1q21.1 syndrome:
A perspective on structural variant detection & the evolutionary profile of associated protein domains

Final Master Project

Paula España Bonilla

Master in Bioinformatics and Biostatistics (UOC -UB)

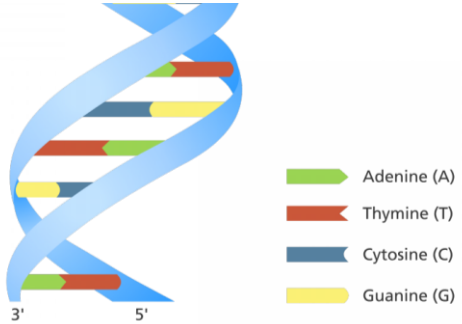
Area 3 – Subarea 8: Analysis and integration of omics data

Cedric Boeckx

Andreu Paytuví

January, 2021

Evolution of sequencing techniques



Low-throughput

High-throughput

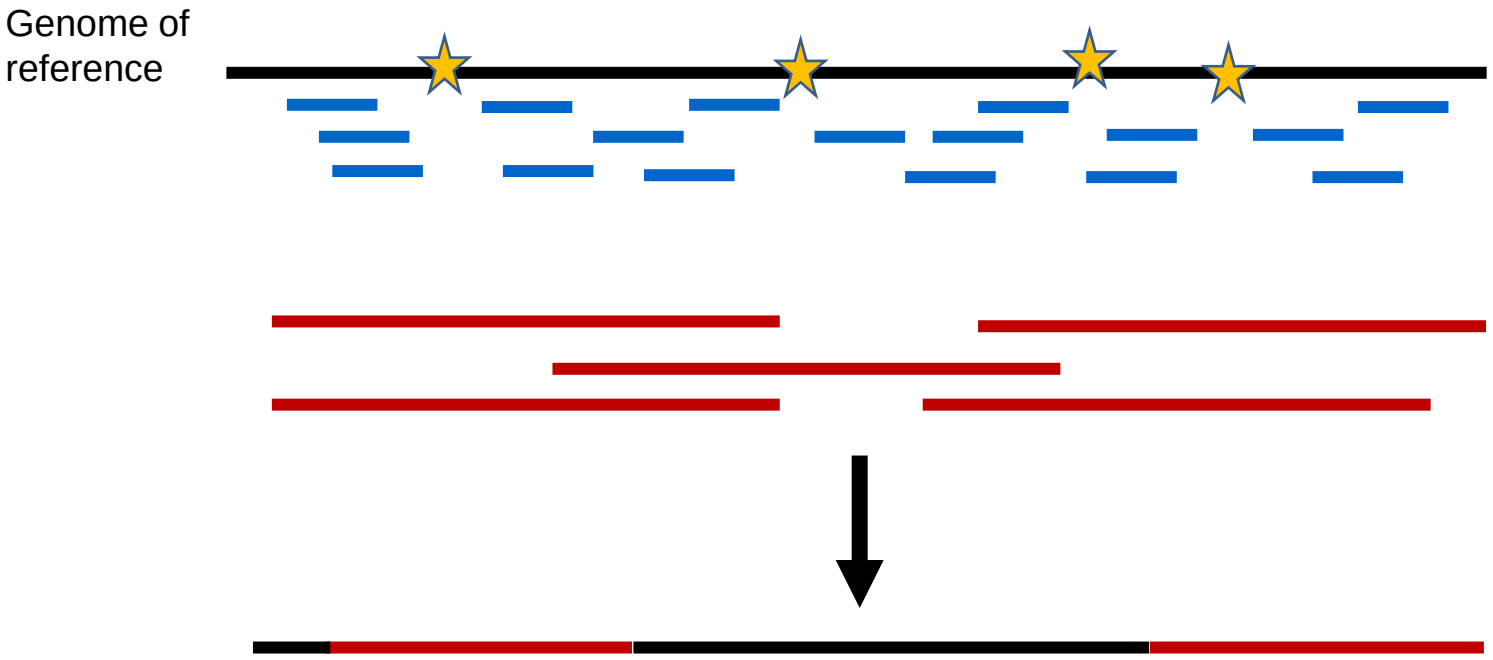
Sanger-based methods → Second Generation Sequencing → Third Generation Sequencing

Illumina

PacBio

ONT

Short vs Long reads



Repetitive & variable regions

Short reads

Long reads

PacBio HiFi reads can sequence structural variants

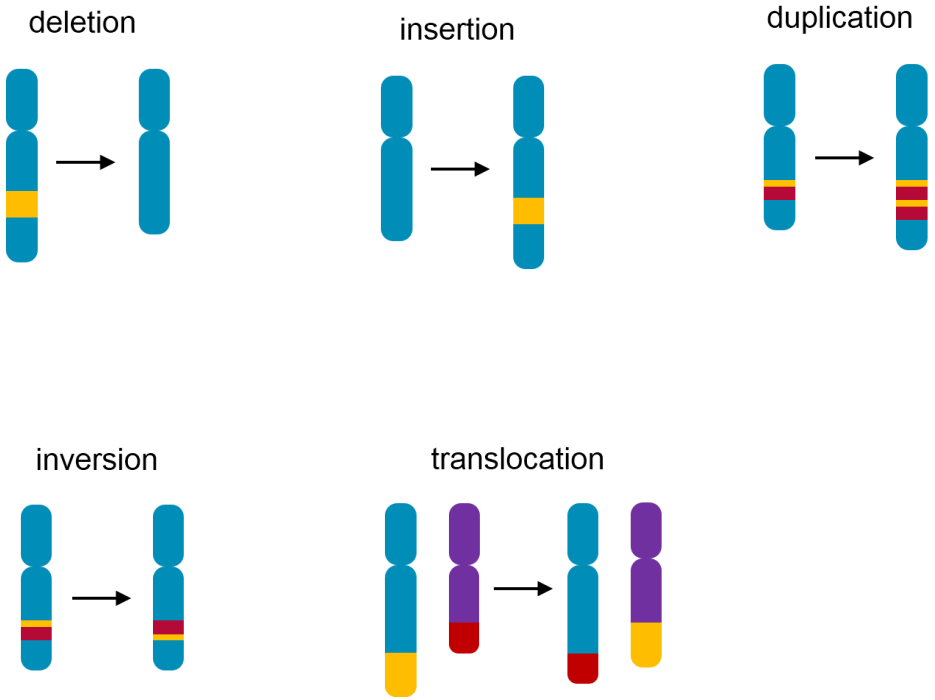
PacBio sequencing

DNA insert 1 -100 kb

Long reads

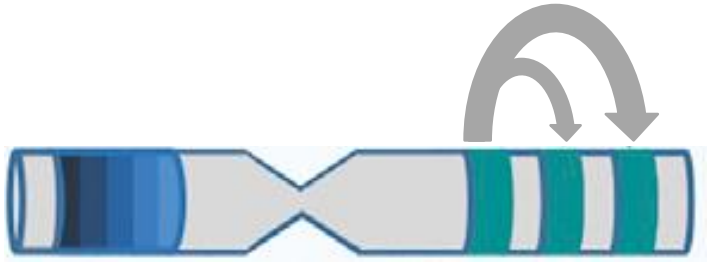
99% accuracy

Structural variants



Hickey, 2017

1q21.1 syndrome



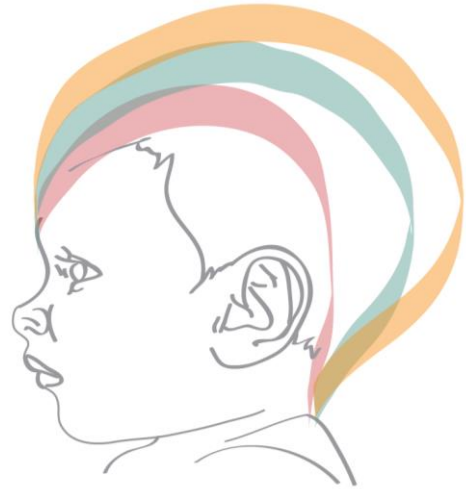
1q21

1q21.1 Microdeletion

Microcephaly
Schizophrenia

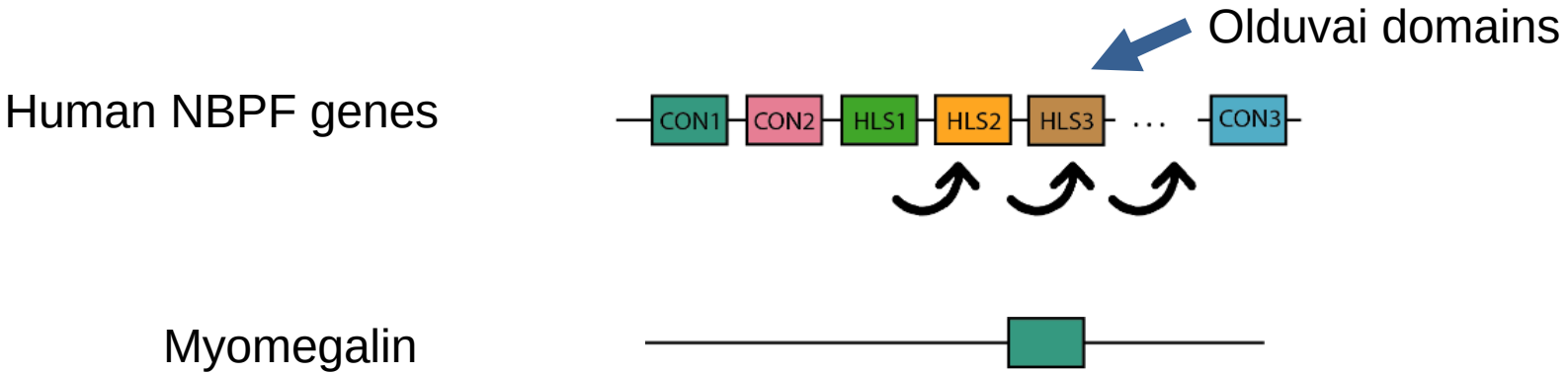
1q21.1 Microduplication

Macrocephaly
ASD



NBPF genes: implications in 1q21.1 syndrome

	Non-primate mammals	Primates	Human
			
Total Nº of NBPF genes	3	12	18
Total Nº of Olduvai domains	~10	~47	~300

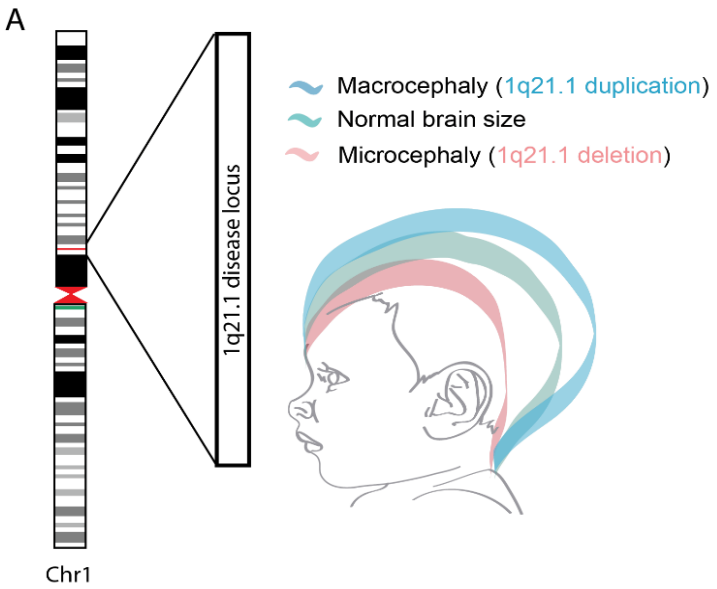


Research proposal

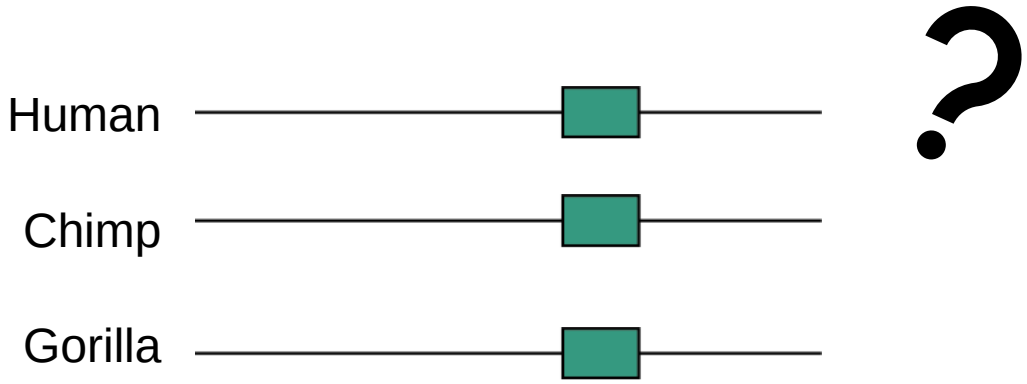
Specific gene affectation in patients

SJD
Sant Joan de Déu
Barcelona · Hospital

9 patients with microdeletion
2 patients with microduplication



Changes in myomegalin sequence in evolution



Objectives

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

Methods

Sequencing Data Analysis	**Structural Variant simulation**	SVGen
	Read simulation	Wgsim
	Bioinformatic pipeline for SV detection	Pbmm2, pbsv
Phylogeny of Olduvai domains	**Olduvai domain alignment**	MUSCLE
	Phylogenetic tree	PhyML

Results I. Genomes simulation

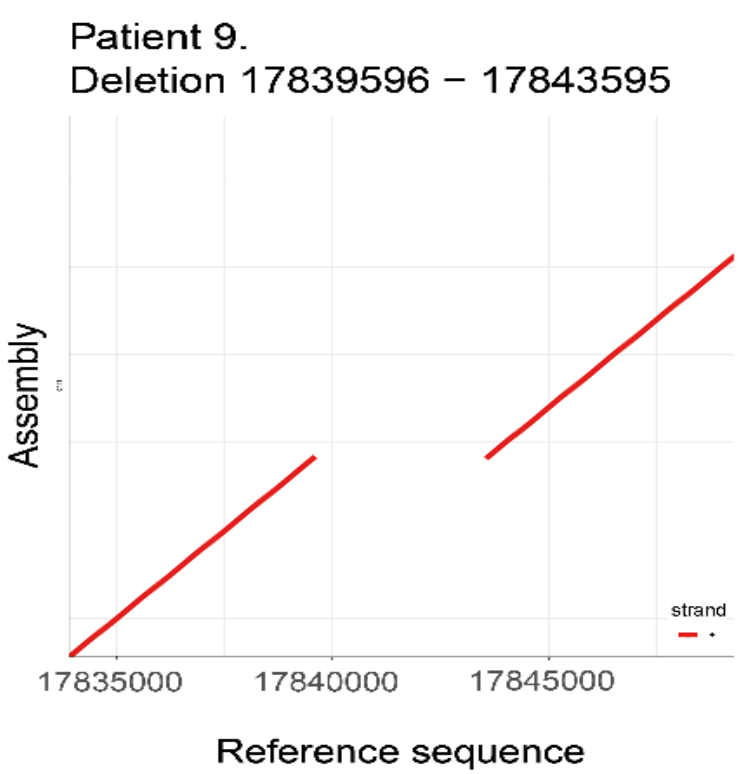
Structural variants coordinates
SVGen

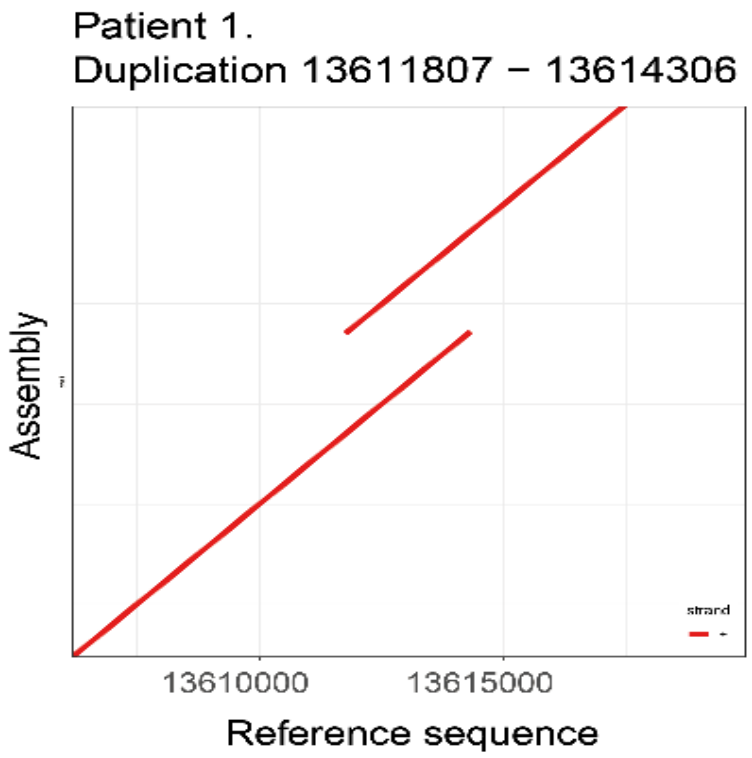
→

Insert of SVs
SVGen

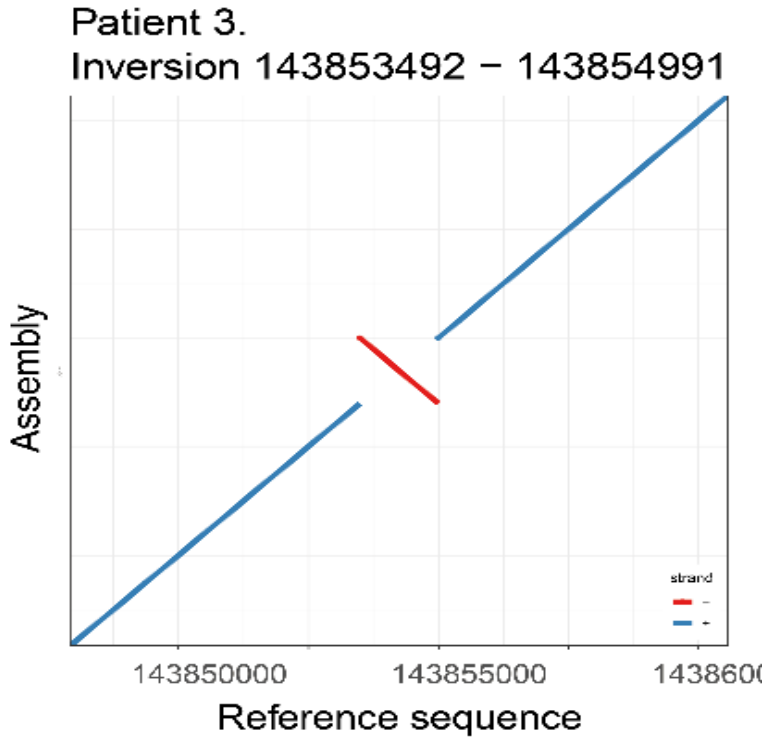
→

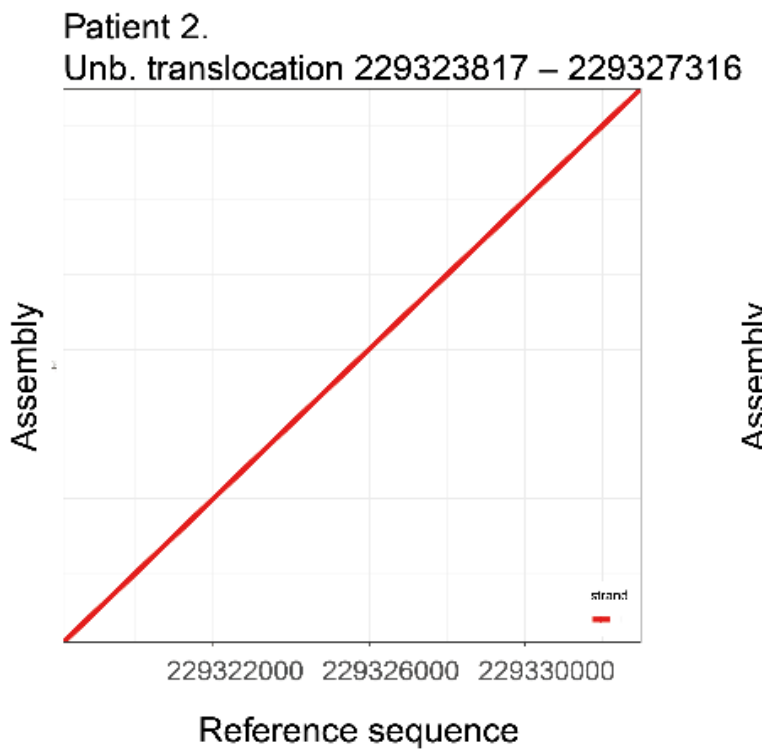
Genome alignment to reference
MUMmer





Results I. Genomes simulation





Results II. Read simulation

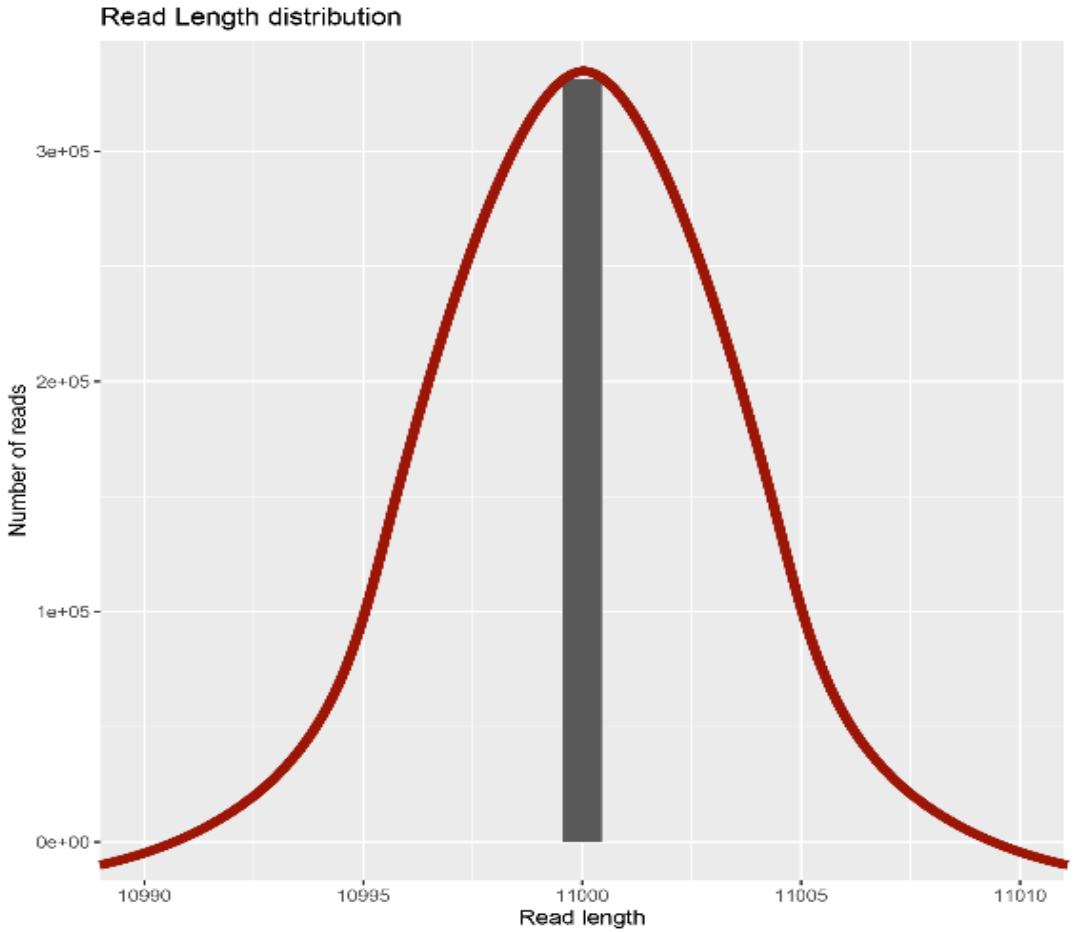
Simulation of PacBio HiFi reads
Wgsim

```
@chr1_62148042_62159041_1:0:0_107:2:0_0/1
GGTTTATTTTTTGAGACTGAGTCTCACTCTGTTGCCCAGGGGAGTACAGTGA
+
5555555555555555555555555555555555555
```

ASCII characters

331364 reads

99% base call accuracy



Results I. Read mapping & structural variant call

Read mapping to the reference
pbmm2

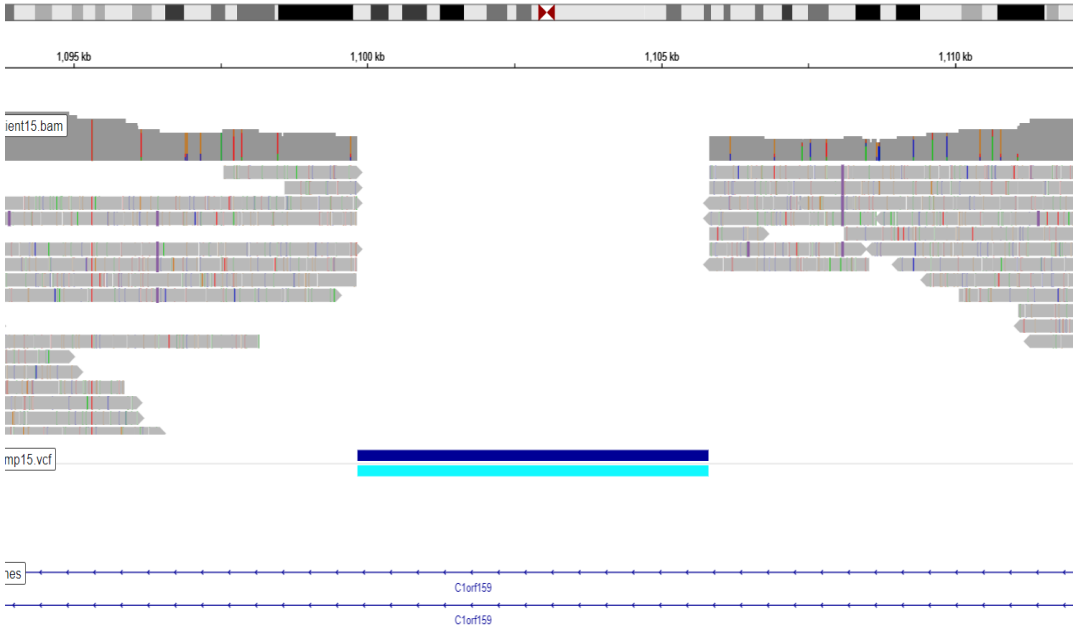
→

Structural variants call
pbsv

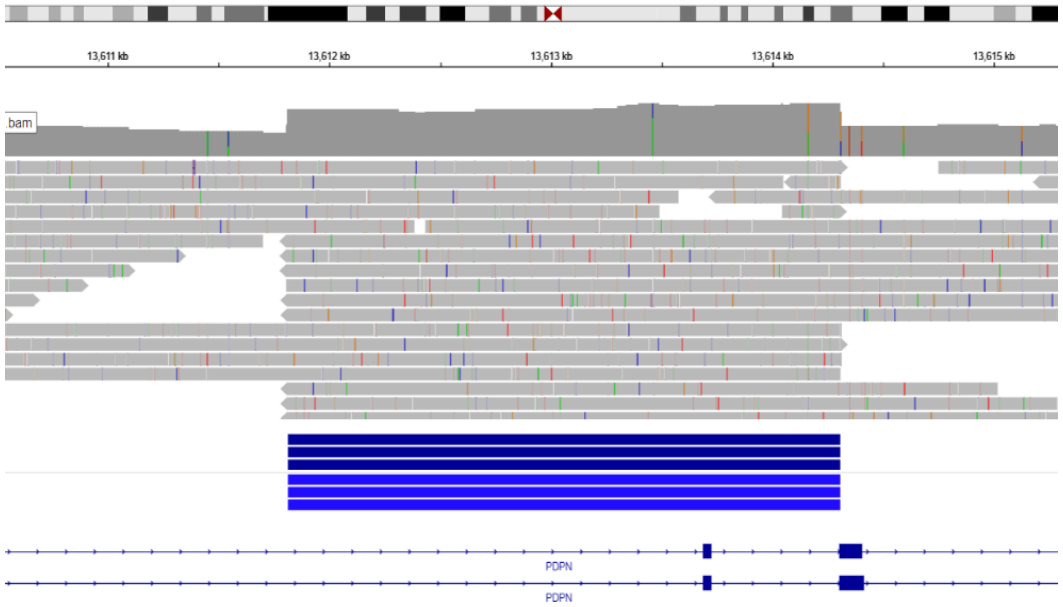
→

Visualization
IGV

Deletion



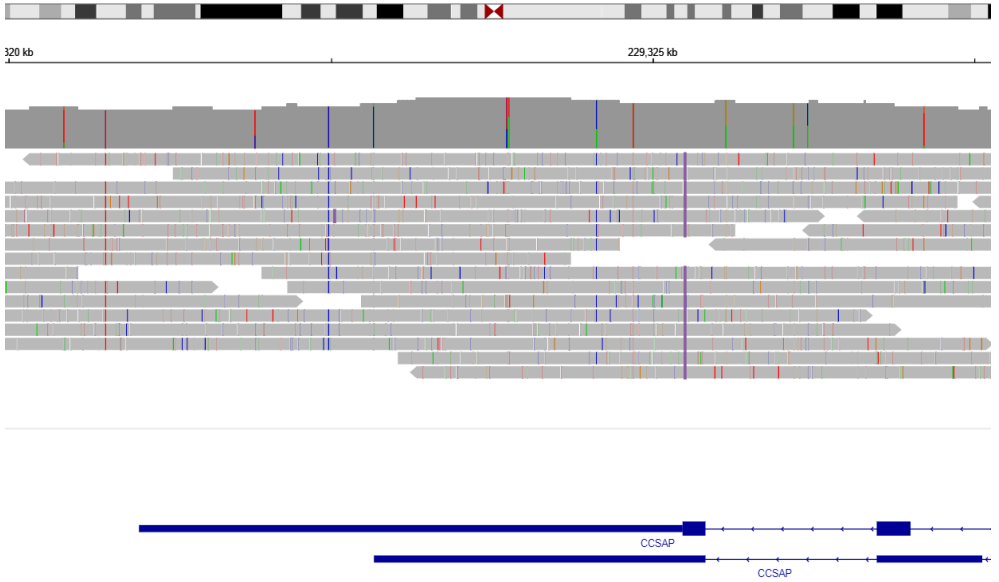
Duplication



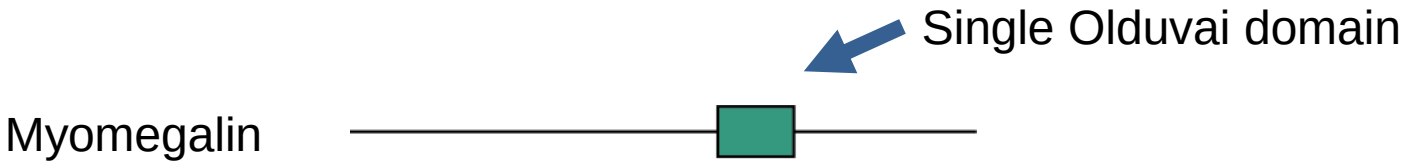
Results I. Read mapping & structural variant call

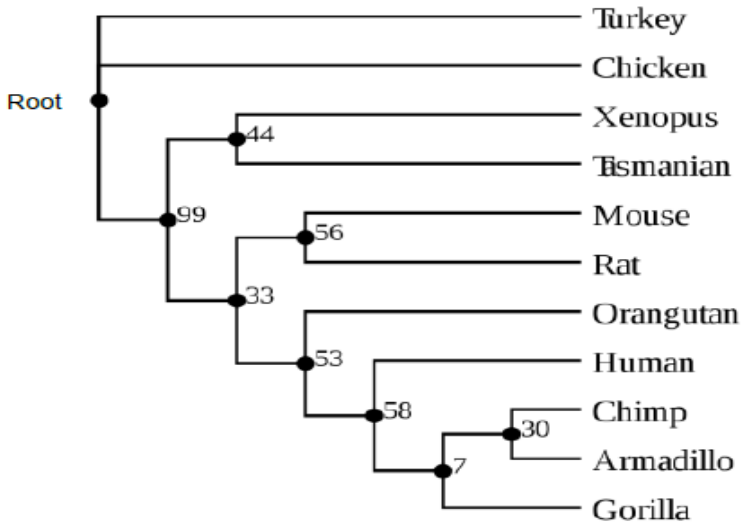
Inversion

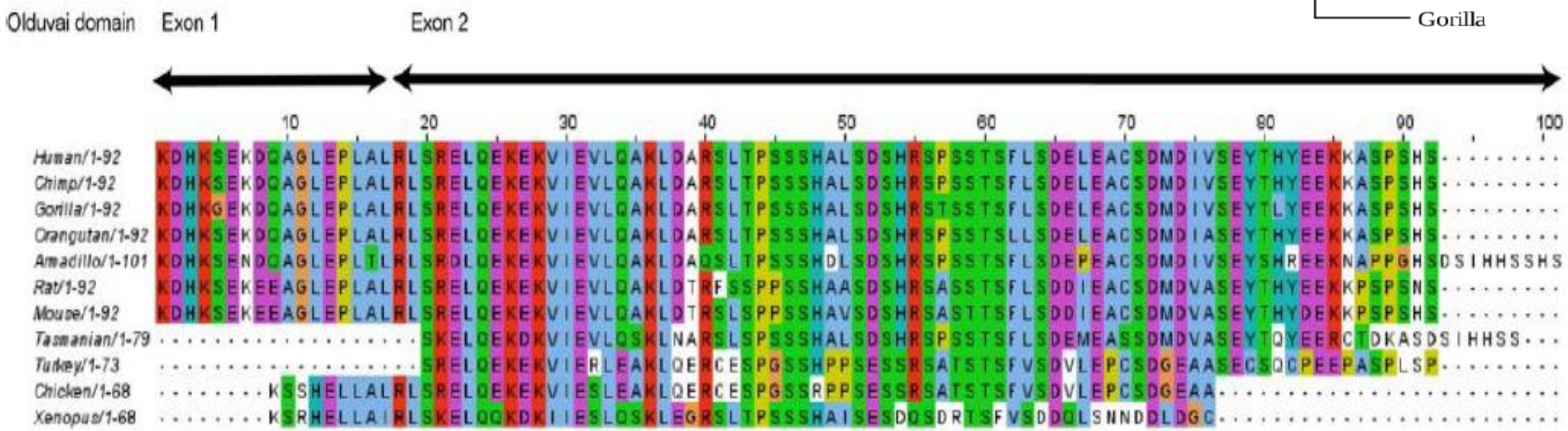
Translocation



Results II. Olduvai domain in myomegalin

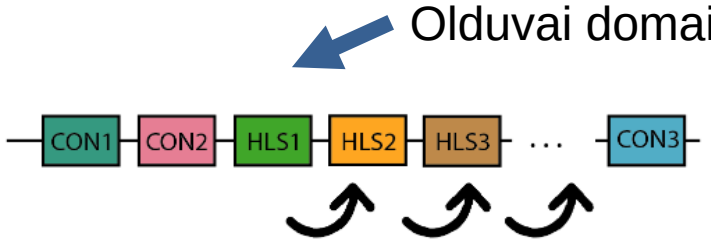


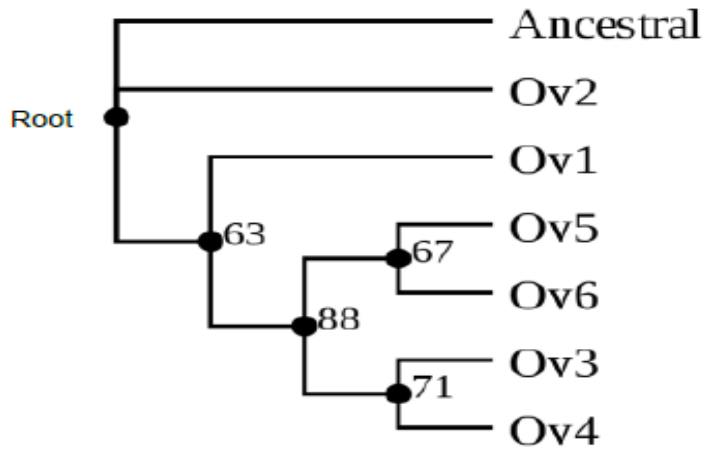


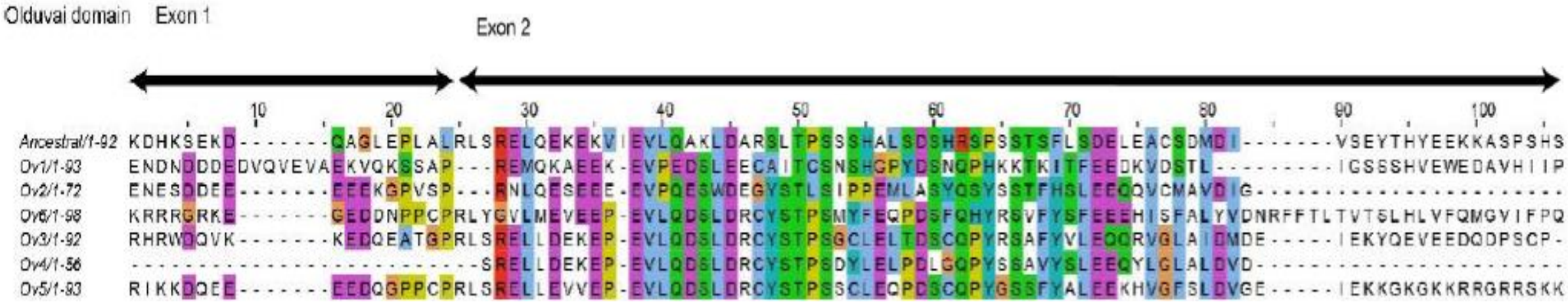


Results II. Olduvai domain in NBPF genes

Human NBPF genes







Concluding remarks

Objective 1:

- SVGen was appropriate for structural variant simulation
- Wgsim produces PacBio reads – can be improved
- Dot plots and IGV are suitable for visualisation of known SVs
- Read mapping is a cheap and effective option
- Structural variant call detected the inserted SVs

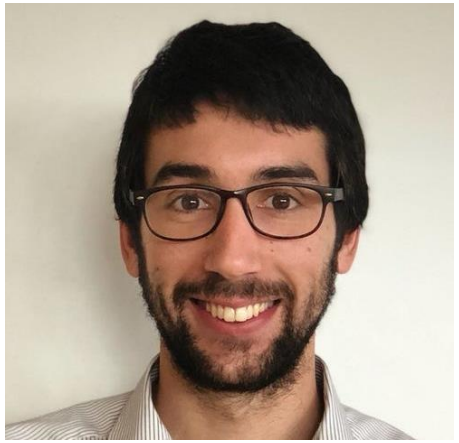
Objective 2:

- Myomegalin and Olduvai domains are highly conserved
- NBPF have diverged from their ancestral copy

Acknowledgements

Murielle Saade

Cedric Boeckx



Andreu Paytuví