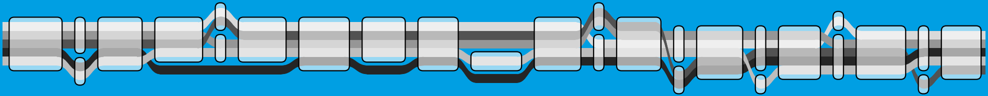


Integrating **structural variants** in genomic studies of rare and complex diseases with **long-read sequencing** and **pangenomes**

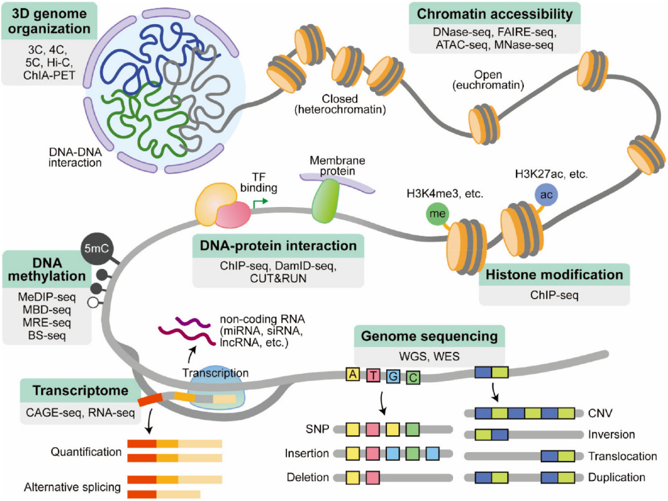
Jean Monlong

irsd
Institut de recherche en santé digestive

Inserm
La science pour la santé
From science to health

JOBIM
11/07/2025

Understanding functional impact of genomic variation



Dev. Reprod. 2023; 27(1):9-24 DOI: 10.12717/DR.2023.27.1.9

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**)

GAT**C**AGC

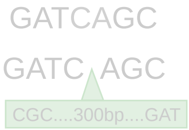
GAT**G**AGC

Insertion-deletion polymorphisms (**INDELs**)

GAT**CA**GC

GAT - - GC

Structural variants (**SVs**)



GATCAGC

GATC AGC

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

Different types of genomic variants

Single-nucleotide polymorphisms (**SNPs**)

GAT**C**AGC

GAT**G**AGC

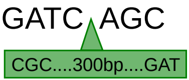
Insertion-deletion polymorphisms (**INDELs**)

GAT**CA**GC

GAT **- -** GC

Structural variants (**SVs**)

GATCAGC

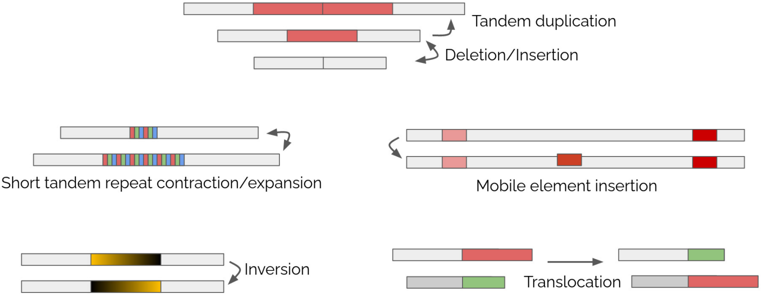


GATC AGC

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

Structural variants (SVs) come in diverse shapes and sizes

Variant size: from 50 bases to megabases.



Tandem duplication

Deletion/Insertion

Short tandem repeat contraction/expansion

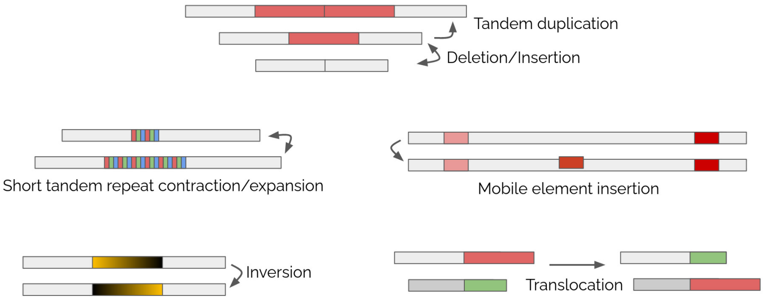
Mobile element insertion

Inversion

Translocation

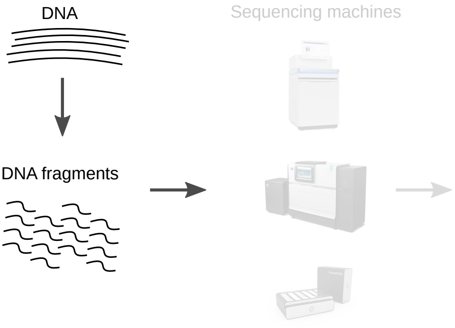
Structural variants (SVs) come in diverse shapes and sizes

Variant size: from 50 bases to megabases.



- High functional impact
- Involved in rare and common diseases, and cancers.
- **Hard to detect**

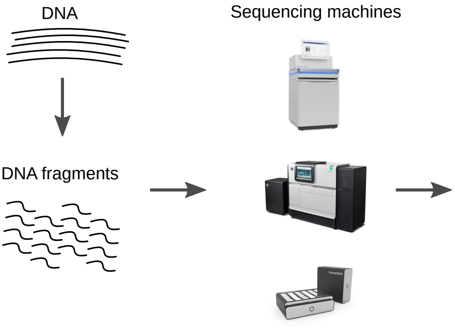
Genome sequencing



File (~100-300 Gb)

```
@ERR903030.219 HWI-D00574:82:C6L01ANXX:3:1101:3953:1913/1
AGCTCTTATTATTTTGAAATATGTCCCATCAATACCTAATTTATTGAGAGTTTTTAGCATGAAGGGTTGTTGAAT
+
<<A0?FGGGGGGGGGGEGGGGGGGGGGGGGGGG: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
GATGGGGTTTCACATTGGCCAGGCTGGTCTCAAACTCCTAACCTCAAGGGATCCACCCACCTCGGCCTCCCAAAG
+
<<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
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGC>FFGGGEF@EDFGCEGFGGGGGGGGEGGGGCFB9EG
...
...
```

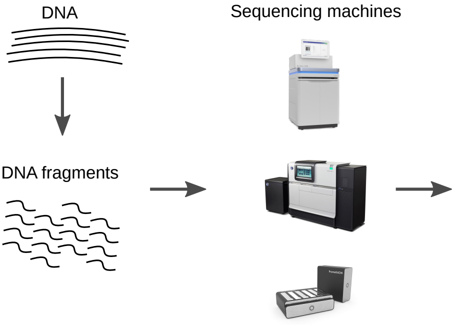
Genome sequencing



File (~100-300 Gb)

```
@ERR903030.219 HWI-D00574:82:C6L01ANXX:3:1101:3953:1913/1
AGCTCTTATTATTTTGAAATATGTCCCATCAATACCTAATTTATTGAGAGTTTTTAGCATGAAGGGTTGTTGAAT
+
<<A0?FGGGGGGGGGGEGGGGGGGGGGGGGGGG:CFCGDFGG>FG1F@0:FGGGGGF@>FG1=F@FFGGGGFGGC
@ERR903030.220 HWI-D00574:82:C6L01ANXX:3:1101:3863:1914/1
ACCATGAAAGACAGGTGTTAGAATCAGTACAAGAAGCAACAGGGAGCCATTGCATTTTGAGCATTTGTTGTATCA
+
<ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGG
@ERR903030.221 HWI-D00574:82:C6L01ANXX:3:1101:3906:1914/1
GATGGGGTTTCACATTGGCCAGGCTGGTCTCAAACTCCTAACCTCAAGGGATCCACCCACCTCGGCCTCCCAAAG
+
<<AAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGG>
@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
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGC>FFGGGEF@EDFGCEGFGGGGGGGGEGGGGCFB9EG
...
...
```

Genome sequencing



File (~100-300 Gb)

```
@ERR903030.219 HWI-D00574:82:C6L01ANXX:3:1101:3953:1913/1
AGCTCTTATTATTTTGAAATATGTCCCATCAATACCTAATTTATTGAGAGTTTTTAGCATGAAGGGTTGTTGAAT
+
<<A0?FGGGGGGGGGGGGEGGGGGGGGGGGGGG: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
GATGGGGTTTCACATTGGCCAGGCTGGTCTCAAACTCCTAACCTCAAGGGATCCACCCACCTCGGCCTCCCAAAG
+
<<AAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGG>
@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
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@ERR903030.224 HWI-D00574:82:C6L01ANXX:3:1101:3920:1929/1
ATTCCATTCGATTCCATTCGATGATTCCATTGGATTCCATTCGATGATGATTCCATTCAAATCCATTCGATAATT
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGC>FFGGGEF@EDFGCEGFGGGGGGGEGGGGCFB9EG
...
...
```

Sequencing reads

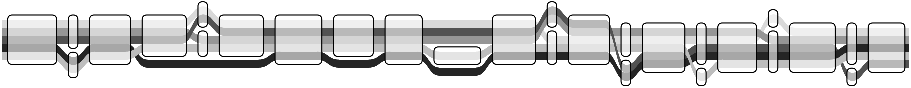
- **Short:** 150-250 bp (current tech)
- **Long:** 10,000s-100,000s bp (new tech. $$$)

Outline - Studying structural variants

Short-read sequencing, pangenomes, and complex diseases

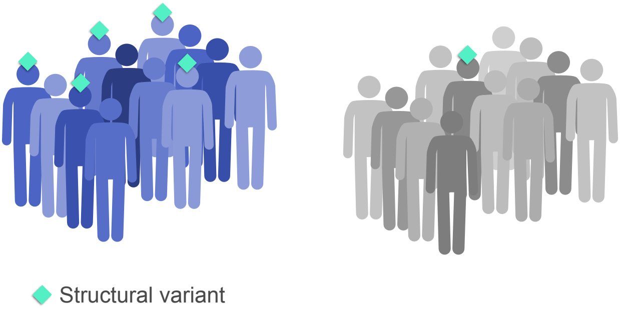
Long-read sequencing and rare diseases

Pangenomes meet long-read sequencing



Short-read sequencing, pangenomes, and complex diseases

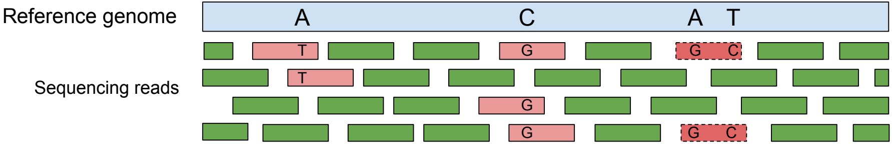
Common variants associated with a complex disease



Goal

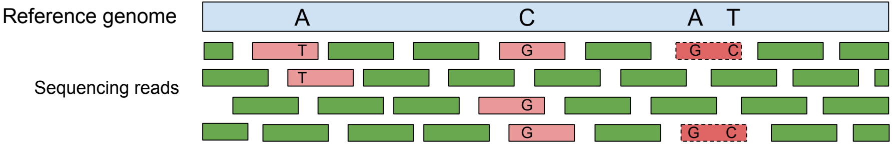
Genotype a comprehensive catalog of common variants across a large cohort.

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.

Aligning reads to a reference genome

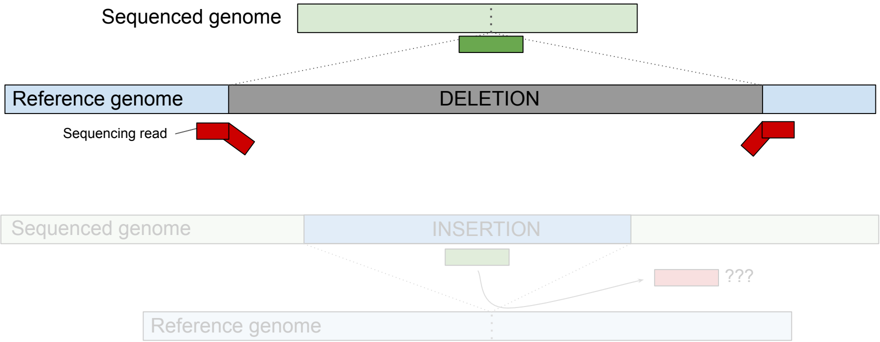


Assuming the reads are correctly placed, small variants are identified as recurrent differences between reads and the reference genome.

Variants can be missed, resulting in **reference bias**.

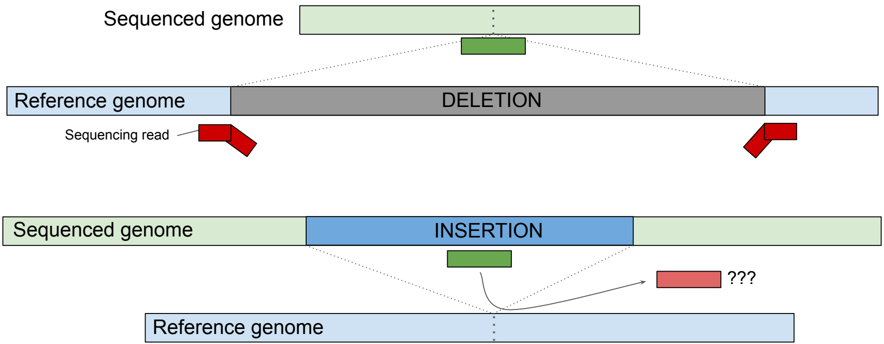
The challenges of structural variant detection

Around breakpoints, short sequencing reads are hard to map on the reference genome.



The challenges of structural variant detection

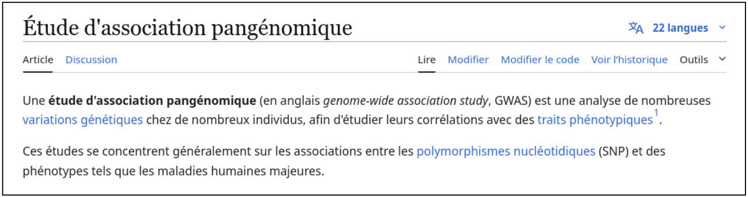
Around breakpoints, short sequencing reads are hard to map on the reference genome.



Pangenomics to the rescue. Which pangenomics?

Pangenomics to the rescue. Which pangenomics?

Not “Genome-wide association studies” in French.



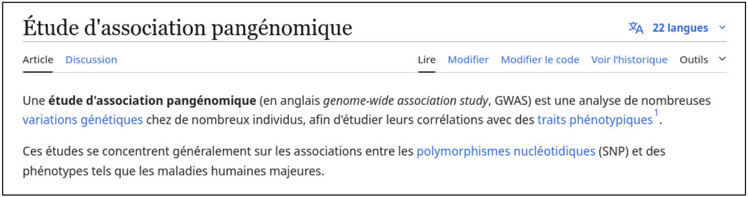
Étude d'association pangénomique

Une **étude d'association pangénomique** (en anglais *genome-wide association study*, GWAS) est une analyse de nombreuses variations génétiques chez de nombreux individus, afin d'étudier leurs corrélations avec des traits phénotypiques[1].

Ces études se concentrent généralement sur les associations entre les polymorphismes nucléotidiques (SNP) et des phénotypes tels que les maladies humaines majeures.

Pangenomics to the rescue. Which pangenomics?

Not “Genome-wide association studies” in French.

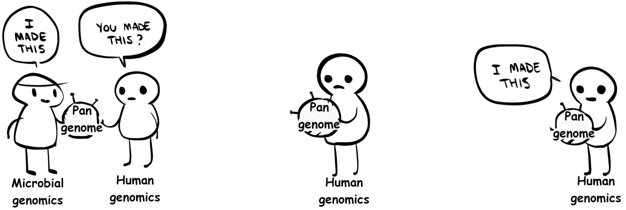


Étude d'association pangénomique

Une **étude d'association pangénomique** (en anglais *genome-wide association study*, GWAS) est une analyse de nombreuses variations génétiques chez de nombreux individus, afin d'étudier leurs corrélations avec des traits phénotypiques[1].

Ces études se concentrent généralement sur les associations entre les polymorphismes nucléotidiques (SNP) et des phénotypes tels que les maladies humaines majeures.

Also not exactly the set of **genes** from all strains within a clade, like in microbial pangenome.

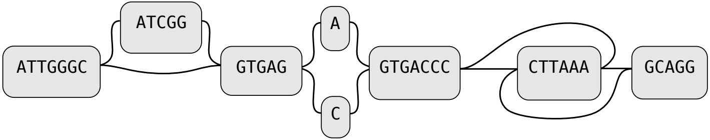


Pangenomes represent genetic diversity succinctly

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

ATTGGGC**ATCGG**GTGAG**A**GTGACCC**TTTAAG**GCAGG

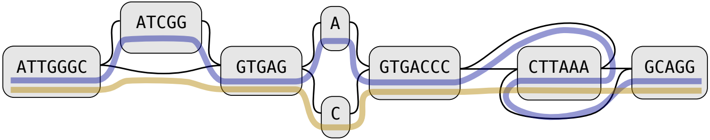
ATTGGGC**-----**GTGAG**C**GTGACCC**CTTAAA**GCAGG



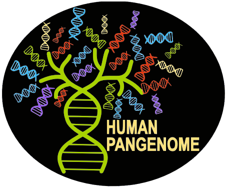
Pangenomes represent genetic diversity succinctly

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

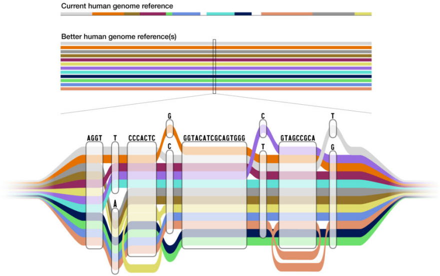
```
ATTGGGCATCGGGTGAGAGTGACCCTTTAAGGCAGG
ATTGGGC-----GTGAGCGTGACCCCTTAAAGCAGG
```



Building a Human pangenome reference



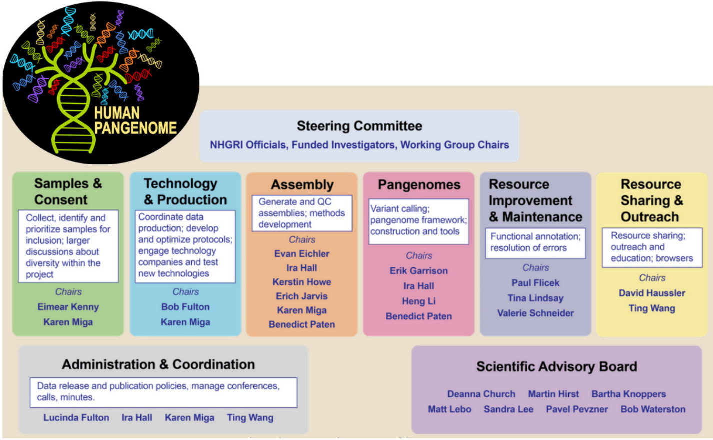
- Human Pangenome Reference Consortium (**HPRC**)
- Latest sequencing technologies for 350 diverse individuals
- Pangenome containing a comprehensive catalog of (structural) variants


Current human genome reference
Better human genome reference(s)
AGGT
T
A
CCCACTC
G
C
GGTACATCGCAGTGGG
C
T
GTAGCCGCA
T
G

Liao*, Asri*, Ebler*, et al. Nature 2023

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

Building a Human pangenome reference, a team effort



Steering Committee

NHGRI Officials, Funded Investigators, Working Group Chairs

Samples & Consent

Collect, identify and prioritize samples for inclusion; larger discussions about diversity within the project

Chairs

Eimear Kenny
Karen Miga

Technology & Production

Coordinate data production; develop and optimize protocols; engage technology companies and test new technologies

Chairs

Bob Fulton
Karen Miga

Assembly

Generate and QC assemblies; methods development

Chairs

Evan Eichler
Ira Hall
Kerstin Howe
Erich Jarvis
Karen Miga
Benedict Paten

Pangenomes

Variant calling; pangenome framework; construction and tools

Chairs

Erik Garrison
Ira Hall
Heng Li
Benedict Paten

Resource Improvement & Maintenance

Functional annotation; resolution of errors

Chairs

Paul Flicek
Tina Lindsay
Valerie Schneider

Resource Sharing & Outreach

Resource sharing; outreach and education; browsers

Chairs

David Haussler
Ting Wang

Administration & Coordination

Data release and publication policies, manage conferences, calls, minutes.

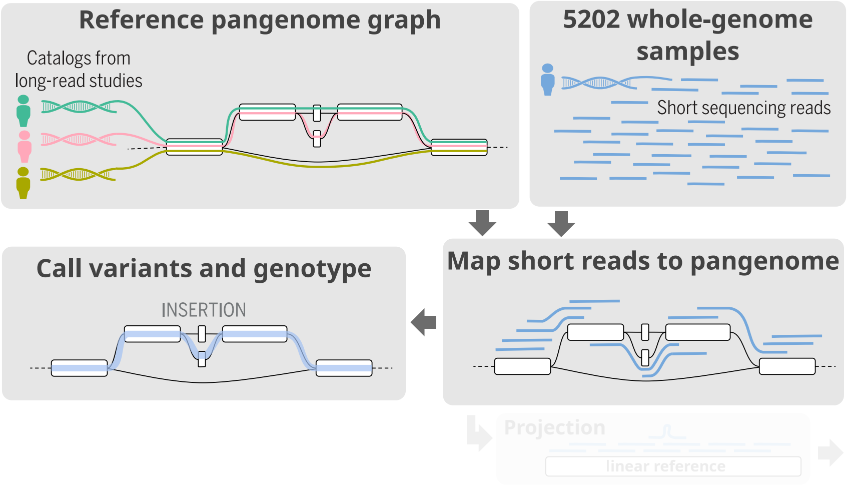
Lucinda Fulton Ira Hall Karen Miga Ting Wang

Scientific Advisory Board

Deanna Church Martin Hirst Bartha Knoppers
Matt Lebo Sandra Lee Pavel Pevzner Bob Waterston

Check out the latest data at: `https://data.humanpangenome.org`

Short-read mapping and structural variant genotyping



Reference pangenome graph

Catalogs from long-read studies

5202 whole-genome samples

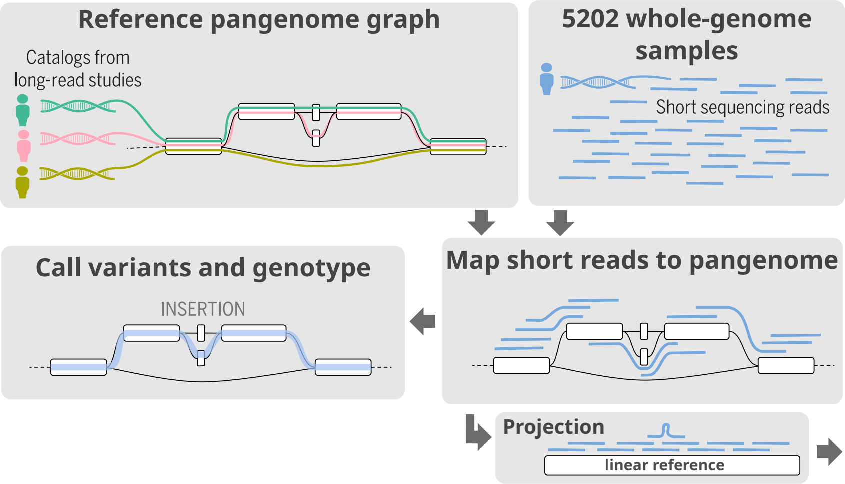
Short sequencing reads

Map short reads to pangenome

Call variants and genotype

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

Short-read mapping and structural variant genotyping



Reference pangenome graph

Catalogs from long-read studies

5202 whole-genome samples

Short sequencing reads

Map short reads to pangenome

Call variants and genotype

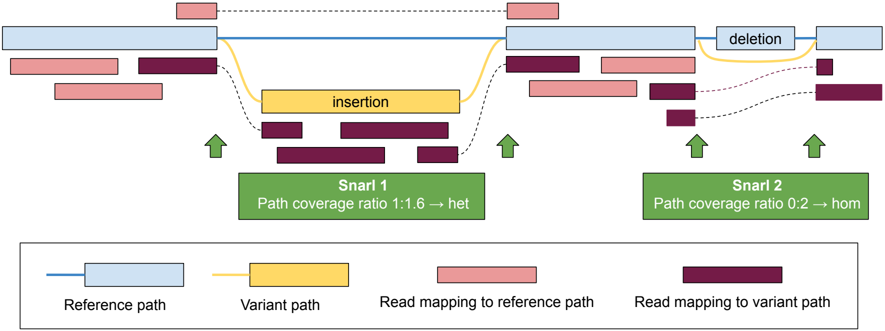
INSERTION

Projection

linear reference

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

Genotyping structural variation from pangenomic mapping

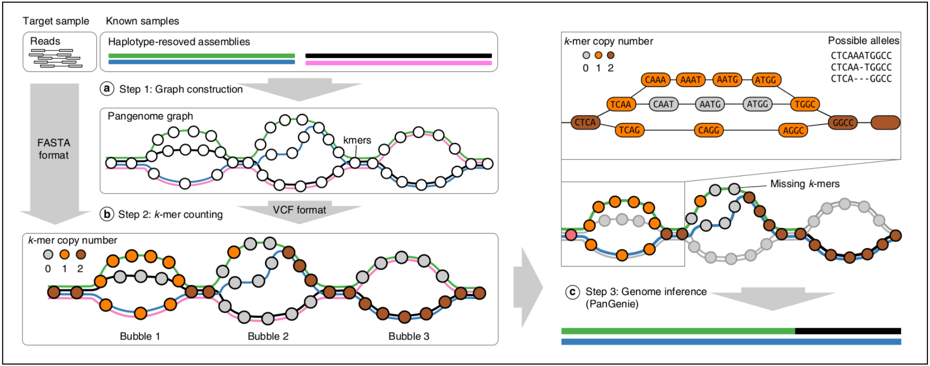

deletion
insertion
Snarl 1
Path coverage ratio 1:1.6 → het
Snarl 2
Path coverage ratio 0:2 → hom
Reference path
Variant path
Read mapping to reference path
Read mapping to variant path

`https://github.com/vgteam/vg`

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

Genotyping structural variation from phased variants

PanGenie uses k-mer and haplotype information to genotype SVs.



```
https://github.com/eblerjana/PanGenie
```

Ebler et al. Nature Genetics 2022

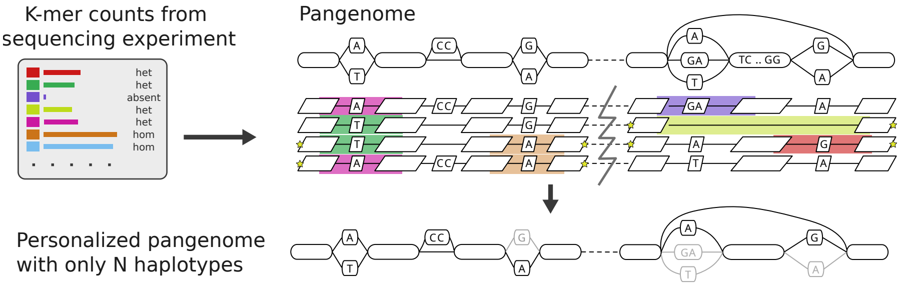
Personalized pangenomes with haplotype sampling

With pangenomes becoming larger, analysis can suffer.

Personalized pangenomes with haplotype sampling

With pangenomes becoming larger, analysis can suffer.

One solution: k-mer-guided "down-sampling" of the full pangenome.

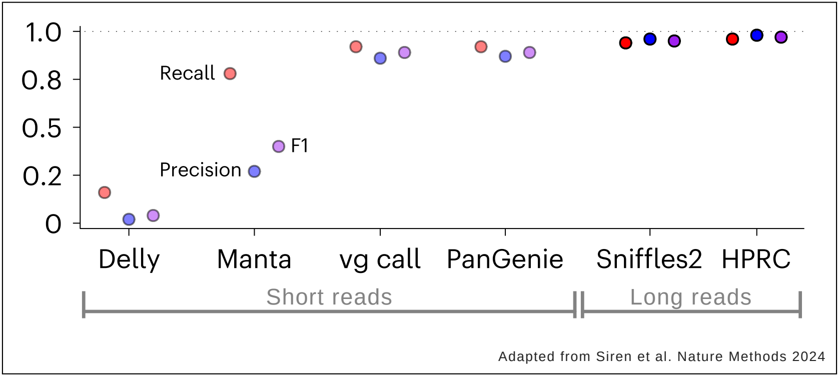

K-mer counts from sequencing experiment

Pangenome

Personalized pangenome with only N haplotypes

Sirén et al. Nature Methods 2024

Structural variant genotyping performance



Adapted from Siren et al. Nature Methods 2024

*vg call and PanGenie using the “personalized pangenome” approach.

Current state of pangenomics tools

Construction

Read mapping

Genotyping

Complex variants

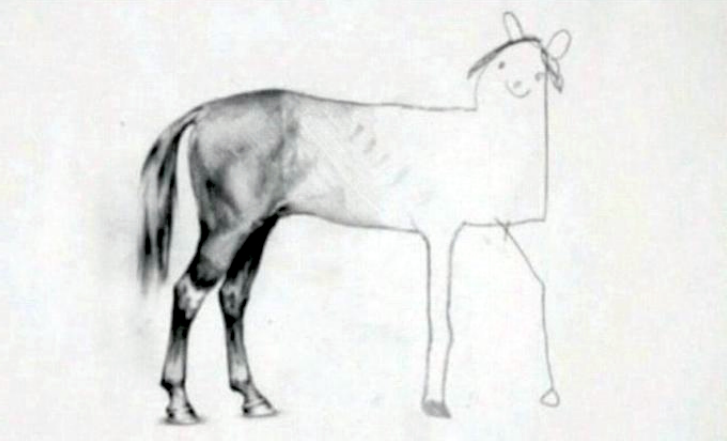
Functional genomics

Visualization

Annotation

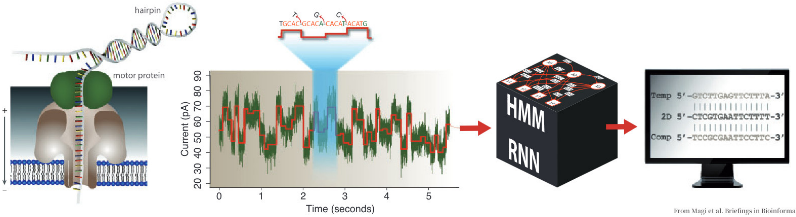
Association tests

Multi-species



Long-read sequencing and rare diseases

Long-read sequencing with Oxford Nanopore Technologies

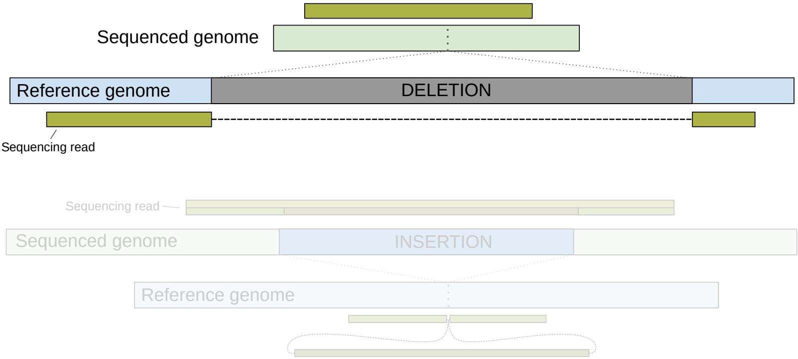


From Magi et al. Briefings in Bioinformatics

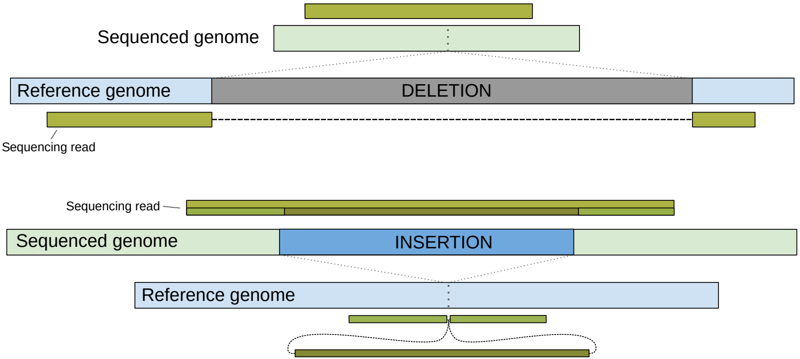
As the DNA (or RNA) fragment passes through the pore, the current changes and is decoded to predict nucleotides.

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

Longer reads improve structural variant detection



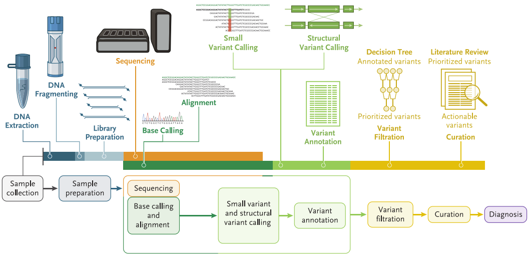
Longer reads improve structural variant detection



Oxford Nanopore is portable (space!) and fast

Oxford Nanopore 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



Ultrarapid Genome Sequencing Pipeline

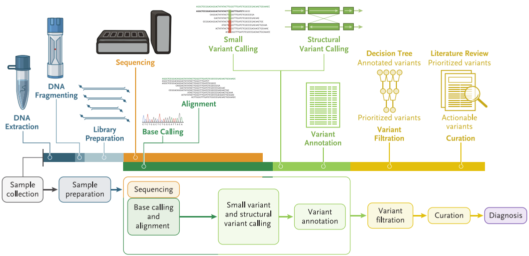
Gorzynski et al. N. Engl. J. Med. 2022

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

Oxford Nanopore 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

Ultrarapid Genome Sequencing Pipeline


DNA Extraction
DNA Fragmenting
Library Preparation
Sequencing
Base Calling
Alignment
Small Variant Calling
Structural Variant Calling
Variant Annotation
Decision Tree
Annotated variants
Prioritized variants
Variant Filtration
Literature Review
Prioritized variants
Actionable variants
Curation

Sample collection → Sample preparation → Sequencing
Base calling and alignment → Small variant and structural variant calling → Variant annotation → Variant filtration → Curation → Diagnosis

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

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

"Fastest DNA sequencing technique": 5h2m

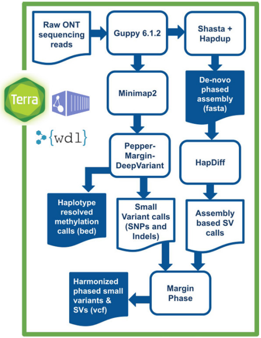
Cost-efficient Nanopore pipeline

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



Raw ONT sequencing reads

Guppy 6.1.2

Shasta + Hapdup

Minimap2

De-novo phased assembly (fasta)

Pepper-Margin-DeepVariant

HapDiff

Haplotype resolved methylation calls (bed)

Small Variant calls (SNPs and Indels)

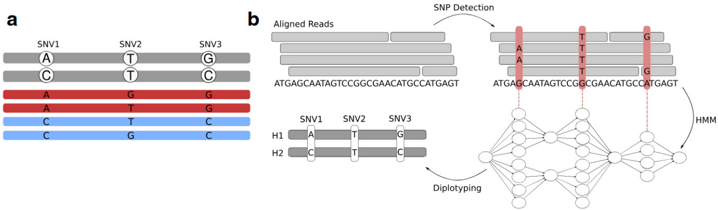
Assembly based SV calls

Harmonized phased small variants & SVs (vcf)

Margin Phase

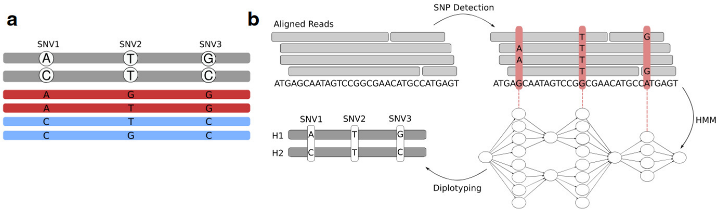
Kolmogorov*, Billingsley*, et al. Nature Methods 2023

Under the hood: phased variants and methylation calls

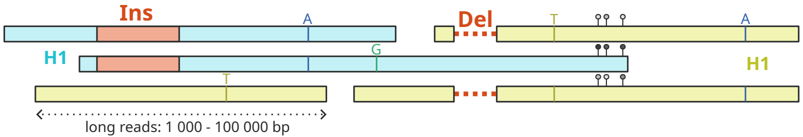


Reads are **haplo-tagged** using information across heterozygous sites with Margin (Ebler*, Haukness*, Pesout*, et al. Genome Biology 2019).

Under the hood: phased variants and methylation calls



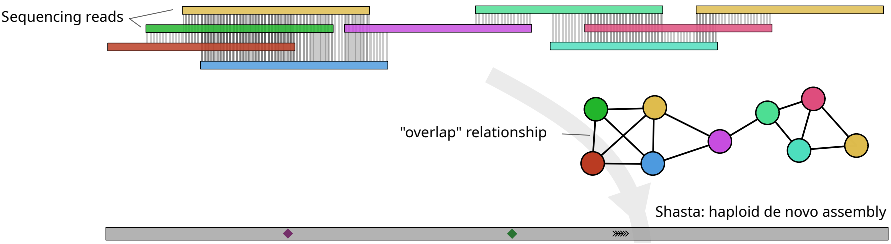
Reads are **haplo-tagged** using information across heterozygous sites with Margin (Ebler*, Haukness*, Pesout*, et al. Genome Biology 2019).



Phased small variants (DeepVariant) and methylation calls (ModKit)

Under the hood: *de novo* genome assembly polishing

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

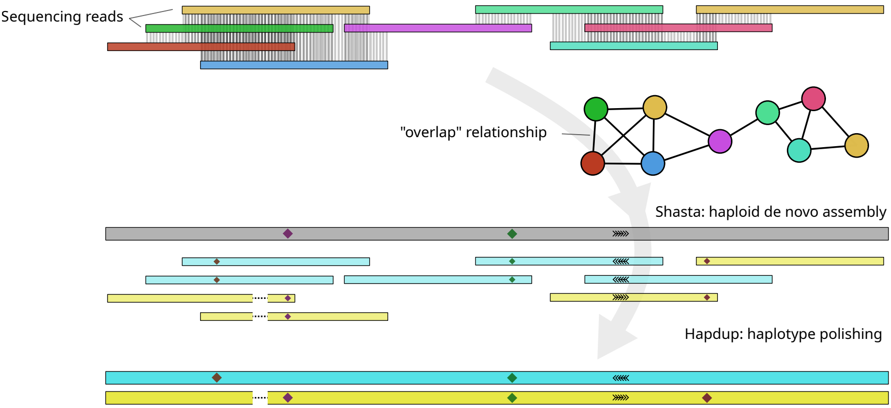


Shafin*, Pesout*, Lorig-Roach*, Haukness*, Olsen*, et al. Nat. Biotechnol. 2020

Kolmogorov*, Billingsley*, et al. Nature Methods 2023

Under the hood: *de novo* genome assembly polishing

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



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

Kolmogorov*, Billingsley*, et al. Nature Methods 2023

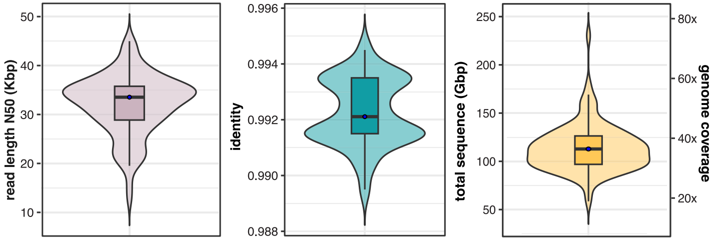
Application to a cohort of rare disease patients

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).

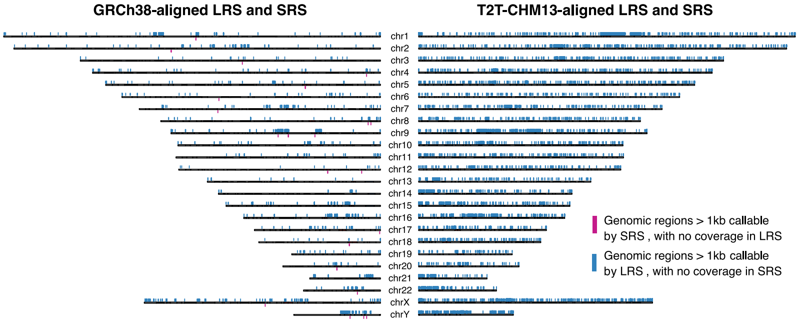


Negi et al. AJHG 2025

Better coverage of confidently mapped reads

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

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

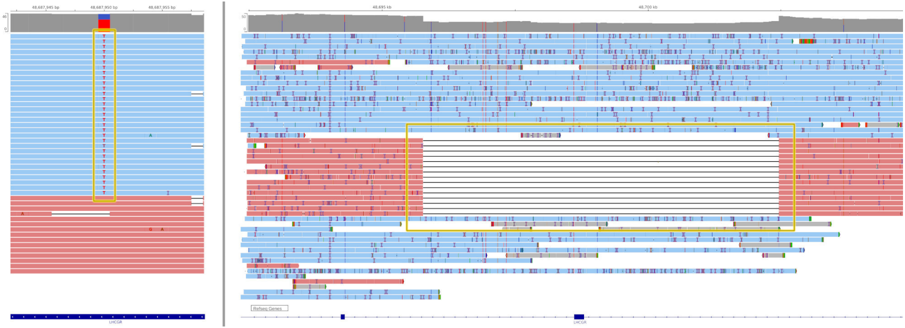


Negi et al. AJHG 2025

Resolving compound heterozygous variants

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

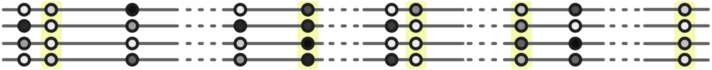
- Coding SNV on haplotype 1 (left, blue reads)
- ∼7 Kbp deletion of an exon on haplotype 2 (right, red reads)



Negi et al. AJHG 2025

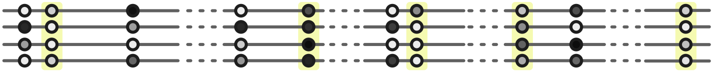
Known episignature of Coffin-Siris syndrome 1

Episignature: methylation pattern, across 10-100s of sites, associated with disease.



Known episignature of Coffin-Siris syndrome 1

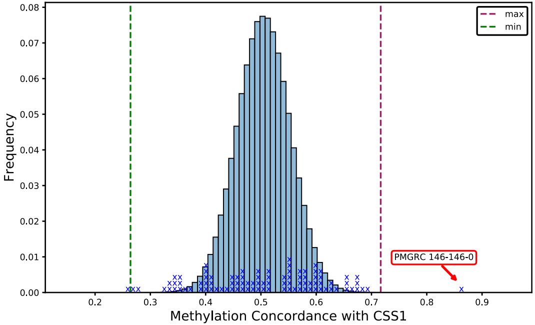
Episignature: methylation pattern, across 10-100s of sites, associated with disease.



One patient with suspected Coffin-Siris syndrome 1.

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

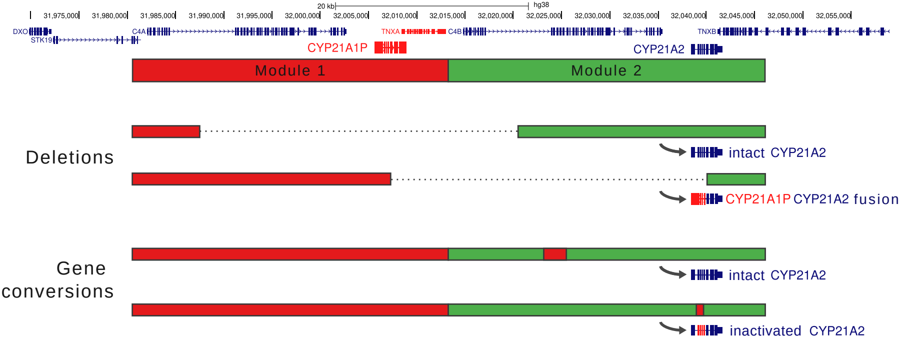
Negi et al. AJHG 2025



Pangenomes meet long-read sequencing

Challenging RCCX modules in the HLA region

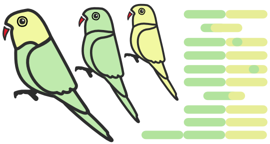
- Tandem-duplication of ∼30 Kbp genetic *module* (99% similar).
- CYP21A1P pseudogene and **CYP21A2 gene**.
- Variants cause congenital adrenal hyperplasia (recessive).

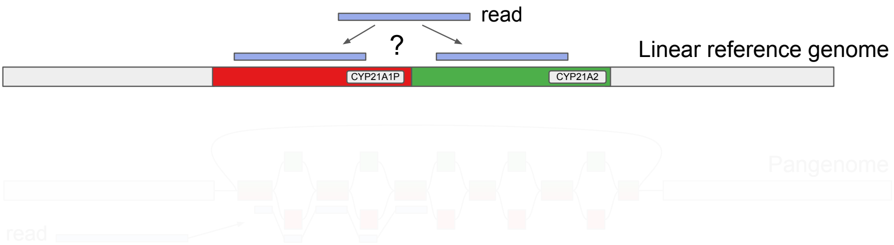


Parakit: paralog toolkit using collapsed pangenomes

Goal

Address multi-mapping confusion by mapping to a **collapsed pangenome** and by analyzing the alignment profile.



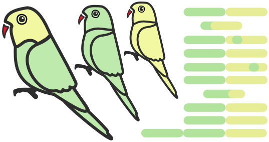


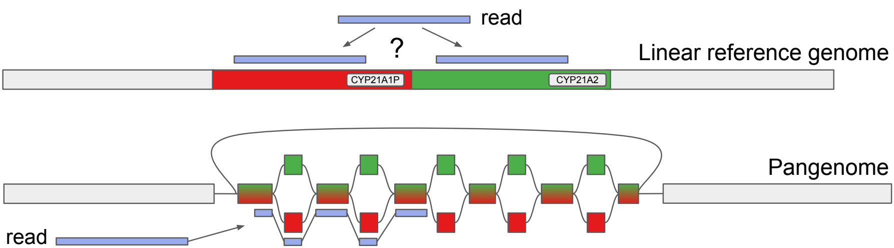
`https://github.com/jmonlong/parakit` Monlong et al. medRxiv 2025

Parakit: paralog toolkit using collapsed pangenomes

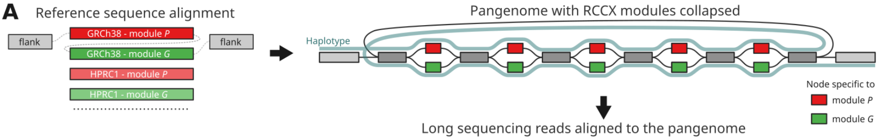
Goal

Address multi-mapping confusion by mapping to a **collapsed pangenome** and by analyzing the alignment profile.



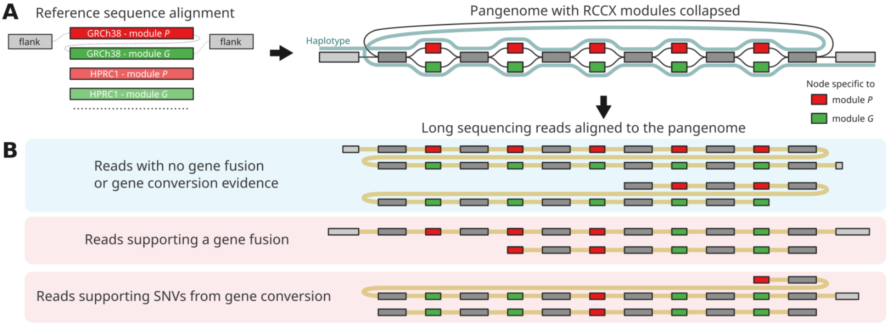


`https://github.com/jmonlong/parakit` Monlong et al. medRxiv 2025



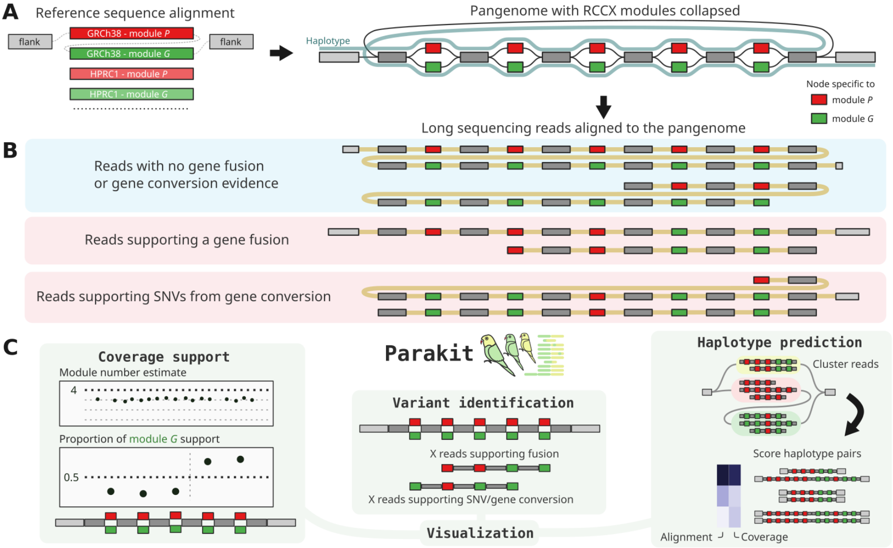
Parakit

`https://github.com/jmonlong/parakit` Monlong et al. medRxiv 2025


A
Reference sequence alignment
flank
GRCh38 - module P
GRCh38 - module G
HPRC1 - module P
HPRC1 - module G
flank
Pangenome with RCCX modules collapsed
Haplotype
Node specific to
module P
module G
Long sequencing reads aligned to the pangenome
B
Reads with no gene fusion
or gene conversion evidence
Reads supporting a gene fusion
Reads supporting SNVs from gene conversion

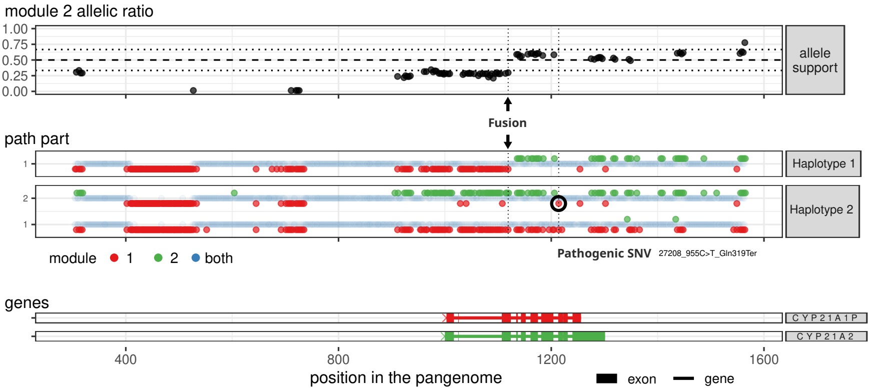
Parakit

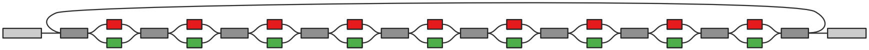
`https://github.com/jmonlong/parakit` Monlong et al. medRxiv 2025


A
Reference sequence alignment
flank
GRCh38 - module P
GRCh38 - module G
flank
HPRC1 - module P
HPRC1 - module G
Pangenome with RCCX modules collapsed
Haplotype
Node specific to
module P
module G
Long sequencing reads aligned to the pangenome
B
Reads with no gene fusion or gene conversion evidence
Reads supporting a gene fusion
Reads supporting SNVs from gene conversion
C
Coverage support
Module number estimate
4
Proportion of module G support
0.5
Parakit
Variant identification
X reads supporting fusion
X reads supporting SNV/gene conversion
Haplotype prediction
Cluster reads
Score haplotype pairs
Alignment
Coverage
Visualization

`https://github.com/jmonlong/parakit` Monlong et al. medRxiv 2025

Example: patients with a gene fusion and pathogenic SNV



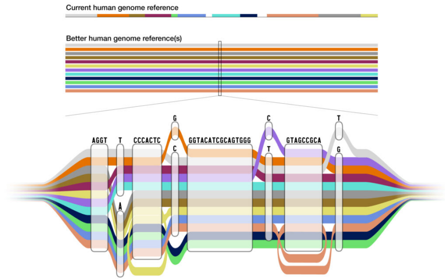


`https://github.com/jmonlong/parakit` Monlong et al. medRxiv 2025

Conclusions

Two approaches to integrate structural variants into genomic studies:

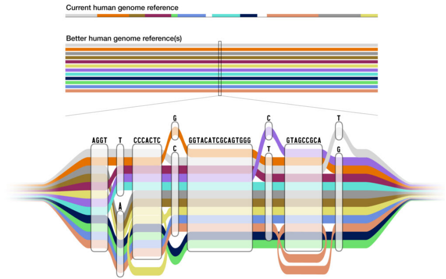
Genotyping with **pangenomes** from **short-read sequencing** data, e.g. for **genome-wide association studies**.


Current human genome reference
Better human genome reference(s)
AGGT
T
A
CCCACTC
G
C
GGTACATCGCAGTGGG
C
T
GTAGCCGCA
T
G

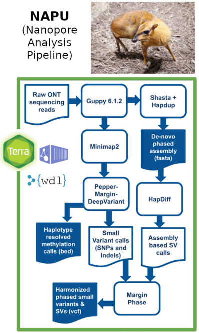
Conclusions

Two approaches to integrate structural variants into genomic studies:

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


Current human genome reference
Better human genome reference(s)
AGGT
T
CCCACTC
G
C
GGTACATCGCAGTGGG
C
T
GTAGCCGCA
T
G
A

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



Raw ONT sequencing reads → Guppy 6.1.2 → Shasta + Hapdup

Guppy 6.1.2 → Minimap2 → Pepper-Margin-DeepVariant

Shasta + Hapdup → De-novo phased assembly (fasta) → HapDiff → Assembly based SV calls

Pepper-Margin-DeepVariant → Haplotype resolved methylation calls (bed)

Pepper-Margin-DeepVariant → Small Variant calls (SNPs and Indels)

Small Variant calls (SNPs and Indels), Assembly based SV calls → Margin Phase → Harmonized phased small variants & SVs (vcf)

New bioinformatics avenues

Methods, tools, benchmark, and analysis needed !

Pangenomes with haplotype-resolved near-complete genomes.

Single-molecule long read sequencing (nanopore, PacBio).

New bioinformatics avenues

Methods, tools, benchmark, and analysis needed !

Pangenomes with haplotype-resolved near-complete genomes.

Single-molecule long read sequencing (nanopore, PacBio).

- Complex variants
- Repeat-rich regions
- Association studies
- Functional genomics
- Epigenomics
- ...

Acknowledgments

Univ. California, Santa Cruz

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

INSERM IRSD

- Sarah Djebali
- Matis Alias-Bagarre

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

anr agence nationale de la recherche

UC SANTA CRUZ
Genomics Institute

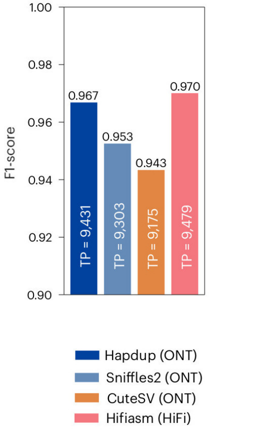
HUMAN PANGENOME

Children's National

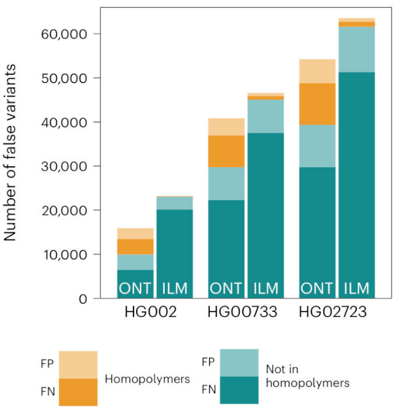
GREGoR consortium

Better calls for both small and structural variants...

SV concordance with GIAB HG002 benchmark

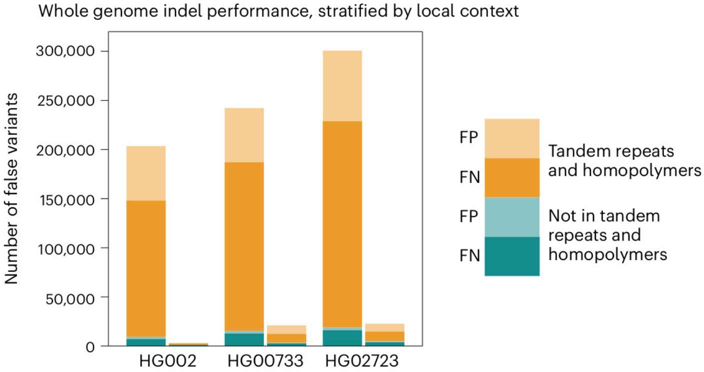


Whole genome SNP performance, stratified by local context



Kolmogorov*, Billingsley*, et al. Nature Methods 2023

...except for indels in homopolymers



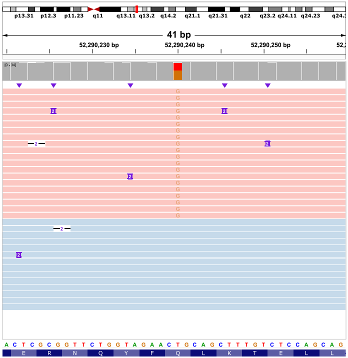
Note: Results above are for the R9 chemistry. The new R10 chemistry has lower error rate and better (indel) calling performance.

Kolmogorov*, Billingsley*, et al. Nature Methods 2023

Small variants found by long-reads only

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

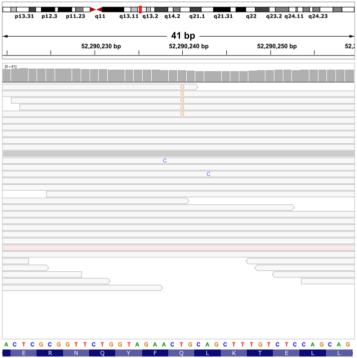
chr12:52,290,220-52,290,259



p13.31 p12.3 p11.23 q11 q13.11 q13.2 q14.2 q21.1 q21.31 q22 q23.2 q24.11 q24.23 q24.3
41 bp
52,290,230 bp
52,290,240 bp
52,290,250 bp
52,2
[0 - 34]
G
2
A C T C G C G G T T C T G G T A G A A C T G C A G C T T T G T C T C C A G C A G
E R N Q Y F Q L K T E L L

KRT86

long-reads



p13.31 p12.3 p11.23 q11 q13.11 q13.2 q14.2 q21.1 q21.31 q22 q23.2 q24.11 q24.23
41 bp
52,290,230 bp
52,290,240 bp
52,290,250 bp
52,2
[0 - 41]
G
C
A C T C G C G G T T C T G G T A G A A C T G C A G C T T T G T C T C C A G C A G
E R N Q Y F Q L K T E L L

KRT86

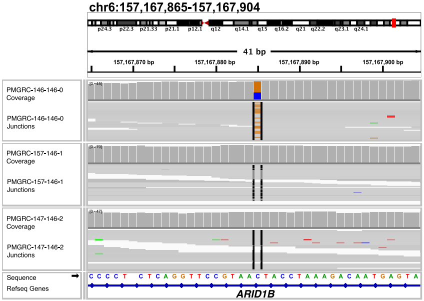
short-reads

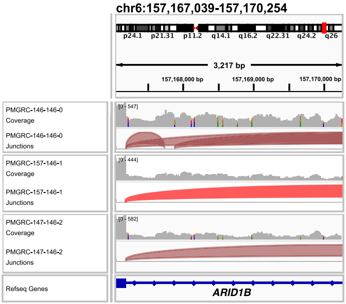
Negi et al. AJHG 2025

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).


chr6:157,167,865-157,167,904
p24.3 p22.3 p21.33 p21.1 p12.1 q12 q14.1 q15 q16.2 q21 q22.2 q23.1 q24.1
41 bp
157,167,870 bp
157,167,880 bp
157,167,890 bp
157,167,900 bp
PMGRC-146-146-0 Coverage
PMGRC-146-146-0 Junctions
[0 - 45]
PMGRC-157-146-1 Coverage
PMGRC-157-146-1 Junctions
[0 - 70]
PMGRC-147-146-2 Coverage
PMGRC-147-146-2 Junctions
[0 - 47]
Sequence
Refseq Genes
C C C C T C T C A G G T T C C G T A A C T A C C T A A A G A C A A T G A G T A
ARID1B



Negi et al. AJHG 2025