Integrating structural variants in genomic studies with long-read sequencing and pangenomes

Jean Monlong
23/05/2025

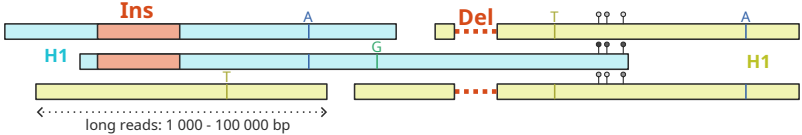
CRCT Bioinformatics Seminar

Outline

Introduction: genome sequencing and genomic variants

Structural variants and rare disease with long-read sequencing

Structural variants and complex disease with pangenomes



Introduction: genome sequencing and genomic variants

Different types of genomic variants

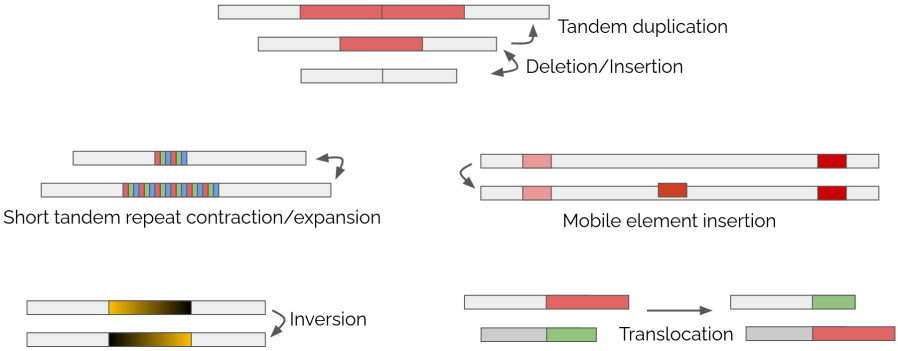
Different types of genomic variants

Single-nucleotide polymorphisms (**SNPs**)	Insertion-deletion polymorphisms (**INDELs**)	Structural variants (**SVs**)
GAT**C**AGC	GAT**CA**GC	GATCAGC
GAT**G**AGC	GAT - - GC	GATC AGC
		CGC....300bp....GAT

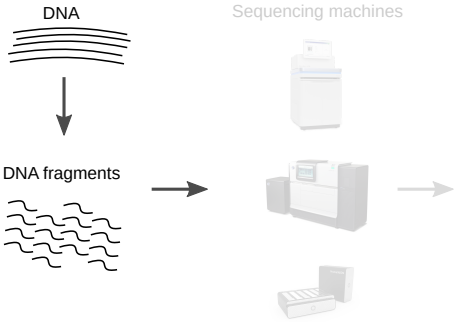
Different types of genomic variants

Single-nucleotide polymorphisms (**SNPs**)	Insertion-deletion polymorphisms (**INDELs**)	Structural variants (**SVs**)
GAT**C**AGC	GAT**CA**GC	GATCAGC
GAT**G**AGC	GAT - - GC	GATC AGC

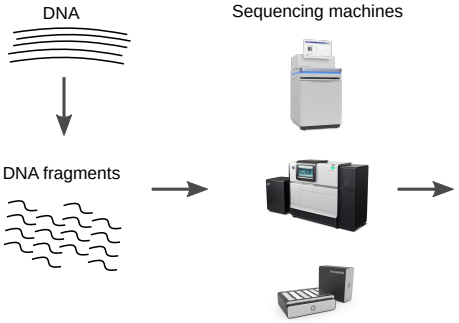
Structural variants (SVs) come in diverse shapes and sizes



Genome sequencing



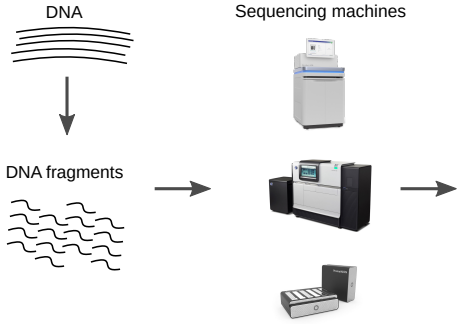
Genome sequencing



File (~100-300 Gb)

```
@ERR903030.219 HWI-D00574:82:C6L01ANXX:3:1101:3953:1913/1
AGCTCTTATTATTTTGAAATATGTCCCATCAATACCTAATTTATTGAGAGTTTTTAGCATGAAGGGTTGTTGAAT
+
<<A07FGGGGGGGGGGGEGGGGGGGGGGGGGGG:CFCGDFGG>FG1F@0:FGGGGGF@>FG1=F0FFGGGGFGGC
@ERR903030.220 HWI-D00574:82:C6L01ANXX:3:1101:3863:1914/1
ACCATGAAAGACAGGTGTTAGAATCAGTACAAGAAGCAACAGGGAGCCATTGCATTTTGAGCATTTGTTGTATCA
+
<ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@ERR903030.221 HWI-D00574:82:C6L01ANXX:3:1101:3906:1914/1
GATGGGGTTTCACATTGGCCAGGCTGGTCTCAAACTCCTAACCTCAAGGGATCCACCCACCTCGGCCTCCCAAAG
+
<<AAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG>
@ERR903030.222 HWI-D00574:82:C6L01ANXX:3:1101:3833:1922/1
TACTGAAGAATCAGTGTCAGTTTTGTTAGTTGTGATAATGACATTCTGCTAAGCTAAAGTATAGAGGGGAGAAAT
+
<<BBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@ERR903030.223 HWI-D00574:82:C6L01ANXX:3:1101:3962:1927/1
AAAAAATGAACTAAAAATGCATTAAGACCAAATGTAATACCTAAAAATGTAAAACTTTTAGAAGGAAACATAGG
+
<ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@ERR903030.224 HWI-D00574:82:C6L01ANXX:3:1101:3920:1929/1
ATTCCATTCGATTCCATTCGATGATTCCATTGGATTCCATTCGATGATGATTCCATTCAAATCCATTCGATAATT
+
<ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
...
...
```

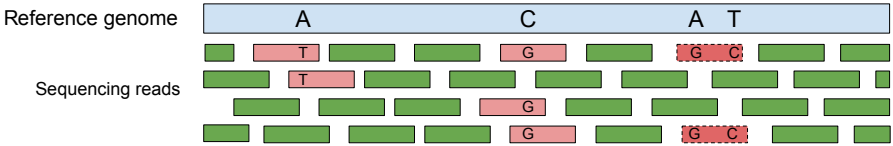
Genome sequencing



Sequencing reads

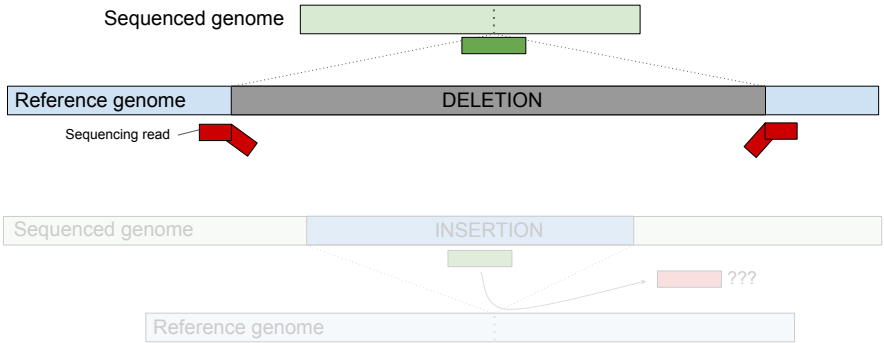
- 150-250 bp (current tech)
- 10,000s-100,000s bp (new tech. $$$)

Aligning reads to a reference genome

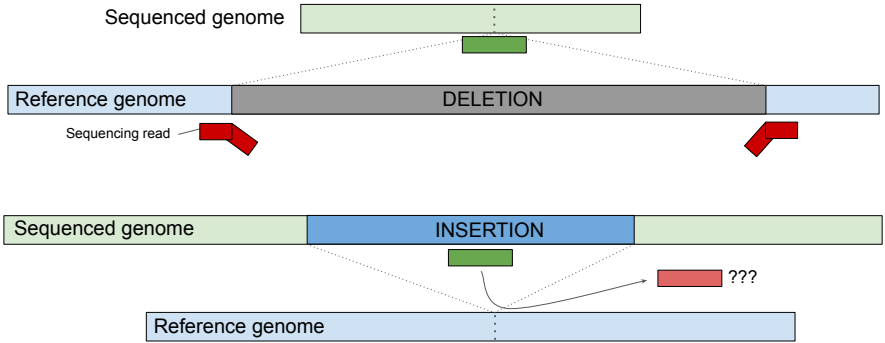


Assuming the reads are correctly placed, small variants are identified as recurrent differences between reads and the reference genome.

The challenges of structural variant detection

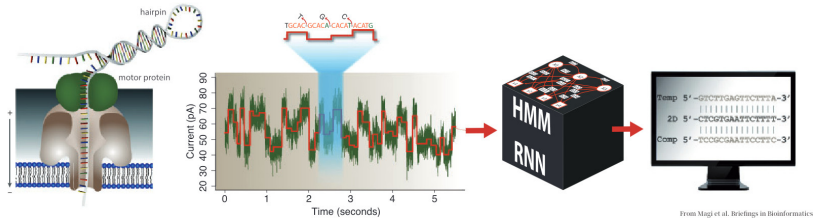


The challenges of structural variant detection



Structural variants and rare disease with long-read sequencing

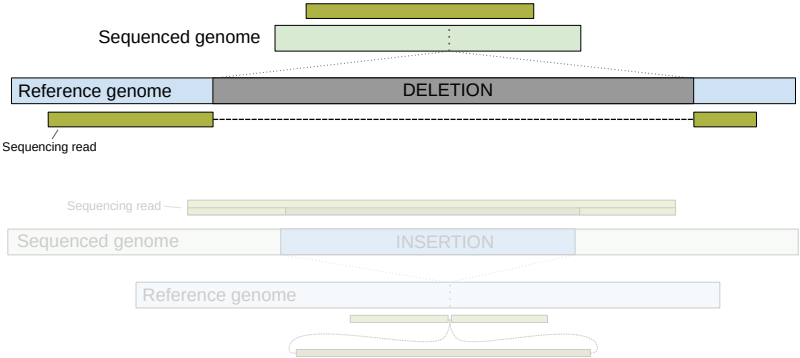
Long-read sequencing with Oxford Nanopore Technologies



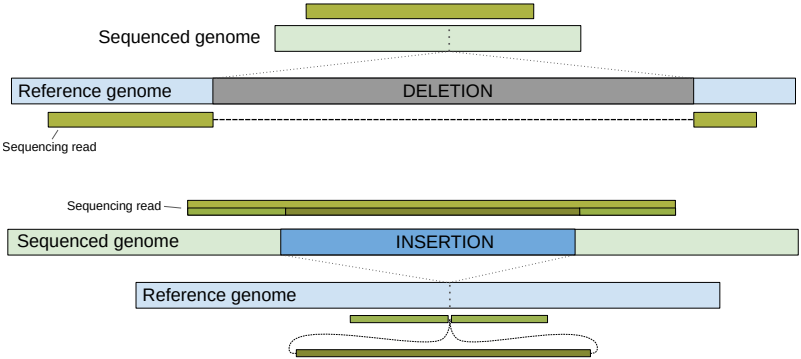
As the DNA (or RNA) fragment passes through the pore, the current changes and is decoded to predict nucleotides.

Reads length of 1,000s-100,000s of nucleotides.

Longer reads improve structural variant detection

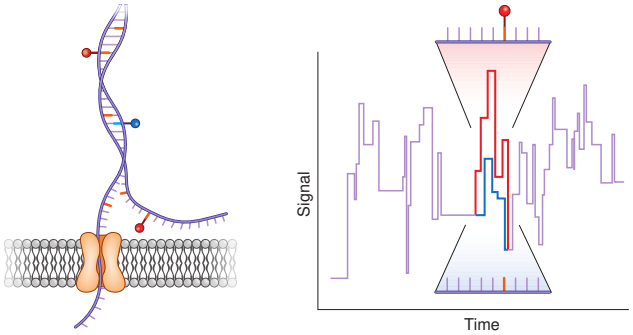


Longer reads improve structural variant detection



Nanopore sequencing can detect DNA/RNA modifications

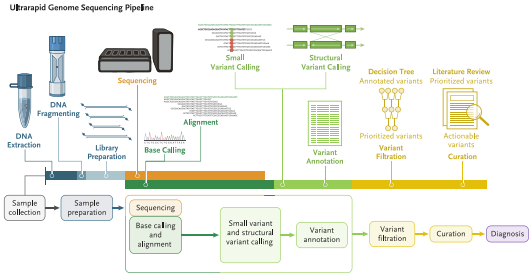
- **5-methylcysosine** (5mC) for DNA/RNA
- 4-methylcysosine (4mC) for DNA
- N^6-Methyladenine (6mA) for DNA/RNA



Schatz, Nature Methods 2023

ONT is portable (space!) and fast

- Sequence as fast as possible
- Get a genomic diagnosis quick
- E.g. for newborns with suspicion of a rare genetic disease

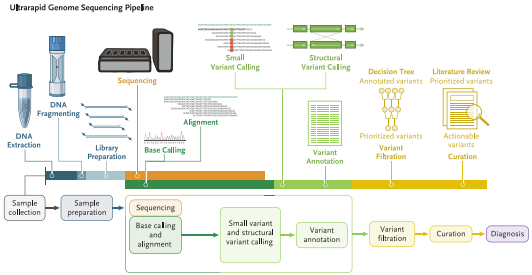


Gorzynski et al. N. Engl. J. Med. 2022

Goenka, Gorzynski, Shafin, et al. Nat. Biotechnol. 2022

ONT is portable (space!) and fast

- Sequence as fast as possible
- Get a genomic diagnosis quick
- E.g. for newborns with suspicion of a rare genetic disease



Gorzynski et al. N. Engl. J. Med. 2022

Goenka, Gorzynski, Shafin, et al. Nat. Biotechnol. 2022

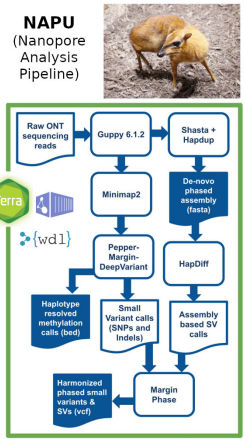
"Fastest DNA sequencing technique": 5h2m

Cost-efficient Nanopore pipeline

- Only **one flow-cell** of Nanopore
- ~30X coverage with 30 Kbp N50 reads

Cost-efficient Nanopore pipeline

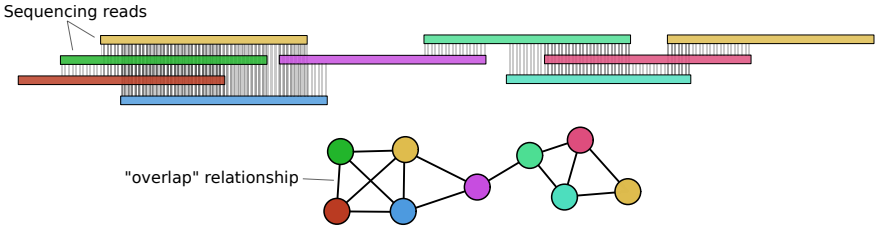
- Only **one flow-cell** of Nanopore
- ~30X coverage with 30 Kbp N50 reads
- Nanopore Analysis Pipeline (U?) to get haplotype resolved:
  1. small variants (SNPs/indels)
  2. structural variants
  3. *de novo* assembly
  4. methylation marks



Kolmogorov, Billingsley, et al. Nature Methods 2023

Longer reads enable *de novo* genome assembly

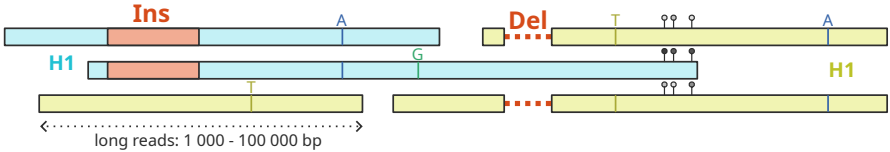
Reconstructs genomes without reference bias, hence better able to identify complex variants (e.g. combination of deletion/inversion)



The Shasta assembler is an overlap-layout-consensus assembler for Nanopore reads.

Shafin, Pesout, Lorig-Roach, Haukness, Olsen, et al. Nat. Biotechnol. 2020

Phased variants and methylation calls

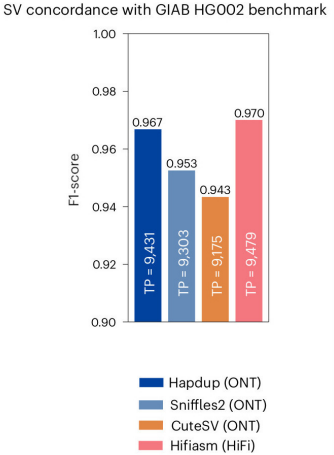


Reads are **haplo-tagged** using information across heterozygous sites.

- Phased structural variants with Hapdup
- Phased small variants with DeepVariant
- Phased methylation calls with ModKit

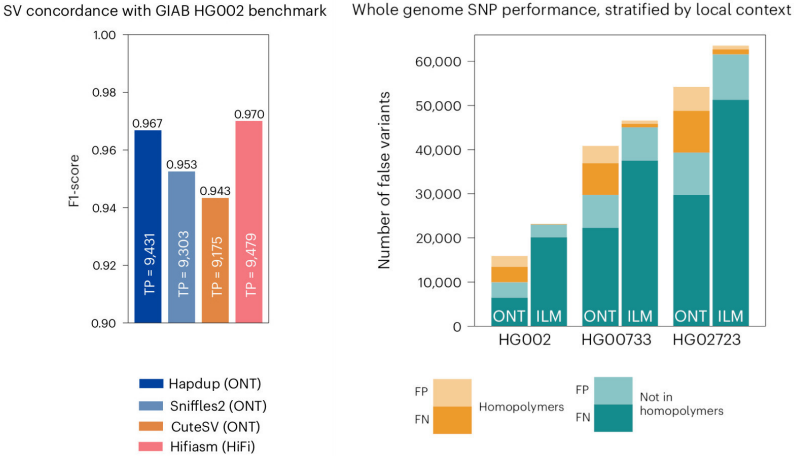
Kolmogorov, Billingsley, et al. Nature Methods 2023

Better calls for both small and structural variants...



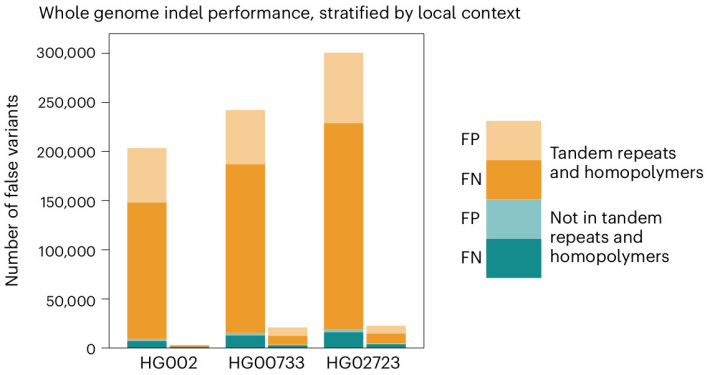
Kolmogorov, Billingsley, et al. Nature Methods 2023

Better calls for both small and structural variants...



Kolmogorov, Billingsley, et al. Nature Methods 2023

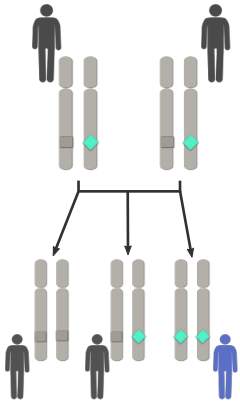
...except for indels in homopolymers



Note: Results above are for the R9 chemistry. The new R10 chemistry has lower error rate and better (indel) calling performance.

Kolmogorov, Billingsley, et al. Nature Methods 2023

Pathogenic variants in undiagnosed rare disease patients



Goal

- Identify as many variants as possible
- All types, all sizes, across the whole genome

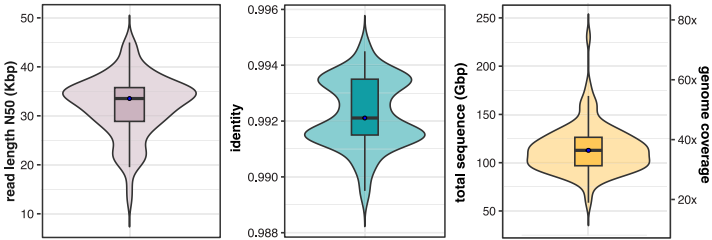
Application to a cohort of rare disease patients

Chan Zuckerberg Initiative

Children's National

GREGoR consortium

42 probands and 56 unaffected family members, sequenced with one-flowcell of ONT long-read sequencing (R10).

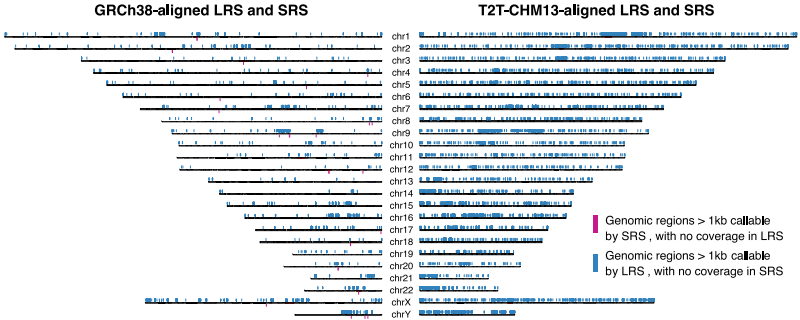


Negi et al. AJHG 2025

Better coverage of confidently mapped reads

More of the CHM13-T2T genome covered with at least 10x.

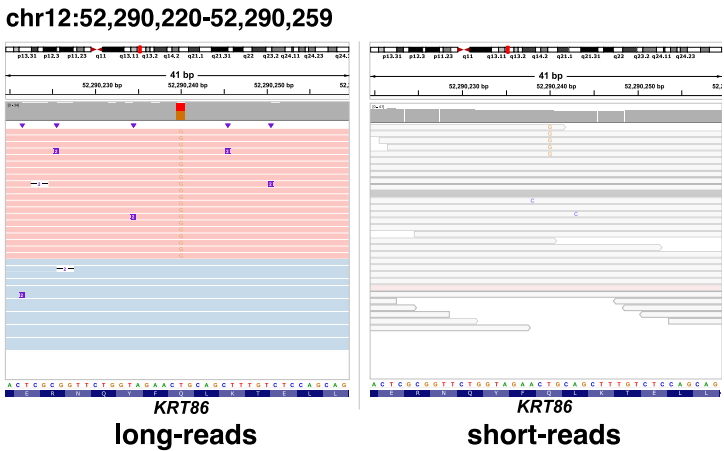
- **93.99%** (LRS) vs. 88.27% (SRS)



Negi et al. AJHG 2025

Small variants found by long-reads only

Missense mutation in *KRT86* disease gene (monilethrix) invisible with short reads.



Negi et al. AJHG 2025

Compound heterozygous variants thanks to phasing information

In *LHCGR* gene, associated with Leydig cell hypoplasia:

- Coding SNV on haplotype 1 (left, blue reads)
- ~7 Kbp deletion of an exon on haplotype 2 (right, red reads)

Negi et al. AJHG 2025

Collaboration with Genetic Unit

Dr. Julie Plaisancié

CHU Toulouse

Centre de Référence des Anomalies Rares en Génétique Ophtalmologiques

- Patients with severe and bilateral ocular phenotypes, e.g. microphthalmia (small eye) or anophthalmia (no eye).

Collaboration with Genetic Unit

Dr. Julie Plaisancié

CHU Toulouse

Centre de Référence des Anomalies Rares en Génétique Ophtalmologiques

- Patients with severe and bilateral ocular phenotypes, e.g. microphthalmia (small eye) or anophthalmia (no eye).
- Subset of patients **still undiagnosed after whole-genome short-read sequencing**.

Collaboration with Genetic Unit

Hôpitaux de Toulouse

Dr. Julie Plaisancié

CHU Toulouse

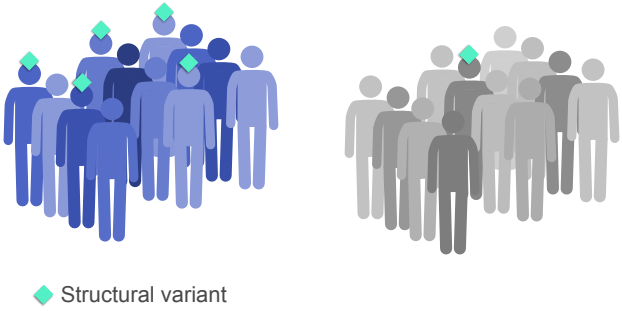
Centre de Référence des Anomalies Rares en Génétique Ophtalmologiques

Genotoul GeT

- Patients with severe and bilateral ocular phenotypes, e.g. microphthalmia (small eye) or anophthalmia (no eye).
- Subset of patients **still undiagnosed after whole-genome short-read sequencing**.
- Goal: test long-read sequencing and NAPU to solve those cases.
  - Ongoing: Nanopore sequencing for 8 patients at Genotoul/GeT-PlaGe.

Structural variants and complex disease with pangenomes

Common variants associated with a complex disease



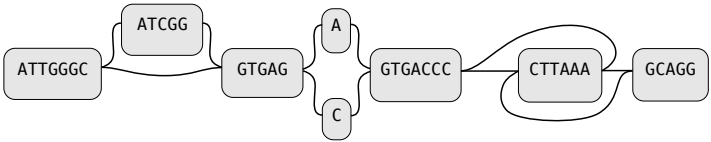
Goal

Genotype a comprehensive catalog of common variants across a large cohort.

Pangenomes represent genetic diversity succinctly

A pangenome represents a **collection of genomes** and the genetic variants among them.

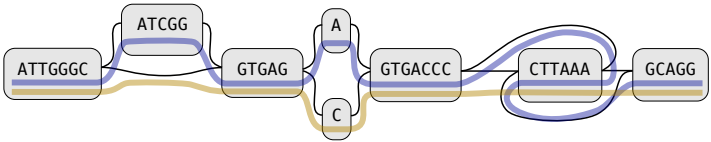
ATTGGGC**ATCGG**GTGAG**A**GTGACCC**TTTAAG**GCAGG
ATTGGGC**-----**GTGAG**C**GTGACCC**CTTAAA**GCAGG



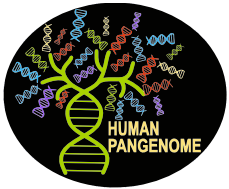
Pangenomes represent genetic diversity succinctly

A pangenome represents a **collection of genomes** and the genetic variants among them.

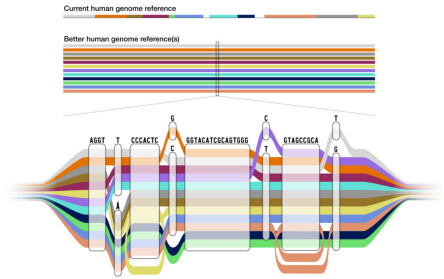
ATTGGGC**ATCGG**GTGAG**A**GTGACCC**TTTAAG**GCAGG
ATTGGGC**-----**GTGAG**C**GTGACCC**CTTAAA**GCAGG



Building a Human pangenome reference



- Human Pangenome Reference Consortium (HPRC)
- Latest sequencing technologies for 350 diverse individuals
- Pangenome containing a comprehensive catalog of (structural) variants



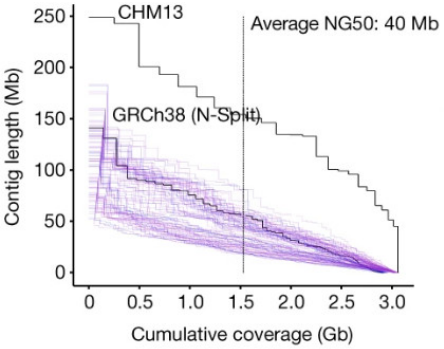
Liao, Asri, Ebler, et al. Nature 2023

Hickey, Monlong, et al. Nat. Biotechnol. 2023

Building a Human pangenome reference, a team effort

Year 1: 47 phased diploid assemblies of high quality

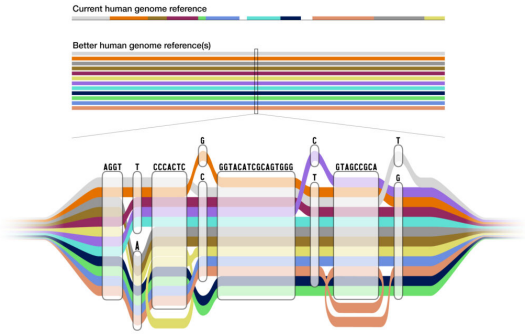
- GRCh38 (bottom black line): latest official reference genome
- CHM13 (top black line): recent complete telomere-to-telomere genome
- HPRC assemblies (light blue lines)

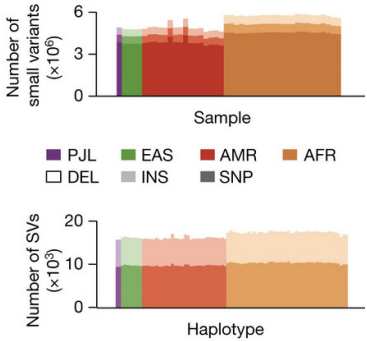


Liao, Asri, Ebler, et al. Nature 2023

Year 1: pangenome(s) from 47 phased diploid assemblies

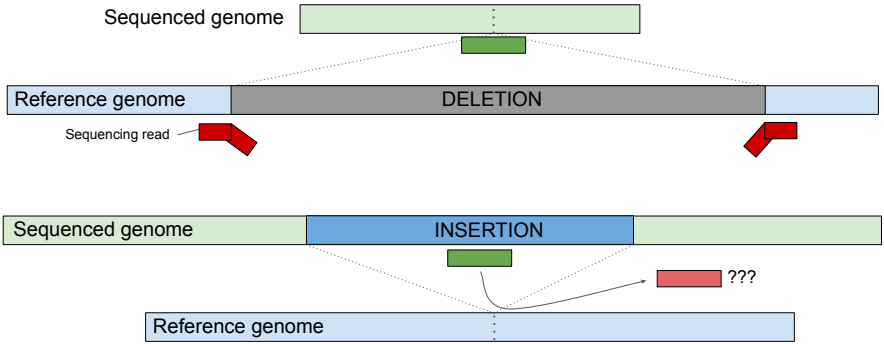
https://github.com/human-pangenomics/hpp_pangenome_resources



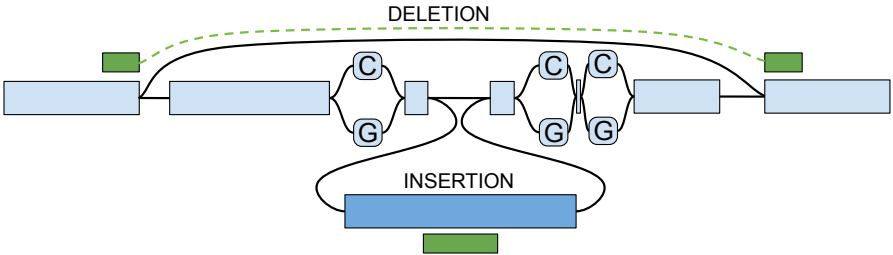


Liao, Asri, Ebler, et al. Nature 2023

Remember the challenges of structural variant detection



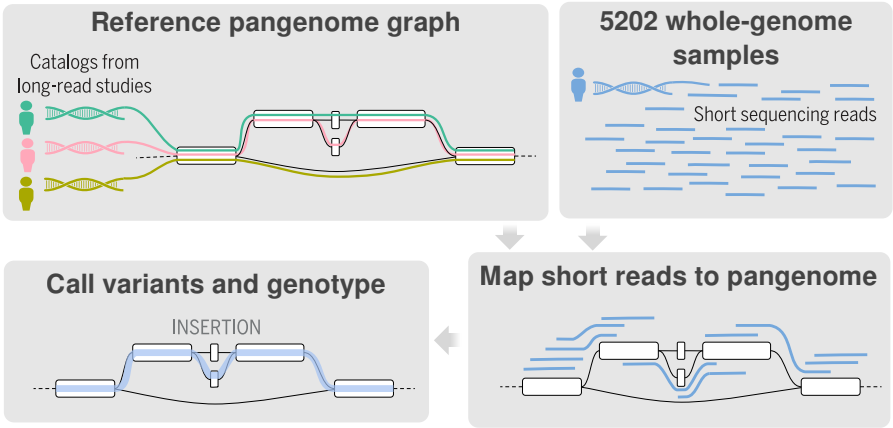
Aligning reads to a reference pangenome to genotype structural variants



Hickey, Heller, Monlong, et al. Genome Biology 2020

Siren, Monlong, Chang, Novak, Eizenga, et al. Science 2021

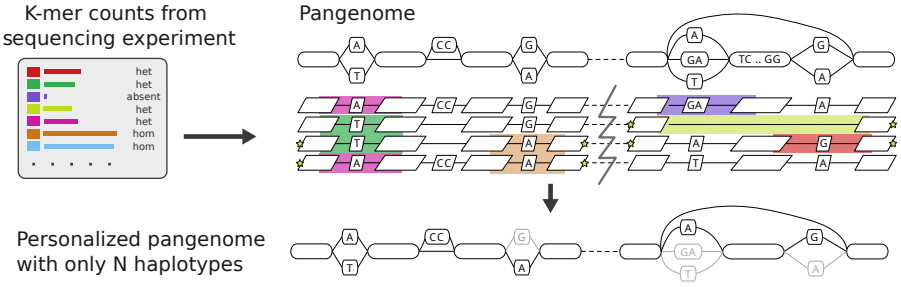
Short-read mapping and structural variant genotyping



Siren*, **Monlong***, Chang*, Novak*, Eizenga*, et al. Science 2021

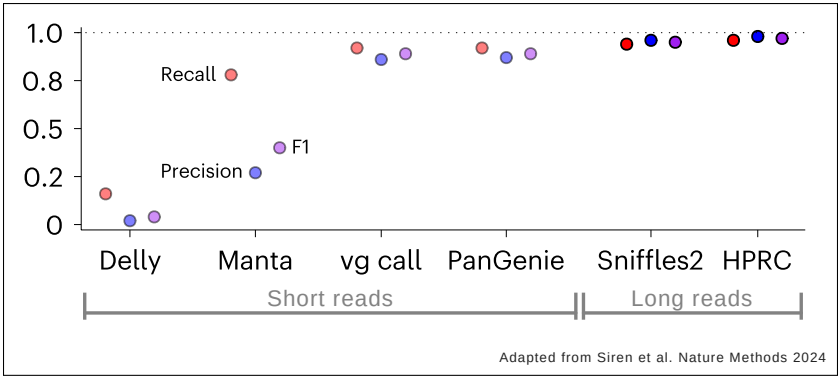
Personalized pangenomes with haplotype sampling

K-mer-guided “down-sampling” of the full pangenome.



Sirén et al. Nature Methods 2024

Structural variant genotyping performance



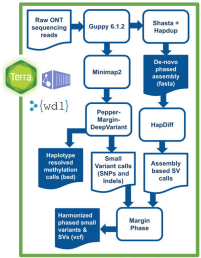
*vg call and Pangenie using the latest "personalized pangenome" mapping approach from *Sirén et al. Nature Methods 2024*.

Conclusions

Two approaches to integrate structural variants into genomic studies:

Cost-effective **long-read sequencing** using nanopore technologies to help solve undiagnosed **rare disease** cases.

NAPU
(Nanopore Analysis Pipeline)

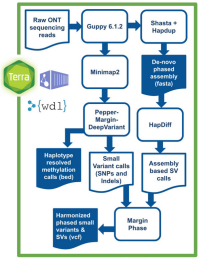


Conclusions

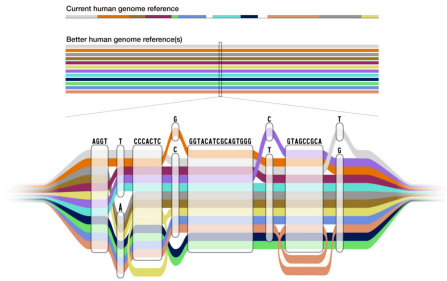
Two approaches to integrate structural variants into genomic studies:

Cost-effective **long-read sequencing** using nanopore technologies to help solve undiagnosed **rare disease** cases.

NAPU
(Nanopore Analysis Pipeline)



Genotyping with **pangenomes** from **short-read sequencing** data, e.g. for **genome-wide association studies**.



Acknowledgments

Univ. California, Santa Cruz

- **Benedict Paten**
- **Shloka Negi**
- **Karen Miga**
- **Glenn Hickey**
- Brandy McNulty
- Melissa Meredith
- Paolo Carnevali
- Trevor Pesout
- Kishwar Shafin
- Mira Mastoras
- Mobin Asri
- Adam Novak
- Xian Chang
- Jordan Eizenga

IRSD

- Sarah Djebali
- Hélène Coppin
- Marie-Paule Roth
- Delphine Meynard

NIH

- **Mikhail Kolmogorov**
- Cornelis Blauwendraat
- Kimberley Billingsley
- Pilar Alvarez Jerez

Broad Institute

- Anne O'Donnell-Luria
- Sarah Stenton
- Melanie O'Leary

Univ. California, Irvine

- Emmanuèle Délot
- Eric Vilain

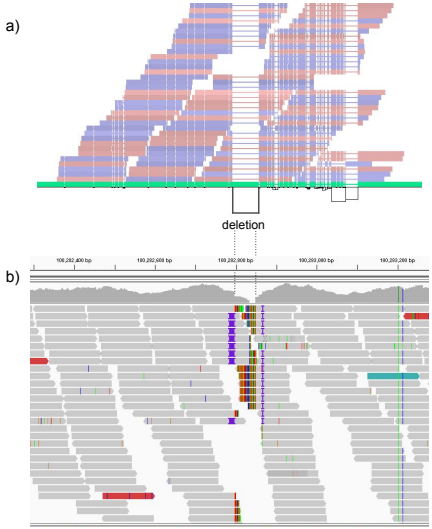
Children's National Research Institute

- Seth Berger
- Paolo Canigiula

CHU Toulouse

- Julie Plaisancié

Example of a deletion



Reads are correctly aligned "through" the deletion on the pangenome.

Many reads are aligned to the linear reference with the end unaligned (soft-clipped).

Hickey, Heller, Monlong, et al. Genome Biology 2020

Long-Read Somatic Variant Calling

Severus: somatic complex and haplotype-specific SVs

"takes advantage of long-read phasing and uses the breakpoint graph framework to model complex chromosomal rearrangements."
Keskus et al. Nature Biotech 2025

Long-Read Somatic Variant Calling

Severus: somatic complex and haplotype-specific SVs

"takes advantage of long-read phasing and uses the breakpoint graph framework to model complex chromosomal rearrangements."
Keskus et al. Nature Biotech 2025

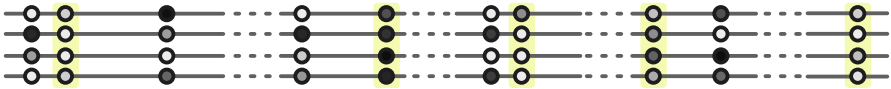
Pipeline from the Kolmogorov Lab

`https://github.com/KolmogorovLab/longread_somatic_nf`

- Alignment with minimap2
- Small variant calling with Clair3
- Phasing with longphase
- Somatic SV calling with Severus
- CNA calling with Wakhan
- Somatic small variant calling with DeepSomatic

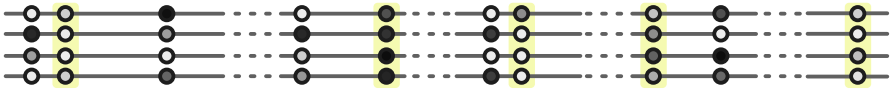
Episignatures of disease

Methylation pattern, across 10-100s of sites, associated with disease.

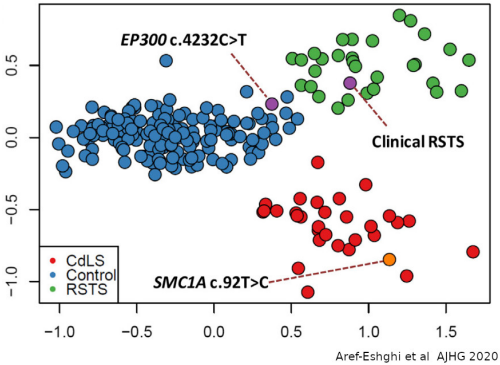


Episignatures of disease

Methylation pattern, across 10-100s of sites, associated with disease.



Aref-Eshghi et al. (AJHG 2020) found an episignature with 34 genetic syndromes, from blood samples using methylation arrays.

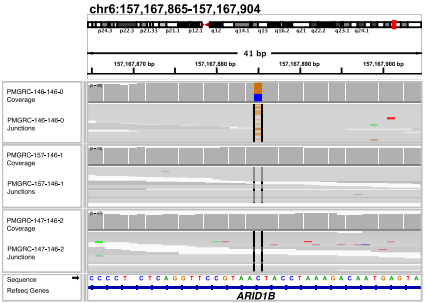


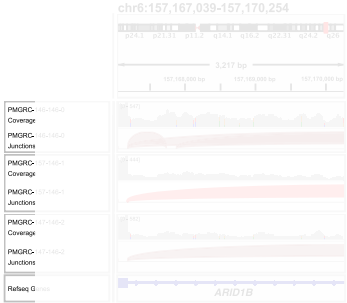
Aref-Eshghi et al AJHG 2020

Patient with complex neurodevelopmental phenotype

Variant of Uncertain Significance SNV in *ARID1B* gene (Coffin-Siris syndrome 1?).

- *De novo*, SRS and LRS, new splice site predicted *in silico* (SpliceAI).



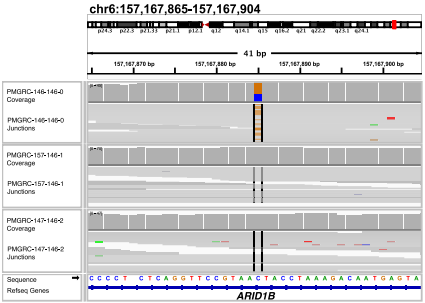


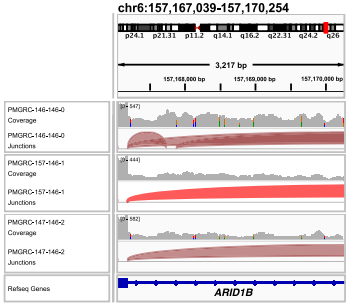
Negi et al. AJHG 2025

Patient with complex neurodevelopmental phenotype

Variant of Uncertain Significance SNV in *ARID1B* gene (Coffin-Siris syndrome 1?).

- *De novo*, SRS and LRS, new splice site predicted *in silico* (SpliceAI).





Negi et al. AJHG 2025

Known episignature of Coffin-Siris syndrome 1

- 106 differentially methylated CpG sites from Aref-Eshghi et al.

Known episignature of Coffin-Siris syndrome 1

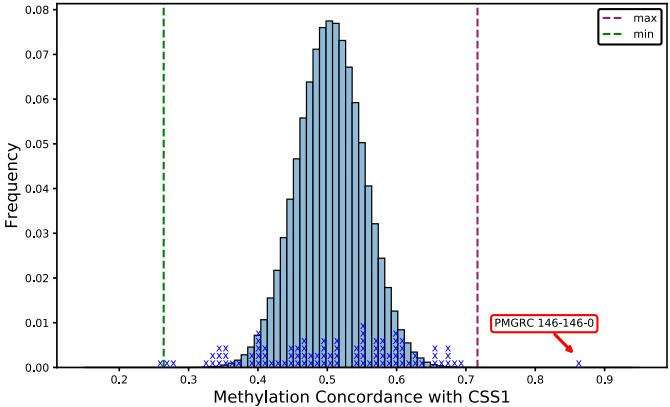
- 106 differentially methylated CpG sites from Aref-Eshghi et al.
- Count sites hyper/hypo-methylated consistently with known episignature.

Known episignature of Coffin-Siris syndrome 1

- 106 differentially methylated CpG sites from Aref-Eshghi et al.
- Count sites hyper/hypo-methylated consistently with known episignature.
- Significance by permuting sites across samples.

Known episignature of Coffin-Siris syndrome 1

- 106 differentially methylated CpG sites from Aref-Eshghi et al.
- Count sites hyper/hypo-methylated consistently with known episignature.
- Significance by permuting sites across samples.



Negi et al. AJHG 2025