KU LEUVEN

Analysis of long read data at the Flemish Super Computer

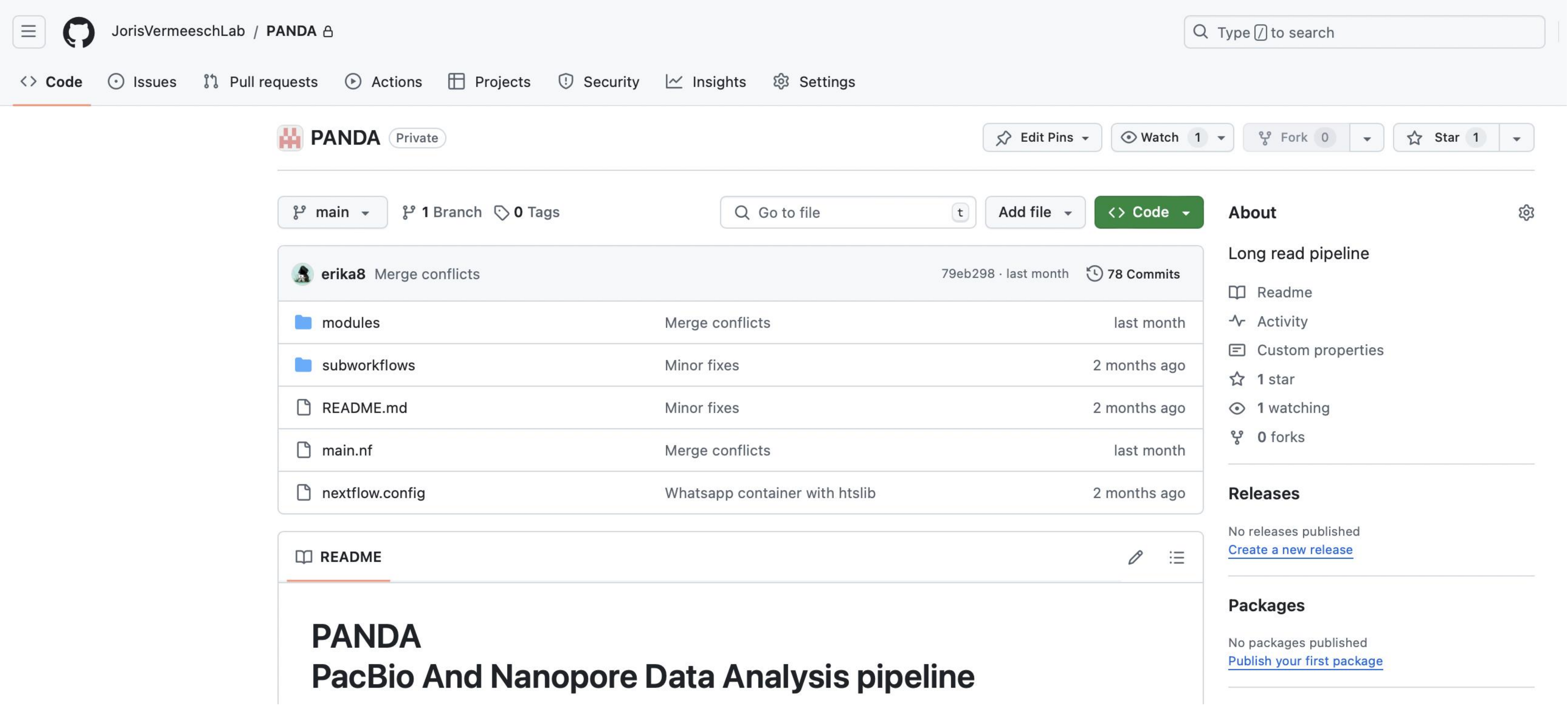
Erika Souche

Laboratory for Cytogenetics and Genome Research

Laboratory for Cytogenetics and Genome Research

- Aims
  - Implement genomic technologies to improve genetic diagnostic testing
  - Map the causes and mechanisms underlying chromosomal rearrangements
  - Improve preimplantation, prenatal and postnatal genetic testing of rare diseases
  - Use liquid biopsies as a biomarker to map genic and non-genic diseases

One nextflow pipeline



erika8 Merge conflicts 79eb298 · last month 78 Commits

modules	Merge conflicts	last month
subworkflows	Minor fixes	2 months ago
README.md	Minor fixes	2 months ago
main.nf	Merge conflicts	last month
nextflow.config	Whatsapp container with htslib	2 months ago

About

Long read pipeline

Readme
Activity
Custom properties
1 star
1 watching
0 forks

Releases

No releases published
Create a new release

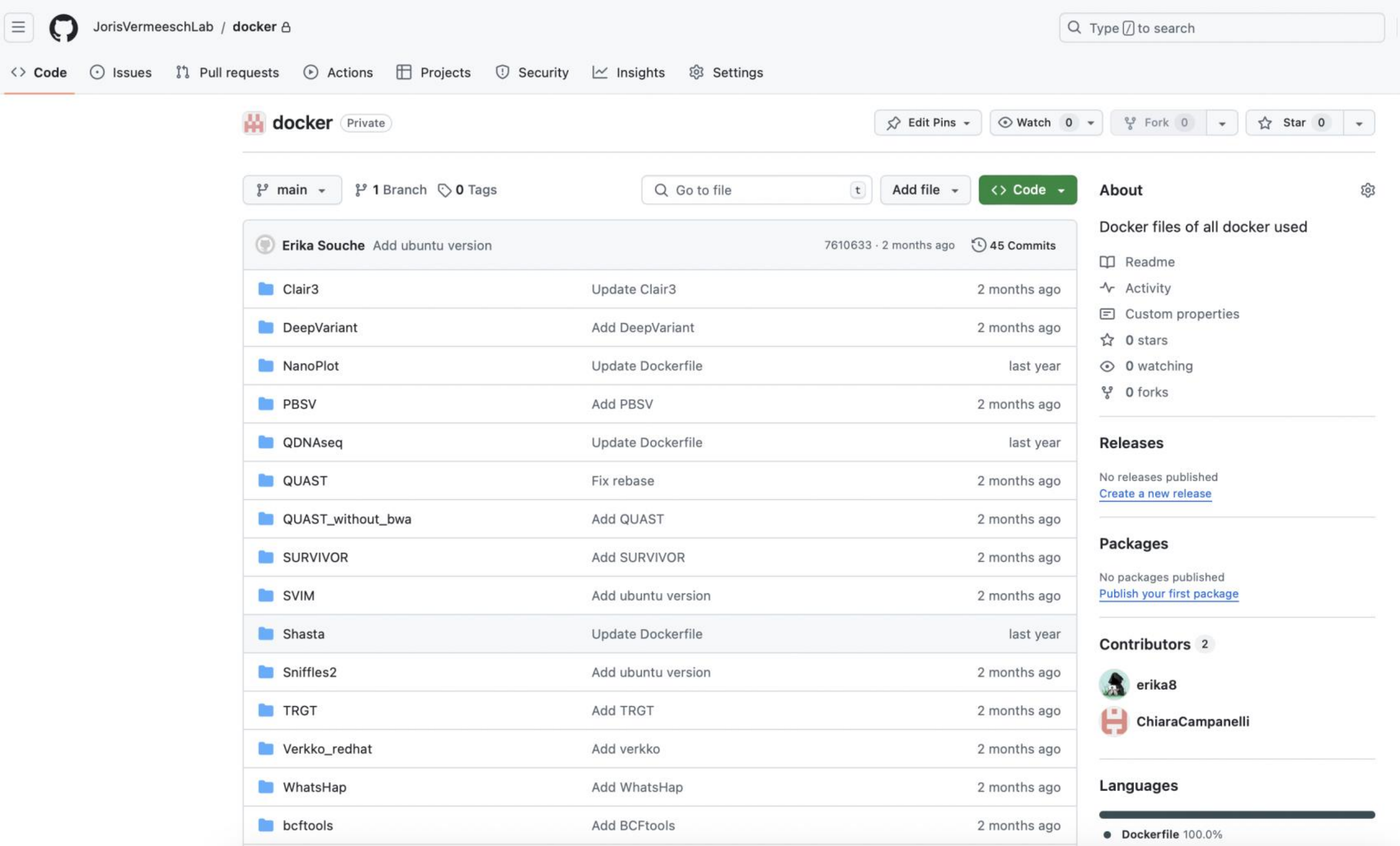
Packages

No packages published
Publish your first package

README

PANDA
PacBio And Nanopore Data Analysis pipeline

Use of external software



Erika Souche Add ubuntu version 7610633 · 2 months ago 45 Commits

Clair3	Update Clair3	2 months ago
DeepVariant	Add DeepVariant	2 months ago
NanoPlot	Update Dockerfile	last year
PBSV	Add PBSV	2 months ago
QDNAseq	Update Dockerfile	last year
QUAST	Fix rebase	2 months ago
QUAST_without_bwa	Add QUAST	2 months ago
SURVIVOR	Add SURVIVOR	2 months ago
SVIM	Add ubuntu version	2 months ago
Shasta	Update Dockerfile	last year
Sniffles2	Add ubuntu version	2 months ago
TRGT	Add TRGT	2 months ago
Verkko_redhat	Add verkko	2 months ago
WhatsHap	Add WhatsHap	2 months ago
bcftools	Add BCFtools	2 months ago

About

Docker files of all docker used

- Readme
- Activity
- Custom properties
- 0 stars
- 0 watching
- 0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

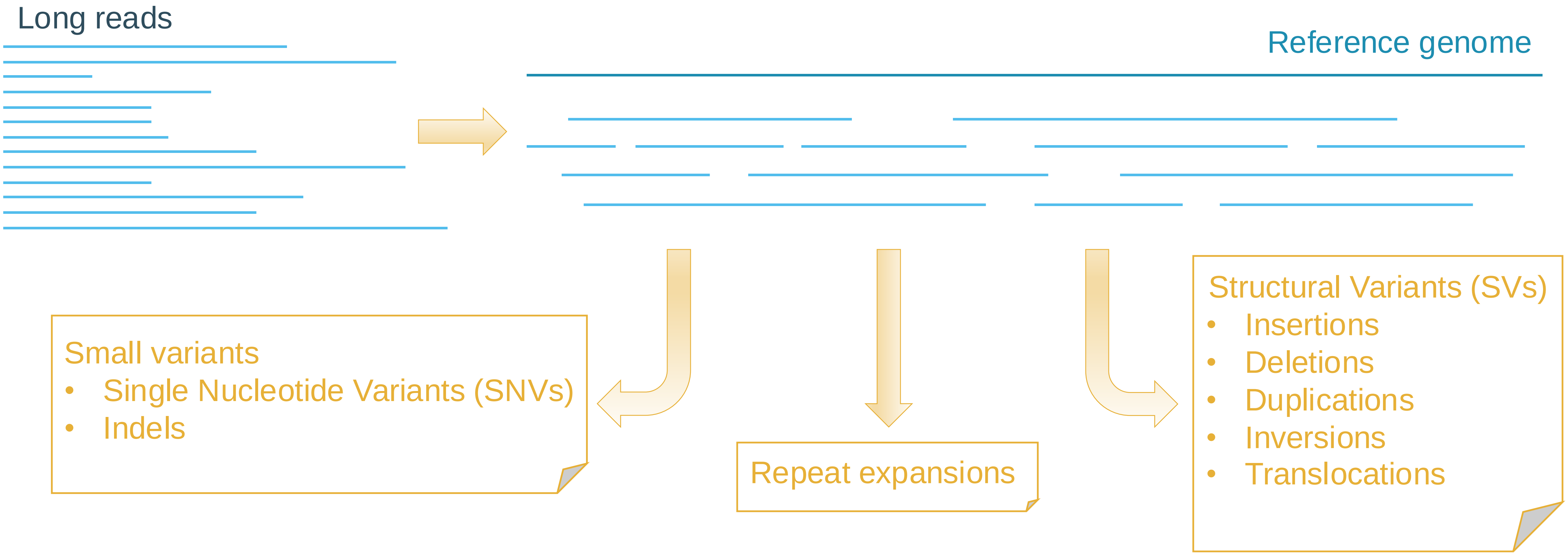
Contributors 2

- erika8
- ChiaraCampanelli

Languages

- Dockerfile 100.0%

Variant calling from mapping



Long reads

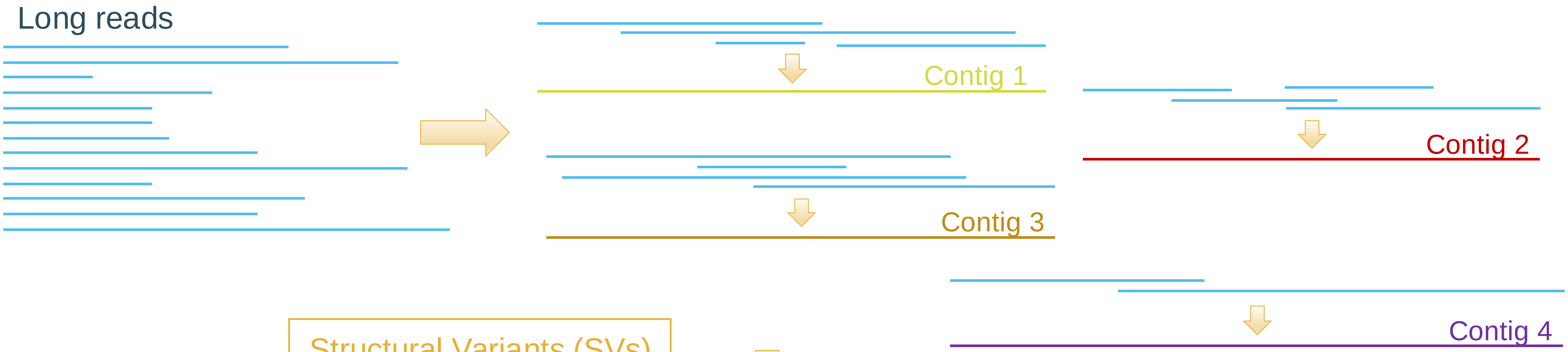
Reference genome

Small variants
- Single Nucleotide Variants (SNVs)
- Indels

Repeat expansions

Structural Variants (SVs)
- Insertions
- Deletions
- Duplications
- Inversions
- Translocations

Variant calling from *de novo* assembly



Structural Variants (SVs)

- Insertions
- Deletions
- Duplications
- Inversions
- Translocations

Project 1: Developmental disorders

- Whole Exome Sequencing / Whole Genome Sequencing (WGS)
  - ~ 10-15 % Copy Number Variants (CNVs)
  - ~ 25-30 % SNVs & indels

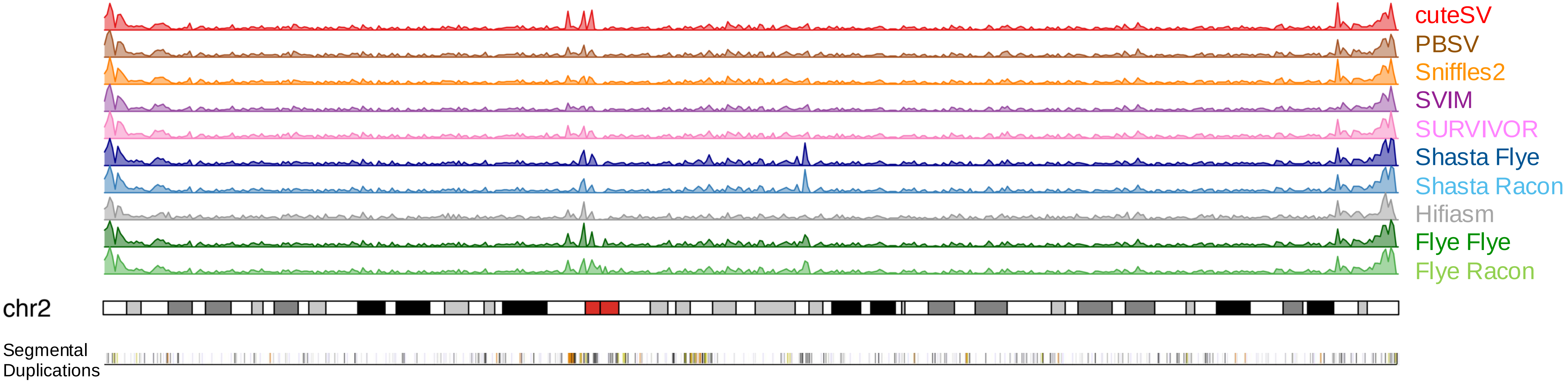
Project 1: Developmental disorders

- Whole Exome Sequencing / Whole Genome Sequencing (WGS)
  - ~ 10-15 % Copy Number Variants (CNVs)
  - ~ 25-30 % SNVs & indels

- ONT sequencing of unsolved trios
  - 27 trios
  - Structural variants (SVs)

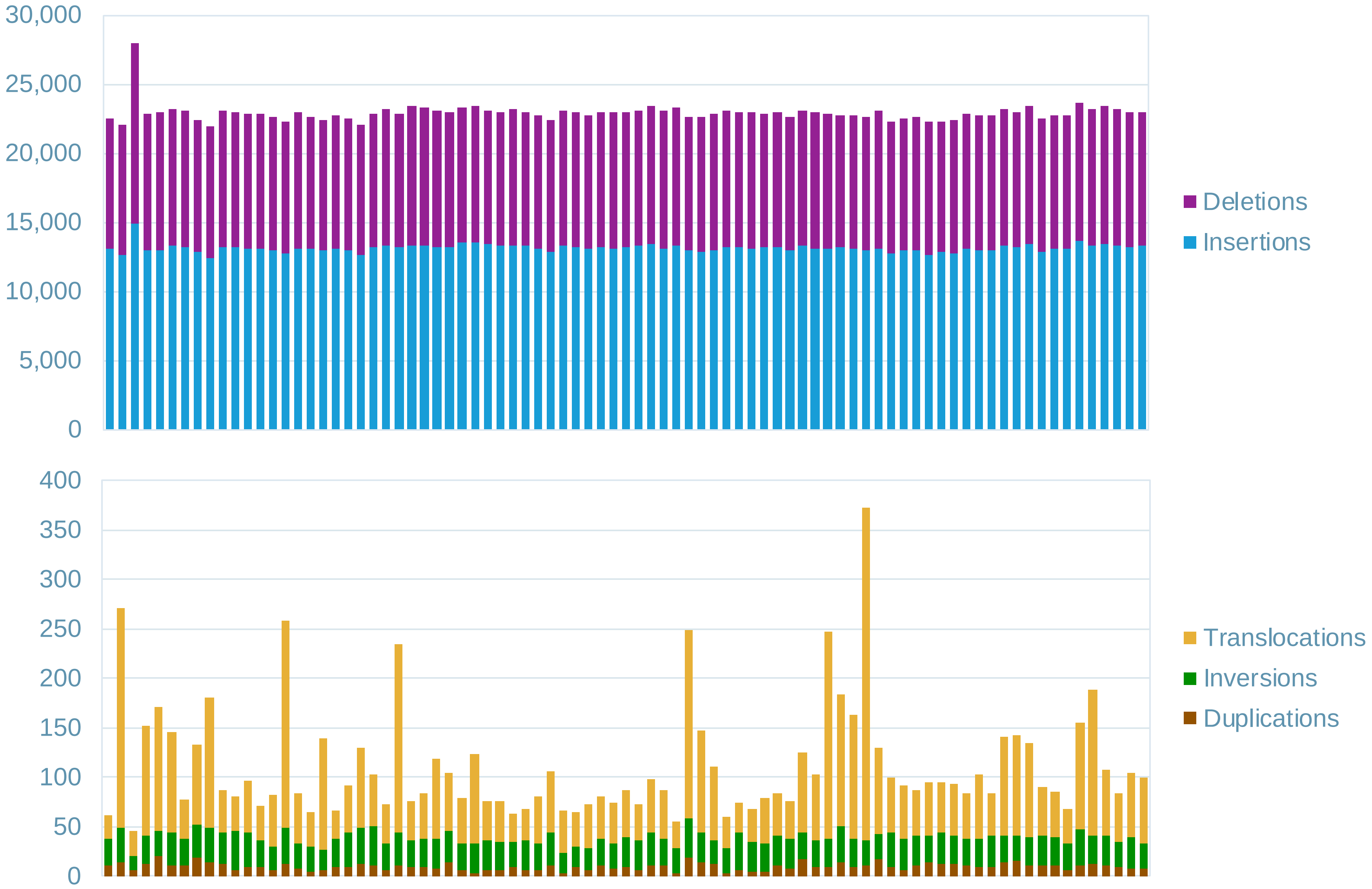
SVs

~ 23,000 SVs per individual



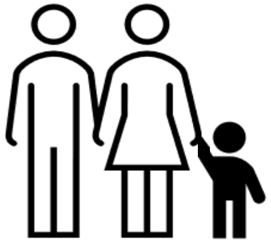
SVs

~ 23,000 SVs per individual



SVs

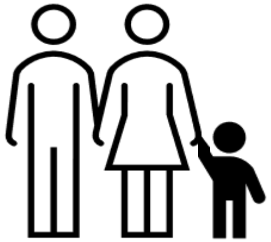
~ 23,000 SVs per individual



~ 100 *de novo* SVs

SVs

~ 23,000 SVs per individual

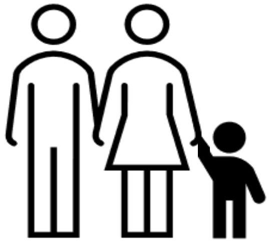


~ 100 *de novo* SVs

~ 7 rare *de novo* SVs

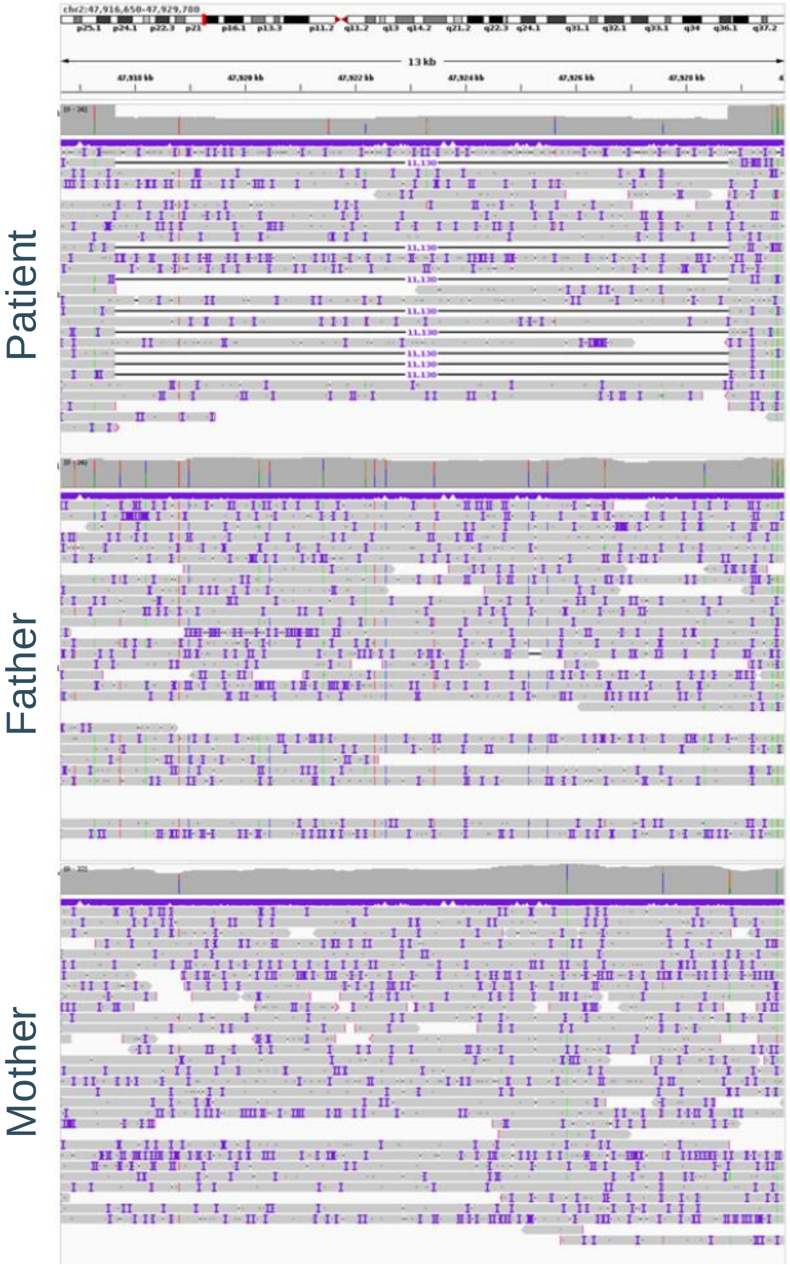
SVs

~ 23,000 SVs per individual



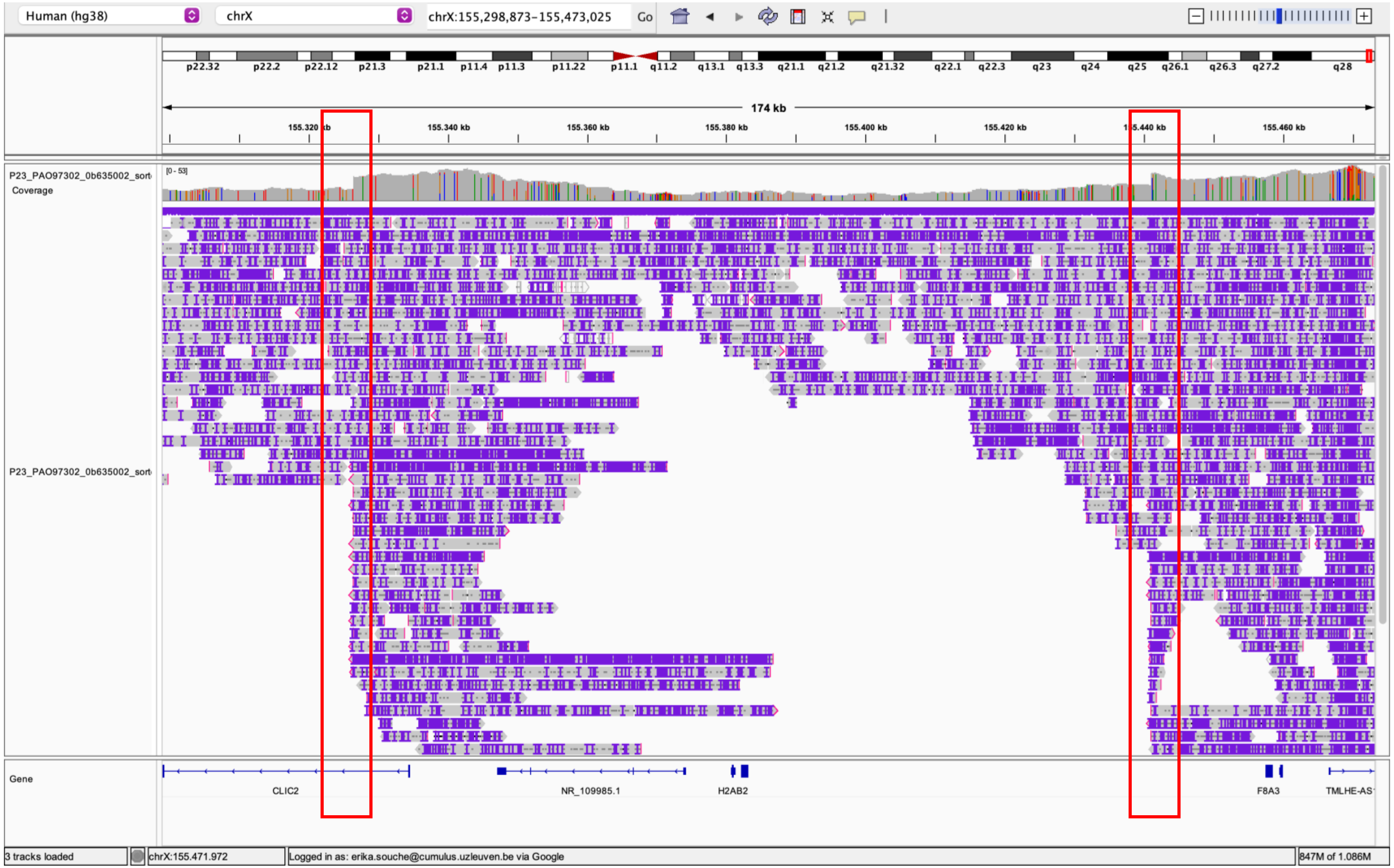
~ 100 *de novo* SVs

~ 7 rare *de novo* SVs



Complex SVs

- Inversion on chrX in a male patient



Human (hg38)

chrX

chrX:155,298,873-155,473,025

Go

3 tracks loaded

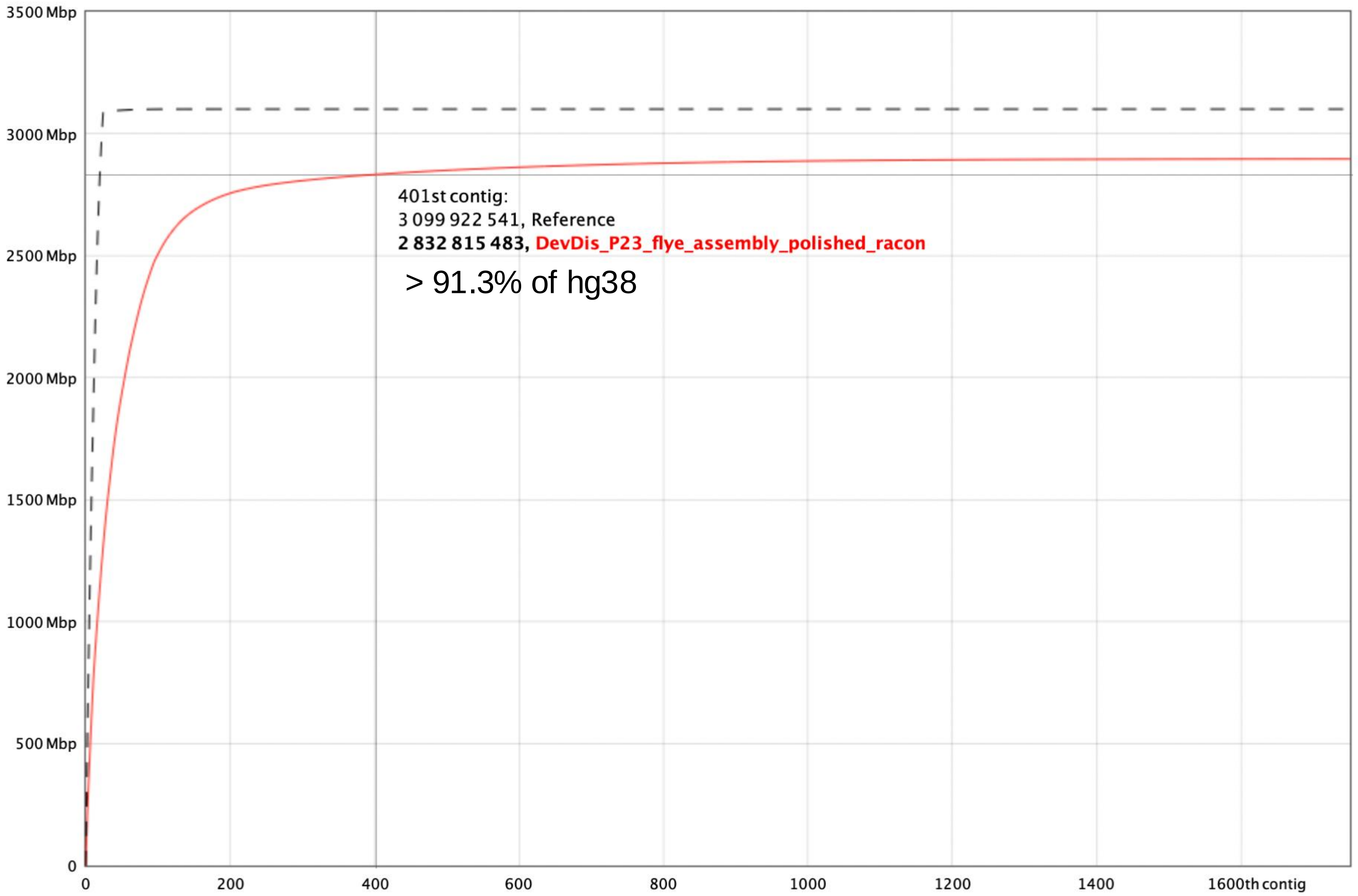
chrX:155.471.972

Logged in as: erika.souche@cumulus.uzleuven.be via Google

847M of 1.086M

Complex SVs

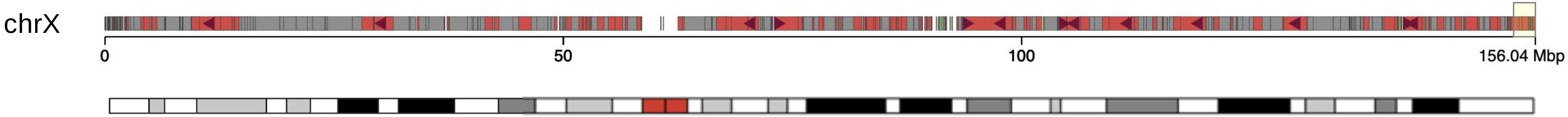
- *de novo* assembly



> 91.3% of hg38

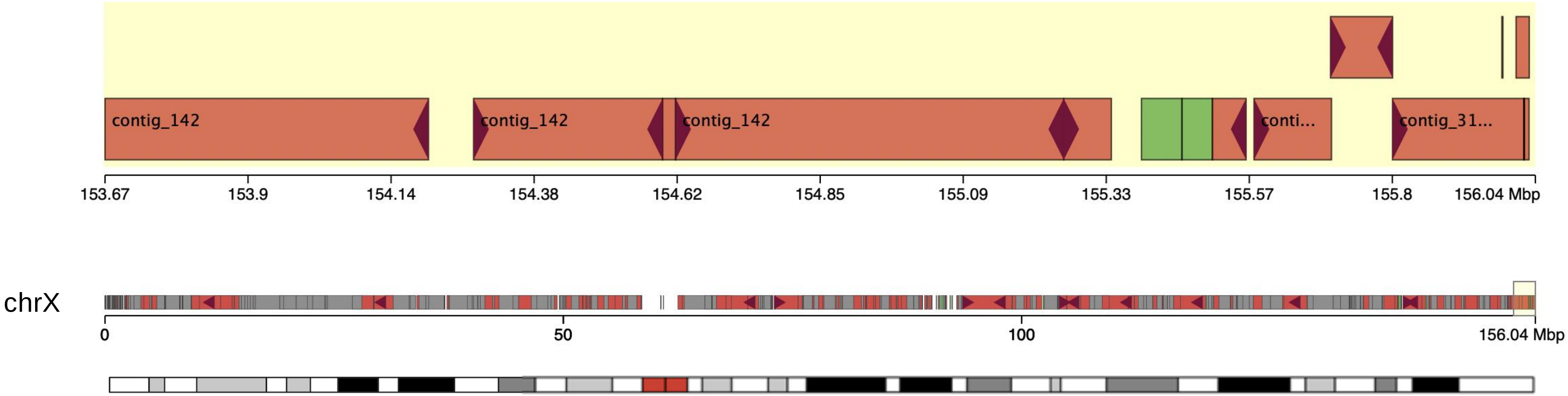
Complex SVs

- *de novo* assembly
- Mapping of contigs to hg38



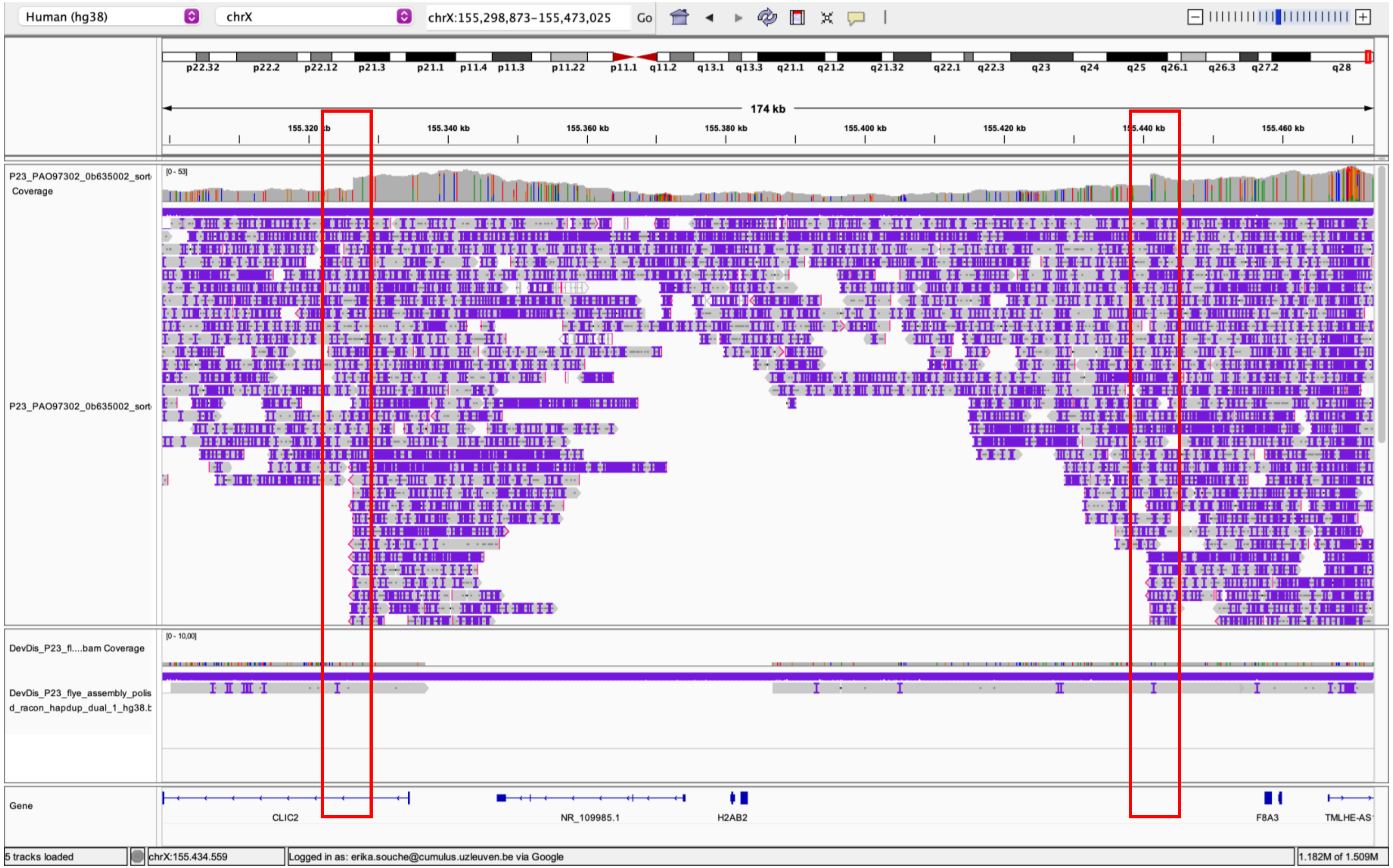
Complex SVs

- *de novo* assembly
- Mapping of contigs to hg38



Complex SVs

- SV locus not covered by *de novo* assembly



Project 1: Developmental disorders

- ONT sequencing of unsolved trios
    - Structural variants (SVs)
    - Repeat expansions (STRs)
    - Epigenic modifications
    - Mosaïcism
    - …

Project 1: Developmental disorders

- ONT sequencing of unsolved trios
  - Structural variants (SVs)
  - Repeat expansions (STRs)
  - Epigenic modifications
  - Mosaïcism
  - …
- Need for longer-term storage for active data

Project 1: Developmental disorders

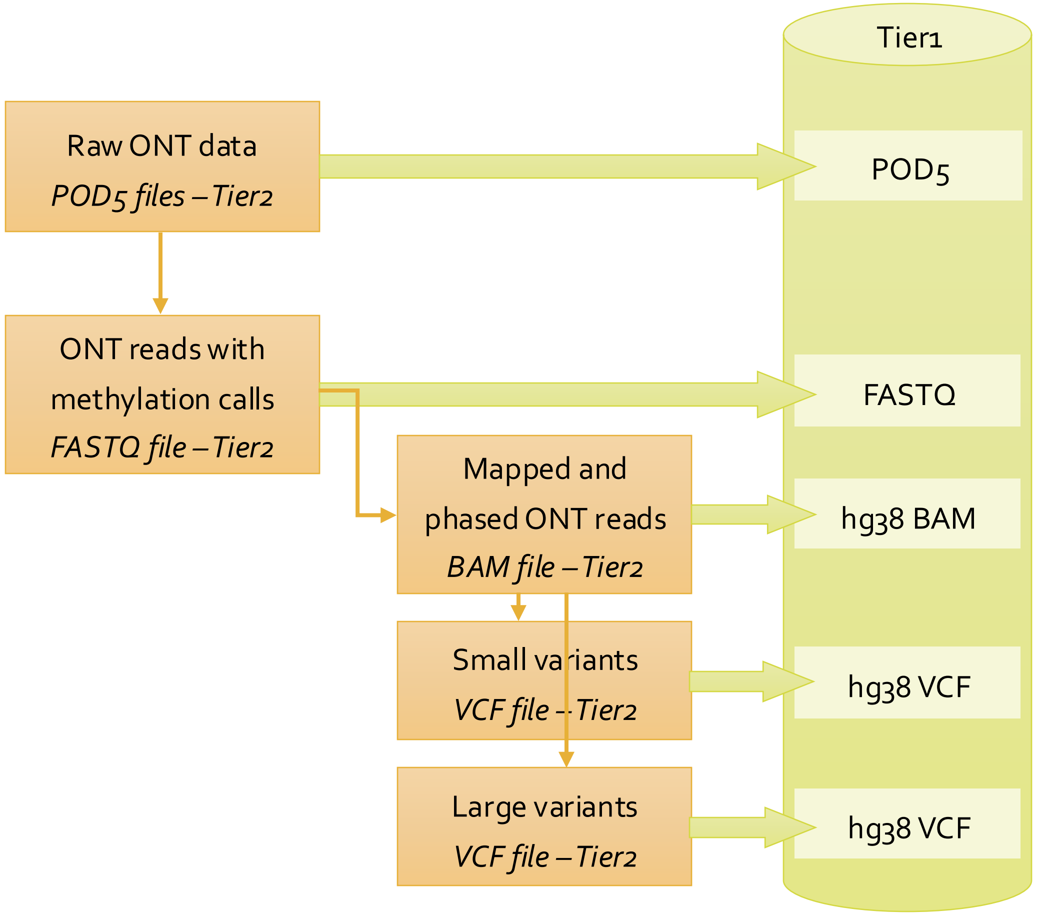
- ONT sequencing of unsolved trios
  - Structural variants (SVs)
  - Repeat expansions (STRs)
  - Epigenic modifications
  - Mosaïcism
  - …
- Need for longer-term storage for active data

Tier1 data

Tier1 data

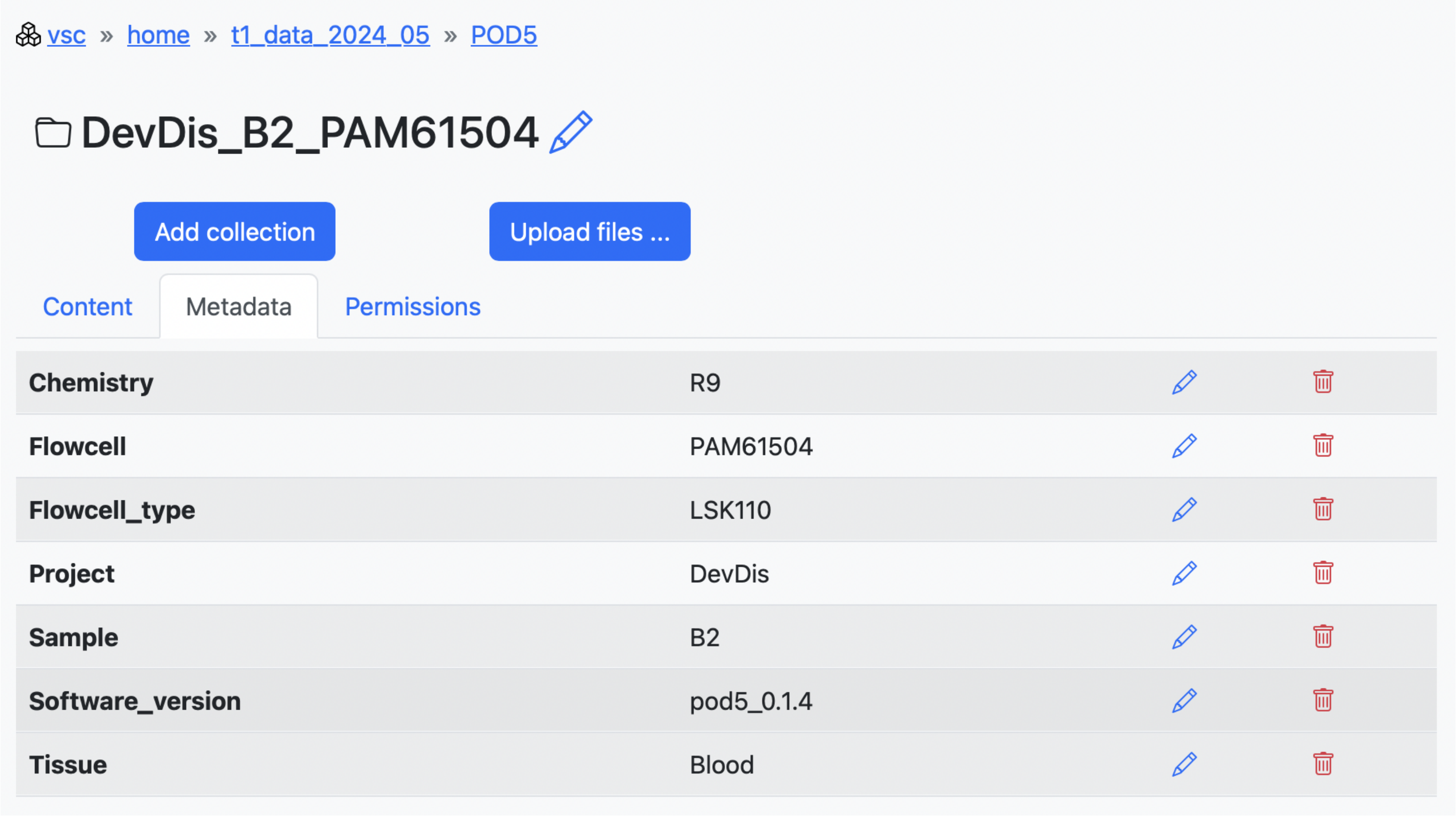
- Active data
- Data workflow
- Data structure

=> More sustainable for research laboratories



Tier1 data

- Metadata
- 4-year plan
  - Number of samples
  - Resource estimation
- Exit strategy

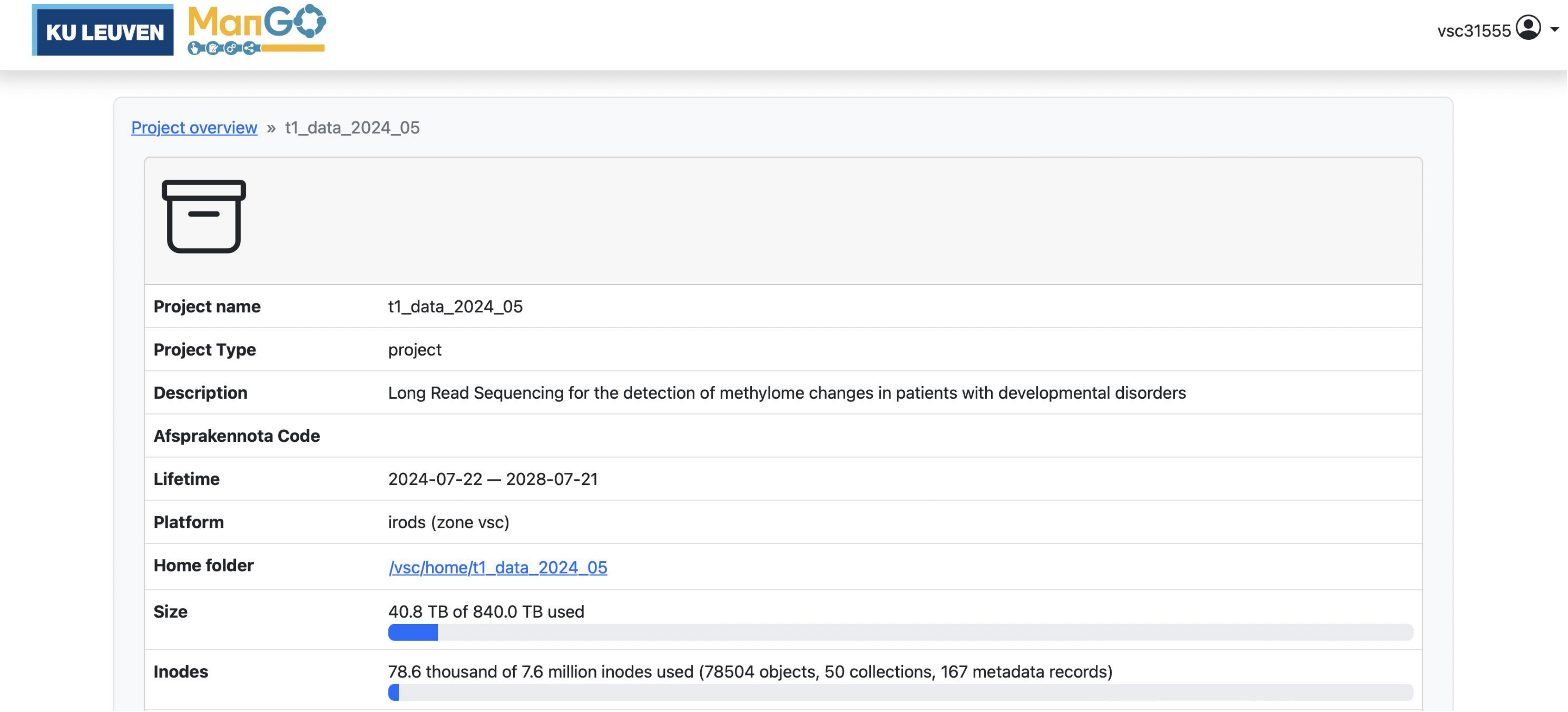


vsc » home » t1_data_2024_05 » POD5

DevDis_B2_PAM61504

Chemistry	R9
Flowcell	PAM61504
Flowcell_type	LSK110
Project	DevDis
Sample	B2
Software_version	pod5_0.1.4
Tissue	Blood

Resource monitoring

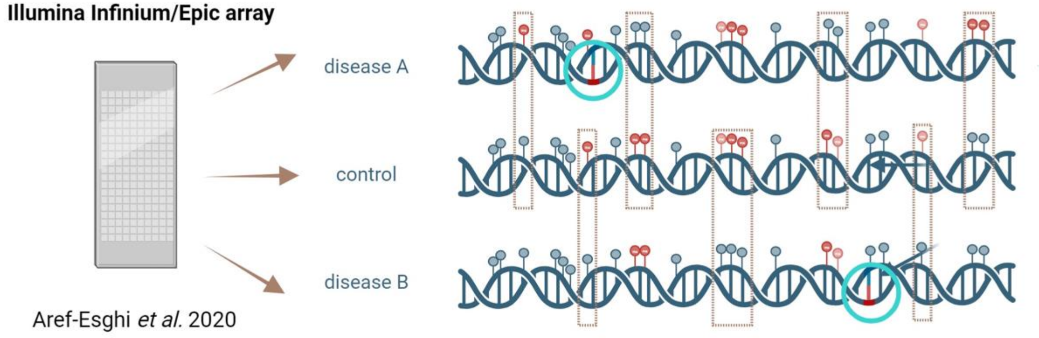


Project overview » t1_data_2024_05

Project name	t1_data_2024_05
Project Type	project
Description	Long Read Sequencing for the detection of methylome changes in patients with developmental disorders
Afsprakennota Code	
Lifetime	2024-07-22 — 2028-07-21
Platform	irods (zone vsc)
Home folder	/vsc/home/t1_data_2024_05
Size	40.8 TB of 840.0 TB used
Inodes	78.6 thousand of 7.6 million inodes used (78504 objects, 50 collections, 167 metadata records)

Project 2: Episignature

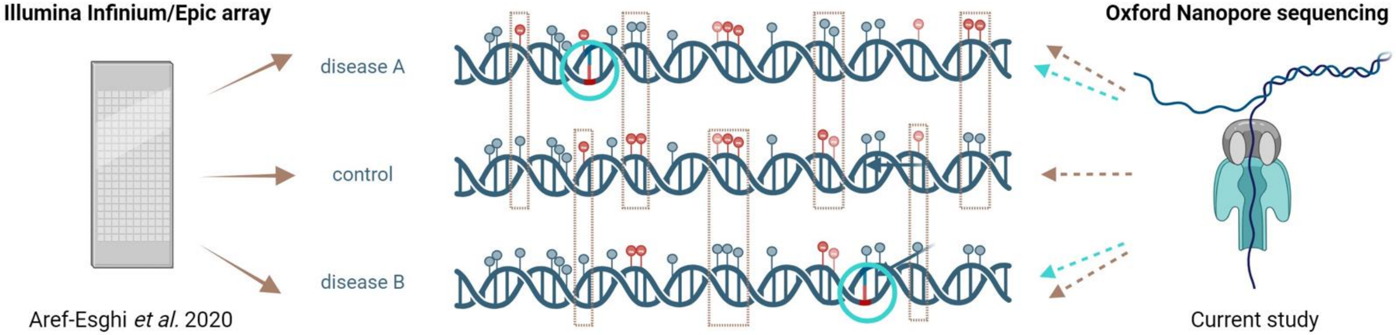
- Neurodevelopmental disorders with methylation disturbances



Talk “Episignature in patients with developmental disorders” by Benjamin Huremagic

Project 2: Episignature

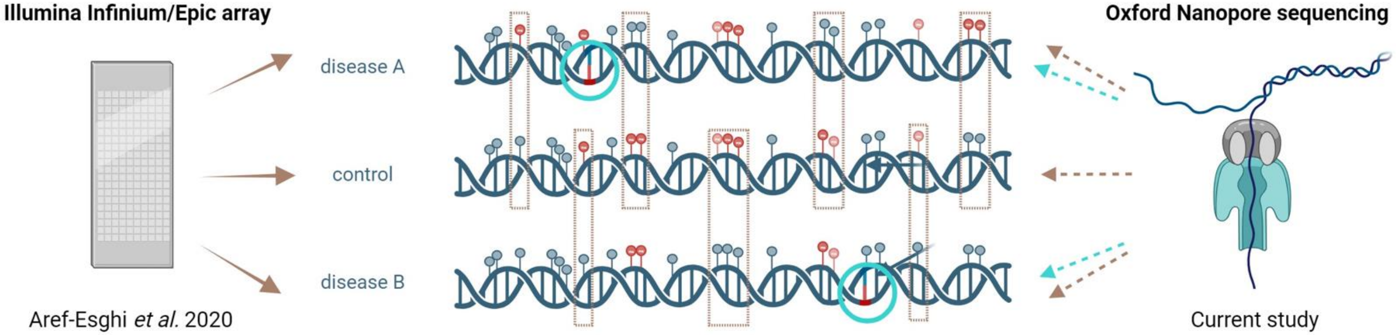
- Neurodevelopmental disorders with methylation disturbances
- Proof of concept
  - 20 patients
  - 40 control samples



Talk "Episignature in patients with developmental disorders" by Benjamin Huremagic

Project 2: Episignature

- Neurodevelopmental disorders with methylation disturbances
- Proof of concept
  - 20 patients
  - 40 control samples



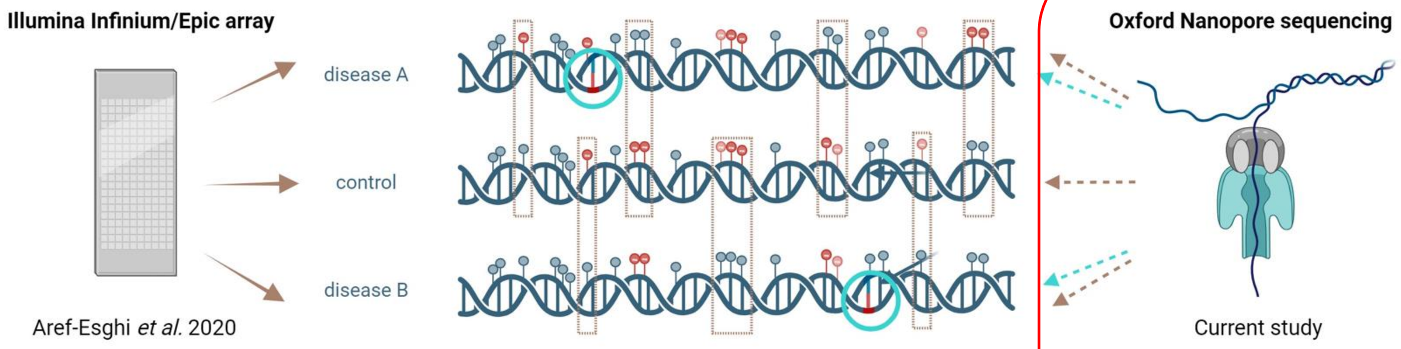
- Extend model

  => Processing of 210 samples

Talk “Episignature in patients with developmental disorders” by Benjamin Huremagic

Project 2: Episignature

- Neurodevelopmental disorders with methylation disturbances
- Proof of concept
  - 20 patients
  - 40 control samples



Aref-Esghi *et al.* 2020

Need GPU processing to get methylation calls

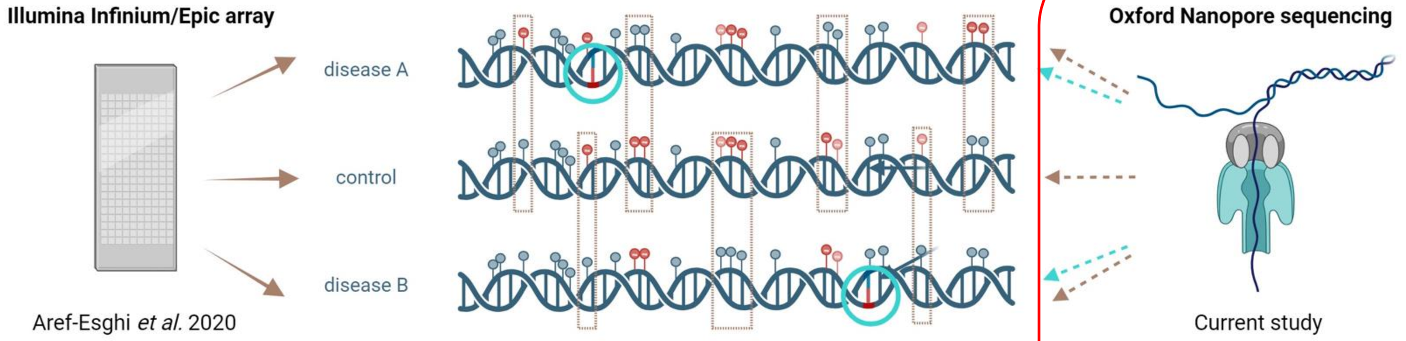
- Extend model

  => Processing of 210 samples

Talk “Episignature in patients with developmental disorders” by Benjamin Huremagic

Project 2: Episignature

- Neurodevelopmental disorders with methylation disturbances
- Proof of concept
  - 20 patients
  - 40 control samples



Aref-Esghi *et al.* 2020

Current study

Need GPU processing to get methylation calls

- Extend model

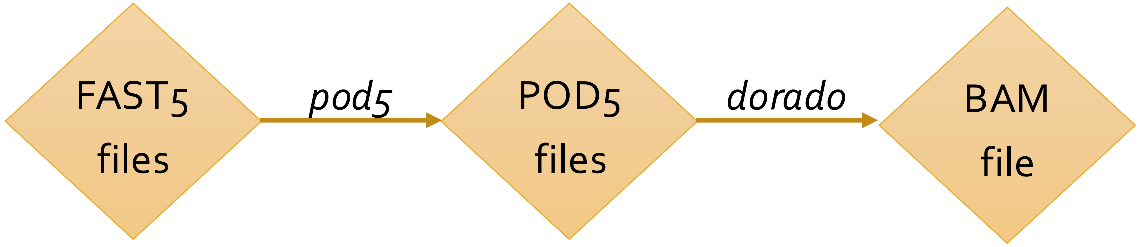
  => Processing of 210 samples

Tier1 compute

Talk “Episignature in patients with developmental disorders” by Benjamin Huremagic

Tier1 compute

- Pipeline description



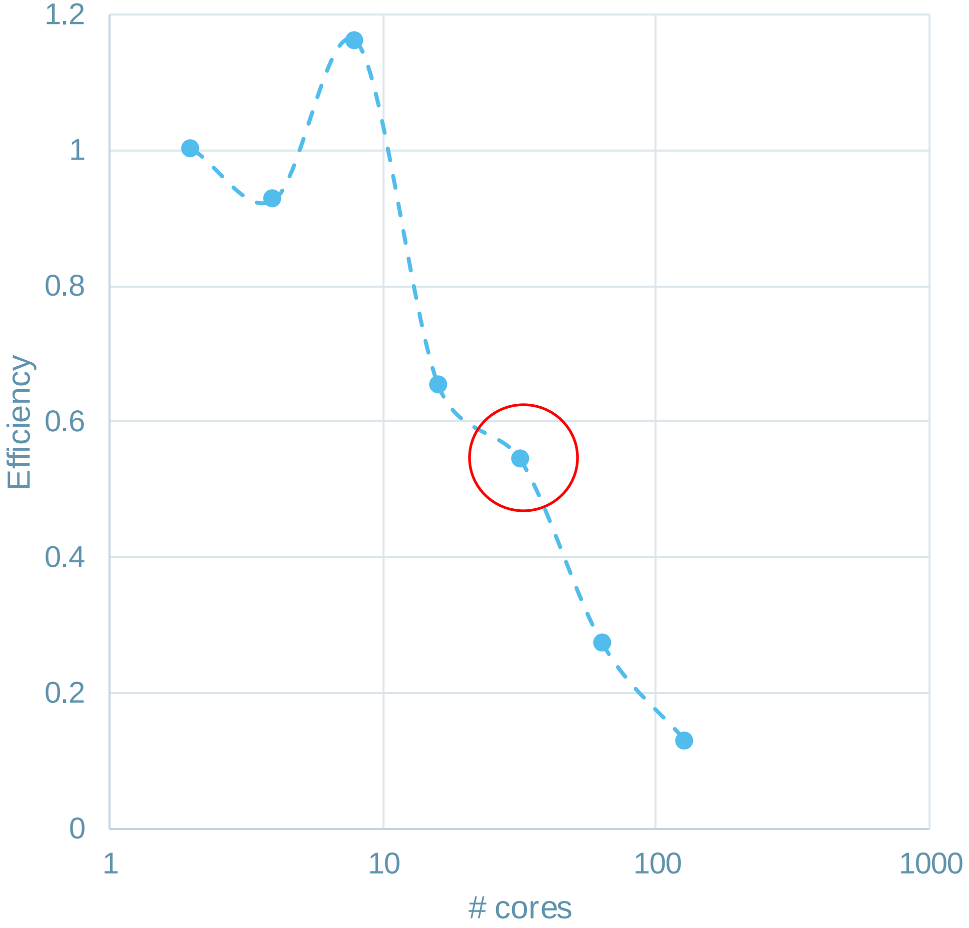
- Resource optimisation

  => Starting grant

Resource optimisation

- FAST5 to POD5

Number of nodes	Total number of cores	Wall clock time (s)	Speed-up (w.r.t. baseline)	Efficiency
1	1	16128	Not applicable	Not applicable
1	2	12566	1	1
1	4	6778	1.85	0.93
1	8	2708	4.64	1.16
1	16	2407	5.22	0.65
1	32	1445	8.70	0.54
1	64	1449	8.67	0.27
1	128	1548	8.12	0.13



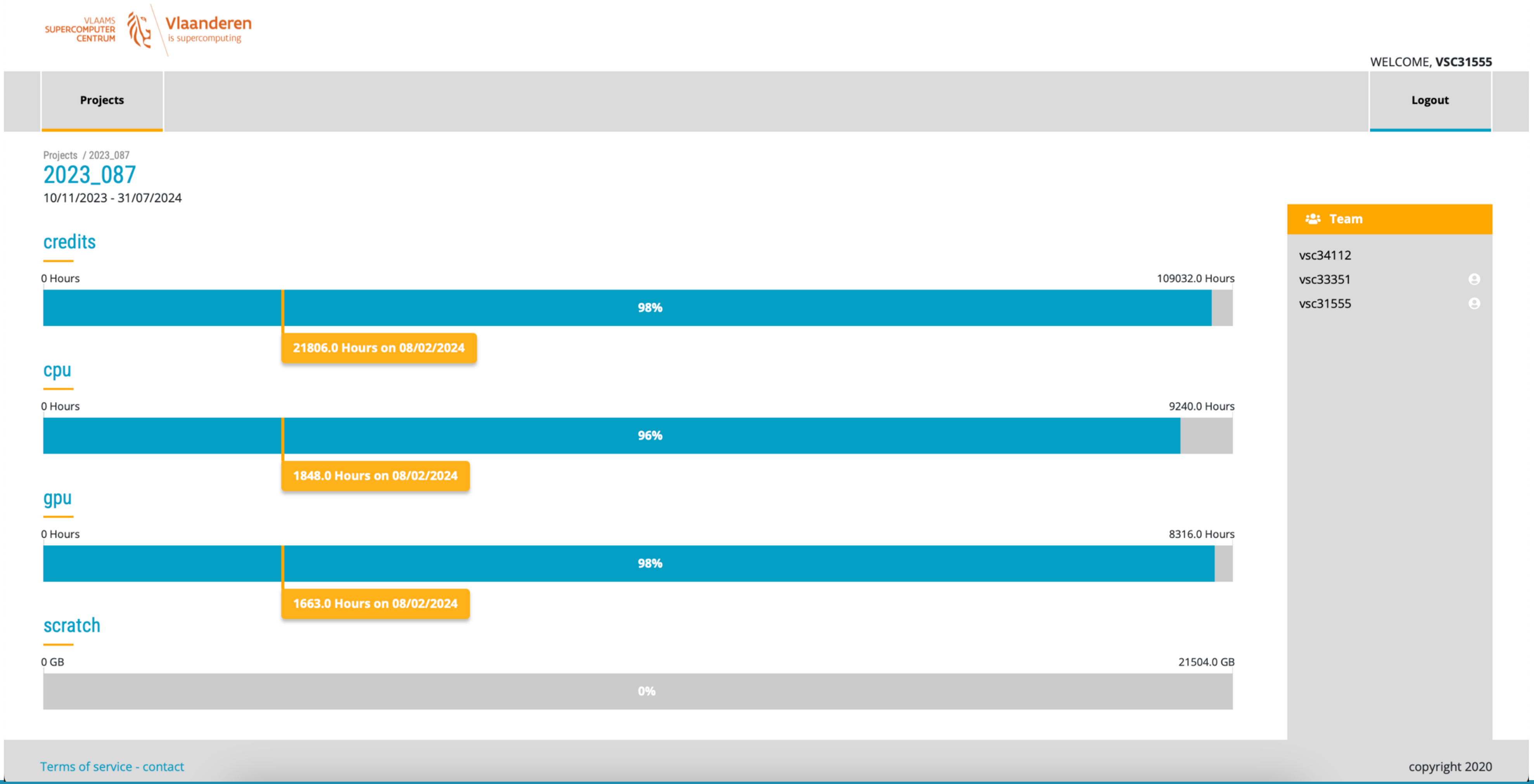
Resource estimation

- Based on
  - Results of efficiency tests
  - Number of samples to process

Tier1 compute

9,240 CPU hours
8,316 GPU hours
21 TB scratch volume

Resource monitoring



WELCOME, **VSC31555**

Projects / 2023_087

2023_087

10/11/2023 - 31/07/2024

credits

0 Hours — 109032.0 Hours — 98% — 21806.0 Hours on 08/02/2024

cpu

0 Hours — 9240.0 Hours — 96% — 1848.0 Hours on 08/02/2024

gpu

0 Hours — 8316.0 Hours — 98% — 1663.0 Hours on 08/02/2024

scratch

0 GB — 21504.0 GB — 0%

Team

- vsc34112
- vsc33351
- vsc31555

Terms of service - contact

copyright 2020

Acknowledgements

Mathilde Geysens

Benjamin Huremagic

Chiara Campanelli

Greet Peeters

Senne Meynants

Natalia Olszewska

Kris Van Den Bogaert

Joris Vermeesch

Qiang Fu

Tatjana Jatsenko

Marta Sousa Santos

Kate Elizabeth Stanley

Olga Tsuiko

Stefania Tuveri

Yan Zhao

Jonas Demeulemeester

VLAAMS
SUPERCOMPUTER
CENTRUM

Vlaanderen
is supercomputing

Tools used

Mapping

Minimap2

Quality control

mosdepth

NanoPlot

QUAST

de novo assembly

shasta

flye

racon

hifiasm

Verkko

HapDup

Structural variant calling

PBSV

SVIM

Sniffles2

cuteSV

QDNAseq

DipDiff

SURVIVOR

Repeat expansion calling

straglr

TRGT

Small variant calling

Clair3

DeepVariant

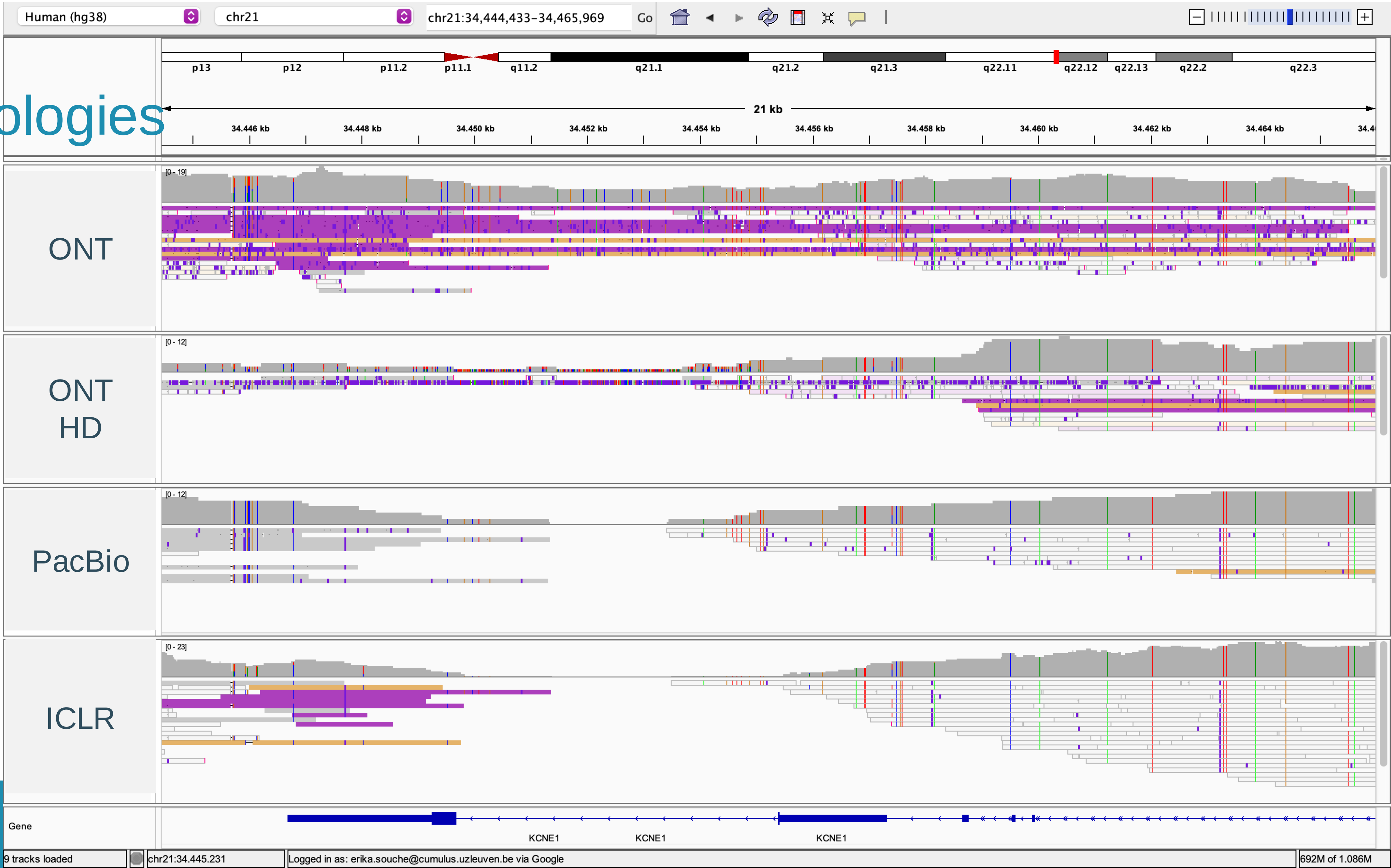
WhatsHap

Data footprint

	Format	ONT	PacBio	ICLR
Raw data	FAST5/POD5 FASTQ.gz	1 TB 100 GB	- 50 GB	- 40 GB *
Aligned reads	BAM CRAM	100 GB 50 GB	50 GB 15 GB	25 GB 5 GB
Small variants	VCF.gz	100 MB	100 MB	100 MB
Structural variants	VCF	50 MB	50 MB	50 MB
Repeat expansions	VCF	10 MB	10 MB	10 MB
De novo assembly	FASTA	3 GB	3 GB	3 GB

* After processing by Illumina

All technologies


Human (hg38)

chr21

chr21:34,444,433-34,465,969

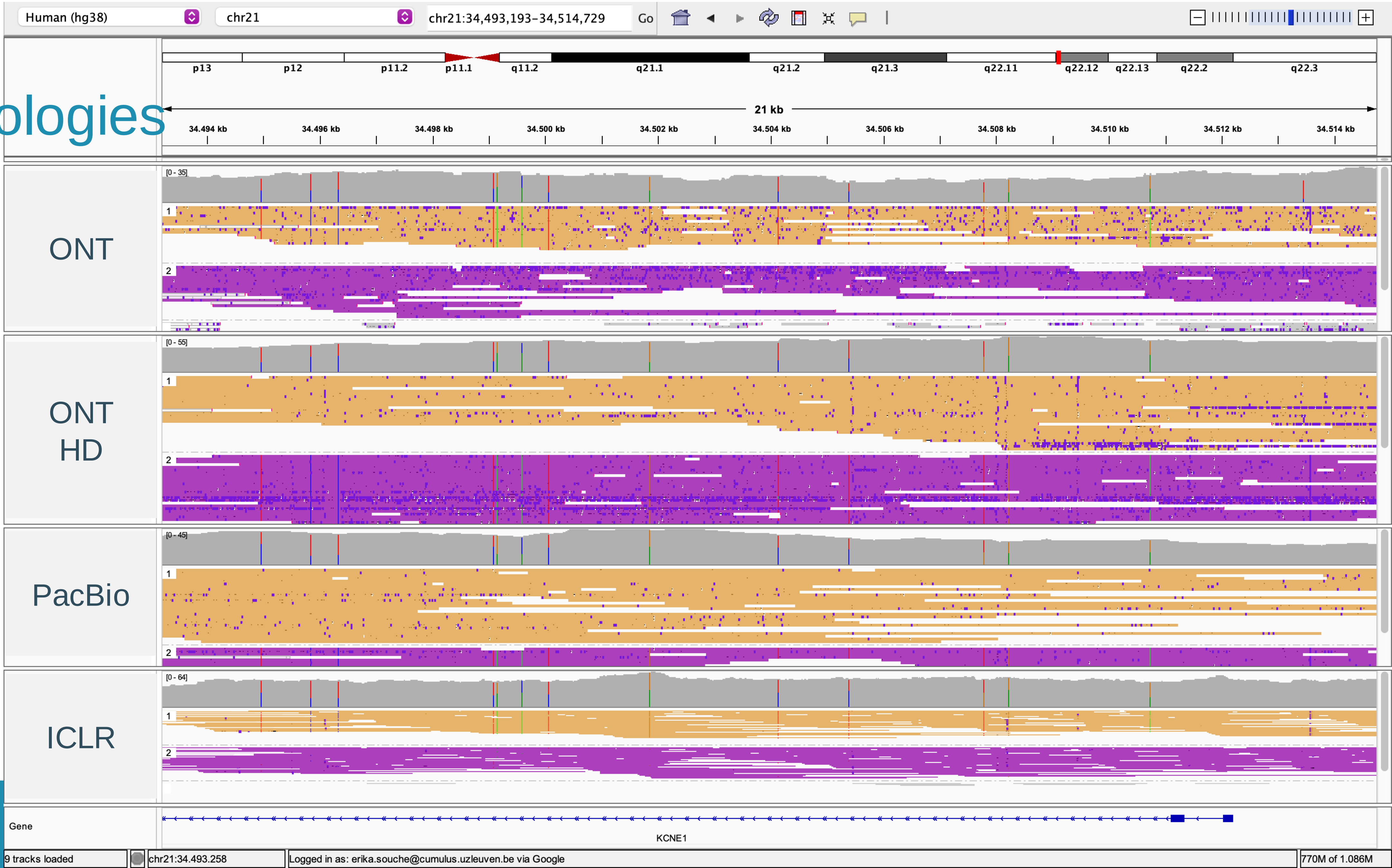
9 tracks loaded

chr21:34.445.231

Logged in as: erika.souche@cumulus.uzleuven.be via Google

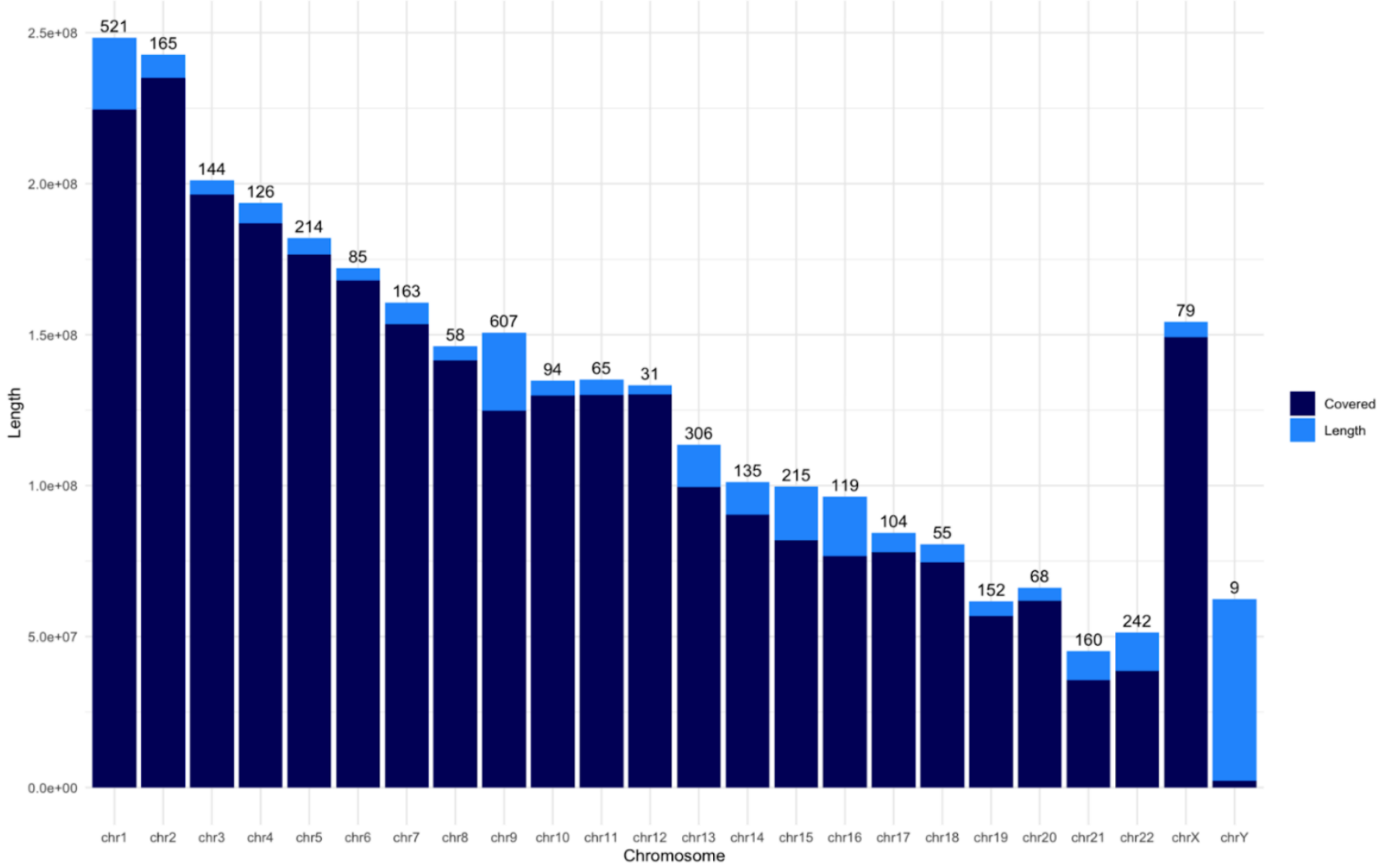
692M of 1.086M

All technologies



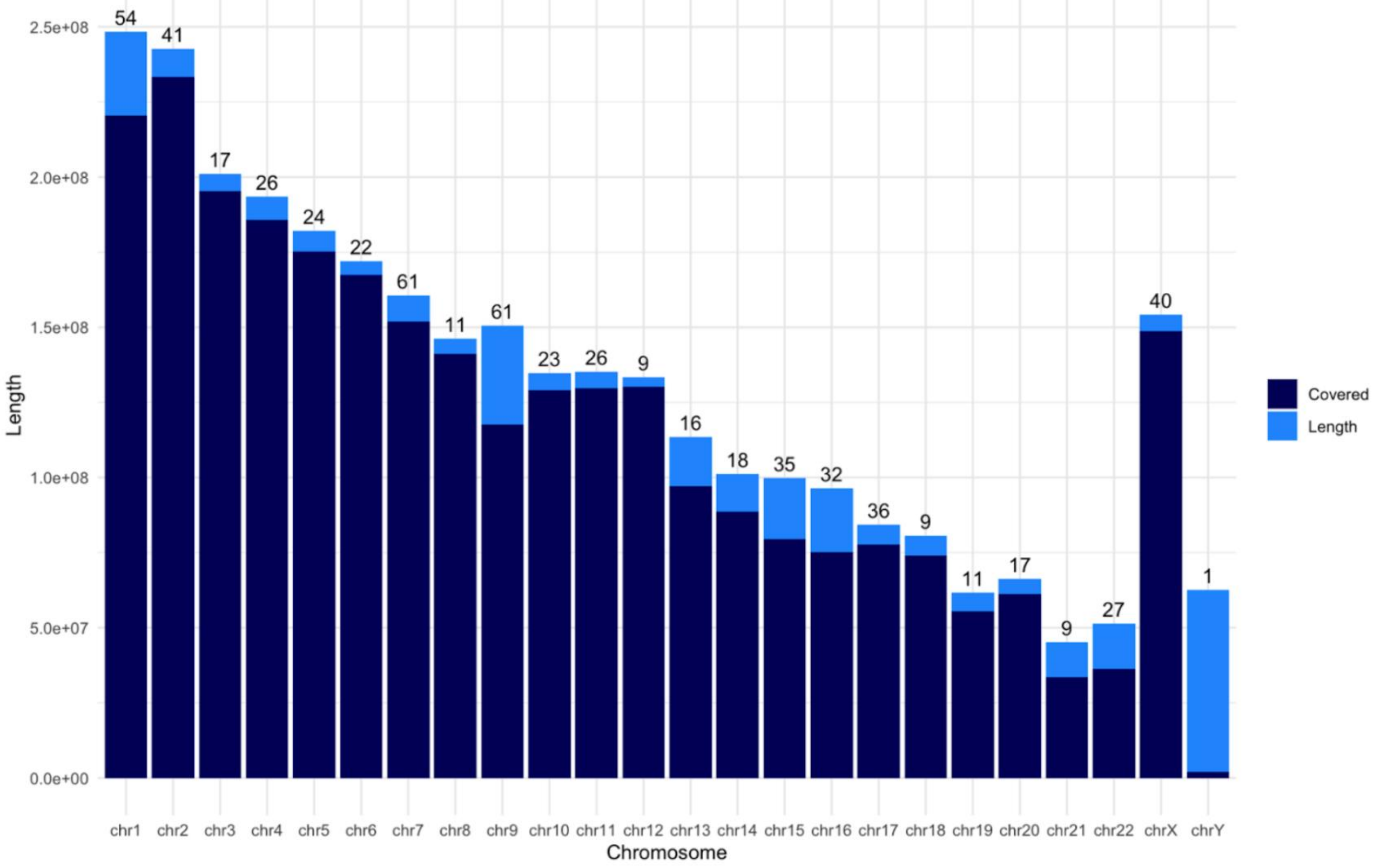
de novo assembly

- 2,727 contigs
- ~86.7% of T2T covered

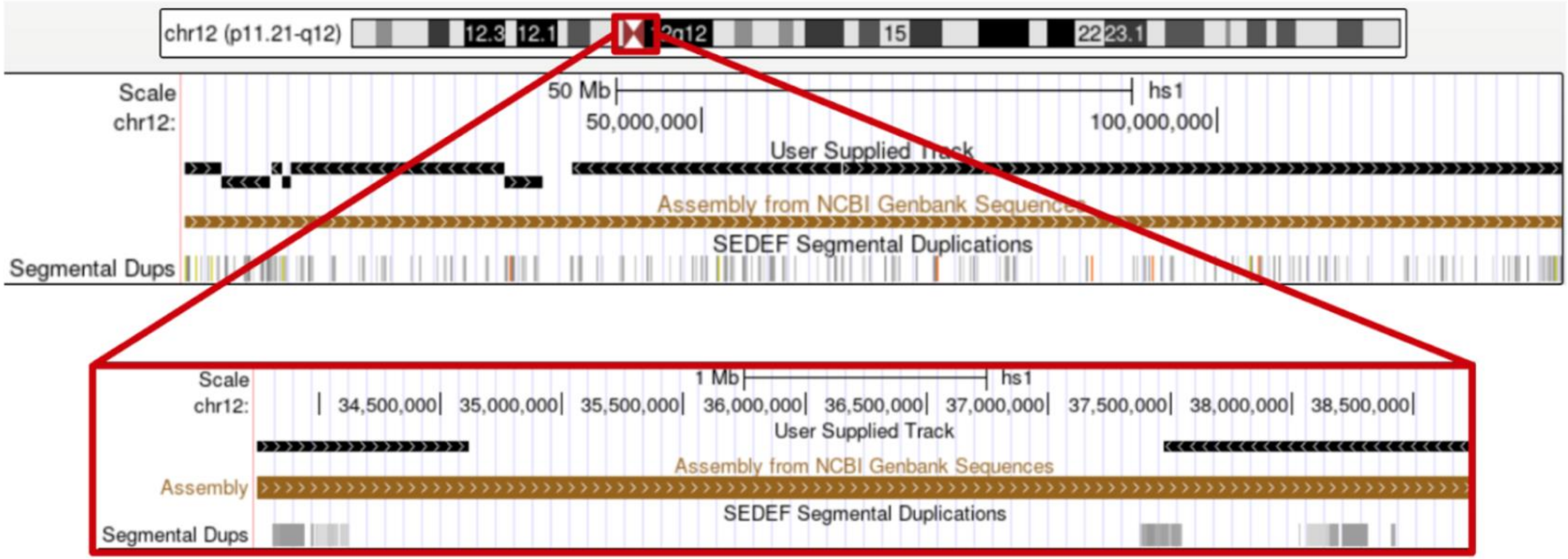


de novo assembly

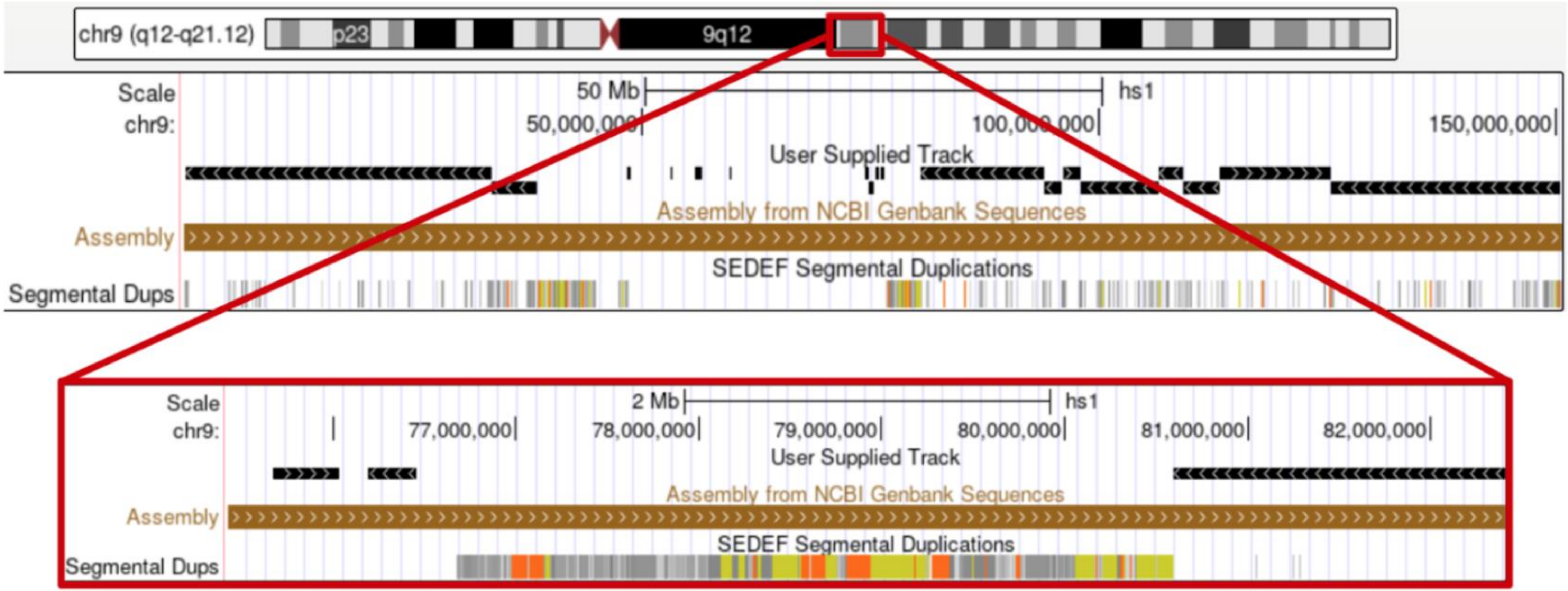
- 626 contigs > 100 kb
- ~86.67% of T2T covered



de novo assembly



de novo assembly


chr9 (q12-q21.12)
p23
9q12
Scale
chr9:
50 Mb
hs1
50,000,000
100,000,000
150,000,000
User Supplied Track
Assembly from NCBI Genbank Sequences
Assembly
SEDEF Segmental Duplications
Segmental Dups
Scale
chr9:
2 Mb
hs1
77,000,000
78,000,000
79,000,000
80,000,000
81,000,000
82,000,000
User Supplied Track
Assembly from NCBI Genbank Sequences
Assembly
SEDEF Segmental Duplications
Segmental Dups