts271

61. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456: 53–59. doi:10.1038/nature07517

62. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009;19: 1124–1132. doi:10.1101/gr.088013.108

63. Ajay SS, Parker SCJ, Abaan HO, Fajardo KVF, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome Res. 2011;21: 1498–1505. doi:10.1101/gr.123638.111

64. Krøigård AB, Thomassen M, Lænkholm A-V, Kruse TA, Larsen MJ. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. PLoS One. 2016;11: e0151664. doi:10.1371/journal.pone.0151664

65. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011;29: 908–914. doi:10.1038/nbt.1975

66. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. 2011;21: 940–951. doi:10.1101/gr.117259.110

67. Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C, et al. A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet. 2012;8: e1002944. doi:10.1371/journal.pgen.1002944

68. Marschall T, Costa IG, Canzar S, Bauer M, Klau GW, Schliep A, et al. CLEVER: clique-enumerating variant finder. Bioinformatics. 2012;28: 2875–2882. doi:10.1093/bioinformatics/bts566

69. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12: 363–376. doi:10.1038/nrg2958

70. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526: 75–81. doi:10.1038/nature15394

71. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. Nat Biotechnol. 2009;27: 455–457. doi:10.1038/nbt0509-455

72. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6: 677–681. doi:10.1038/nmeth.1363

73. Sindi S, Helman E, Bashir A, Raphael BJ. A geometric approach for classification and comparison of structural variants. Bioinformatics. 2009;25: i222–30. doi:10.1093/bioinformatics/btp208

74. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28: i333–i339. doi:10.1093/bioinformatics/bts378

75. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. 2010;20: 623–635. doi:10.1101/gr.102970.109

76. Lee S, Hormozdiari F, Alkan C, Brudno M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat Methods. 2009;6: 473–474. doi:10.1038/nmeth.f.256

77. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 2009;10: R23. doi:10.1186/gb-2009-10-2-r23

78. Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, et al. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics. 2010;26: 1277–1283. doi:10.1093/bioinformatics/btq152

79. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach

to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25: 2865–2871. doi:10.1093/bioinformatics/btp394

80. Li S, Li R, Li H, Lu J, Li Y, Bolund L, et al. SOAPindel: efficient identification of indels from short paired reads. Genome Res. 2013;23: 195–200. doi:10.1101/gr.132480.111

81. Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, Mark K, et al. Detection of structural variants and indels within exome data. Nat Methods. 2011;9: 176–178. doi:10.1038/nmeth.1810

82. Lam HYK, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol. 2010;28: 47–55. doi:10.1038/nbt.1600

83. Abo RP, Ducar M, Garcia EP, Thorner AR, Rojas-Rudilla V, Lin L, et al. BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. Nucleic Acids Res. 2015;43: e19. doi:10.1093/nar/gku1211

84. Sherry ST. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29: 308–311. doi:10.1093/nar/29.1.308

85. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2016;45: D840–D845. doi:10.1093/nar/gkw971

86. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . Taylor & Francis; 2012;6: 80–92. doi:10.4161/fly.19695

87. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. academic.oup.com; 2010;38: e164. doi:10.1093/nar/gkq603

88. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4: 1073–1081. doi:10.1038/nprot.2009.86

89. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. Wiley Online Library; 2013;Chapter 7: Unit7.20. doi:10.1002/0471142905.hg0720s76

90. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. nature.com; 2010;7: 575–576. doi:10.1038/nmeth0810-575

91. Griffith M, Griffith OL. HGMD (Human Gene Mutation Database). Dictionary of Bioinformatics and Computational Biology. 2004. doi:10.1002/9780471650126.dob0942

92. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. academic.oup.com; 2005;33: D514–7. doi:10.1093/nar/gki033

93. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. academic.oup.com; 2014;42: D980–5. doi:10.1093/nar/gkt1113

94. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014;47: 11.12.1–34. doi:10.1002/0471250953.bi1112s47

95. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. academic.oup.com; 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

96. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. ncbi.nlm.nih.gov; 2012;3: 35. doi:10.3389/fgene.2012.00035

97. Baird PA, Anderson TW, Newcombe HB, Lowry RB. Genetic disorders in children and young adults: a population study. Am J Hum Genet. ncbi.nlm.nih.gov; 1988;42: 677–693. Available: https://www.ncbi.nlm.nih.gov/pubmed/3358420

98. Robinson PN, Krawitz P, Mundlos S. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin Genet. Wiley Online Library; 2011;80: 127–132. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1399-0004.2011.01713.x/full

99. Ku C-S, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. Hum Genet. Springer; 2011;129: 351–370. doi:10.1007/s00439-011-0964-2

100. Lalonde E, Albrecht S, Ha K, Jacob K. Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Mutat Res. Wiley Online Library; 2010; Available: http://onlinelibrary.wiley.com/doi/10.1002/humu.21293/full

101. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. nature.com; 2010;42: 790–793. doi:10.1038/ng.646

102. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40: D54–6. doi:10.1093/nar/gkr854

103. Whiteford N. An analysis of the feasibility of short read sequencing. Nucleic Acids Res. 2005;33: e171–e171. doi:10.1093/nar/gni170

104. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of

read trimming effects on Illumina NGS data analysis. PLoS One. 2013;8: e85024. doi:10.1371/journal.pone.0085024

105. Holtgrewe M, Emde A-K, Weese D, Reinert K. A novel and well-defined benchmarking method for second generation read mapping. BMC Bioinformatics. bmcbioinformatics.biomedcentral. …; 2011;12: 210. doi:10.1186/1471-2105-12-210

106. Data. GATK | Resource Bundle [Internet]. [cited 18 Jul 2017]. Available: https://software.broadinstitute.org/gatk/download/bundle

107. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS One. 2012;7: e52249. doi:10.1371/journal.pone.0052249

108. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. 2013. pp. 11.10.1–11.10.33. doi:10.1002/0471250953.bi1110s43

109. Teer JK, Green ED, Mullikin JC, Biesecker LG. VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer. Bioinformatics. 2012;28: 599–600. doi:10.1093/bioinformatics/btr711

110. Görmez Z, Bakir-Gungor B, Sağıroğlu MŞ. HomSI: a homozygous stretch identifier from next-generation sequencing data. Bioinformatics. 2013;30: 445–447. doi:10.1093/bioinformatics/btt686

111. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733–45. doi:10.1093/nar/gkv1189

112. Exome Variant Server. In: Exome Variant Server [Internet]. [cited 19 Aug 2013]. Available: http://http://evs.gs.washington.edu/EVS/

113. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536: 285–291. doi:10.1038/nature19057

114. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47: 435–444. doi:10.1038/ng.3247

115. Orphanet. In: The portal for rare diseases and orphan drugs [Internet]. [cited 19 Aug 2013]. Available: http://www.orpha.net/

116. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15: 1034–1050. doi:10.1101/gr.3715005

117. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting

genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102

118. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32: 894–899. doi:10.1002/humu.21517

119. Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31: 3812–3814. doi:10.1093/nar/gkg509

120. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46: 310–315. doi:10.1038/ng.2892

121. Horaitis O, Talbot CC Jr, Phommarinh M, Phillips KM, Cotton RGH. A database of locus-specific databases. Nat Genet. 2007;39: 425. doi:10.1038/ng0407-425

122. Alanay Y, Ergüner B, Utine E, Haçariz O, Kiper POS, Taşkıran EZ, et al. TMCO1 deficiency causes autosomal recessive cerebrofaciothoracic dysplasia. Am J Med Genet A. 2014;164A: 291–304. doi:10.1002/ajmg.a.36248

123. Caglayan AO, Per H, Akgumus G, Gumus H, Baranoski J, Canpolat M, et al. Whole-exome sequencing identified a patient with TMCO1 defect syndrome and expands the phenotic spectrum. Clin Genet. 2013;84: 394–395. doi:10.1111/cge.12088

124. Xin B, Puffenberger EG, Turben S, Tan H, Zhou A, Wang H. Homozygous frameshift mutation in TMCO1 causes a syndrome with craniofacial dysmorphism, skeletal anomalies, and mental retardation. Proc Natl Acad Sci U S A. 2010;107: 258–263. doi:10.1073/pnas.0908457107

125. Bayrakli F, Guclu B, Yakicier C, Balaban H, Kartal U, Erguner B, et al. Mutation in MEOX1 gene causes a recessive Klippel-Feil syndrome subtype. BMC Genet. 2013;14: 95. doi:10.1186/1471-2156-14-95

126. Saga Y. The mechanism of somite formation in mice. Curr Opin Genet Dev. 2012;22: 331–338. doi:10.1016/j.gde.2012.05.004

127. Vasta V, Ng SB, Turner EH, Shendure J, Hahn SH. Next generation sequence analysis for mitochondrial disorders. Genome Med. 2009;1: 100. doi:10.1186/gm100

128. Wills-Karp M. Faculty of 1000 evaluation for Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma [Internet]. F1000 - Post-publication peer review of the biomedical literature. 2007. doi:10.3410/f.1089136.542319

129. Mohamed JY, Faqeih E, Alsiddiky A, Alshammari MJ, Ibrahim NA, Alkuraya FS. Mutations in MEOX1, encoding mesenchyme homeobox 1, cause Klippel-Feil anomaly. Am J Hum Genet. 2013;92: 157–161. doi:10.1016/j.ajhg.2012.11.016

130. Mankoo BS, Skuntz S, Harrigan I, Grigorieva E, Candia A, Wright CVE, et al. The concerted action of Meox homeobox genes is required upstream of genetic

pathways essential for the formation, patterning and differentiation of somites. Development. 2003;130: 4655–4664. doi:10.1242/dev.00687

131. Bayrakli F, Poyrazoglu HG, Yuksel S, Yakicier C, Erguner B, Sagiroglu MS, et al. Hereditary spastic paraplegia with recessive trait caused by mutation in KLC4 gene. J Hum Genet. 2015;60: 763–768. doi:10.1038/jhg.2015.109

132. Zhu H, Lee HY, Tong Y, Hong B-S, Kim K-P, Shen Y, et al. Crystal structures of the tetratricopeptide repeat domains of kinesin light chains: insight into cargo recognition mechanisms. PLoS One. 2012;7: e33943. doi:10.1371/journal.pone.0033943

133. Gindhart JG Jr, Desai CJ, Beushausen S, Zinn K, Goldstein LS. Kinesin light chains are essential for axonal transport in Drosophila. J Cell Biol. 1998;141: 443–454. Available: https://www.ncbi.nlm.nih.gov/pubmed/9548722

134. Rahman A, Kamal A, Roberts EA, Goldstein LS. Defective kinesin heavy chain behavior in mouse kinesin light chain mutants. J Cell Biol. 1999;146: 1277–1288. Available: https://www.ncbi.nlm.nih.gov/pubmed/10491391

135. Sahin Y, Güngör O, Gormez Z, Demirci H, Ergüner B, Güngör G, et al. Exome sequencing identifies a novel homozygous CLN8 mutation in a Turkish family with Northern epilepsy. Acta Neurol Belg. 2017;117: 159–167. doi:10.1007/s13760-016-0721-3

136. Herva R, Tyynelä J, Hirvasniemi A, Syrjäkallio-Ylitalo M, Haltia M. Northern epilepsy: a novel form of neuronal ceroid-lipofuscinosis. Brain Pathol. 2000;10: 215–222. Available: https://www.ncbi.nlm.nih.gov/pubmed/10764041

137. Vantaggiato C, Redaelli F, Falcone S, Perrotta C, Tonelli A, Bondioni S, et al. A novel CLN8 mutation in late-infantile-onset neuronal ceroid lipofuscinosis (LINCL) reveals aspects of CLN8 neurobiological function. Hum Mutat. 2009;30: 1104–1116. doi:10.1002/humu.21012

138. Reinhardt K, Grapp M, Schlachter K, Brück W, Gärtner J, Steinfeld R. Novel CLN8 mutations confirm the clinical and ethnic diversity of late infantile neuronal ceroid lipofuscinosis. Clin Genet. 2010;77: 79–85. doi:10.1111/j.1399-0004.2009.01285.x

139. Mahajnah M, Zelnik N. Phenotypic heterogeneity in consanguineous patients with a common CLN8 mutation. Pediatr Neurol. 2012;47: 303–305. doi:10.1016/j.pediatrneurol.2012.05.016

140. Ranta S, Topcu M, Tegelberg S, Tan H, Ustübütün A, Saatci I, et al. Variant late infantile neuronal ceroid lipofuscinosis in a subset of Turkish patients is allelic to Northern epilepsy. Hum Mutat. 2004;23: 300–305. doi:10.1002/humu.20018

141. Cannelli N, Cassandrini D, Bertini E, Striano P, Fusco L, Gaggero R, et al. Novel mutations in CLN8 in Italian variant late infantile neuronal ceroid lipofuscinosis: Another genetic hit in the Mediterranean. Neurogenetics. 2006;7: 111–117. doi:10.1007/s10048-005-0024-y

142. Stein L. Creating a bioinformatics nation. Nature. 2002;417: 119–120.

doi:10.1038/417119a

143. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2014;15: 256–278. doi:10.1093/bib/bbs086

144. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013;369: 1502–1511. doi:10.1056/NEJMoa1306555

145. Brogna S, Wen J. Nonsense-mediated mRNA decay (NMD) mechanisms. Nat Struct Mol Biol. 2009;16: 107–113. doi:10.1038/nsmb.1550

146. Pereverzev AP, Gurskaya NG, Ermakova GV, Kudryavtseva EI, Markina NM, Kotlobay AA, et al. Method for quantitative analysis of nonsense-mediated mRNA decay at the single cell level. Sci Rep. 2015;5: 7729. doi:10.1038/srep07729

147. Chong Z, Ruan J, Gao M, Zhou W, Chen T, Fan X, et al. novoBreak: local assembly for breakpoint detection in cancer genomes. Nat Methods. 2017;14: 65–67. doi:10.1038/nmeth.4084

148. Chen K, Chen L, Fan X, Wallis J, Ding L, Weinstock G. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. Genome Res. 2014;24: 310–317. doi:10.1101/gr.162883.113

149. Gymrek M. PyBamView: a browser-based application for viewing short read alignments. Bioinformatics. 2014;30: 3405–3407. doi:10.1093/bioinformatics/btu565

150. Edmonson MN, Zhang J, Yan C, Finney RP, Meerzaman DM, Buetow KH. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. Bioinformatics. 2011;27: 865–866. doi:10.1093/bioinformatics/btr032

151. Spies N, Zook JM, Salit M, Sidow A. svviz: a read viewer for validating structural variants. Bioinformatics. 2015;31: 3994–3996. doi:10.1093/bioinformatics/btv478

152. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

153. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7: 203–214. doi:10.1089/10665270050081478

154. Cetinkaya A, Xiong JR, Vargel İ, Kösemehmetoğlu K, Canter Hİ, Gerdan ÖF, et al. Loss-of-Function Mutations in ELMO2 Cause Intraosseous Vascular Malformation by Impeding RAC1 Signaling. Am J Hum Genet. 2016;99: 299–317. doi:10.1016/j.ajhg.2016.06.008

155. Parikh H, Mohiyuddin M, Lam HYK, Iyer H, Chen D, Pratt M, et al. svclassify: a method to establish benchmark structural variant calls. BMC Genomics. 2016;17: 64. doi:10.1186/s12864-016-2366-2

156. Russell J, Cohn R. International Hapmap Project [Internet]. Book on Demand

Limited; 2012. Available: https://books.google.com/books/about/International_Hapmap_Project.html?hl=&id=4Bt5MAEACAAJ

157. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011;39: D19–21. doi:10.1093/nar/gkq1019

158. Bartenhagen C, Dugas M. Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms. Brief Bioinform. 2016;17: 51–62. doi:10.1093/bib/bbv028

159. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007;5: e254. doi:10.1371/journal.pbio.0050254

## Appendix A Supplementary Figures & Tables

**Table A1:** High confidence BreakDancer SV calls at 20:45,021K-45,040K.

| Chr1 | Pos1 | Orientation 1 | Chr 2 | Pos2 | Orientation 2 | Type | Size | Score | Num Reads |
|---|---|---|---|---|---|---|---|---|---|
| 20 | 45021083 | 3+2- | 20 | 45021281 | 640+568- | DEL | 341 | 99 | 221 |
| 20 | 45024714 | 640+568- | 20 | 45031300 | 653+310- | INV | 7155 | 99 | 287 |
| 20 | 45024714 | 640+568- | 20 | 45037323 | 7+213- | INV | 14009 | 99 | 132 |
| 20 | 45024714 | 640+568- | 20 | 45037534 | 0+28- | INV | 13470 | 99 | 19 |
| 20 | 45025280 | 25+21- | 20 | 45025346 | 114+150- | DEL | 347 | 99 | 68 |
| 20 | 45031300 | 653+310- | 20 | 45037323 | 7+213- | INV | 5827 | 99 | 60 |
| 20 | 45031300 | 653+310- | 20 | 45037534 | 0+28- | INV | 6006 | 74 | 7 |

**Table A2:** BreaKmer output showing the breakpoints at 20:45,021K-45,040K.

| Target_Name | SV_subtype | All_genomic_breakpoints | Target_genomic_breakpoints |
|---|---|---|---|
| ELMO2 | trl | chr20:45031259,chr3:87987195 | chr20:45031259 |
| ELMO2 | tandem_dup | chr20:45031259,chr20:45022911 | chr20:45031259,chr20:45022911 |
| ELMO2 | trl | chr20:45031259,chr9:113150090 | chr20:45031259 |
| ELMO2 | inversion | chr20:45037123,chr20:45031202 | chr20:45037123,chr20:45031202 |
| ELMO2 | inversion | chr20:45037123,chr20:45031202 | chr20:45037123,chr20:45031202 |
| ELMO2 | inversion | chr20:45037123,chr20:45031202 | chr20:45037123,chr20:45031202 |
| ELMO2 | inversion | chr20:45037123,chr20:45031202 | chr20:45037123,chr20:45031202 |
| ELMO2 | inversion | chr20:45031193,chr20:45023171 | chr20:45031193,chr20:45023171 |
| ELMO2 | inversion | chr20:45031193,chr20:45023171 | chr20:45031193,chr20:45023171 |
| ELMO2 | inversion | chr20:45026874,chr20:45026625 | chr20:45026874,chr20:45026625 |
| ELMO2 | inversion | chr20:45024428,chr20:45024084 | chr20:45024428,chr20:45024084 |
| ELMO2 | inversion | chr20:45024428,chr20:45024084 | chr20:45024428,chr20:45024084 |
| ELMO2 | trl | chr20:45024263,chr15:22792297 | chr20:45024263 |
| ELMO2 | inversion | chr20:45023171,chr20:45031193 | chr20:45023171,chr20:45031193 |
| ELMO2 | inversion | chr20:45023171,chr20:45031193 | chr20:45023171,chr20:45031193 |

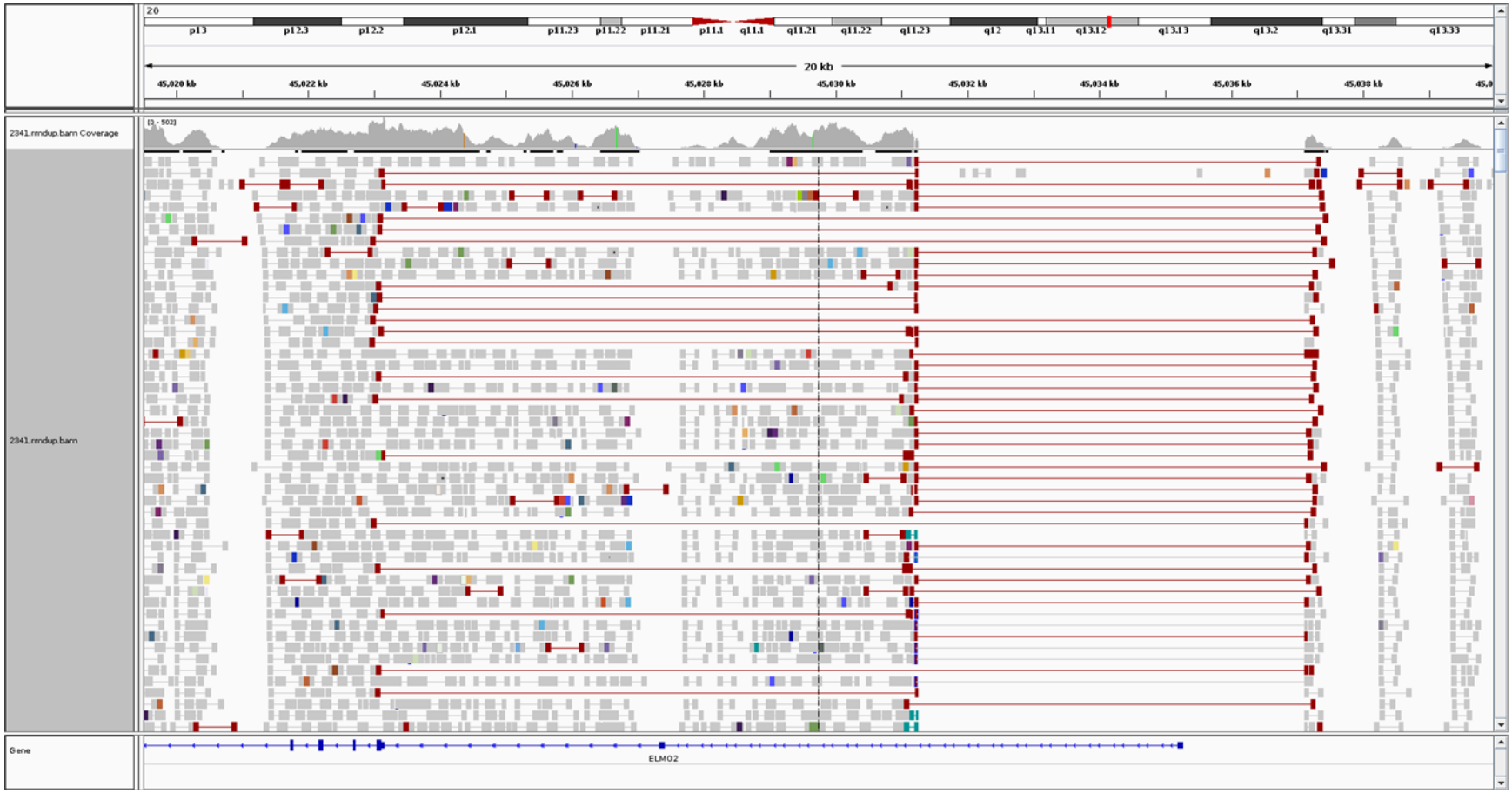


**Figure A1:** IGV image showing complex rearrangement affecting the ELMO2 gene from the aligned short read file. Discordant pair-end reads are shown in red.


**Figure A2:** IGV image of ELMO2 region after substituting the affected reference sequence with the assembled contig prioritized by SVMap. Compared to Figure S1, the number of discordant reads decreased to normal levels of the DNA library.

| Score | Expect | Identities | Gaps | Strand |
|---|---|---|---|---|
| 825 bits(914) | 0.0 | 461/462(99%) | 1/462(0%) | Plus/Plus |

```
Query  2407  TTTTTGGAAATGAGAGGAAACAGAAACTGAGATGGAAAATACAGATGAACTTCAGAATTA  2466
             ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     TTTTTGGAA-TGAGAGGAAACAGAAACTGAGATGGAAAATACAGATGAACTTCAGAATTA  59

Query  2467  AACTTTAGAAAACCGTGATTGCAGCCGTGTCTGTGTTTTTGTCTCGCAGAATTAGAGCCC  2526
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  60    AACTTTAGAAAACCGTGATTGCAGCCGTGTCTGTGTTTTTGTCTCGCAGAATTAGAGCCC  119

Query  2527  ATTGGGAACGATGCCACCACCGTCAGACATTGTCAAAGTGGCCATTGAGTGGCCAGGTGC  2586
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  120   ATTGGGAACGATGCCACCACCGTCAGACATTGTCAAAGTGGCCATTGAGTGGCCAGGTGC  179

Query  2587  TAACGCCCAGCTCCTTGAAATCGACCAGGTATGCTCCTGAAGTGAGAAGCAGTGGTTCAA  2646
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  180   TAACGCCCAGCTCCTTGAAATCGACCAGGTATGCTCCTGAAGTGAGAAGCAGTGGTTCAA  239

Query  2647  GGAAAGGCACCTGGGGAGTGCATGGCAGAGGACATCTTGAGGGATGGGGACCACCGGCAT  2706
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  240   GGAAAGGCACCTGGGGAGTGCATGGCAGAGGACATCTTGAGGGATGGGGACCACCGGCAT  299

Query  2707  CAAGAGTAAGAACGAGCAACAGGAAGGCTAAGCTTTGGGCTTACAACTTAAGCTTAGTTG  2766
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  300   CAAGAGTAAGAACGAGCAACAGGAAGGCTAAGCTTTGGGCTTACAACTTAAGCTTAGTTG  359

Query  2767  TAACCCAGGACTCTTCCCAGCACACCACACTCCCTCCCTACCCAAGCTCCCAGTGGGCAA  2826
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  360   TAACCCAGGACTCTTCCCAGCACACCACACTCCCTCCCTACCCAAGCTCCCAGTGGGCAA  419

Query  2827  CTAGGTTCAGGTGCCGTGTTTTCAGGCTGGAATCAAATAGGA  2868
              |||||||||||||||||||||||||||||||||||||||||
Sbjct  420   CTAGGTTCAGGTGCCGTGTTTTCAGGCTGGAATCAAATAGGA  461
```

**Figure A3:** Pairwise alignment of the assembled contig (query) prioritized by SVMap against the Sanger sequence (sbjct) of the region affected by the complex rearrangement.

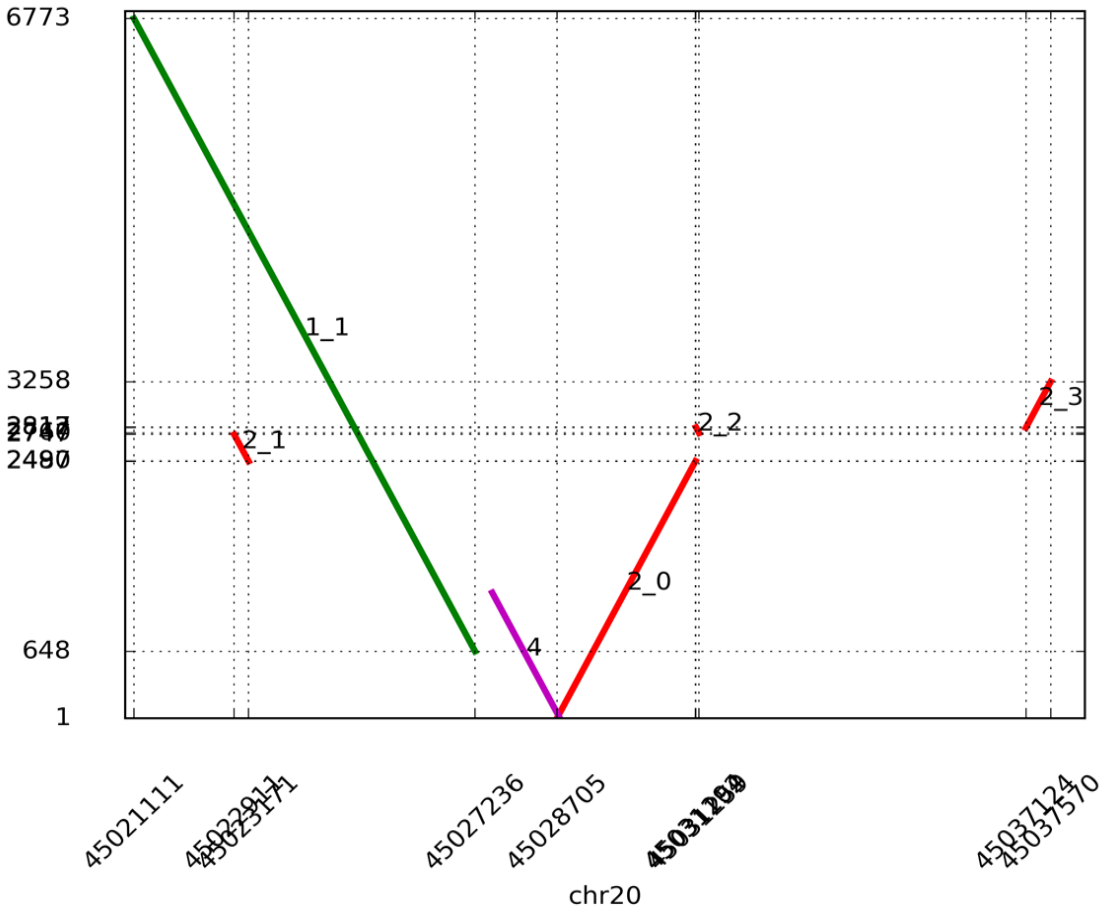


**Figure A4:** The plot showing the local hits selected from the ELMO2 alignment of ELMO2 assembly. The hits labeled "2_0, 2_1, 2_2, 2_3" belong to the contig "2" which scored highest among the other contigs. The contig 2 covers complete scope of the complex rearrangement.