1q21.1 syndrome:

A perspective on Structural Variant detection & the evolutionary profile of associated protein domains

Final Master Project

Paula España Bonilla

Master's in bioinformatics and Biostatistics (UOC -UB)

Area 3 – Subarea 8: Analysis and integration of omics data

Cedric Boeckx

Andreu Paytuví

January 2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*1q21.1 syndrome: A perspective on Structural Variant detection & the evolutionary profile of associated protein domains*
Nombre del autor:	*Paula España Bonilla*
Nombre del consultor/a:	*Cedric Boeckx*
Nombre del PRA:	*Andreu Paytuví*
Fecha de entrega (mm/aaaa):	*01/2021*
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Área 3 – Subárea 8: Análisis e integración de datos ómicos*
Idioma del trabajo:	*Inglés*
Palabras clave	*Long-reads, Olduvai, simulation*

Resumen del Trabajo (máximo 250 palabras):

Pese a que el síndrome 1q21.1 ha sido descrito clínicamente, sus causas genéticas se desconocen. Esta región es altamente repetitiva y por ello requiere tecnologías de secuenciación basadas en *long reads* para poder genotipar a sus pacientes de forma precisa. Además, 1q21.1 comprende múltiples duplicaciones segmentarias, tales como los dominios Olduvai, asociados con macro- y microcefalia. El primer objetivo de este estudio es diseñar un *pipeline* de detección de variantes estructurales a partir de secuenciación con *long reads*. El proyecto incluye la simulación de genomas con diversas variantes estructurales y la simulación de *reads* de secuenciación que permitan validar la aplicación de este *pipeline*. El segundo objetivo es analizar la filogenia de los dominios Olduvai durante la evolución. Los resultados muestran que el diseño del *pipeline* es exitoso y permite la identificación de las variantes estructurales simuladas. Este *pipeline* puede ser mejorado y aplicado a pacientes reales. Además, los dominios Olduvai se encuentran altamente conservados en su secuencia ancestral, mientras que sufren diversificación de secuencias en regiones de duplicación segmentaria. El estudio filogenético indica que los dominios Olduvai juegan un rol relevante en diversas especies. En definitiva, este estudio es un enfoque introductorio al análisis de variantes estructurales a partir de *long reads* y también permite resaltar la importancia de los dominios Olduvai en la evolución de primates a humanos.

Abstract (in English, 250 words or less):

While 1q21.1 syndrome has been described from clinical aspects, the underlaying genetic mechanisms remain to be discovered. Due to the repetitive nature of the region, patients with this syndrome require long read sequencing strategies for accurate mapping of the region. The region is enriched in segmental duplications such as Olduvai domains that are associated to the phenotype of macro- and microcephaly. The first objective of this study is to design a pipeline for structural variant call in long read sequencing data. The project includes structural variant and read simulation for pipeline evaluation. The second objective is to analyse the phylogeny of Olduvai domains during evolution. The results indicate correct structural variant simulation and generation of the reads. The pipeline design is successful in the identification of simulated structural variants. Moreover, the evolutionary profile of Olduvai domains indicates high conservation in the ancestral copy and posterior diversification of the sequence in segmental duplicated regions. The study constitutes a preliminary approach on structural variation identification from long read datasets and highlights Olduvai domains as important actors in evolution from primates to humans.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Context and justification of the Project

1.1.1. General description:

The 1q21 region is a highly repetitive region of the human genome, divided into 3 parts: 1q21.1, 1q21.2, 1q21.3. Clinical reports on patients with 1q21.1 deletion show increased prevalence of microcephaly, whereby 1q21.1 duplication carriers describe macrocephaly and Autism Spectrum Disorder features (Bernier et al., 2016; Mefford et al., 2008). 1q21 region is highly enriched in the Neuroblastoma Breakpoint Family (NBPF) of genes. This family contains the so called 'Olduvai domains' (also referred as DUF1220 or NBPF domains), which show human-specific amplification during evolution (Dumas et al., 2012; Popesco et al., 2006).

Hospital de Sant Joan de Déu contains 9 patients diagnosed with 1q21 microdeletions and microcephalic phenotype, and 2 macrocephalic patients who exhibit 1q21 microduplications. The genome of these patients will be sequenced by PacBio long reads technology for a research project in Institut de Biologia Molecular de Barcelona (IBMB-CSIC).

The main scope of this project is to develop an analytic pipeline to process the PacBio data and assess the specific changes produced in the genome of these patients. Since the PacBio data is not available yet, this project will include a sample simulation of the patients. As a secondary consideration, the project will study Olduvai domain phylogeny across species.

1.1.2. Justification of the work:

On the one hand, the lack of cost-effective high-throughput techniques has limited the study of 1q21.1 in the past. Recent studies have identified novel candidates in this region as potentially implicated genes in brain development, such as NOTCH2NL genes (Lodewijk et al., 2020). In addition, other papers show the importance of structural variants in human-specific brain development (Florio et al., 2015). Thereby, the study of 1q21 region constitutes an interesting topic for evolutionary, clinical and biological fields.

On the other hand, the biomedical field requires multidisciplinary expertise to produce experimental data, analyse this data and provide it with biological and clinical significance. For instance, sequencing technologies produce big amounts of information which need to be processed accurately. In this sense, this project aims to produce a simple and effective tool to process and assemble PacBio sequencing data from Whole Genome Sequencing of 1q21 patients. This project will allow to perform a robust assembly of the genome of the patients in the future. In addition, Olduvai domain phylogeny can provide additional knowledge on the evolutionary history of this region. Overall, the genetic characterisation of the patients constitutes a useful source of information for clinical, biological and evolutionary characterisation.

1.2. Objectives

1.2.1. General objectives:

A. To develop an analytic pipeline for PacBio data analysis of sample patients.

B. To analyse the phylogeny of Olduvai domains among multiple species.

1.2.2. Specific objectives:

A. To develop an analytic pipeline for PacBio data analysis of sample patients.

   a. To simulate patient genomes with SV regions and generate derived PacBio reads

   b. To develop the bioinformatic pipeline for PacBio data analysis.

B. To analyse the phylogeny of Olduvai domains among multiple species.

   a. To align Olduvai sequences from different species

   b. To create a phylogenetic tree

1.3. Strategy and Methodology

The first objective of our project is to develop a set of useful analytic tools to detect different SV in the genome of our patients after PacBio sequencing. The creation of this pipeline is limited by the fact that genome sequencing of the patients is still in process. Therefore, the best strategy is to simulate the genomes of the patients and simulate derived PacBio reads from these genomes. This approach will allow (1) to have data for the analysis and (2) to test the efficiency of our method.

Data simulation is a popular strategy to benchmark computational methods, which provides a big number of data and a controlled environment with predefined parameters (Escalona et al., 2016). Multiple software for long-read data simulation have been developed, such as PBSIM2, PaSS or SimLoRD (Ono et al., 2020; Stöcker et al., 2016; Zhang et al., 2019). In addition, other programmes like SVEngine have focused on simulating data with embedded structural variations (Xia et al., 2018). However, for our project we plan to use SVGen, a simulation tool that exhibits a wide range of options for single nucleotide variations, structural variations, and it is able to produce PacBio-derived long read datasets (LA Lima, H Yang, C Dong, n.d.).

Data simulation is followed by genome assembly. Long read assembly programs are different from their short-read counterparts because they must adapt to the increased read length and error rate. Assembly approaches can be divided into three methods - hybrid, direct and hierarchical – depending on the number of rounds of read alignment and dependence on complementary information (Koren et al., 2017).

For our analysis, the best strategy is the hierarchical method, which does not require a secondary technology and uses multiple rounds of read overlapping and correction. Likewise, the software to use is Canu, a sequence assembler that creates complete eukaryotic assemblies from PacBio and Nanopore sequencing (Miga et al., 2020). As a final step, the resulting contigs will be annotated using MAKER (Cantarel et al., 2008). This software annotates and masks repetitive elements in the genome. It basically uses two methods: ab initio gene prediction and alignment with ribonucleic acid and protein datasets. This methods result in the identification of our sequences in the human genome (Campbell et al., 2014).

The second objective of our project is to describe Olduvai domain sequence evolution across species. In order to assess sequence similarity, the best approach is to use is multiple sequence alignment between Olduvai sequences. Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences are obtained. There are several sequence alignment methods for multiple sequences, which are suitable for different types of sequences (i.e., Clustal Omega, Kalign, MAFFT) (Chenna et al., 2003). The software MUSCLE (multiple sequence comparison by log-expectation) is an accurate, fast tool for medium-size protein sequences. Thereby, MUSCLE will be appropriate for our alignment (Edgar, 2004). The protein sequences will be retrieved from Uniprot (Consortium, 2019).

Finally, a phylogenetic tree can decipher the molecular evolution of Olduvai domains. Maximum likelihood phylogenies provide the most successful advances in tree building methods but require big computational power. However, the heuristic of PHYML software implements fast and simple estimations of maximum likelihood phylogenies that can be performed in a standard personal computer. Thus, PHYML will be applied in our project (Guindon et al., 2005).

1.4. Work Planification

The development of the project is divided in 7 PECs, which are the different assignments that are done between September the 16th, 2020 and January the 20th, 2021.

- PEC0: Definition of the project. The topic of the project and the strategy to follow were defined.
- PEC1: Work plan. Justification of the project, declaration of general and specific objectives and timeline planification.
- PEC2: Development of the project I. Objective one was produced, which includes all the phylogenetic analysis. Objective 2 was started by producing the structural variant simulation.
- PEC3: Development of the project I. Read simulation and the development of the pipeline were created (Objective 2).
- PEC4: Memory writing. Graphs and figures were created. The memory was written.
- PEC5a: Virtual presentation. The virtual presentation will explain the objectives, methodology and results of this project. The presentation will be created and recorded in Power Point.
- PEC5b: Public Defense. The defense is prepared and questions or comments from the tribunal are answered.

The computational resources used in this project include a Windows 10 personal computer with 16 GB of memory. Since the analytic pipeline requires increased computational power and a Linux environment, Google Cloud Compute Engine was required. A n1-highmem-4 Ubuntu, 20.04 LTS virtual machine was created, with 4 vCPUs and 26 GB of memory. This environment required 300 $ of computational power that were provided for free to the newly created user.

The planification of the project was adjusted to PECs deadlines and it can be visualized in a Gantt diagram as follows:

Final Masters Project	16/09/20	23/09/20	29/09/20	06/10/20	14/10/20	21/10/20	28/10/20	04/11/20	11/11/20	18/11/20	25/11/20	02/12/20	09/12/20	15/12/20	22/12/20	29/12/20	06/01/21	13/01/21
PEC 0 – Definition of the project	🟩	🟩																
PEC 1 – Work plan			🟩	🟩														
PEC2 and PEC3 – Project execution					🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩					
Objective 2: Phylogeny analysis					🟧	🟧												
MUSCLE alignment					▨													
Phylogenetic tree						▨												
Objective 1: Sequencing pipeline							🟧	🟧	🟧	🟧	🟧	🟧	🟧					
Sample simulation – SVGen –							▨											
Read assembly – Canu –								▨	▨									
Sequence annotation – MAKER –										▨	▨							
Analysis and interpretation of the results												▨	▨					
PEC 4 – Report writing														🟩	🟩	🟩		
PEC 5 – Presentation																	🟩	🟩
PEC 5a – Elaboration of the presentation																	▩	
PEC 5b – Public Defense																		▩

Table 1. Gantt diagram with the TFM planification

1.4.1. Milestones

The following milestones represent the crucial parts of the project that need to be done in order to achieve our objectives. Each landmark is important to develop the next one. Consequently, a delay in one milestone can affect the timeline of the next parts of the project. The following work plan can be useful to pace the rhythm of the project:

Objective 1:

- MS1. Sample simulation
- MS2. Canu assembly
- MS3. MAKER annotation
- MS4. Evaluation of the model

Objective 2:

- MS1. MUSCLE alignment
- MS2. Phylogenetic tree building

1.4.2. Risk analysis

On the first place, common problems when using different software is to download and install them properly. This process can produce errors or incompatibilities that require time to be solved.

On the second place, programming skills can also be a limiting factor to produce results. Since my profile is not enriched in informatics background, the project can experience difficulties to apply the software.

Finally, the available computing power can also be a problem for statistical analysis, alignments or the use of some of the software. This problem could be solved by implementing a computing cloud.

1.5. Summary of the products

1.5.1. Work plan:

This document includes a justification of the reasons to perform this project. It also includes the objectives and work planification of the process.

1.5.2. Report:

The final report constitutes the most complete document of the project, that will include a state-of-the-art of the subject, the objectives, methods of the project and finally the results and conclusions found.

1.5.3. Resulting product:

The resulting product of this project will include the report with the detailed information on the work done, as well as a useful pipeline to analyse PacBio sequencing data.

1.5.4. Virtual presentation:

The virtual presentation of this project will explain the highlights of the project in a clear, concise and brief manner.

1.5.5. Self-assessment:

The self-assessment will be a useful tool to explain the development of the project and evaluate different aspects from a personal point of view.

1.6. Description of other chapters of the memory

The memory will include the following parts:

1.6.1. Introduction:

Description of 1q21.1 syndrome, the application of long-read sequencing in structural variant detection and the contribution of Olduvai domains to the disease.

1.6.2. Materials and Methods:

Concise explanation of the methods used for structural variant and read simulation, design of the pipeline and methods applied for Olduvai phylogenetic studies.

1.6.3. Results:

Description of the results obtained from the application of the methods.

1.6.4. Discussion:

Comments on the general meaning of the results obtained and description of the limitations and possible improvements.

1.6.5. Conclusions:

Analysis of the retrieved from the study and fulfilment of the objectives.

1.6.6. Bibliography:

List of resources consulted in this project.

2. PROJECT MEMORY

Introduction

In 1953, Linus Pauling published its famous article “A Proposed Structure for the Nucleic Acids” in the Proceeding of the National Academy of Sciences. The article proposed a model of deoxyribonucleic acid (DNA) structure composed of a tripe helix. Even though this theory was incorrect, it was essential to inspire Watson and Crick to the discovery of the double helix DNA structure. By going back in history, we realize that science is made up of big mistakes like this one. The story that I will describe in this report is also filled with mistakes that eventually can lead one to find the right answer. Without further delay, I invite the readers to enjoy the beautiful story of sequencing.

A brief history of sequencing

In the 1950s, Watson and Crick revolutionized science with the discovery of the structure of DNA (Watson & Crick, 1953). As the double helix DNA was deciphered, its sequence was still a mystery to science. The research study of this project involves sequencing techniques. For this purpose, this report will firstly describe the evolution of sequencing methods along history.

In 1977, Alan Coulson and Frederick Sanger implemented a sequencing protocol that would change biology forever. Their method was based on a Polymerase Chain Reaction (PCR) that incorporated both usual nucleotides and radiolabelled dideoxynucleotides (ddNTPs), which lack the 3’ hydroxyl group and produce a stop in elongation. PCR fragments were separated by gel electrophoresis and allowed the reconstruction of the sequence by radio signals, leading to the development of the now widely known Sanger method. Since then, the Sanger method has been improved by using fluorescent ddNTPs and capillary-based electrophoresis, and it is still used for small-scale sequencing experiments (Giani et al., 2020; Heather & Chain, 2016; Kumar et al., 2019; Mardis, 2007). While Sanger sequencing is considered the first sequencing method, it is limited for the study of large DNA sequences.

In parallel to the discovery of radio- and fluorescently labelled dNTPs, researchers developed a luminescent method based on pyro-phosphate synthesis. This method is based on the production of pyrophosphate by the addition of each type of dNTP during chain elongation. Pyrophosphate is converted into adenosine triphosphate (ATP) by ATP sulfurylase, and ATP is substrate for luciferase, thus producing light proportional to the amount of pyrophosphate. Since then, other high-throughput techniques were developed, giving rise to Next Generation Sequencing (NGS) technologies, also known as Second Generation Sequencing (SGS). NGS have produced a paradigm shift by allowing huge amounts of DNA to be sequenced in a really short time. Nowadays, Illumina technology stands as the main representative of the NGS methods. It can sequence billions of reads of about 300 base-pair (bp) from different kind of genomes in a short time (Heather & Chain, 2016; Logsdon et al., 2020).

As a result of this explosion of techniques, the genomes of many species have been sequenced, starting by a virus called phiX174 in 1977 to the first completed human genome sequence in 2003 (Giani et al., 2020; Heather & Chain, 2016; Kumar et al., 2019; Mardis, 2007).

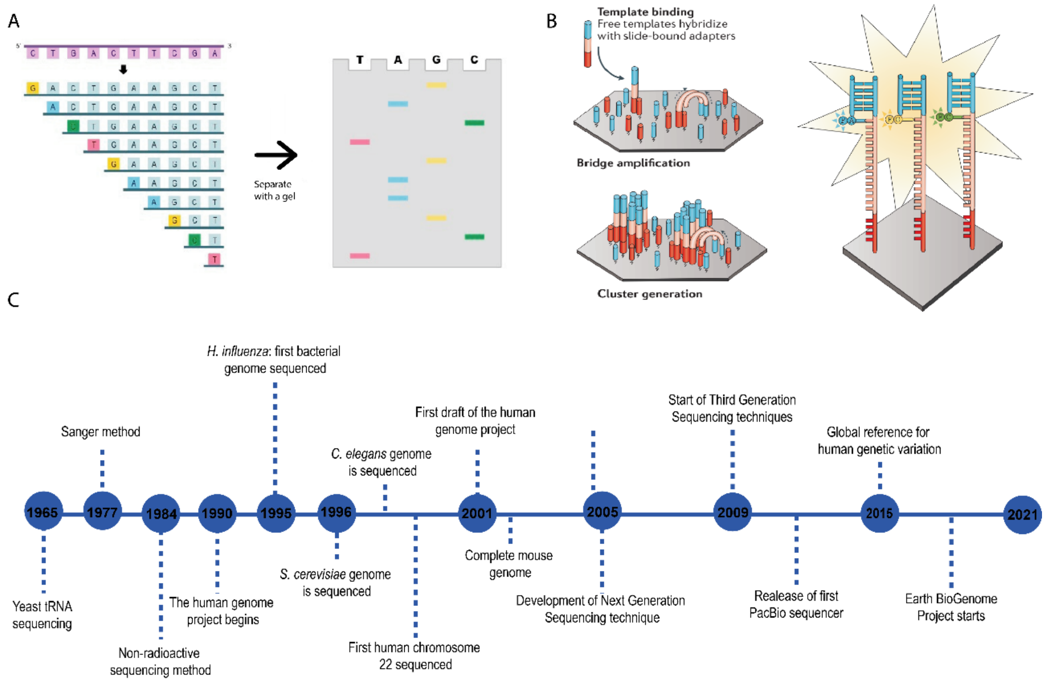

A
5'
C T G A C T T C G A
3
G A C T G A A G C T
A C T G A A G C T
C T G A A G C T
T G A A G C T
G A A G C T
A A G C T
A G C T
G C T
C T
T
Separate with a gel
T
A
G
C
B
Template binding
Free templates hybridize with slide-bound adapters
Bridge amplification
Cluster generation
C
1965
1977
1984
1990
1995
1996
2001
2005
2009
2015
2021
Sanger method
H. influenza: first bacterial genome sequenced
C. elegans genome is sequenced
First draft of the human genome project
Start of Third Generation Sequencing techniques
Global reference for human genetic variation
Yeast tRNA sequencing
Non-radioactive sequencing method
The human genome project begins
S. cerevisiae genome is sequenced
First human chromosome 22 sequenced
Complete mouse genome
Development of Next Generation Sequencing technique
Realease of first PacBio sequencer
Earth BioGenome Project starts

Figure 1. Sequencing milestones. (A) Diagram of Sanger sequencing method. Addition of dideoxynucleotides in a regular PCR stops its elongation and results in multiple fragments of the sequence with a tag in the last nucleotide. Fragments are separated by gel electrophoresis and the sequence is retrieved by reading the tags in each position. (B) Main steps in Illumina SGS. SGS relies on the creation of a library of DNA fragments that are surrounded by adapter sequences. The fragments are bounded on flow cell microwells and amplified by PCR which generates clusters of clonal sequences. Finally, modified nucleotides are added with DNA polymerase and primers. Each nucleotide is blocked by a 3′-O-azidomethyl group and is labelled with a base-specific (thus blocking nucleotide incorporation) and a cleavable fluorophore. Nucleotide incorporation generates specific fluorescence in each cycle that it is recorded to reconstruct the nucleotide sequence. (C) Timeline of major sequencing achievements. Pictures adapted from (Giani et al., 2020; Goodwin et al., 2016).

Third Generation Sequencing came to stay

While Next Generation Sequencing technologies are extremely powerful, they also have some limitations. One major drawback is their short reads. Complex genomes usually contain repetitive sequences that are bigger than the reads, which may lead to misassemblies and gaps in the sequence. That is the reason why some genomes are fragmented into hundreds or thousands of contigs, which are assemblies of the reads by overlapping regions. In addition, short reads do not allow to accurately detect structural variants (SVs) or have limited capacity to discriminate alleles from their respective homolog (Giani et al., 2020; Midha et al., 2019). Structural Variants are defined as deletions, insertions, duplications and inversions of more than 50 base pairs (bp) among human genomes (Ho et al., 2020).

Obviously, the solution to short read sequencing has emerged with long read sequencing technologies, regarded as the Third Generation Sequencing (TGS). The two representatives of TGS are single-molecule real-time (SMRT) by Pacific Biosciences (PacBio) and Nanopore sequencing by Oxford Nanopore Technologies. They differ in their chemistry and sequence detection approaches, which influence their read length, base accuracies and throughput (Logsdon et al., 2020).

In this report we will focus on the PacBio strategy. PacBio SMRT sequencing technology uses a circular DNA molecule - SMRTbell - composed of a double-stranded DNA insert molecule which is flanked with single-stranded DNA adapters. The SMRTbell is loaded on a SMRT Cell, which contains up to 8 million nanophotonic devices for sequencing, called zero-mode waveguides. During the sequencing process, the polymerase (found at the bottom of the well) incorporates fluorescently labelled deoxynucleoside triphosphates into the template. In each incorporation, a laser excites the fluorophore, and a camera records the colour and duration of the emission (Heather & Chain, 2016; Kumar et al., 2019; Logsdon et al., 2020; Midha et al., 2019; van Dijk et al., 2018).

PacBio can generate continuous long reads (CLRs) or high fidelity (HiFi) reads. CLRs can be larger than 30 kb but they contain around 10% of error rate, since the polymerase often can complete only one or a few passes around the template. HiFi reads are larger than 10 kb, which is shorter than CLRs, but are highly accurate (more than 99%) because they allow the polymerase to make several passes around the template, producing a circular consensus sequence (CCSs) (Logsdon et al., 2020).

Despite the different features of each procedure, the multiple developed sequencing strategies make step forward in the understanding of our genomes. A recent study comparing the Illumina and PacBio technologies concluded that PacBio can detect 47%

of the deletions and nearly 78% of insertions that were missed by Illumina (Chaisson et al., 2019). Third Generation Sequencing, by increasing the read length, has reduced the number of contigs and the number of gaps in the assembly (van Dijk et al., 2018). In conclusion, long read sequencing technologies such as PacBio or Oxford Nanopore Technologies are likely to become future benchmark sequencing tools in research and diagnostic platforms (Logsdon et al., 2020; van Dijk et al., 2018).

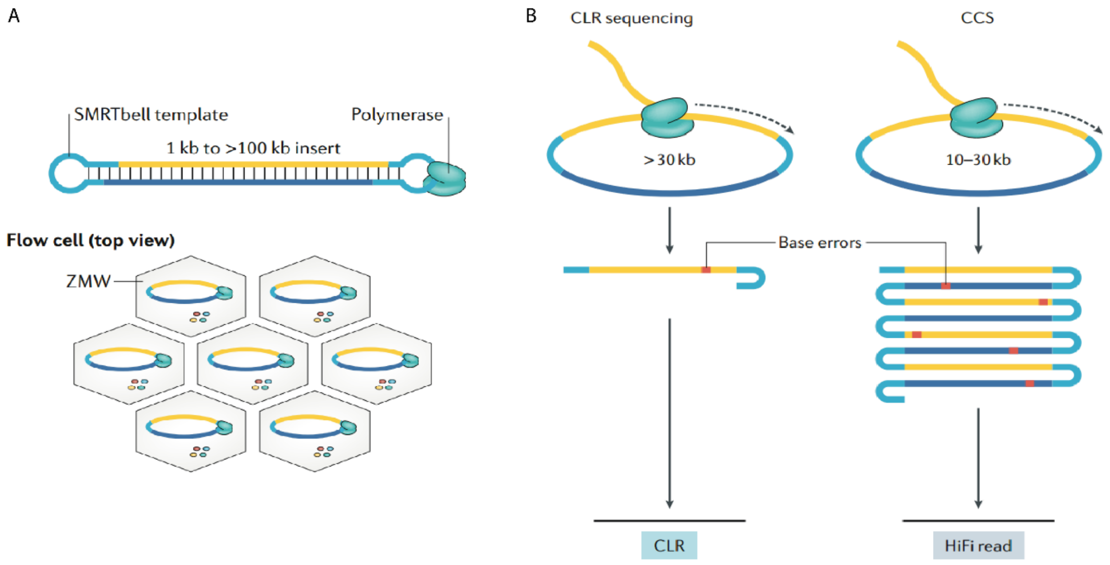


Figure 2. Overview of PacBio sequencing method. (A) PacBio template topology and flow cell view. (B) Comparison of the CLR and CCS sequencing methods. CCS method performs multiple passes through the same region, resulting in higher base accuracy (Logsdon et al., 2020).

SVs contribution to human disease

The use of long-read sequencing platforms has enhanced the study of the role of SVs in human biology with contributions to disease, evolution and population diversity (Ho et al., 2020).

A well-known example in the field of neurobiology is the gene *ARHGAP11B*. This gene araises from a partial duplication of its ancestral gene *ARHGAP11A* and plays a key role in neocortex expansion by amplification of basal progenitors (Heide et al., 2020). In addition, SVs have been associated with many diseases, including intellectual disability, autism spectrum disorders, congenital heart disease, schizophrenia and congenital limb malformation (Spielmann et al., 2018).

The genomes of the great apes are embedded in segmental duplications, defined as duplications of a DNA fragment of high homology along the genome. Recent sequencing studies have allowed to decipher the genomic regions prone to segmental duplications specific to human, which highlight 6 genomic hotspots prone to recurrent Copy Number Variation associated to developmental delay (Dennis et al., 2017). Owing to the new long-read sequencing technologies, such regions are now amenable to study and will enable us to gain insights into their function in the coming years.

1q21.1 syndrome

The region 1q21.1 is one of the six hotspot regions whose copy number variation is associated to developmental delay (Dennis et al., 2017). Pathological phenotypes under its deletion or duplication have been described multiple times and give rise to the 1q21.1 syndrome. Clinical reports on patients with 1q21.1 deletion present multiorganic symptoms, which include dysmorphic facial features, eye abnormalities, intellectual disability and developmental delay. In addition, patients with 1q21.1 duplication exhibit dysmorphic facial features, cardiac abnormalities, autistic features and developmental delay. Clinical reports have shown considerable phenotypic expression variability and incomplete penetrance (Bernier et al., 2016).

However, the striking phenotype that these patients share is the abnormal size of the head. While 1q21.1 deletion carriers are microcephalic (have reduced head size), 1q21.1 duplication patients exhibit macrocephaly (have increase in head size). In addition, deletion cases have been associated to schizophrenia, while duplication cases frequently present Autism Spectrum Disorder features (Bernier et al., 2016; Mefford et al., 2008). Due to the phenotype described above, 1q21.1 Copy Number Variants could be considered as “sister disorders” based on their tendency to exhibit diametrically opposite phenotypes and be caused by duplications versus deletions of the same genomic sequences (Crespi et al., 2009).

The 1q21.1 syndrome-associated phenotype is restricted to the distal part of 1q21.1 region (Mefford et al., 2008). Deletions in the proximal 1q21.1 area have been associated to the thrombocytopenia absent radius syndrome, characterized by reduction in the number of platelets and absence of the radius (Albers et al., 2012). Overall, 1q21.1 syndrome presents a specific phenotype that can be differentiated from other diseases that include the adjacent region.

The clinical description of this syndrome is extensive, but the genetic mechanisms that might cause this phenotype remain to be elucidated. Some studies highlight NOTCH2NL genes, found in the border of 1q21.1, as putative candidates to explain the micro- and macrocephalic traits. NOTCH2NL genes are highly expressed in radial glial cells and seem to play a role in human brain evolution by delaying cortical neuron differentiation in the early human lineage (Fiddes et al., 2018). However, the 1q21.1 region is highly enriched in genes with copy number variants that are putative candidates to cause 1q21.1 syndrome. Defining the genetic changes that produce this phenotype can help to understand the disease.

NBPF genes: putative cause of 1q21.1 syndrome

The 1q21 region is highly enriched in the Neuroblastoma Breakpoint Family (NBPF) of genes, which consist of 22 genes mainly distributed along chromosome 1. The name of the family arises from a neuroblastoma patient which contained a translocation in 1p36 which was truncating NBPF1 gene (Laureys et al., 1990). NBPF genes contain numerous repetitive elements and show high intergenic and intragenic sequence identity, both in coding and noncoding regions. These genes have suffered an exponential expansion from non-human primates to humans, indicating a possible role in the evolution of primates (Vandepoele et al., 2005).

One study suggested that NBPF genes have evolved jointly with NOTCH2NL genes and display a coordinated function (Fiddes et al., 2019). Despite that, the molecular function of NBPF genes has not been described yet. Some articles suggest high expression levels in fetal brain, fetal sympathetic ganglia and particularly in radial glial cells of the developing telencephalon (Diskin et al., 2009; Fiddes et al., 2019). Other articles describe the interaction between NBPF proteins and Chibby, a repressor of the Wnt signalling pathway (Vandepoele et al., 2010), tumor suppressive functions of NBPF1 protein (Andries et al., 2015) or even a possible roles as transcription factors (Zhou et al., 2013). Thus, there is no consensus of the function of NBPF genes.

Interestingly, the NBPF family of genes is characterized by the presence of double-exon tandem repeats, known as Olduvai domains (also referred as DUF1220 or NBPF domains). The name Olduvai refers to the Olduvai Gorge, a source of major anthropological discoveries in East Africa (Sikela & van Roy, 2018). Olduvai domains have been divided into six superclades: CON1/2/3 and HLS1/2/3, depending on their conservation within the primate order and the position order within each gene (O'Bleness et al., 2012). These domains have shown human-specific amplification during evolution, ranging from a single copy in mice, 4 copies in dolphins, 35 in macaques, 99 in gorillas, 125 in chimps up to 272 copies in humans (Dumas et al., 2012; Popesco et al., 2006). Some studies have shown that Olduvai copy number is linearly correlated with increasing cognitive function and brain volume (Davis et al., 2015). Concomitantly, experimental studies have shown an increase in human neural stem cell proliferation under NBPF genes transfection (Keeney et al., 2015). In short, Olduvai domain expansion represent a potential genetic mechanism to enhance volume brain expansion during evolution.

Likewise, the evolutionary source of Olduvai domains in NBPF genes is found in 1q21.2, inside the gene of PDE4DIP as a single copy. PDE4DIP gene comes from duplication of the ancestral CDK5RAP2, which is a known gene causing microcephaly (Wang et al., 2014). Overall, it seems that Olduvai domains inside of NBPF genes are good candidates in the study of 1q21.1 syndrome.

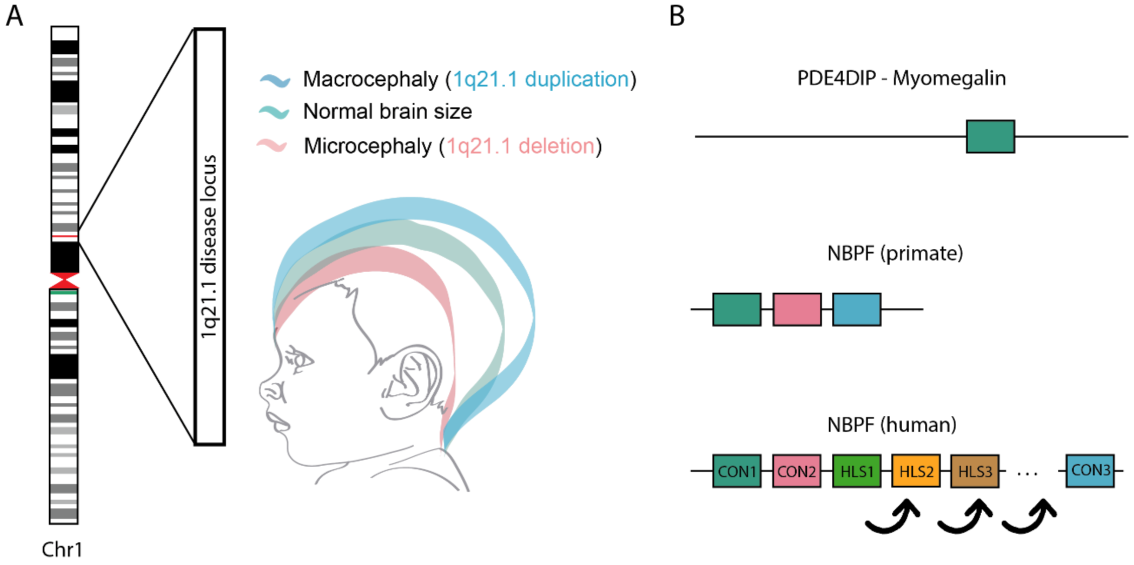


Figure 3. Locus 1q21.1. (A) Scheme representing 1q21.1 syndrome effect in head size. 1q21.1 deletions and duplications produce micro- and acrocephaly, respectively. (B) The protein myomegalin (PDE4DIP gene) contains a single ancestral Olduvai domain. NBPF genes contain tandem repeat amplification of Olduvai domains. The domains are extensively expanded in humans and can be divided into CON1/2/3 and HLS1/2/3 clades (O'Bleness et al., 2012).

Research proposal

The 1q21.1 region seems to be involved in human-specific brain expansion and constitutes an interesting topic for evolutionary, clinical and biological fields. Due to its repetitive nature, 1q21.1 has been difficult to sequence, and remained incorrectly mapped until the last reference genome assembly (GRCh38.p13) (Steinberg et al., 2014). As a result, the genes related to 1q21.1 syndrome have not been identified yet.

Hospital de Sant Joan de Déu counts for 8 patients diagnosed with 1q21 microdeletions and microcephalic phenotype, and 2 macrocephalic patients who exhibit 1q21 microduplications. They were diagnosed by a microarray chip, which does not allow to distinguish accurately which changes have been produced within the region. In order to identify the specific genes associated to the phenotype, deep sequencing of the patient genomes is required. 1q21 regions are not correctly assembled by Illumina technologies and require long-read strategies such as PacBio to efficiently map this region. Nowadays there is a project focused on 1q21.1 syndrome that is preparing the samples for PacBio sequencing.

In addition, the ancestral copy of Olduvai domains is found on the 1q21.2 region, placed in the PDE4DIP gene (Popesco et al., 2006). However, there is no evidence of sequence changes inside of the ancestral Olduvai domain among primates.

Objectives

The objectives of this project and their sub-objectives are described as follows:

1. To develop an analytic pipeline for PacBio data analysis of sample patients
    a. To simulate patient genomes with SV regions and generate derived PacBio reads
    b. To develop the bioinformatic pipeline for PacBio data analysis

2. To analyse the phylogeny of Olduvai domains among multiple species
    a. To align Olduvai sequences from different species
    b. To create a phylogenetic tree

The first objective aims to develop a pipeline of analysis for PacBio derived data in order to identify possible structural variants in the genome of the patients. Since the PacBio data is not available yet, we will generate simulated data that we can use to test our pipeline.

The second objective aims to assess the differences between Olduvai domains in multiple species. This analysis will be useful to understand the evolutionary function of Olduvai domains. For that purpose, we will aim to generate an alignment of the different proteins and the creation of phylogenetic trees.

Materials and Methods

Part 1: SV-calling pipeline

The creation of the genomes and the entire sequencing pipeline was performed with Google Cloud Compute Engine. A n1-highmem-4 Ubuntu, 20.04 LTS virtual machine was created, with 4 vCPUs and 26 GB memory. Software installation is described in (Sup. Information).

Simulation of Genomes with SVs

Genomes of simulated patients were induced by using the hg38 reference genome as a template. SVs include deletions, duplications, inversions and unbalanced translocations along chromosome 1. As an exception, unbalanced translocations were produced between chromosome 1 and another chromosome. This task was performed by SVGen. SVGen is an Open-Source software with an extensive range of options to simulate SVs of various types and to create derived short and long-reads [https://github.com/WGLab/SVGen]. Firstly, SVGen created a table file with 40 coordinates with the corresponding type of SV.

```
python simulate_SV_BED.py \
    --dup_lens SV_lengths.txt \
    --del_lens SV_lengths.txt \
    --inv_lens SV_lengths.txt \
    --unb_trans_lens SV_lengths.txt \
    --chroms 1 \
    --chrom_lens reference/chrom_lengths_hg38.txt \
    --gaps reference/gaps_hg38.txt \
    -o SV_simulation_indel_invtrans.bed
```

Box 1. SVGen command for generation of a table with SV coordinates and types of SVs.

```
counter=0;  cat  SV_simulation_indel_invtrans.bed  |  while  read  LINE;  do
((counter++)); echo $LINE > "patient$counter.bed"; done
```

Box 2. Separation of the different SV coordinates of each line in a separate file. The code creates 40 files, one for each patient.

```
for i in {1..40}; do python2 insert_SVs.py -i reference/hg38/chr1.fa -o
patient$i.fa --chrom_lens reference/chrom_lengths_hg38.txt --chrom 1 --bed
patient$i.bed -v; done
```

Box 3. Command to introduce SVs into the reference genome of each patient, generating 40 FASTA files.

Genomes visualisation

In order to ensure that the SVs were correctly inserted into the reference genome, the MUMmer software was applied for the visualisation of the sequences in Dot plots. MUMmer is an open-source software package for the rapid alignment of very large DNA and amino acid sequences. It contains a wide range of utilities for aligning sequences and creating plots (Kurtz et al., 2004). In the first place, we applied the function NUCmer to produce an alignment of each patient chr1 genome vs hg38 reference genome. Later, the function delta-filter was applied to exclude small non-specific alignments to the sequence. Finally, we applied mummerplot for generating the dot plot for each patient.

```
nohup bash -c 'for i in {1..40}; do ./nucmer --prefix patient1
/home/pebbmc/SVGen/reference/hg38/chr1.fa /home/pebbmc/SVGen/patient$i.fa;
done' &

nohup bash -c 'for i in {1..40}; do ./delta-filter -q -r patient$i.delta >
patient$i.filter; done' &

./mummerplot --prefix patientx -x "[coord1:coord2]" -y "[coord1:coord2]"
patientx.filter
```

Box 4. Command for patient sequence alignment against thè reference sequence. Posterior filtering and plotting.

In addition to mummerplot, an R script was also applied for the generation of the dot plots. The code has been adapted from R functions that are described in Hippocamplus, a blog with some useful Bioinformatic procedures.

[https://jmonlong.github.io/Hippocamplus/2017/09/19/mummerplots-with-ggplot2/]

```
x<-c("dplyr", "magrittr", "GenomicRanges", "knitr","ggplot2","tidyr")

library(x)
```

Box 5. R packages for dot plot generation.

```
readDelta <- function(deltafile){
  lines = scan(deltafile, 'a', sep='\n', quiet=TRUE)
  lines = lines[-1]
  lines.l = strsplit(lines, ' ')
  lines.len = lapply(lines.l, length) %>% as.numeric
  lines.l = lines.l[lines.len != 1]
  lines.len = lines.len[lines.len != 1]
  head.pos = which(lines.len == 4)
  head.id = rep(head.pos, c(head.pos[-1], length(lines.l)+1)-head.pos)
  mat = matrix(as.numeric(unlist(lines.l[lines.len==7])), 7)
  res = as.data.frame(t(mat[1:5,]))
  colnames(res) = c('rs','re','qs','qe','error')
  res$qid = unlist(lapply(lines.l[head.id[lines.len==7]], '[', 2))
  res$rid = unlist(lapply(lines.l[head.id[lines.len==7]], '[', 1)) %>%
gsub('^>', '', .)
  res$strand = ifelse(res$qe-res$qs > 0, '+', '-')
  res
}
mumgp = readDelta("mumplot-example.delta")
```

Box 6. Command function for reading the files.

```
filterMum <- function(df, minl=500, flanks=1e4){
    coord = df %>% filter(abs(re-rs)>minl) %>% group_by(qid, rid) %>%
        summarize(qsL=min(qs)-flanks, qeL=max(qe)+flanks, rs=median(rs)) %>%
        ungroup %>% arrange(desc(rs)) %>%
        mutate(qid=factor(qid, levels=unique(qid))) %>% select(-rs)
    merge(df, coord) %>% filter(qs>qsL, qe<qeL) %>%
        mutate(qid=factor(qid, levels=levels(coord$qid))) %>% select(-qsL, -
qeL)
}
mumgp.filt = filterMum(mumgp, minl=1e4)
```

Box 7. Filtering function for alignments smaller than 500 base pairs.

```
ggplot(mumgp.filt, aes(x=rs, xend=re, y=qs, yend=qe, colour=strand)) +

  geom_segment(size=2) + geom_point(alpha=10) +

  facet_grid(qid~., scales='free', space='free', switch='y') +

  theme_bw() +

  theme(strip.text.y=element_text(angle=180, size=5),
legend.position=c(.99,.01), legend.justification=c(1,0),
strip.background=element_blank(),axis.text.y=element_blank(),
axis.ticks.y=element_blank()) +

  xlab('Reference sequence') + ylab('Assembly') +

  scale_colour_brewer(palette='Set1') +

  coord_cartesian("xlim" = c(Axis), ylim = c(Axis))  +

  ggtitle("Title")
```

Box 8. Command of ggplot function for generating the dot plot. Both axis limits were set around the SV region in each patient.

Simulation of PacBio reads

Reads were created with wgsim software. Wgsim is a tool for simulating sequence reads from a reference genome. It can simulate diploid genomes with single nucleotide polymorfisms and insertion/deletion polymorphisms and simulate reads with uniform substitution sequencing errors [https://github.com/lh3/wgsim/]. Wgsim was applied to generate reads from the simulated genomes created before. The software required the following parameters: coverage (15), mean length (11 000 base pairs), genome size (243 000 000) and read number (331 364). The read number was retrieved by calculating the following formula:

Coverage = (read number * mean length) / genome size

Since the intention was to simulate reads from PacBio HiFi technology, we stablished read length as 11 000, error ratio as 0.01 (which accounts for the base call error in HiFi) and mutation ratio 0.001 (which creates mutations in the reads).

```
nohup bash -c 'for i in {1..40}; do ./wgsim -N331364 -111000 -d0 -S11 - e0.01
-r0.001 /home/pebbmc/SVGen/patient$i.fa patient$i.fq /dev/null; done' &
```

Box 9. Command for generation of wgsim reads.

The reads were posteriorly analysed in terms of number, length and quality. The number of reads in each file was obtained by counting the number of lines of each FASTQ file. Since every read takes 4 lines, the number of lines obtained was divided into 4. The read length distribution was obtained by counting the read length in each read of the files of the patients and creating a table of length frequencies. The distribution was visualized by an R plot.

```
for i in {1..40}; do cat patient$i.fq | wc -l; done
```

Box 10. Command to retrieve the number of lines in each FASTQ file.

```
cat patient1.fq | awk '{if(NR%4==2) print length($1)}' | sort -n | uniq -c >
read_length1.txt
```

Box 11. Command for obtention of read length distribution

```
reads<-read.csv(file="read_length1.txt", sep="", header=FALSE) library(ggplot2)

ggplot(reads, aes(V2, V1)) + geom_bar(stat="identity") + ggtitle("Read Length
distribution") + coord_cartesian(xlim = c(10990, 11010)) + xlab("Read length")
+ ylab("Number of reads")
```

Box 12. Command for read length distribution plot in R.

Read Mapping

Long read strategies are known to produce the lowest amount of contigs after assembly, which facilitates their annotation and analysis. In our case, we applied hiCanu for assembly of the reads. Canu is a fork of the Celera Assembler, designed for high noise single molecule sequencing. hiCanu is a part of this software that is applied specifically to HiFi reads (Nurk et al., 2020).

```
nohup bash -c './canu -p asm -d patient1_hifi genomeSize=243m -pacbiohifi
/home/pebbmc/wgsim/patient1.fq' &
```

Box 13. Canu command for read assembly.

However, this option was computationally expensive for the analysis of 40 patients, given our limited memory resources. An alternative option to the assembly is to map the reads to a reference genome with pbmm2. pbmm2 is a SMRT C++ wrapper for minimap2's C API. Its purpose is to support native PacBio in- and output, provide sets of recommended parameters, generate sorted output on-the-fly, and postprocess alignments (Biosciences, n.d.). The use of this software included two steps: the creation of an index for the reference genome to use and read alignment against the reference.

```
./pbmm2 index /home/pebbmc/SVGen/reference/hg38/chr1.fa chr1.mmi
```

Box 14. Command for alignment indexing.

```
nohup bash -c 'for i in {1..40}; do ./pbmm2 align chr1.mmi --sort
/home/pebbmc/wgsim/patient$i.fq pbmm2.patient$i.bam; done' &
```

Box 15. Command for mapping procedure.

Call of Structural Variants

Read alignment was followed by call of SVs in the sequence. Pbsv is a suite of tools to call and analyse structural variants in haploid and diploid genomes from PacBio single molecule real-time sequencing (SMRT) reads. The algorithm pbsv calls insertions, deletions, inversions, duplications, and translocations.

[https://github.com/PacificBiosciences/pbsv]

Pbsv was applied to the BAM file alignments for SV identification. Pbsv workflow included the discovery of signatures of structural variation, call of structural variants and assignment of genotypes.

```
nohup bash -c 'for i in {1..40}; do ./pbsv discover pbmm2.patient$i.bam
patient$i.svsig.gz; done' &
```

Box 16. Command for signature discovery.

```
nohup bash -c 'for i in {1..40}; do ./pbsv call
/home/pebbmc/SVGen/reference/hg38/chr1.fa patient$i.svsig.gz patient$i.vcf;
done' &
```

Box 17. Command for structural variant call.

Pbsv produced a .VCF file with the coordinates of SVs detected. Imprecise variants were excluded from the list for better interpretation and saved in a .txt file and .vcf file.

```
nohup bash -c 'for i in {1..40}; do sed “/IMPRECISE;/d” patient$i.vcf >
patient_noimp$i.txt; done' &

nohup bash -c 'for i in {1..40}; do sed “/IMPRECISE;/d” patient$i.vcf >
patient_noimp$i.vcf; done' &
```

Box 18. Commands for VCF file filtering.

SVs visualisation

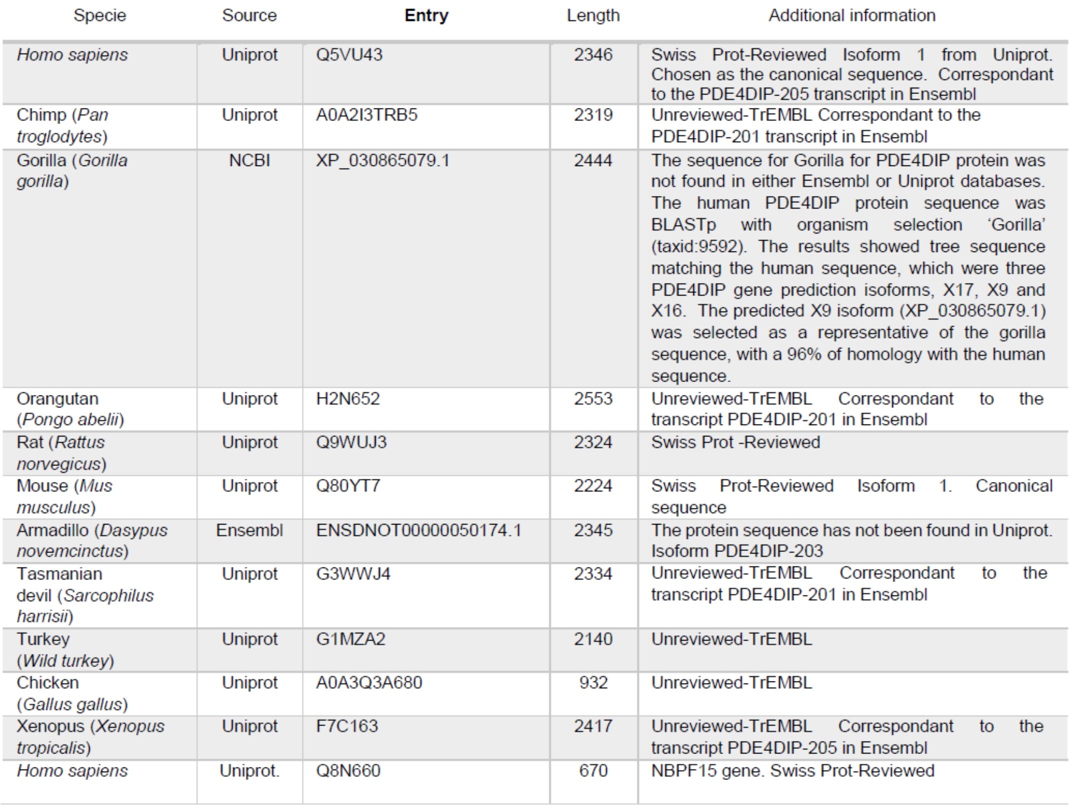
.BAM, .BAM.BAI (index) and .VCF filtered files were exported for visualisation. SV performance was checked in Integrative Genomics Viewer (IGV). IGV is a high-performance, interactive tool for the visual exploration of genomic data (Robinson et al., 2011). From IGV platform [https://igv.org/app/], Human GRCh38/hg38 genome was selected for visualisation, followed by chr1 option. The three previously exported files were loaded for each patient. Chromosome coordinates were adjusted to the SVs detected in the .VCF file.

Part 2: Olduvai domain phylogeny

Selected species and sequence retrieval

The selected organisms in this study were species that belong to the same phylum (chrodata) and class (Mammalia). There were also three outgroup species. The outgroup is defined as a distant group of organisms that serves as a point of comparison for the ingroup and specifically allows for the phylogeny to be rooted. While the outgrup species belonged to the same phylum, they were part of a different class (Amphibia and Aves).

The Olduvai domain of each specie was retrieved from Uniprot (Consortium, 2019) by searching the label "Families and Domains" of the corresponding myomegalin protein (Table 1). We also seek to include the Neanderthal sequence of the Olduvai domain in myomegalin from The Neanderthal Genome Project, based at the department of Evolutionary Genetics at the Max Planck Institute for Evolutionary Anthropology in collaboration with the Neandertal Genome Consortium (Prüfer et al., 2014). However, the Neanderthal genome was finally excluded from the analysis, for two reasons: (1) The Neanderthal Genome exhibits low sequencing coverage and (2) DNA or protein consensus sequences are not available to export, but only mapped reads are available to display.

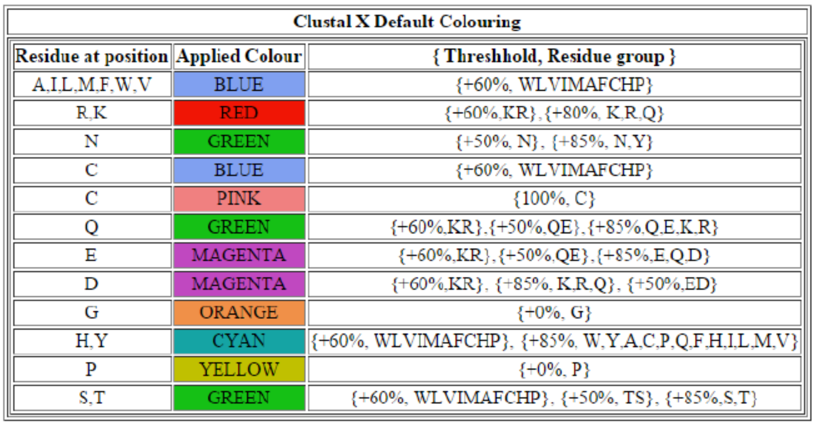


Specie	Source	**Entry**	Length	Additional information
Homo sapiens	Uniprot	Q5VU43	2346	Swiss Prot-Reviewed Isoform 1 from Uniprot. Chosen as the canonical sequence. Correspondant to the PDE4DIP-205 transcript in Ensembl
Chimp (*Pan troglodytes*)	Uniprot	A0A2I3TRB5	2319	Unreviewed-TrEMBL Correspondant to the PDE4DIP-201 transcript in Ensembl
Gorilla (*Gorilla gorilla*)	NCBI	XP_030865079.1	2444	The sequence for Gorilla for PDE4DIP protein was not found in either Ensembl or Uniprot databases. The human PDE4DIP protein sequence was BLASTp with organism selection 'Gorilla' (taxid:9592). The results showed tree sequence matching the human sequence, which were three PDE4DIP gene prediction isoforms, X17, X9 and X16. The predicted X9 isoform (XP_030865079.1) was selected as a representative of the gorilla sequence, with a 96% of homology with the human sequence.
Orangutan (*Pongo abelii*)	Uniprot	H2N652	2553	Unreviewed-TrEMBL Correspondant to the transcript PDE4DIP-201 in Ensembl
Rat (*Rattus norvegicus*)	Uniprot	Q9WUJ3	2324	Swiss Prot -Reviewed
Mouse (*Mus musculus*)	Uniprot	Q80YT7	2224	Swiss Prot-Reviewed Isoform 1. Canonical sequence
Armadillo (*Dasypus novemcinctus*)	Ensembl	ENSDNOT00000050174.1	2345	The protein sequence has not been found in Uniprot. Isoform PDE4DIP-203
Tasmanian devil (*Sarcophilus harrisii*)	Uniprot	G3WWJ4	2334	Unreviewed-TrEMBL Correspondant to the transcript PDE4DIP-201 in Ensembl
Turkey (*Wild turkey*)	Uniprot	G1MZA2	2140	Unreviewed-TrEMBL
Chicken (*Gallus gallus*)	Uniprot	A0A3Q3A680	932	Unreviewed-TrEMBL
Xenopus (*Xenopus tropicalis*)	Uniprot	F7C163	2417	Unreviewed-TrEMBL Correspondant to the transcript PDE4DIP-205 in Ensembl
Homo sapiens	Uniprot.	Q8N660	670	NBPF15 gene. Swiss Prot-Reviewed

Table 2. List of species used in the phylogenetic analysis.

MUSCLE Alignment

MUSCLE (MUltiple Sequence Comparison by Log- Expectation) alignment software was used for the alignment of the sequences [https://www.ebi.ac.uk/Tools/msa/muscle/]. The sequences were uploaded in a .txt file. Two output formats were selected: ClustalW and Phylip Interleaved. Sequence alignments were visualized and edited in Jalview (2.11.1.2) with Clustalx Default Coloring, which represents aminoacidic residues depending on their chemical properties (Table 2). Full myomegalin alignment includes annotations of the Centrosome interaction domain and the Olduvai domain, based on the human sequence.



Clustal X Default Colouring		
Residue at position	Applied Colour	{ Threshhold, Residue group }
A,I,L,M,F,W,V	BLUE	{+60%, WLVIMAFCHP}
R,K	RED	{+60%,KR},{+80%, K,R,Q}
N	GREEN	{+50%, N}, {+85%, N,Y}
C	BLUE	{+60%, WLVIMAFCHP}
C	PINK	{100%, C}
Q	GREEN	{+60%,KR},{+50%,QE},{+85%,Q,E,K,R}
E	MAGENTA	{+60%,KR},{+50%,QE},{+85%,E,Q,D}
D	MAGENTA	{+60%,KR}, {+85%, K,R,Q}, {+50%,ED}
G	ORANGE	{+0%, G}
H,Y	CYAN	{+60%, WLVIMAFCHP}, {+85%, W,Y,A,C,P,Q,F,H,I,L,M,V}
P	YELLOW	{+0%, P}
S,T	GREEN	{+60%, WLVIMAFCHP}, {+50%, TS}, {+85%,S,T}

Table 3. Clustal X Default Coloring format for sequence alignments.

Phylogenetic tree

Phylogenetic trees of the alignments were performed with PhyML (Guindon et al., 2010). The input sequence was the Phylip Interleaved format of the alignment. SMS was chosen as the automatic model selection (Lefort et al., 2017), with AIC (Akaike Information Criterion) selection criterion. Tree searching was done under BIONJ starting tree and Nearest Neighbour Interchange (NNI) type of tree improvement. aLRT-SH-like fast likelihood-based method was used. To determine support values, aLRT-SH was changed into 100 times bootstrap. Tree images were exported from PRESTO (a Phylogenetic tReE viSualisaTiOn).

Results

Part 1: SV-calling pipeline

SVGen can simulate SVs genomes

SVGen was successful in the generation of the genomes with SVs. The output files of SVGen were 40 FASTA files with chr1 genome of each patient, plus a BED file with the SVs coordinates (Table 3).

Patient	Coordinates	Type of SV	Unb. Translocation
1	13611807 - 13614306	Duplication	
2	229323817 – 229327316	Unb. Translocation	4:31945676-31949175
3	143853492 – 143854991	Inversion	
4	30600313 – 31100312	Unb. Translocation	3:91640438-92140437
5	13699584 – 13704583	Unb. Translocation	Y:56907729-56912728
6	31211052 – 31241051	Inversion	
7	31341265 – 31347264	Unb. Translocation	19:25010973-25016972
8	31444261 – 31494260	Unb. Translocation	3:90867961-90917960
9	17839596 – 17843595	Deletion	
10	123028646 – 123029645	Unb. Translocation	20:30927514-30928513
11	225318940 – 225319939	Inversion	
12	17940076 – 18140075	Inversion	
13	229422854 - 229424853	Unb. Translocation	19:24656025-24658024
14	18228010 - 18258009	Duplication	
15	1099819 - 1105818	Deletion	
16	123137860 – 123139359	Unb. Translocation	9:41322046-41323545
17	1202671 - 1222670	Unb. Translocation	5:155876336-155896335
18	3116949 – 3117948	Deletion	
19	1329795 1349794	Duplication	
20	225412394 – 225417393	Deletion	
21	1452820 – 1552819	Deletion	
22	229532879 – 229536378	Deletion	
23	143958073 - 143998072	Deletion	
24	1655169 - 1730168	Duplication	
25	31583959 - 31603958	Inversion	
26	13800496 - 13808495	Deletion	
27	18369150 - 18379149	Inversion	
28	31685757 - 31725756	Inversion	
29	18452944 - 18652943	Duplication	

30	18757769 - 18767768	Deletion	
31	13929996 - 13930995	Duplication	
32	123231605 - 123381604	Duplication	
33	3228275 - 3234274	Inversion	
34	229638360 - 229641859	Inversion	
35	229734871 - 229784870	Duplication	
36	225507520 – 226007519	Duplication	
37	123476031 – 123480030	Unb. Translocation	19:24756223-24760222
38	229889314 – 229964313	Unb. Translocation	21:13062203-13137202
39	3338099 - 3341098	Inversion	
40	31825821 - 31830820	Inversion	

Table 4. List of SV in each patient. The table shows the coordinates and the type of SVs that was generated in each patient. The fourth column shows the coordinates where the chr1 fragment has translocated, which is always in another chromosome.

Dot plots were firstly created by mummerplot (Sup. Figure 1). An R script allowed a better visualisation, with personalized representation settings. The generated SVs are different in type, region and size, which adds variation to the samples. SVs size ranged from 1000 to 500 000 base pairs.

As depicted below, deletions, duplications and inversions along chr1 were introduced into the genomes successfully (Figures 4-6, Sup. Figures 2-4). Unbalanced translocations are shown as deletions in the dot plots since the altered fragments are found in another chromosome. While this “deletion” is found in most of the patients, not all of them display this genotype (Figure 7, Sup. Figure 5). This suggests that SVGen is not always useful in the generation of translocations.

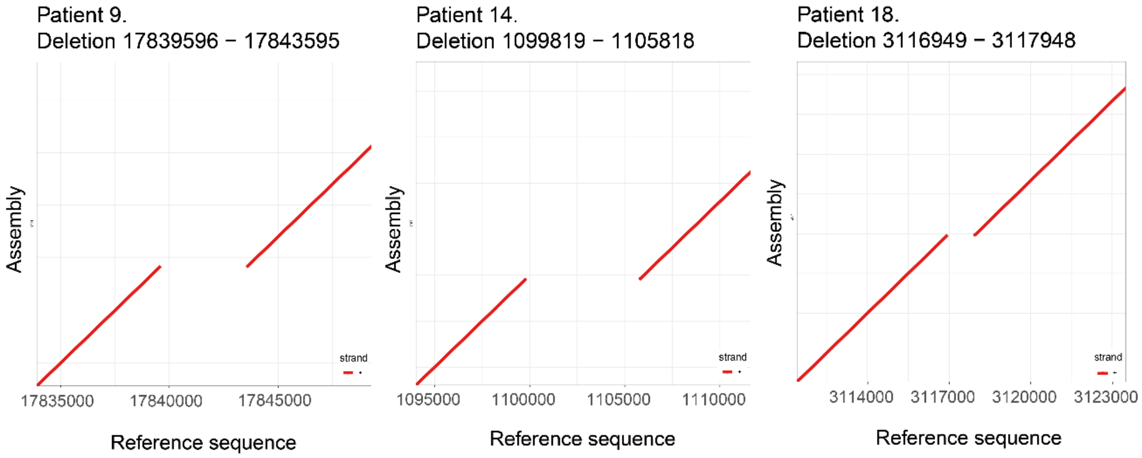


Figure 4. Dot plots of patient deletions. The plots show the lack of a region in the area of the SV.

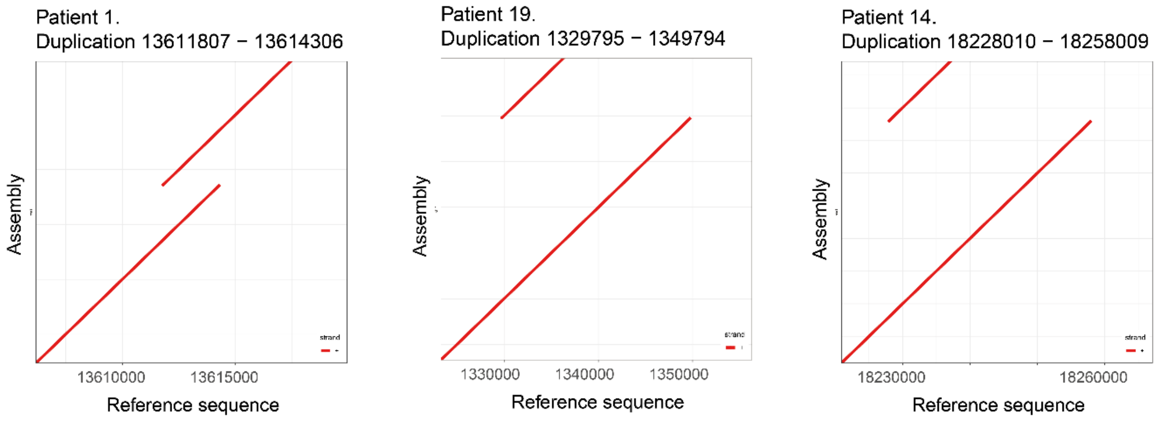


Figure 5. Dot plots of patient duplications. The plots show a double line in the area of the SV, indicating a duplication.

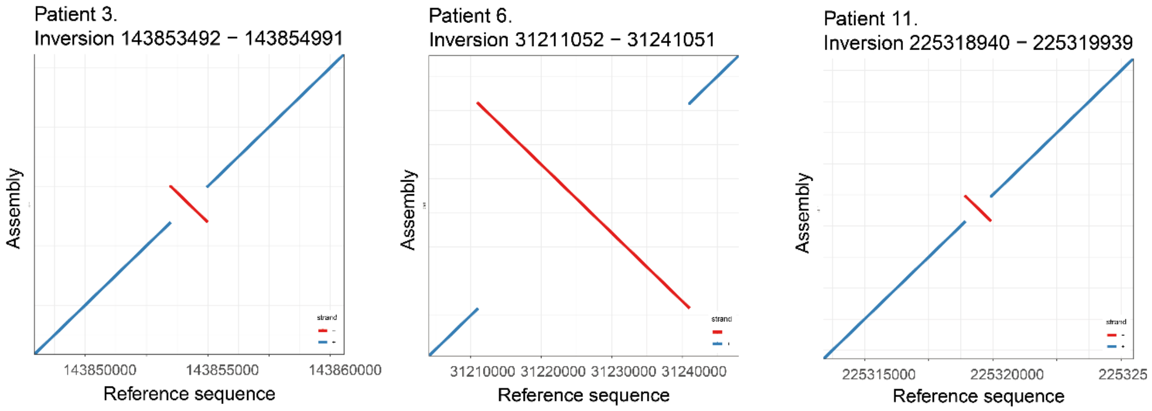


Figure 6. Dot plots of patient inversions. The plots show the inversion of the region in the area of the SV.

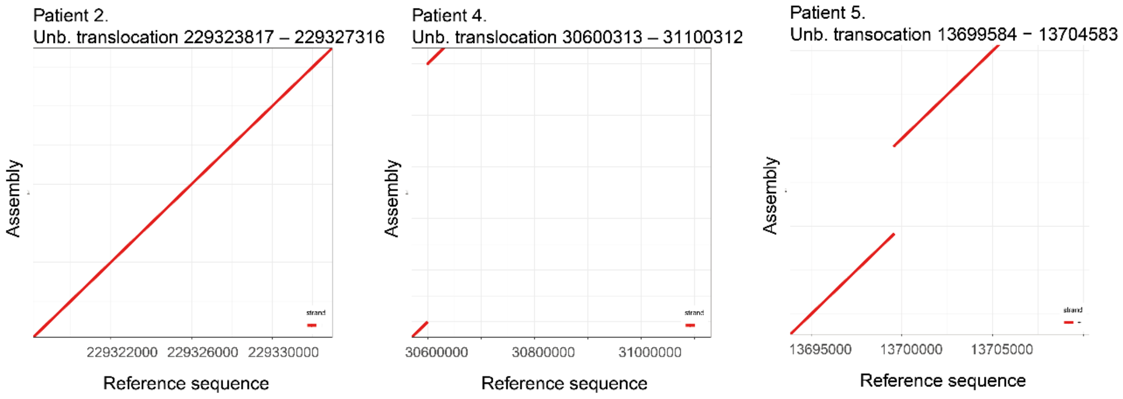


Figure 7. Dot plots of patient unbalanced translocations. The plots show deletions in specific regions in the translocated region. However, not all the translocations have been successful.

Wgsim is a useful tool for read simulation

Our first approach for generating the reads from the simulated patients was to use the software SVGen, already applied in the previous part. Nonetheless, the application of its function ‘create_reads.py resulted in a code error which seems not to allow the script for the generation of long reads (Sup. Information, Sup.Figure 6).

As an alternative, the generation of the reads was done by wgsim software. The output is a FASTQ file for each patient. FASTQ files consist of 4 lines:

1. Sequence identifier with information about the sequencing run and the cluster.
2. Sequence. Base calls A, C, T, G and N.
3. A separator, which is a plus sign (+).
4. The base call quality scores. These scores are encoded using ASCII characters to represent numerical quality scores.

```
@chr1_62148042_62159041_1:0:0_107:2:0_0/1

GGTTTATTTTTTGAGACTGAGTCTCACTCTGTTGCCCAGGGGAGTACAGTGA

+

55555555555555555555555555555555555
```

Figure 8. Example of FASTQ format.

Wgsim was successful for generating sequencing reads from the simulated genomes. Each FASTQ file contained 331364 reads, which was our input parameter. All the reads were exactly 11 000 bp (Figure 9), which was also our desired read length. However, in real sequencing experiments the read length is not a specific number, but a distribution of different lengths that converge to a specific mean. From this result we can observe that wgsim is not able to generate such distribution, which is a limitation of this software.

In addition to the read number and length, the quality of the reads was also checked. The quality score for each sequence is a string of characters, one for each base of the nucleic sequence, used to characterize the probability of misidentification of each base. In FASTQ files, quality scores are encoded into a compact form, which uses only 1 byte per quality value, representing Phred quality scores. Quality scales are different depending on the sequencing technology applied, which hinders the process of read simulation. In our samples, it was observed that the quality scores are always the number 5, which is the ASCII value that represent PHRED quality score 20. A PHRED quality score of 20 represents a base call accuracy of 99%, consistent with our input parameter. While this solution is useful for read generation, it is unrealistic to obtain constant quality values.

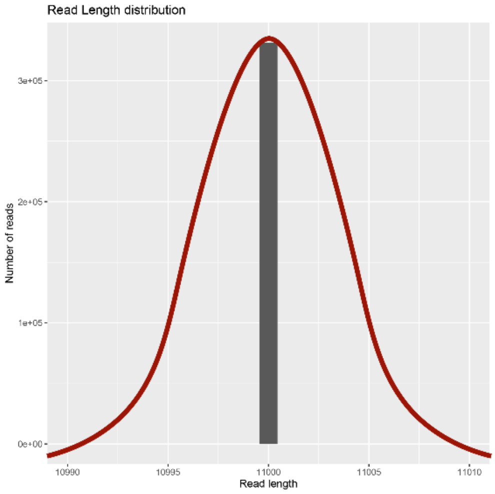


Figure 9. Length distribution of wgsim reads. Wgsim is only able to generate reads from a specific length (black bar). Instead, real sequencing experiments result in a normal distribution of the reads towards a specific lead length (red line).

Read Mapping and Structural Variant call

Once reads were generated and stored in FASTQ files, we sought to detect the embedded SVs. For that purpose, hiCanu was firstly applied to assemble the reads of one genome. However, the process was computationally expensive for 40 patients. The alternative strategy was to map the reads generated for each patient against the reference genome. Mapping was successfully produced with pbmm2. The output of this process for each patient is one .bam file that contains the information regarding the alignment and one .bam.bai file which corresponds to the index file.

Afterwards, pbsv was applied to detect the different SVs in the mapping. The output file of pbsv is a .vcf file for each patient with numerous information about the different SVs that have been detected. Filtered .vcf files can be found in (Supp Files).

As we can see in (Supp. Files 1), our pipeline is able to accurately detect the introduced SVs in the case of deletions, duplications and inversions. The recall is very high, with 28 out of 29 SVs generated were detected by pbsv after filtering of imprecise alterations, including deletions, duplications and inversions. However, the precision is not that high, with 26 out of 55 events approximately. As expected, unbalanced translocations, which were not correctly introduced in the genomes, have not been detected by the software.

In agreement with these results, the SVs detected by pbsv are also found by IGV visualisation. Read coverage in .BAM alignment is zero in areas of deletion, while the reads exhibit double coverage in duplicated regions, compared to the mean coverage. On the contrary, duplicated regions display a double read coverage area. Finally, inversions show a pattern of no read alignment, accompanied by average levels of read coverage, suggesting that the genetic material is present but not aligned to the strand.

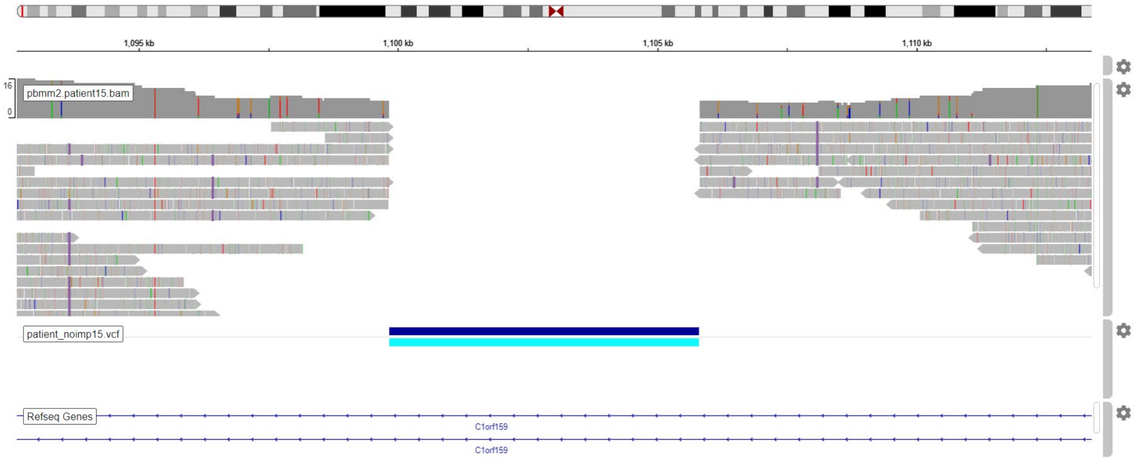


Figure 10. IGV in patient 15. Patient 15 deletion is detected by pbsv. The SV region shows no reads aligned to the region.

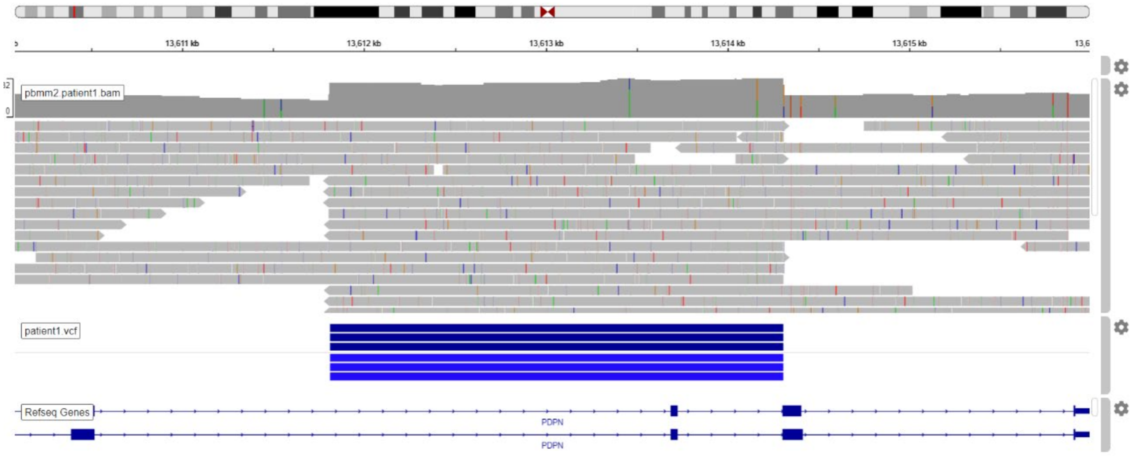


Figure 11. IGV in patient 1. Patient 1 duplication is detected by pbsv. The SV region shows double increase in read coverage.

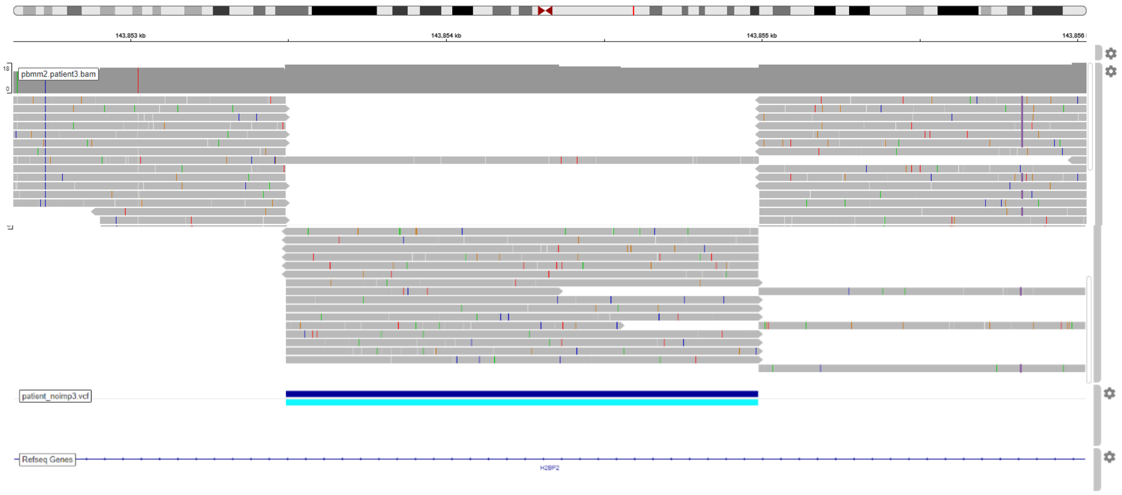


Figure 12. IGV in patient 3. Patient 3 inversion is detected by pbsv. The SV region shows that the coverage is maintained and there are two differentiated tracks of reads.

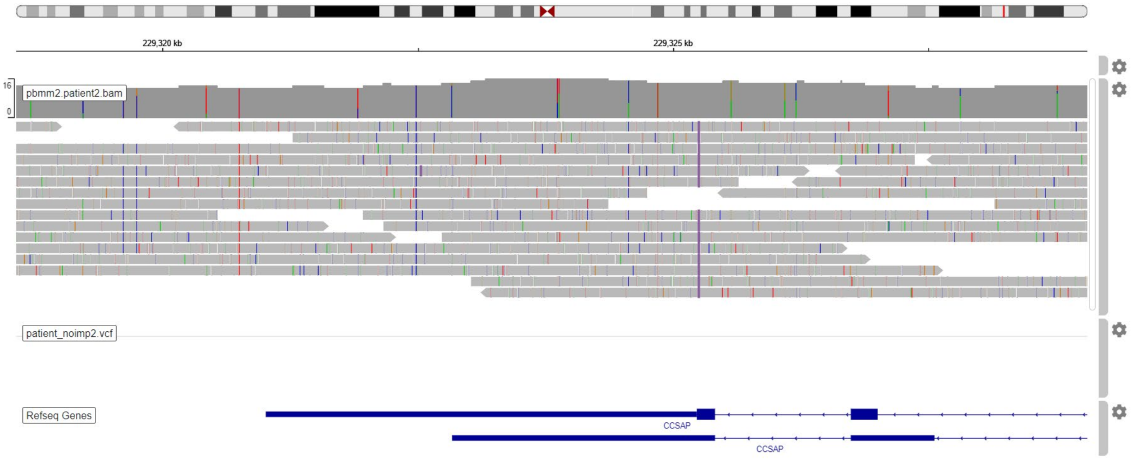


Figure 13. IGV in patient 2. Patient 2 has failed to generate an unbalanced translocation. Concomitant to this, the region shows no SV.

Part 2: Olduvai domain phylogeny

Full myomegalin alignment

Complete myomegalin alignment was performed, aiming to have a suitable comparison between the conservation of the Olduvai domain and the complete protein. Myomegalin protein constitutes a big protein, producing more difficulties to generate alignments of the full sequences. In addition, multiple isoforms have been described. However, for our studies we focus on the canonical amino acid sequence. The complete alignment is found in Myomegalin_Jalview3.pdf file.

From the alignment of full myomegalin, we can firstly describe that the initial part of the sequence of chicken is truncated and firstly appears in the human exon 25. The truncation might be due to a poor annotation of the sequence or a real biological truncation of the protein in chicken. In addition, the sequence from xenopus is larger than any other sequences and tends to insert multiple gaps in the alignment.

Before the start of the human Exon 1, orangutan, gorilla and tasmanian devil exhibit a similar precursor sequence that could be part of a primitive part of myomegalin which has been lost in the human.

The protein displays a precise alignment with gorilla and orangutan. Armadillo, tasmanian devil, mouse and rat sequences present high similarity, including some small gaps at some points of the protein. The chimp, which constitutes the nearest primate to the human in this alignment, shows multiple gaps and discordances to the human sequence in the alignment.

In addition, the outgroup (Turkey, Chicken, Xenopus) presents a higher number of differences from the human sequence. Besides the differences, there are multiple exons with highly conserved regions.

Concomitant to this data, phylogenetic trees of myomegalin show a progressive conservation of the protein sequence among species, where the closest species to human (chimp, gorilla) are placed immediately next to the human lineage; and further species are more distant to the human. The outgroup (chicken, turkey, xenopus) has been placed in the furthest place from the human sequence in our trees. However, the outgroup is not situated in a separate branch from the mammal lineage (as we would primarily expect) probably due to the high conservation of the protein among the entire phylum.

In conclusion, it seems that there is a conserved central core in myomegalin protein that can be found in a wide variety of species, and other regions that have diverged along evolution.

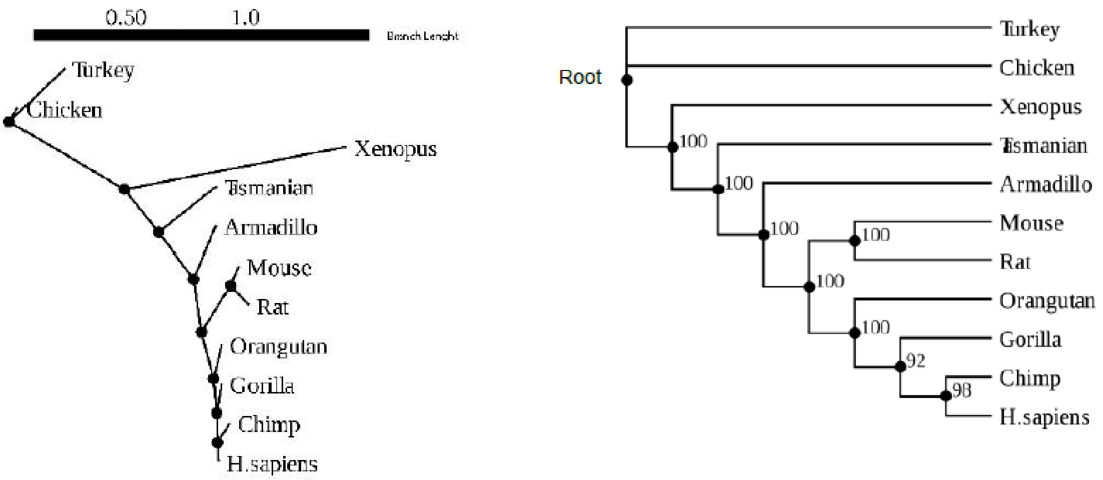


Figure 14. Phylogenetic Tree of Myomegalin sequences. (A) Slanted Phylogram. Branch length represents evolutionary time between two nodes. Unit=substitutions per sequence site (B) Linear Dendogram. Support values at each node describe the number of bootstrap support for 100 analyses, which means the number of times that this branch has been created under the same divisions in 100 trees.

Olduvai domain alignment

Olduvai domains from eleven species were aligned (Figure 15). Non-human primate sequences orangutan, chimp and gorilla exhibited high similarity with the human sequence. Mouse and rat are also really similar to the human sequence, including few amino acidic changes that are common in both species, which is indicative of evolutive proximity between them. Still in mammals, the sequences from armadillo and tasmanian devil show high similarity with the human sequence. However, the tasmanian devil contains a small gap in the initial part of the sequence.

In summary, the alignment follows gradual conservation of Olduvai domains, ranging from a 100% of conservation of Olduvai domain in chimp, an almost perfect alignment in other apes (Gorilla and Orangutan) and big conservation in mammals such as the Armadillo, Mouse and Rat. In the outgroup, turkey, chicken and xenopus lack some parts of the domain which correspond to the beginning and the end. In short, Olduvai domain in myomegalin highly conserved among species, with minor changes produced during evolution.

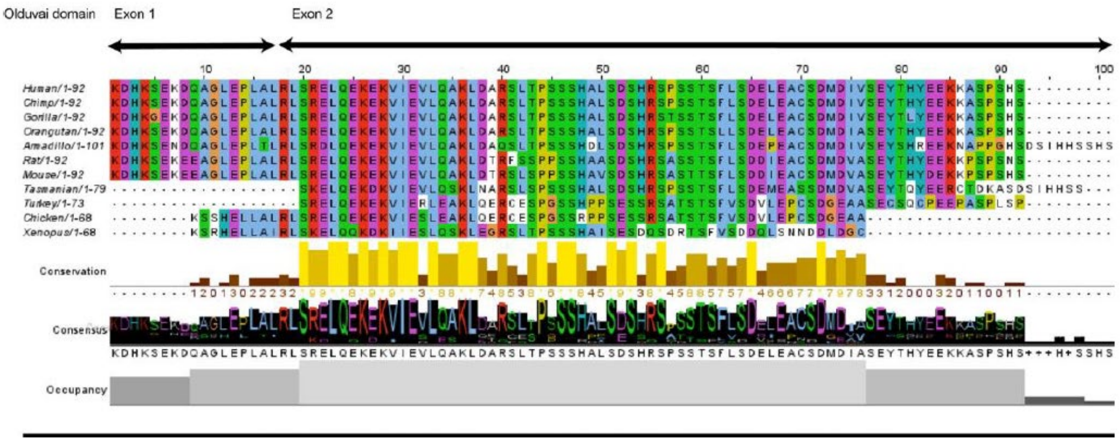


Figure 15. Alignment of Olduvai domains in multiple species. The conservation and quality tracks grade the number of conservation tracks of the alignment in less than 25% of gaps in each column. The quality of the alignment is based on the Blosum62. The consensus shows the % of identity of each amino acid in each column. The occupancy measures the number of aligned positions.

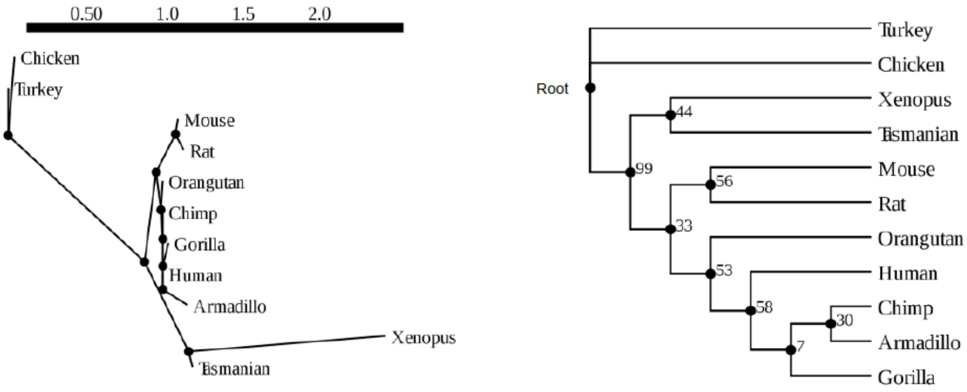


Figure 16. Phylogenetic tree of Olduvai domain sequences from myomegalin in multiple species. (A) Slanted Phylogram with branch lengths. (B) Linear Dendogram with support values.

NBPF15 gene alignment of Olduvai domains

In the previous analysis we have studied the evolution of Olduvai domain sequence from myomegalin in different species. However, another interesting idea is to know how the Olduvai domain has evolved in NBPF genes. To elucidate this, we performed an alignment between the 6 different domains of NBPF15 gene and the ancestral myomegalin Olduvai domain.

As seen in Figure 17, NBPF15 Olduvai domains show high divergence from their ancestral sequence, with only a few amino acids matching the sequence. In addition, Olduvai domains show higher conservation among NBPF15 gene.

CON1/2/3 domains (Olduvai 1/2/6 in NBPF15) are known to be more conserved into the mammal lineage (O'Bleness et al., 2012). These domains seem to be really similar between them in NBPF15, and slightly different from the ancestral Olduvai.

In the other case, HLS1/2/3 domains (Olduvai 3/4/5 in NBPF15) are human-specific domains. They are also very similar between them and they are more distant to the ancestral Olduvai. However, there are not many differences between HLS and CON clades.

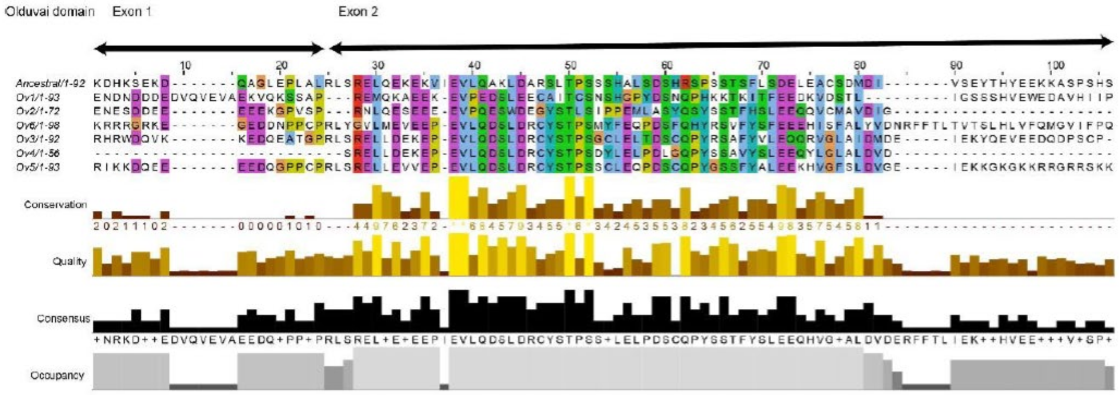


Figure 17. NBPF15 Olduvai domains aligned to the ancestral Olduvai domain in myomegalin.

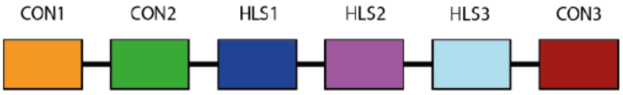


Figure 18. Order of Olduvai domains in NBPF15 gene.

The phylogenetic tree (Figure 19) shows that Olduvai domains 1 and 2 are more similar to the ancestral sequence. On the other hand, Olduvai domains 3, 4, 5 and 6 have evolved in a similar manner. Support values are pretty high, indicative of the consistency of the tree.

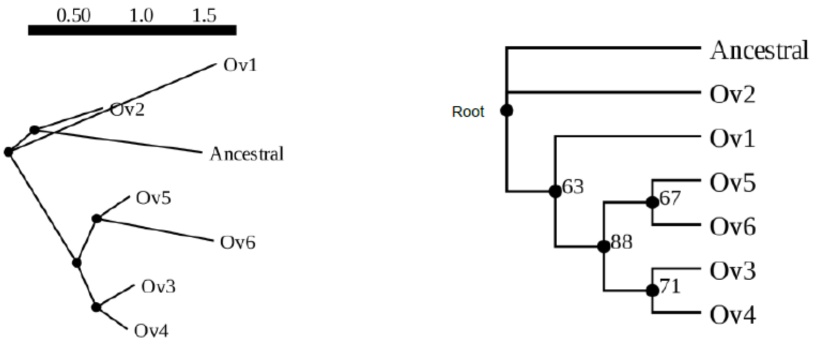


Figure 19. Phylogenetic tree from Olduvai domains in NBPF15 genes and the ancestral Olduvai domain in myomegalin. (A) Slanted Phylogram with branch lengths. (B) Linear Dendogram with support values.

Discussion

Simulation of Structural Variants and derived reads

The development of sequencing analytic tools frequently lacks benchmark methods that can validate its correct function. Although there are some real sequencing datasets available, they are scarce and do not contain concrete variants of research interest. Thus, genome simulators have been improving in the past years to produce sequences that can resemble real genomes to the fullest (Stephens et al., 2016). In the context of this project, SVGen was a successful method to generate different types of SVs, except for unbalanced translocations. The simulation included a wide range of variants in size, type and location. One of the limitations of SVGen is the inability to generate diploid genomes, which approximate better to the human genome and add complexity to the system. Diploid genomes could be applied to the study of heterozygous disease or allele identification procedures. Another improvement that could be done in the simulation is to incorporate the option to personalize structural variant coordinates that might fit better the experimental questions. For instance, SVEngine software (Xia et al., 2018) effectively implements these two options and it would be a good alternative to our strategy. Since there is a wide range of simulation tools, the choice of the best technique can be adjusted to the experimental question.

Likewise, read simulators have also expanded exponentially in the last few years. Some of them include tools for generating both SVs and simulate the reads, such as SVGen, which enables to produce experimental datasets easily. Our strategy firstly included read simulation with SVGen, which was not possible. The study finally implemented wgsim [https://github.com/lh3/wgsim], which was able to generate reads from the input genomes with structural variants. Wgsim has an extensive range of settings that can adapt the reads to a specific coverage, read length, base call error rates and rate of mutations. The resulting reads were useful for the study but had two major limitations: the read length distribution and the quality scores. On the one hand, PacBio HiFi reads exhibit a particular read length distribution that approximates to a normal density curve (Logsdon et al., 2020; Ono et al., 2013), which was not possible to simulate with wgsim. On the other hand, fastq files included a unique Phred quality score of 20. Typically, quality base call values are calculated by Sanger format and encode Phred quality scores, which typically range between 0 and 40 (Sathyanarayanan et al., 2019). In fastq files, quality scores are represented by ASCII characters 33-126. HiFi reads also show a decreasing quality score along the read, with lower values at the end of the sequences (Ono et al., 2013). In our case, simulated reads with Phred quality score of 20 are a good approximation for HiFi reads, with high base call accuracy (around 99%), but a continuous and unaltered Phred quality score along the whole sequence. Alternative methods for a better approximation would be the software PaSS (Zhang et al., 2019) or PBSIM (Ono et al., 2013), which produce read simulation based on real datasets.

Alternatives to genome and read simulation could be the use of public datasets of PacBio HiFi reads. In the literature there are some helpful databases using this technology (Kronenberg et al., 2018; Miga et al., 2020; Wenger et al., 2019). Even though these datasets are limited to study specific genomic changes, they can be used for the design of pipelines of sequencing data in general.

Methods for genome visualisation

Dot plots have been a practical method to visualize the structural variants produced. They are optimal for verification of deletions, duplications, inversions, translocations and for tandem repeats (Cabanettes & Klopp, 2018). A limitation of this graph is the range of genomic coordinates that can be plotted. The structural variants generated in this project were small compared to the huge size of chromosome 1. Therefore, the dot plots of the entire chromosome could not allow to visualize any alteration and needed to be zoomed in for detecting the structural variants. In conclusion, dot plots are a suitable tool only for visualisation, but not for SVs detection purposes.

Moreover, Integrative Genome Viewer (Robinson et al., 2011) offers multiple tools for sequencing data visualisation from a simple online interface. In our project, it has been crucial for visual verification of the pbsv results in an easy manner, allowing us to better understand the changes produced in the genomes and to communicate the results in a graphical design.

Read assembly followed by read mapping is the optimal methodology for SV detection

Long-read sequencing reads are usually corrected, trimmed and assembled into contigs, which are continuous sequences produced by the overlap of multiple reads (Koren et al., 2017). The software that stand for this function in PacBio HiFi experiments are FALCON (Chin et al., 2016) and Canu (Hon et al., 2020). In this project we applied hiCanu, a modification of Canu assembler designed specifically for HiFi reads. We failed in the assembly of the reads by using this method, since the assembly for one chr1 genome was too large. Previous implementation of hiCanu in the literature describe the use of

131 CPU hours for assembly of 150m genome in a 16GB environment. By approximating this data to our assembly, we would obtain 134 CPU hours for the assembly of one patient genome (243m genome and 26 GB environment), which is unsustainable for the assembly of 40 genomes.

Therefore, our strategy included mapping of the reads against a reference genome. This approach is successful for detecting structural variants of different type, location, split-read alignments (where different parts of a read map to distant places in the genome), large indels and coverage alterations. Read mapping is strong in low coverage experiments and able to identify heterozygous alterations. This method is also computationally cheaper than SV detection based on *de novo* assembly. On the contrary, it is limited in the detection of novel sequences (that will not align to the reference) such as large insertions and it requires a resolved good-quality reference genome (Sedlazeck et al., 2018).

In the context of SVs detection, it is often used read assembly followed by contig mapping. This strategy allows a better map of the DNA sequences against the reference and an accurate detection of the multiple types of SVs, but it will be computationally expensive (Sedlazeck et al., 2018). In conclusion, the choice of the best strategy depends on the experimental question.

Structural variants identification

Methods for SV identification based on sequencing technologies rely on read mapping to a reference genome. Software for SV detection, which basically identify alterations along the alignment, have improved in the past years (Sedlazeck et al., 2018). For instance, pbsv was successful in the identification of the inserted SVs. While our model was simple in terms of variation, real genomes may differ a lot more from the reference and give rise to the detection of a big number of variations. Thereby, it is important to develop strategies that allow to distinguish between population SVs and pathogenic alterations. SV detection algorithms will also have to be adapted to the type of SVs of research interest (Sedlazeck et al., 2018).

Olduvai domain plays a potential key role in human brain evolution

Myomegalin and Olduvai domains phylogeny have been interesting for the understanding of Olduvai sequences modification during evolution. Myomegalin seems to be highly conserved in different species. The results also suggest the presence of a proto-Olduvai domain, since they show partial sequences in the myomegalin gene. The high conservation of myomegalin protein and the Olduvai domain implies that they might be playing important biological functions in a wide range of species. In addition, the divergence between the ancestral domain and the domains found in NBPF genes seem to indicate a diversification of the role that this domain might have. These results are also consistent with the current knowledge about the gene NBPF15: CON1/2/3 are regarded as closer sequences from the ancestral copy, compared with HLS1/2/3 sequences that should be more diversified and human-specific (O'Bleness et al., 2012). The analysis constitutes a first step in the understanding of Olduvai domain evolution.

Next steps in the analysis of Olduvai domain could make use of the PacBio sequencing data, aiming to detect which are the differences of this sequences and repeat number in the population. Moreover, recent evolutionary studies are implementing a strategy based on the mapping sequencing data from different species to the human reference genome (Kronenberg et al., 2018; Sulovari et al., 2019). This method is successful for comparing the copy number variants of each species as well as observe the evolutionary divergence of the sequences. The major limitation of this study is the high cost of sequencing multiple species. Future perspectives need to potentiate freely available datasets for their wide-spread use in science.

3. CONCLUSION

Overall, the proposed objectives of this study were fulfilled. The first objective was accomplished by generating a pipeline of long-read sequencing analysis. Genomes simulation of patients with structural variants was complete. The simulation of the reads was accurate, although some of the parameters could resemble to PacBio HiFi reads even more. In addition, the design of the pipeline was successful and allowed the detection of the introduced structural variants after a complete analysis of the reads.

The second objective was also achieved, and we obtained an elaborated analysis of the phylogeny of Olduvai domains. We reach both specific objectives, which include the alignment of the domain in different species and creation of the phylogenetic trees. The study also generated two more alignments for the complete myomegalin protein and the different clades of Olduvai domains, which gave a broader perspective on Olduvai phylogeny.

Regarding the planification of the TFM, first objective of the study has been challenging due to the limited time and experience in the field. The methodology used in this part has changed substantially from the planification, which has produced a delay in the schedule after genomes simulation. The cause of this change has been that some software did not work as it was predicted. However, new methods were appropriate and finally permitted to achieve the expected results. A plausible solution for a better simulation of PacBio reads would be the use of another read simulator, more specialized for this kind of reads. The second part (objective) of the study has been developed under the predicted schedule and methodology. Results were obtained as it was expected and the timelines allowed to generate some complementary results.

This project has been a good opportunity to learn how to analyse long-read sequencing data and structural variant call. This first introduction to sequencing analysis will be extremely useful to set the basis for real patient analysis in the context of 1q21.1 syndrome. Future perspectives will seek to implement a wider range of techniques and perfectionate the pipeline that has been generated in this TFM. In particular, *de novo* assembly will be attempted in a smaller set of patients. Moreover, the TFM has permitted to observe the resemblance of Olduvai domains in different clades and species. It has been noted that evolutionary analysis is complex and, in this case, will require the integration of more data for further insight.

4. GLOSSARY

Olduvai domain:

Protein domain found in myomegalin and NBPF proteins. It shows human lineage-specific increase in copy number and appears to be involved in human brain evolution.

NBPF genes:

Neuroblastoma breaking point family genes are a family of genes that are highly primate specific and have been associated to autism and schizophrenia, among other neurological disease.

Myomegalin:

Myomegalin (Phosphodiesterase 4D-interacting protein) is a protein encoded by PDE4DIP gene in humans, found in the 1q21.2 locus. Its function is related to the microtubules and the centrosome. One of its protein domains is the Olduvai domain.

Structural variant:

Deletion, insertion, duplication and inversion of more than 50 base pairs (bp) among human genomes.

Read:

Sequence of base pairs corresponding to all or part of a single DNA fragment.

Contig:

Series of overlapping DNA reads used to reconstruct the original DNA sequence of a chromosome or a region of a chromosome.

FASTQ:

Text-based format for storing sequencing reads and its corresponding quality scores.

Multiple sequence alignment:

Similarity comparison between three or more biological sequences that can be protein, DNA, or RNA. From them, homology between the sequences and phylogenetic analysis can be retrieved.

5. BIBLIOGRAPHY

Albers, C. A., Paul, D. S., Schulze, H., Freson, K., Stephens, J. C., Smethurst, P. A., Jolley, J. D., Cvejic, A., Kostadima, M., Bertone, P., Breuning, M. H., Debili, N., Deloukas, P., Favier, R., Fiedler, J., Hobbs, C. M., Huang, N., Hurles, M. E., Kiddle, G., … Ghevaert, C. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nature Genetics*, *44*(4), 435–439. https://doi.org/10.1038/ng.1083

Andries, V., Vandepoele, K., Staes, K., Berx, G., Bogaert, P., Van Isterdael, G., Ginneberge, D., Parthoens, E., Vandenbussche, J., Gevaert, K., & van Roy, F. (2015). NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest. *BMC Cancer*, *15*, 391. https://doi.org/10.1186/s12885-015-1408-5

Bernier, R., Steinman, K. J., Reilly, B., Wallace, A. S., Sherr, E. H., Pojman, N., Mefford, H. C., Gerdts, J., Earl, R., Hanson, E., Goin-Kochel, R. P., Berry, L., Kanne, S., Snyder, L. G., Spence, S., Ramocki, M. B., Evans, D. W., Spiro, J. E., Martin, C. L., … Wenegrat, J. (2016). Clinical phenotype of the recurrent 1q21.1 copy-number variant. *Genetics in Medicine*, *18*(4), 341–349. https://doi.org/10.1038/gim.2015.78

Biosciences, P. (n.d.). *SMRT ® Tools Reference Guide*. 1–98.

Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *2018*(6). https://doi.org/10.7717/peerj.4958

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. In *Current Protocols in Bioinformatics* (Vol. 2014, Issue December). https://doi.org/10.1002/0471250953.bi0411s48

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1), 1–16. https://doi.org/10.1038/s41467-018-08148-z

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., & Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, *31*(13), 3497–3500. https://doi.org/10.1093/nar/gkg500

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., & Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, *13*(12), 1050–1054. https://doi.org/10.1038/nmeth.4035

Consortium, T. U. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Crespi, B., Summers, K., & Dorus, S. (2009). Genomic sister-disorders of neurodevelopment: An evolutionary approach. *Evolutionary Applications*, *2*(1), 81–100. https://doi.org/10.1111/j.1752-4571.2008.00056.x

Davis, J. M., Searles, V. B., Anderson, N., Keeney, J., Raznahan, A., Horwood, L. J., Fergusson, D. M., Kennedy, M. A., Giedd, J., & Sikela, J. M. (2015). DUF1220 copy number is linearly associated with increased cognitive function as measured by total IQ and mathematical aptitude scores. *Human Genetics*, *134*(1), 67–75. https://doi.org/10.1007/s00439-014-1489-2

Dennis, M. Y., Harshman, L., Nelson, B. J., Penn, O., Cantsilieris, S., Huddleston, J., Antonacci, F., Penewit, K., Denman, L., Raja, A., Baker, C., Mark, K., Malig, M., Janke, N., Espinoza, C., Stessman, H. A. F., Nuttle, X., Hoekzema, K., Lindsay-Graves, T. A., … Eichler, E. E. (2017). The evolution and population diversity of human-specific segmental duplications. *Nature Ecology and Evolution*, *1*(3), 1–10. https://doi.org/10.1038/s41559-016-0069

Diskin, S. J., Hou, C., Glessner, J. T., Attiyeh, E. F., Laudenslager, M., Bosse, K., Cole, K., Mossé, Y. P., Wood, A., Lynch, J. E., Pecor, K., Diamond, M., Winter, C., Wang, K., Kim, C., Geiger, E. A., McGrady, P. W., Blakemore, A. I. F., London, W. B., … Maris, J. M. (2009). Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, *459*(7249), 987–991. https://doi.org/10.1038/nature08035

Dumas, L. J., O'bleness, M. S., Davis, J. M., Dickens, C. M., Anderson, N., Keeney, J. G., Jackson, J., Sikela, M., Raznahan, A., Giedd, J., Rapoport, J., Nagamani, S. S. C., Erez, A., Brunetti-Pierri, N., Sugalski, R., Lupski, J. R., Fingerlin, T., Cheung, S. W., & Sikela, J. M. (2012). DUF1220-domain copy number implicated in human brain-size pathology and evolution. *American Journal of Human Genetics*, *91*(3), 444–454. https://doi.org/10.1016/j.ajhg.2012.07.016

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Escalona, M., Rocha, S., & Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature Reviews Genetics*, *17*(8), 459–469. https://doi.org/10.1038/nrg.2016.57

Fiddes, I. T., Lodewijk, G. A., Mooring, M., Bosworth, C. M., Ewing, A. D., Mantalas, G. L., Novak, A. M., van den Bout, A., Bishara, A., Rosenkrantz, J. L., Lorig-Roach, R., Field, A. R., Haeussler, M., Russo, L., Bhaduri, A., Nowakowski, T. J., Pollen, A. A., Dougherty, M. L., Nuttle, X., … Haussler, D. (2018). Human-Specific <em>NOTCH2NL</em> Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell*, *173*(6), 1356-1369.e22. https://doi.org/10.1016/j.cell.2018.03.051

Fiddes, I. T., Pollen, A. A., Davis, J. M., & Sikela, J. M. (2019). Paired involvement of human-specific Olduvai domains and NOTCH2NL genes in human brain evolution. *Human Genetics*, *138*(7), 715–721. https://doi.org/10.1007/s00439-019-02018-4

Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., Haffner, C., Sykes, A., Wong, F. K., Peters, J., Guhr, E., Klemroth, S., Prüfer, K., Kelso, J., Naumann, R., Nüsslein, I., Dahl, A., Lachmann, R., Pääbo, S., & Huttner, W. B. (2015). Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science*, *347*(6229), 1465–1470. https://doi.org/10.1126/science.aaa1975

Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, *18*, 9–19. https://doi.org/10.1016/j.csbj.2019.11.002

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Guindon, S., Lethiec, F., Duroux, P., & Gascuel, O. (2005). PHYML Online - A web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research*, *33*(SUPPL. 2), 557–559. https://doi.org/10.1093/nar/gki352

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Heide, M., Haffner, C., Murayama, A., Kurotaki, Y., Shinohara, H., Okano, H., Sasaki, E., & Huttner, W. B. (2020). Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. *Science*, *369*(6503), 546–550. https://doi.org/10.1126/science.abb2401

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, *21*(3), 171–189. https://doi.org/10.1038/s41576-019-0180-9

Hon, T., Mars, K., Young, G., Tsai, Y. C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, *7*(1), 1–22. https://doi.org/10.1038/s41597-020-00743-4

Keeney, J. G., Davis, J. M., Siegenthaler, J., Post, M. D., Nielsen, B. S., Hopkins, W. D., & Sikela, J. M. (2015). DUF1220 protein domains drive proliferation in human neural stem cells and are associated with increased cortical volume in anthropoid primates. *Brain Structure and Function*, *220*(5), 3053–3060. https://doi.org/10.1007/s00429-014-0814-9

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., Munson, K. M., Hastie, A. R., Diekhans, M., Hormozdiari, F., Lorusso, N., Hoekzema, K., Qiu, R., Clark, K., Raja, A., … Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science*, *360*(6393), eaar6343. https://doi.org/10.1126/science.aar6343

Kumar, K. R., Cowley, M. J., & Davis, R. L. (2019). Next-Generation Sequencing and Emerging

Technologies. *Seminars in Thrombosis and Hemostasis*, *45*(7), 661–673. https://doi.org/10.1055/s-0039-1688446

Kurtz, S., Shumway, M., Antonescu, C., Salzberg, S. L., Phillippy, A., Smoot, M., Delcher, A. L., & Delcher, A. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), 12. http://www.tigr.org/software/mummer.

LA Lima, H Yang, C Dong, K. W. Svg. (n.d.). *Simulation of structural variants in next-generation sequencing data.*

Laureys, G., Speleman, F., Opdenakker, G., Benoit, Y., & Leroy, J. (1990). Constitutional translocation t(1;17)(p36;q12–21) in a patient with neuroblastoma. *Genes, Chromosomes and Cancer*, *2*(3), 252–254. https://doi.org/https://doi.org/10.1002/gcc.2870020315

Lefort, V., Longueville, J.-E., & Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution*, *34*(9), 2422–2424. https://doi.org/10.1093/molbev/msx149

Lodewijk, G. A., Fernandes, D. P., Vretzakis, I., Savage, J. E., & Jacobs, F. M. J. (2020). Evolution of Human Brain Size-Associated NOTCH2NL Genes Proceeds toward Reduced Protein Levels. *Molecular Biology and Evolution*, *37*(9), 2531–2548. https://doi.org/10.1093/molbev/msaa104

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(October). https://doi.org/10.1038/s41576-020-0236-x

Mardis, E. R. (2007). *Commentary A brief history of (DNA sequencing) time*. *october*, 63108. https://doi.org/10.1038/mrg2240

Mefford, H., Sharp, A., Baker, C., Itsara, A., Jiang, Z., Buysse, K., Huang, S., Maloney, V., Crolla, J., Baralle, D., Collins, A., Mercer, C., Norga, K., De Ravel, T., Devriendt, K., Bongers, E., De Leeuw, N., Reardon, W., Gimelli, S., … Eichler, E. (2008). Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *New England Journal of Medicine*, *359*(16), 1685–1699. https://doi.org/10.1056/NEJMoa0805384

Midha, M. K., Wu, M., & Chiu, K. P. (2019). Long-read sequencing in deciphering human genetics to a greater depth. *Human Genetics*, *138*(11–12), 1201–1215. https://doi.org/10.1007/s00439-019-02064-y

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., … Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), 79–84. https://doi.org/10.1038/s41586-020-2547-7

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., Miga, K. H., Eichler, E. E., Phillippy, A. M., & Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research*, *30*(9), 1291–1305. https://doi.org/10.1101/gr.263566.120

O'Bleness, M. S., Michael Dickens, C., Dumas, L. J., Kehrer-Sawatzki, H., Wyckoff, G. J., & Sikela, J. M. (2012). Evolutionary history and genome organization of duf1220 protein domains. *G3: Genes, Genomes, Genetics*, *2*(9), 977–986. https://doi.org/10.1534/g3.112.003061

Ono, Y., Asai, K., & Hamada, M. (2013). PBSIM: PacBio reads simulator—toward accurate genome assembly. *Bioinformatics*, *29*(1), 119–121. https://doi.org/10.1093/bioinformatics/bts649

Ono, Y., Asai, K., & Hamada, M. (2020). *Sequence analysis PBSIM2 : a simulator for long read sequencers with a novel generative model of quality scores*. 1–7.

Popesco, M. C., MacLaren, E. J., Hopkins, J., Dumas, L., Cox, M., Meltesen, L., McGavran, L., Wyckoff, G. J., & Sikela, J. M. (2006). Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science*, *313*(5791), 1304–1307. https://doi.org/10.1126/science.1127980

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., … Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49. https://doi.org/10.1038/nature12886

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Sathyanarayanan, A., Manda, S., Poojary, M., & Nagaraj, S. H. (2019). *Exome Sequencing Data Analysis* (S. Ranganathan, M. Gribskov, K. Nakai, & C. B. T.-E. of B. and C. B. Schönbach (Eds.); pp. 164–175). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-809633-8.20094-0

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, *19*(6), 329–346. https://doi.org/10.1038/s41576-018-0003-4

Sikela, J. M., & van Roy, F. (2018). Changing the name of the NBPF/DUF1220 domain to the Olduvai domain [version 2; peer review: 3 approved]. *F1000Research*, *6*(2185). https://doi.org/10.12688/f1000research.13586.2

Spielmann, M., Lupiáñez, D. G., & Mundlos, S. (2018). Structural variation in the 3D genome. *Nature Reviews Genetics*, *19*(7), 453–467. https://doi.org/10.1038/s41576-018-0007-0

Steinberg, K. M., Schneider, V. A., Graves-Lindsay, T. A., Fulton, R. S., Agarwala, R., Huddleston, J., Shiryev, S. A., Morgulis, A., Surti, U., Warren, W. C., Church, D. M., Eichler, E. E., & Wilson, R. K. (2014). Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Research*, *24*(12), 2066–2076. https://doi.org/10.1101/gr.180893.114

Stephens, Z. D., Hudson, M. E., Mainzer, L. S., Taschuk, M., Weber, M. R., & Iyer, R. K. (2016). Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. *PloS One*, *11*(11), e0167047–e0167047. https://doi.org/10.1371/journal.pone.0167047

Stöcker, B. K., Köster, J., & Rahmann, S. (2016). SimLoRD: Simulation of Long Read Data. *Bioinformatics*, *32*(17), 2704–2706. https://doi.org/10.1093/bioinformatics/btw286

Sulovari, A., Li, R., Audano, P. A., Porubsky, D., Vollger, M. R., Logsdon, G. A., Warren, W. C., Pollen, A. A., Chaisson, M. J. P., & Eichler, E. E. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(46), 23243–23253. https://doi.org/10.1073/pnas.1912175116

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, *34*(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Vandepoele, K., Staes, K., Andries, V., & van Roy, F. (2010). Chibby interacts with NBPF1 and clusterin, two candidate tumor suppressors linked to neuroblastoma. *Experimental Cell Research*, *316*(7), 1225–1233. https://doi.org/10.1016/j.yexcr.2010.01.019

Vandepoele, K., Van Roy, N., Staes, K., Speleman, F., & Van Roy, F. (2005). A novel gene family NBPF: Intricate structure generated by gene duplications during primate evolution. *Molecular Biology and Evolution*, *22*(11), 2265–2274. https://doi.org/10.1093/molbev/msi222

Wang, Z., Zhang, C., & Qi, R. Z. (2014). A newly identified myomegalin isoform functions in Golgi microtubule organization and ER-Golgi transport. *Journal of Cell Science*, *127*(22), 4904–4917. https://doi.org/10.1242/jcs.155408

WATSON, J. D., & CRICK, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. *BioRxiv*, 519025. https://doi.org/10.1101/519025

Xia, L. C., Ai, D., Lee, H., Andor, N., Li, C., Zhang, N. R., & Ji, H. P. (2018). SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. *GigaScience*, *7*(7). https://doi.org/10.1093/gigascience/giy081

Zhang, W., Jia, B., & Wei, C. (2019). PaSS: A sequencing simulator for PacBio sequencing. *BMC Bioinformatics*, *20*(1), 1–7. https://doi.org/10.1186/s12859-019-2901-7

Zhou, F., Xing, Y., Xu, X., Yang, Y., Zhang, J., Ma, Z., & Wang, J. (2013). NBPF is a potential DNA-binding transcription factor that is directly regulated by NF-κB. *The International Journal of Biochemistry & Cell Biology*, *45*(11), 2479–2490. https://doi.org/https://doi.org/10.1016/j.biocel.2013.07.022

6. SUPPLEMENTARY INFORMATION

Software installation

This section describes the require commands to setup of the different software that have been used. In general terms, they need to be downloaded, installed and made executable. Other required packages for its function are also installed. Firstly, system and repositories are updated, and python and git are installed.

```
sudo apt update
sudo apt upgrade -y
sudo apt install python python3 python2 -y python2
sudo apt install git -y
```

Installation of SVGen:

```
wget https://github.com/WGLab/SVGen/archive/master.zip
sudo apt install unzip
unzip master.zip
mv SVGen-master SVGen
cd SVGen
./download_and_format_database.sh hg38
```

Installation of wgsim:

```
git init
git clone https://github.com/lh3/wgsim.git
cd wgsim
#Making executable wgsim.c
sudo apt install gcc -y
sudo apt install libz-dev -y
gcc -g -O2 -Wall -o wgsim wgsim.c -lz -lm
sudo apt install samtools -y
```

Installation of canu:

```
wget https://github.com/marbl/canu/releases/download/v2.1.1/canu-2.1.1.tar.xz
tar -xJf canu-2.1.1.tar.xz
cd canu-2.1.1/src
sudo apt install make -y
sudo apt install g++ -y
make -j 8
sudo apt install default-jre -y
```

Installation of MUMer4:

```
wget https://github.com/mummer4/mummer/releases/download/v4.0.0rc1/mummer-4.0.0rc1.tar.gz
tar -xf mummer-4.0.0rc1.tar.gz
mv mummer-4.0.0rc1 mummer
cd mummer
./configure --enable-python-binding=$/home/angelaespa32/mummer
make
sudo make install
sudo apt install gnuplot -y #necessary for mummerplot
```

Conda installation. Conda is an open-source package management system and environment management system.

```
curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
sh Miniconda3-latest-Linux-x86_64.sh
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
```

Installation of pbmm2:

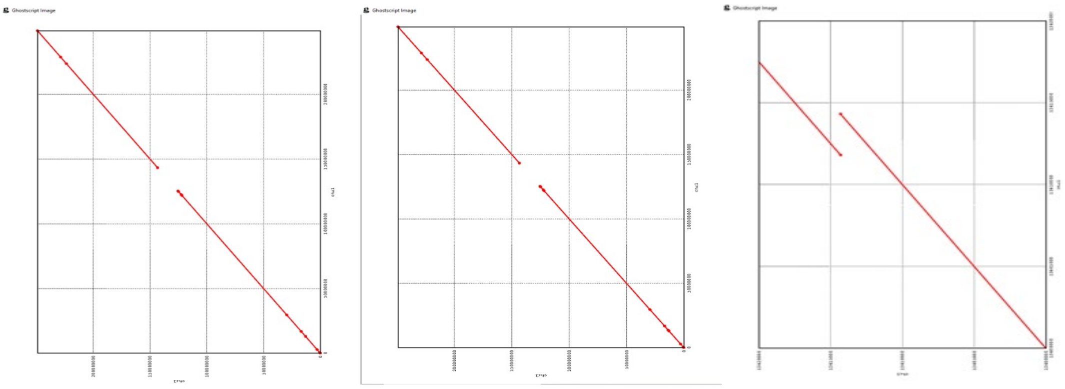
```
cd miniconda3/bin
./conda install -c bioconda pbmm2
```

Installation of pbsv:

```
cd miniconda3/bin
./conda install -c bioconda pbsv
```

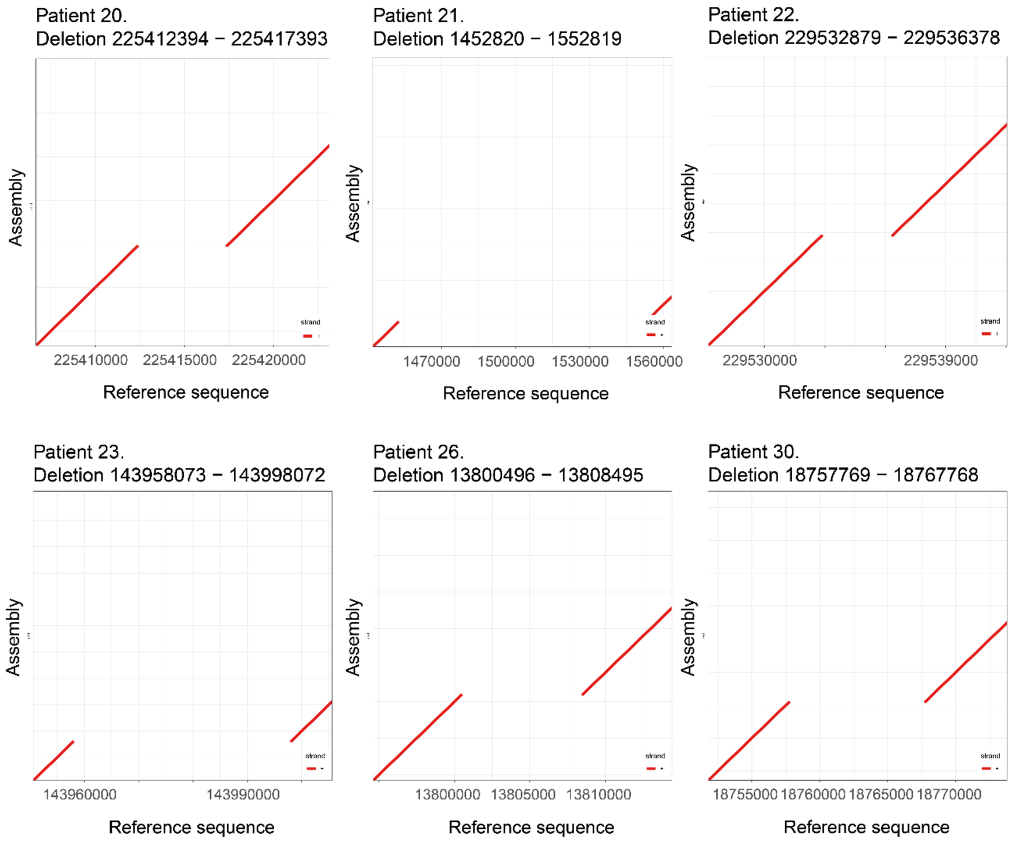
Dot plots with mummerplot (MUMmer)

Dot plots generated with MUMmer are effective to visualize the generated structural variants. They allow to zoom in a specific region, but the graphical parameters can not be modified.



Sup. Figure 1. Dot plots generated with MUMmer. (A) Dot plot of aligned reference assembly. (B) Dot plot of patient 1, carrier of a duplication (C) Zoomed in dot plot of patient1 flanking the duplicated region.

Dot plots with R



Patient 20.
Deletion 225412394 − 225417393

Patient 21.
Deletion 1452820 − 1552819

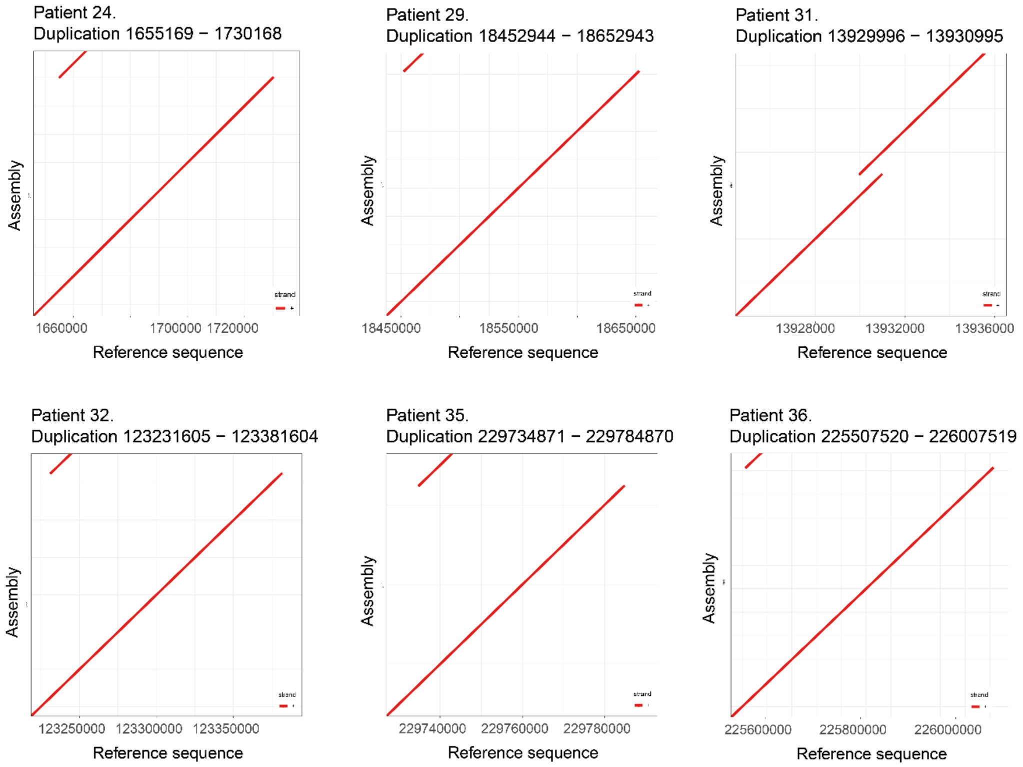
Patient 22.
Deletion 229532879 − 229536378

Patient 23.
Deletion 143958073 − 143998072

Patient 26.
Deletion 13800496 − 13808495

Patient 30.
Deletion 18757769 − 18767768

Sup figure 2. Dot plots of patients with deleted regions, created by R.



Patient 24.
Duplication 1655169 − 1730168

Patient 29.
Duplication 18452944 − 18652943

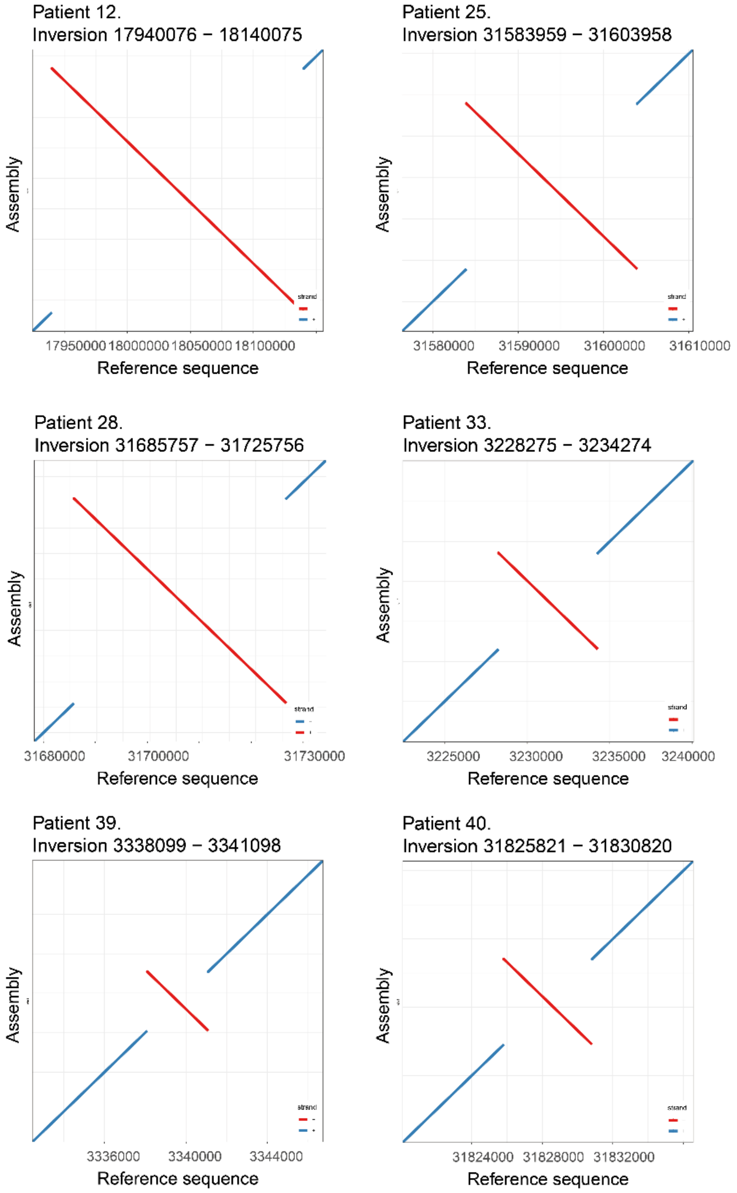
Patient 31.
Duplication 13929996 − 13930995

Patient 32.
Duplication 123231605 − 123381604

Patient 35.
Duplication 229734871 − 229784870

Patient 36.
Duplication 225507520 − 226007519

Sup figure 3. Dot plots of patients with duplicated regions, created by R.



Patient 12.
Inversion 17940076 − 18140075

Patient 25.
Inversion 31583959 − 31603958

Patient 27.
Inversion 18369150 − 18379149

Patient 28.
Inversion 31685757 − 31725756

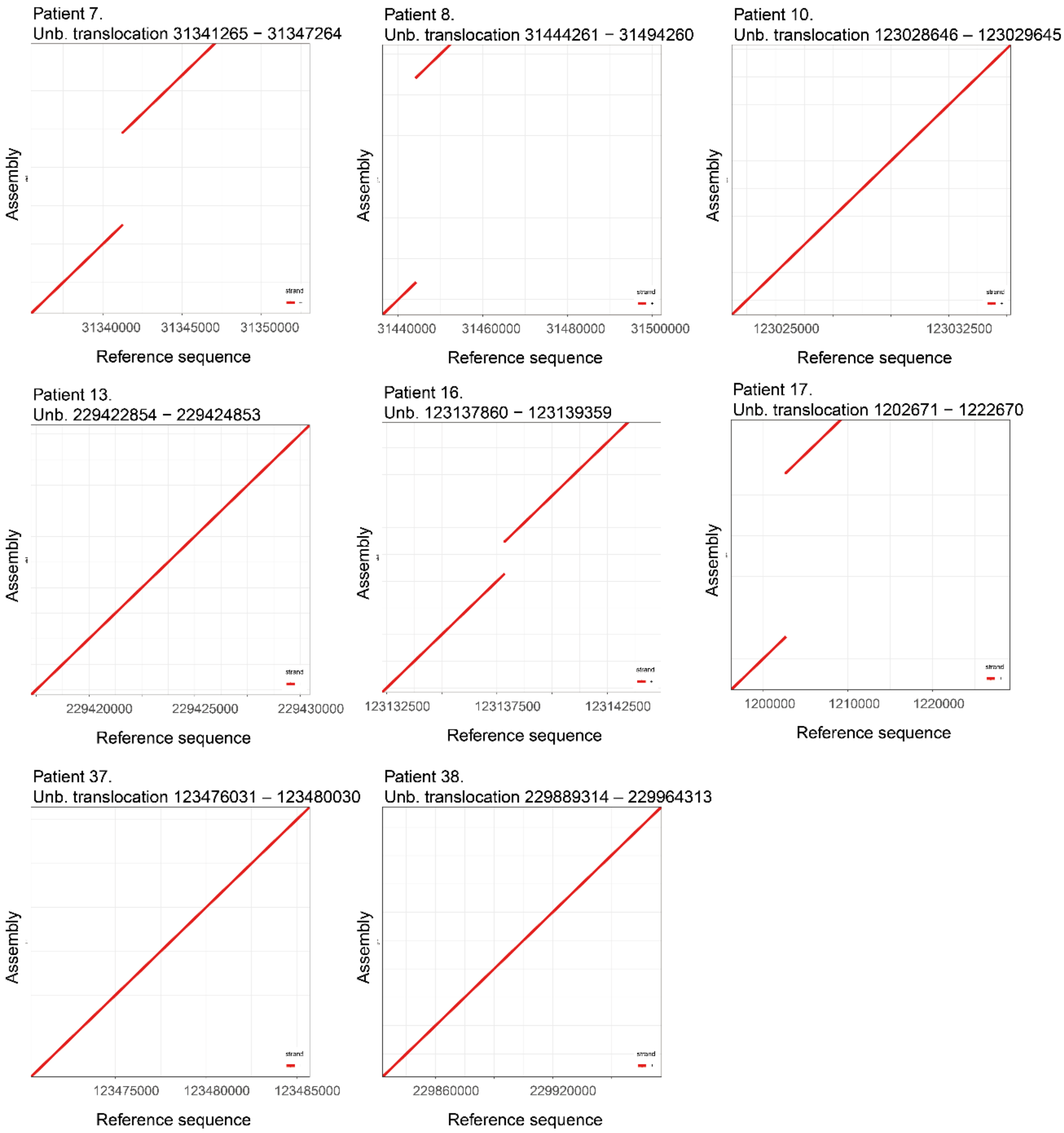
Patient 33.
Inversion 3228275 − 3234274

Patient 34.
Inversion 229638360 − 229641859

Patient 39.
Inversion 3338099 − 3341098

Patient 40.
Inversion 31825821 − 31830820

Sup figure 4. Dot plots of patients with inverted regions, created by R.



Patient 7.
Unb. translocation 31341265 – 31347264

Patient 8.
Unb. translocation 31444261 – 31494260

Patient 10.
Unb. translocation 123028646 – 123029645

Patient 13.
Unb. 229422854 – 229424853

Patient 16.
Unb. 123137860 – 123139359

Patient 17.
Unb. translocation 1202671 – 1222670

Patient 37.
Unb. translocation 123476031 – 123480030

Patient 38.
Unb. translocation 229889314 – 229964313

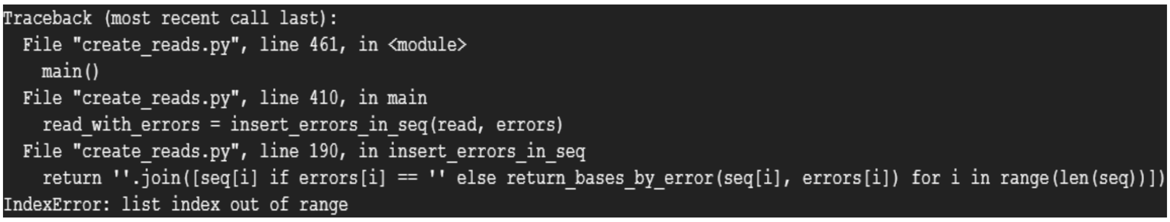
Sup figure 5. Dot plots of patients with unbalanced translocations, created by R.

SVGen generation of long-reads

The following code was applied for the generation of long-reads derived from our genomes:

```
python2 create_reads.py \
-i patient1.fa \
-o patient1.fq \
--cov 10 \
--read_len 11000 \
--snp_rate 0.001 \
--del_rate 0.0001 \
--ins_rate 0.0001
```

The code resulted in a code error that did not allow to produce the reads. However, when setting the reads to short-reads generation, this error did not appear, and the creation of the reads was performed. From the error description, it seems that the problem is found in the indexes of a specific function “create_reads.py”, which introduces errors in the read sequences.



```
Traceback (most recent call last):
  File "create_reads.py", line 461, in <module>
    main()
  File "create_reads.py", line 410, in main
    read_with_errors = insert_errors_in_seq(read, errors)
  File "create_reads.py", line 190, in insert_errors_in_seq
    return ''.join([seq[i] if errors[i] == '' else return_bases_by_error(seq[i], errors[i]) for i in range(len(seq))])
IndexError: list index out of range
```

Sup. Figure 6. Error code in SVGen generation of long-reads.