Comprehensive haplotype-resolved view of genomic variation and methylation with long-read nanopore sequencing

Jean Monlong
14/03/2025

Symposium FHU-G4 Génomique

irsd
Institut de recherche en santé digestive

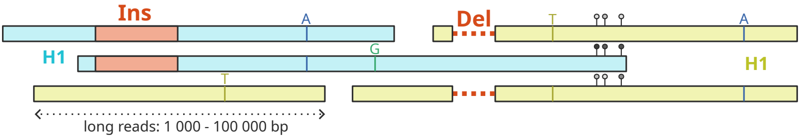
Inserm
La science pour la santé
From science to health

Outline

Introduction: genomic variants, DNA methylation, long-read sequencing

Napu computational pipeline

Application to a cohort of rare disease patients



Introduction: genomic variants, DNA methylation, long-read sequencing

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**)

GATCAGC
GATGAGC

Insertion-deletion polymorphisms (**INDELs**)

GATCAGC
GAT - - GC

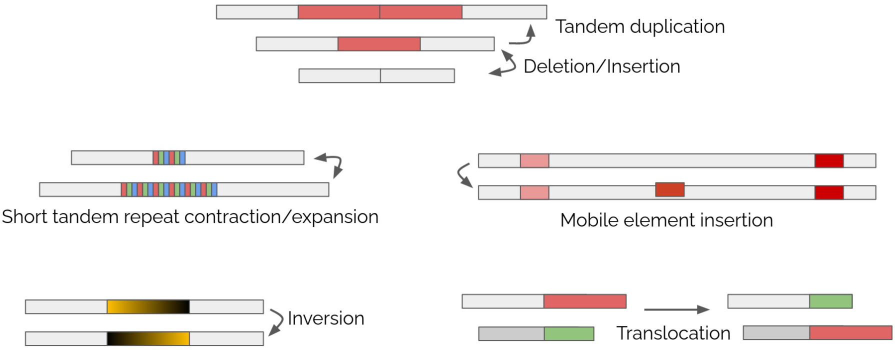
Structural variants (**SVs**)

GATCAGC
GATC AGC
CGC....300bp....GAT

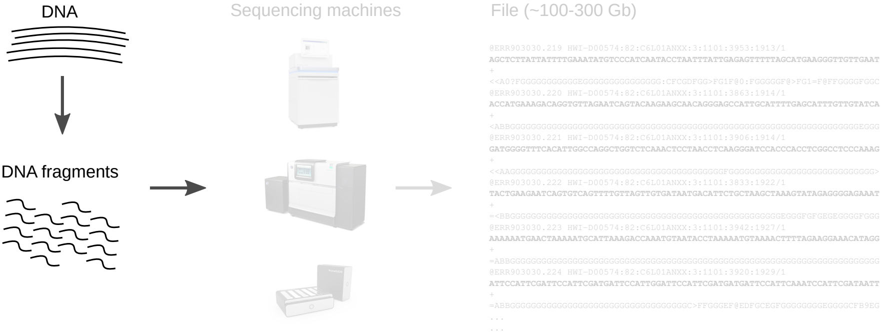
Different types of genomic variants

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
		CGC....300bp....GAT

Structural variants (SVs) come in diverse shapes and sizes



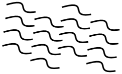
Genome sequencing



Genome sequencing

DNA

DNA fragments

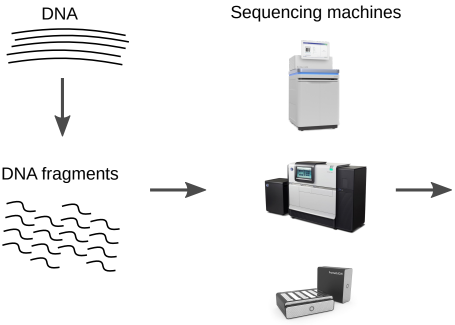


Sequencing machines

File (~100-300 Gb)

```
@ERR903030.219 HWI-D00574:82:C6L01ANXX:3:1101:3953:1913/1
AGCTCTTATTATTTTGAAATATGTCCCATCAATACCTAATTTATTGAGAGTTTTTAGCATGAAGGGTTGTTGAAT
+
<<A0?FGGGGGGGGGGGEGGGGGGGGGGGGGGG:CFCGDFGG>FG1F@0:FGGGGGF@>FG1=F@FFGGGGFGGC
@ERR903030.220 HWI-D00574:82:C6L01ANXX:3:1101:3863:1914/1
ACCATGAAAGACAGGTGTTAGAATCAGTACAAGAAGCAACAGGGAGCCATTGCATTTTGAGCATTTGTTGTATCA
+
<ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGG
@ERR903030.221 HWI-D00574:82:C6L01ANXX:3:1101:3906:1914/1
GATGGGGTTTCACATTGGCCAGGCTGGTCTCAAACTCCTAACCTCAAGGATCCACCCACCTCGGCCTCCCAAAG
+
<<AAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGG>
@ERR903030.222 HWI-D00574:82:C6L01ANXX:3:1101:3833:1922/1
TACTGAAGAATCAGTGTCAGTTTTGTTAGTTGTGATAATGACATTCTGCTAAGCTAAAGTATAGAGGGGAGAAAT
+
=<BBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGFGFGEGEGGGGFGGG
@ERR903030.223 HWI-D00574:82:C6L01ANXX:3:1101:3942:1927/1
AAAAAATGAACTAAAAATGCATTAAAGACCAAATGTAATACCTAAAAATGTAAAACTTTTAGAAGGAAACATAGG
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@ERR903030.224 HWI-D00574:82:C6L01ANXX:3:1101:3920:1929/1
ATTCCATTCGATTCCATTCGATGATTCCATTGGATTCCATTCGATGATGATTCCATTCAAATCCATTCGATAATT
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGC>FFGGGEF@EDFGCEGFGGGGGGGGEGGGGCFB9EG
...
...
```

Genome sequencing



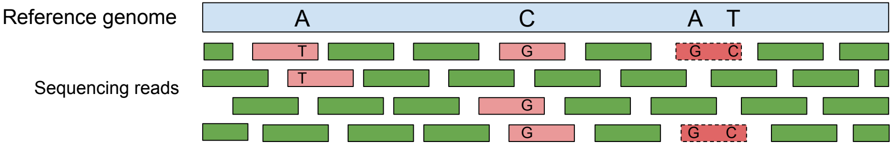
File (~100-300 Gb)

```
@ERR903030.219 HWI-D00574:82:C6L01ANXX:3:1101:3953:1913/1
AGCTCTTATTATTTTGAAATATGTCCCATCAATACCTAATTTATTGAGAGTTTTTAGCATGAAGGGTTGTTGAAT
+
<<A0?FGGGGGGGGEGGGGGGGGGGGGGGGGGG:CFCGDFGG>FG1F@0:FGGGGGF@>FG1=F@FFGGGGFGGC
@ERR903030.220 HWI-D00574:82:C6L01ANXX:3:1101:3863:1914/1
ACCATGAAAGACAGGTGTTAGAATCAGTACAAGAAGCAACAGGGAGCCATTGCATTTTGAGCATTTGTTGTATCA
+
<ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGG
@ERR903030.221 HWI-D00574:82:C6L01ANXX:3:1101:3906:1914/1
GATGGGGTTTCACATTGGCCAGGCTGGTCTCAAACTCCTAACCTCAAGGATCCACCCACCTCGGCCTCCCAAAG
+
<<AAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGG>
@ERR903030.222 HWI-D00574:82:C6L01ANXX:3:1101:3833:1922/1
TACTGAAGAATCAGTGTCAGTTTTGTTAGTTGTGATAATGACATTCTGCTAAGCTAAAGTATAGAGGGGAGAAAT
+
=<BBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGFGFGEGEGGGGFGGG
@ERR903030.223 HWI-D00574:82:C6L01ANXX:3:1101:3942:1927/1
AAAAAATGAACTAAAAATGCATTAAAGACCAAATGTAATACCTAAAAATGTAAAACTTTTAGAAGGAAACATAGG
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@ERR903030.224 HWI-D00574:82:C6L01ANXX:3:1101:3920:1929/1
ATTCCATTCGATTCCATTCGATGATTCCATTGGATTCCATTCGATGATGATTCCATTCAAATCCATTCGATAATT
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGC>FFGGGEF@EDFGCEGFGGGGGGEGGGGGGCFB9EG
...
...
```

Sequencing reads

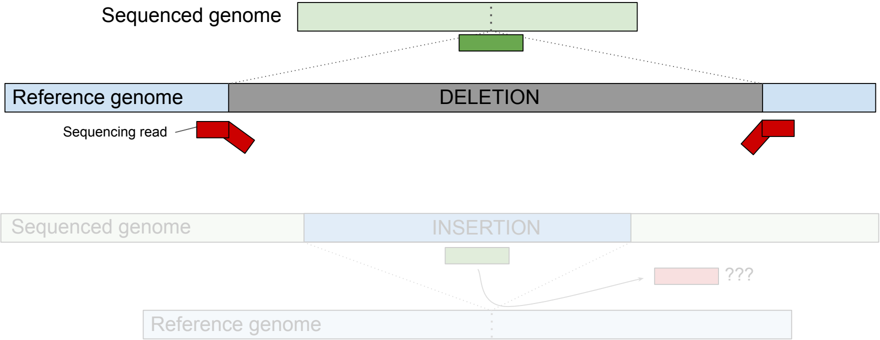
- 150-250 bp (current tech)
- 10,000s-100,000s bp (new tech. $$$)

Aligning reads to a reference genome

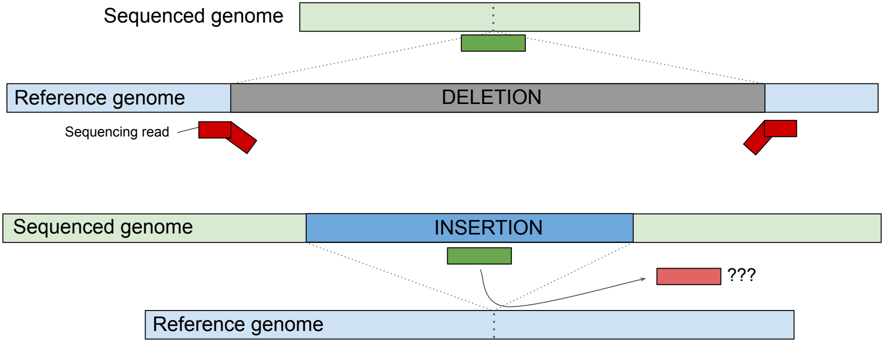


Assuming the reads are correctly placed, small variants are identified as recurrent differences between reads and the reference genome.

The challenges of structural variant detection

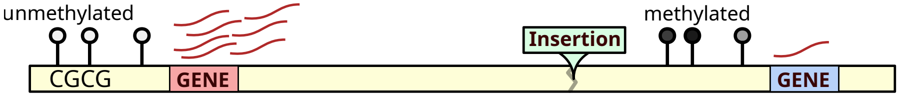


The challenges of structural variant detection



DNA methylation

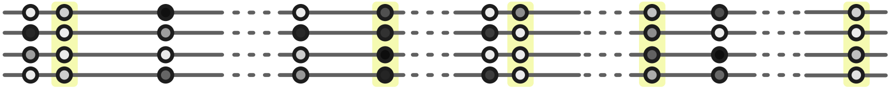
Epigenetic modification of the DNA, e.g. 5m-cytosine at CpG sites.
More promoter methylation → less transcription.



Aberrant methylation patterns can cause diseases (e.g. *FMR1* in FXS).

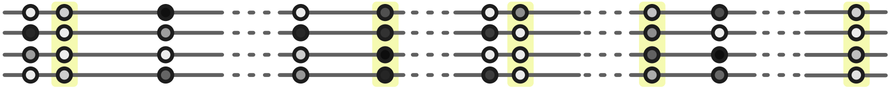
Episignatures of disease

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

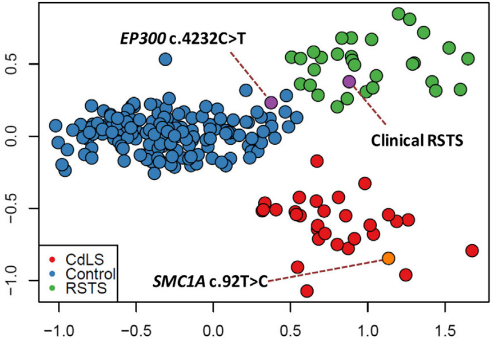


Episignatures of disease

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

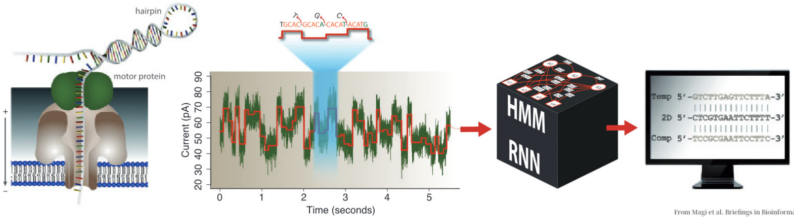


Aref-Eshghi et al. (AJHG 2020) found an episignature with 34 genetic syndromes, from blood samples using methylation arrays.



Aref-Eshghi et al AJHG 2020

Long-read sequencing with Oxford Nanopore Technologies

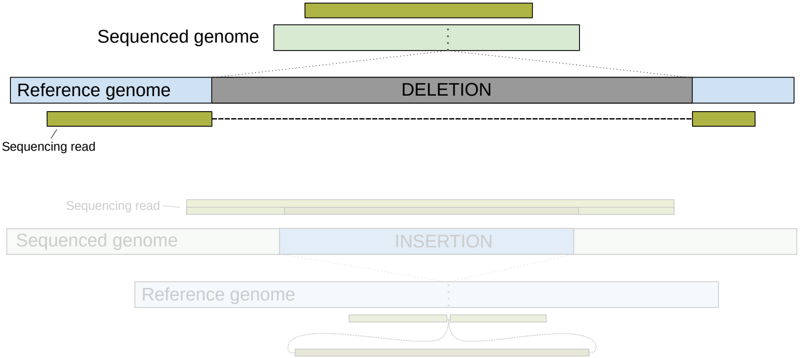


From Magi et al. Briefings in Bioinformatics

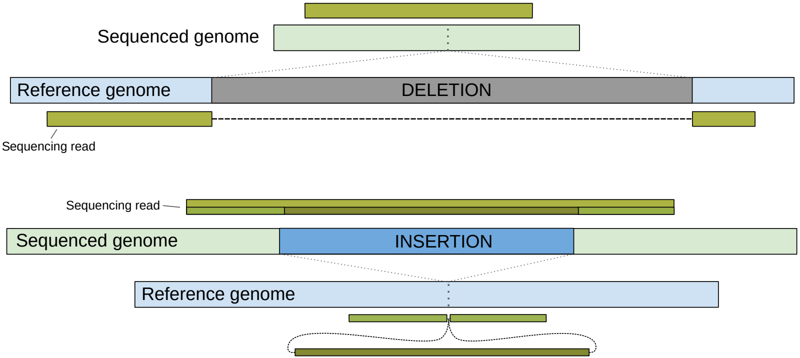
As the DNA (or RNA) fragment passes through the pore, the current changes and is decoded to predict nucleotides.

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

Longer reads improve structural variant detection

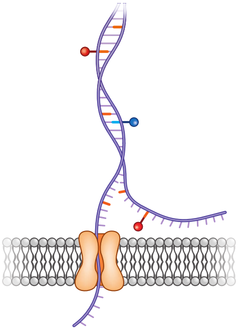


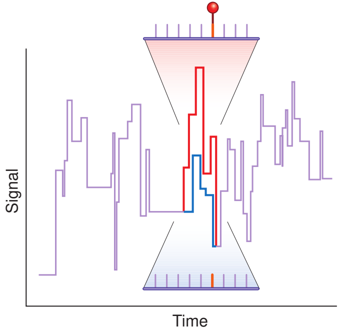
Longer reads improve structural variant detection



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

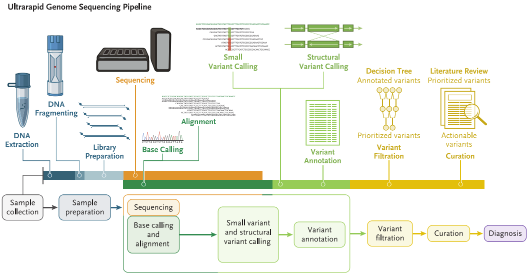




Schatz, Nature Methods 2023

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

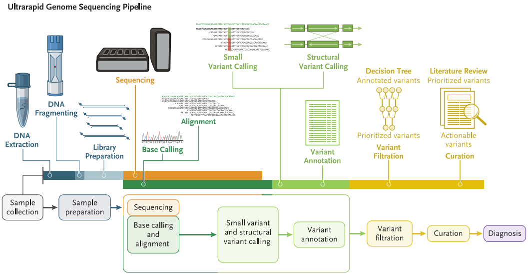


Gorzynski et al. N. Engl. J. Med. 2022

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

Napu computational pipeline

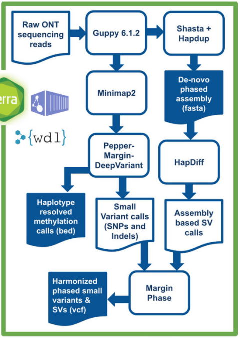
Cost-efficient Nanopore pipeline

- Only **one flow-cell** of Nanopore
- $\sim$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

NAPU
(Nanopore Analysis Pipeline)

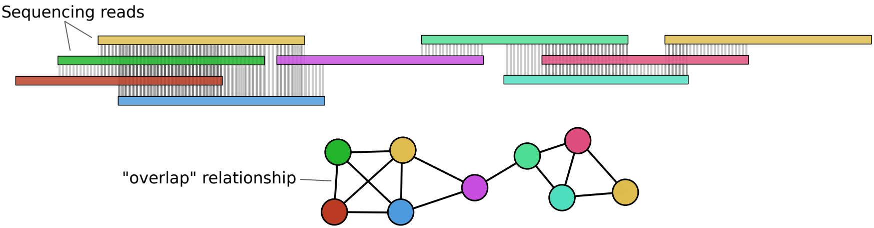


Terra {wdl}

Kolmogorov, Billingsley, et al. Nature Methods 2023

Longer reads enable *de novo* genome assembly

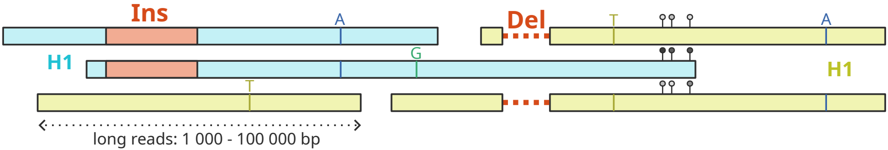
Reconstructs genomes without reference bias, hence better able to identify complex variants (e.g. combination of deletion/inversion)



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

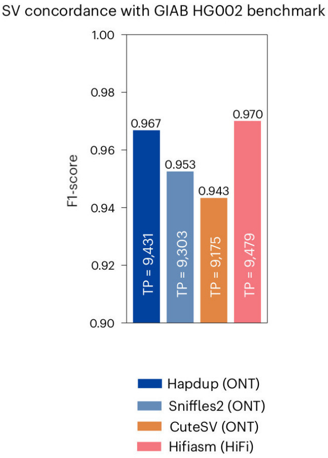


Reads are **haplo-tagged** using information across heterozygous sites.

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

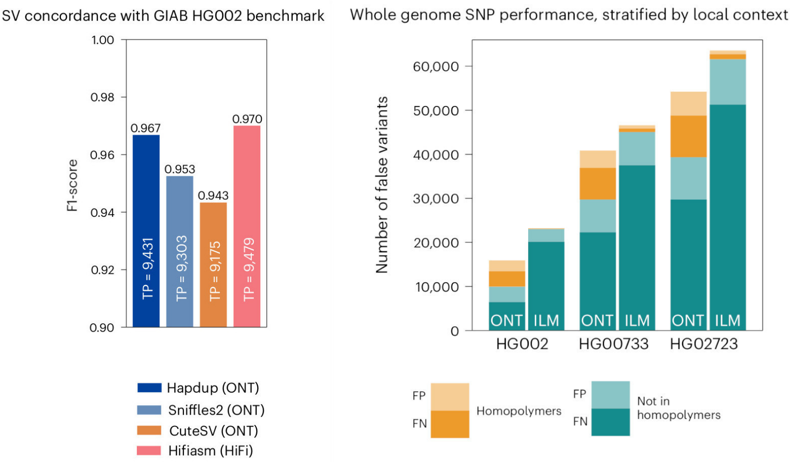
Kolmogorov, Billingsley, et al. Nature Methods 2023

Better calls for both small and structural variants...



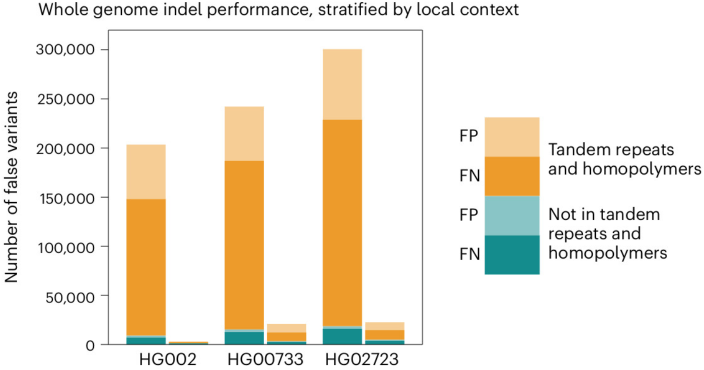
Kolmogorov, Billingsley, et al. Nature Methods 2023

Better calls for both small and structural variants…



Kolmogorov, Billingsley, et al. Nature Methods 2023

...except for indels in homopolymers

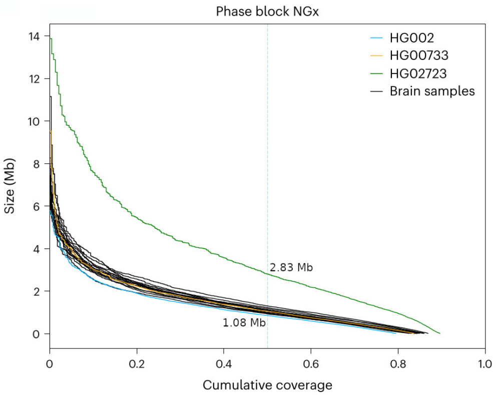


Note: Results above are for the R9 chemistry. The new R10 chemistry has lower error rate and better (indel) calling performance.

Kolmogorov, Billingsley, et al. Nature Methods 2023

Variants and methylation phased in Mbp-long blocks

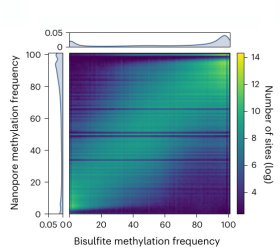
Small variants, structural variants, methylation marks are homogenized into megabase-long phase blocks.



Kolmogorov, Billingsley, et al. Nature Methods 2023

Haplotype-resolved methylation at CpG sites

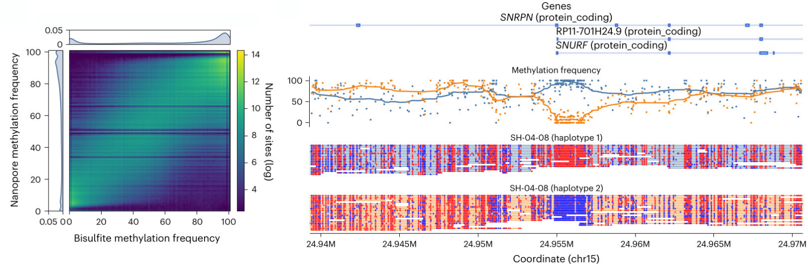
- Good concordance with short-read bisulfite sequencing.
- Haplotype-specific methylation patterns.



Kolmogorov, Billingsley, et al. Nature Methods 2023

Haplotype-resolved methylation at CpG sites

- Good concordance with short-read bisulfite sequencing.
- Haplotype-specific methylation patterns.



Kolmogorov, Billingsley, et al. Nature Methods 2023

Application to a cohort of rare disease patients

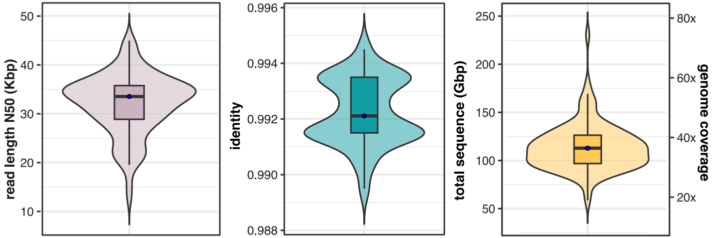
Application to a cohort of rare disease patients

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).

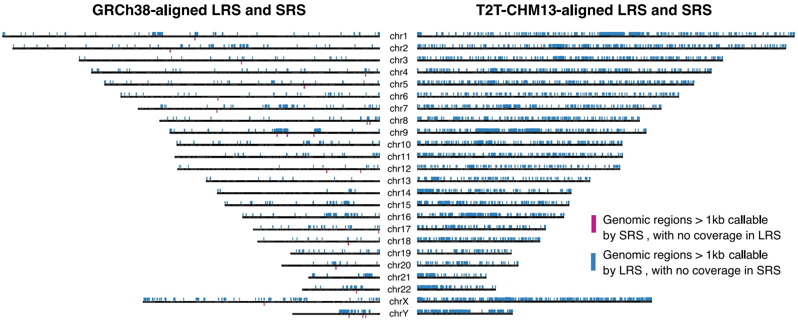


Negi et al. AJHG 2025

Better coverage of confidently mapped reads

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

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

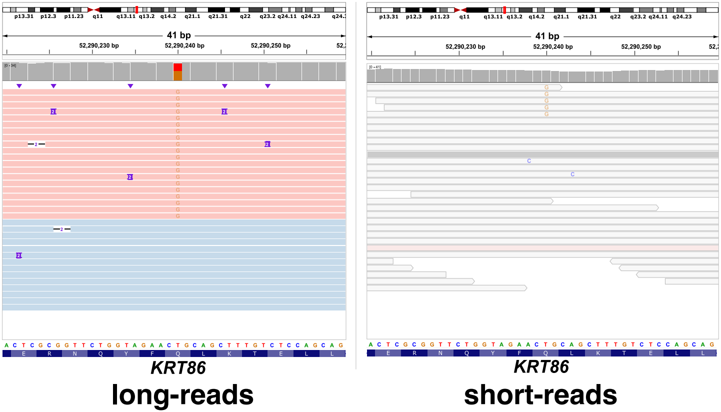


Negi et al. AJHG 2025

Small variants found by long-reads only

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

chr12:52,290,220-52,290,259

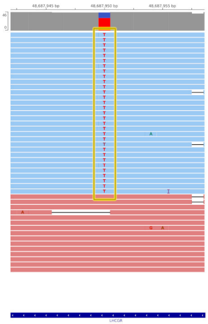


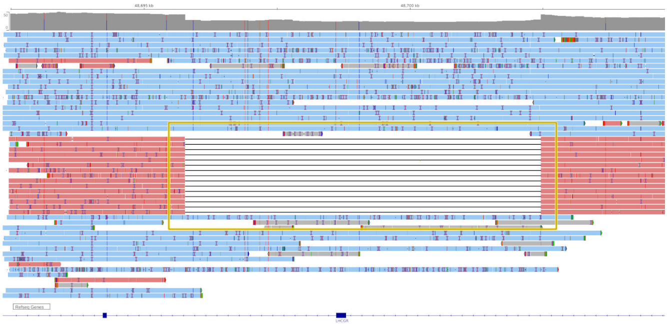
Negi et al. AJHG 2025

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)



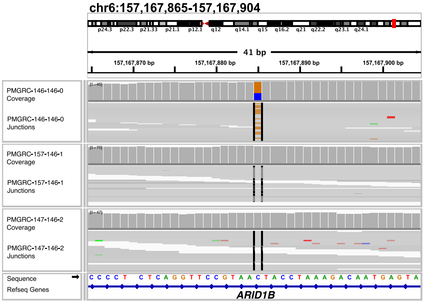


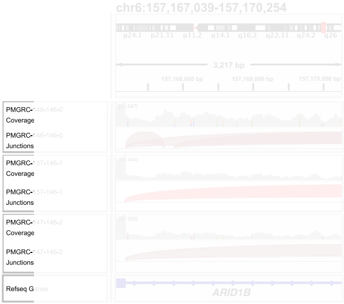
Negi et al. AJHG 2025

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).



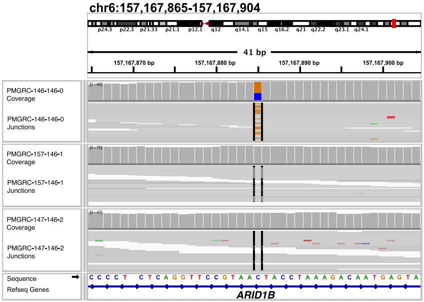


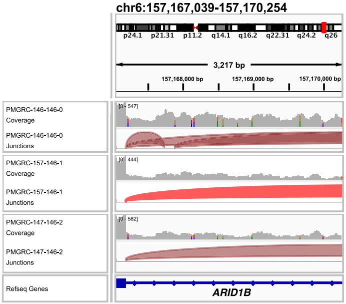
Negi et al. AJHG 2025

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

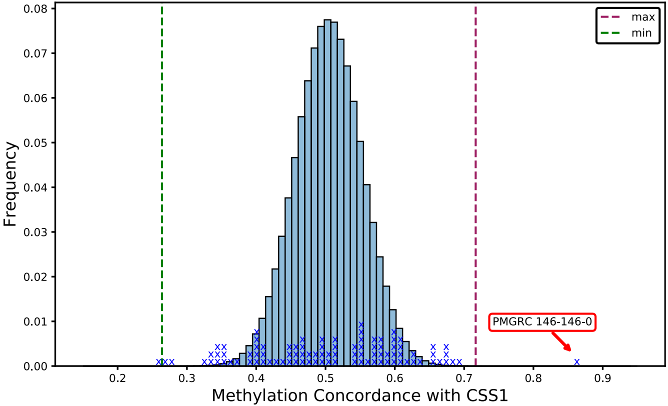
- 106 differentially methylated CpG sites from Aref-Eshghi et al.
- Count sites hyper/hypo-methylated consistently with known episignature.

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

Large-scale study of brain samples

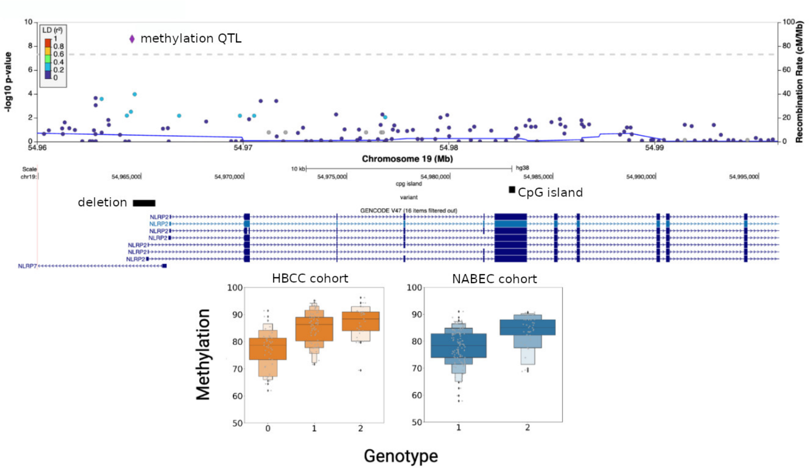
New resource: 351 brain control samples sequenced in the NIH's Center of Alzheimer's and Related Dementias (CARD) long-read sequencing initiative.

- 234,905 SVs
- >800 SV expression QTLs
- >2000 SV methylation QTLs

National Institutes of Health
Center for Alzheimer's and Related Dementias

Billingsley, Meredith, Daida, et al. bioRxiv 2024

Example of a SV methylation QTL

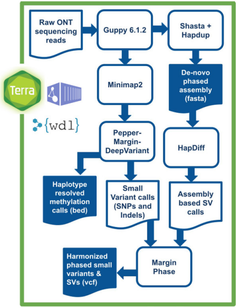


Billingsley, Meredith, Daida, et al. bioRxiv 2024

Take-home message

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

NAPU
(Nanopore Analysis Pipeline)



Haplotype-resolved

- small variants (SNPs/indels)
- structural variants
- *de novo* assembly
- methylation marks

Kolmogorov, Billingsley, et al. Nature Methods 2023

Negi et al. AJHG 2025

Acknowledgments

Univ. California, Santa Cruz

- **Benedict Paten**
- **Shloka Negi**
- **Karen Miga**
- Brandy McNulty
- Melissa Meredith
- Paolo Carnevali
- Trevor Pesout
- Kishwar Shafin
- Ivo Violich
- Joshua Gardner
- Sara M. O'Rourke
- Mira Mastoras
- Mobin Asri

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

NIH National Institutes of Health
Center for Alzheimer's and Related Dementias

Chan Zuckerberg Initiative

UC SANTA CRUZ
Genomics Institute

Children's National

GREGoR consortium