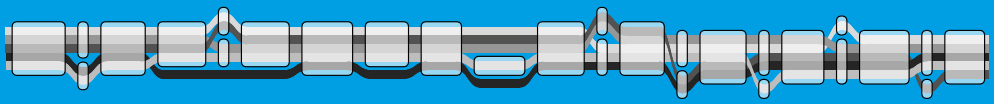


Intérêt d'une référence du pangénome humain pour l'étude des variants structuraux

Jean Monlong

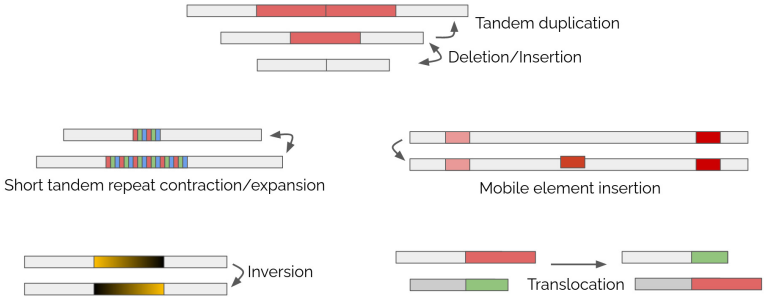
COLLOQUE ACLF
01/10/2025

irsd
Institut de recherche en santé digestive

Inserm
La science pour la santé
From science to health

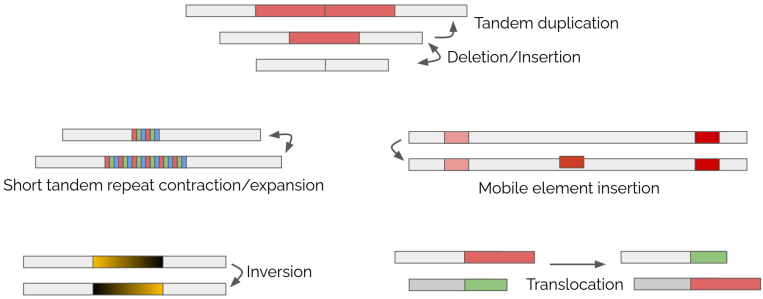
Structural variants (SVs) come in diverse shapes and sizes

Variant size: from 50 bases to megabases.



Structural variants (SVs) come in diverse shapes and sizes

Variant size: from 50 bases to megabases.

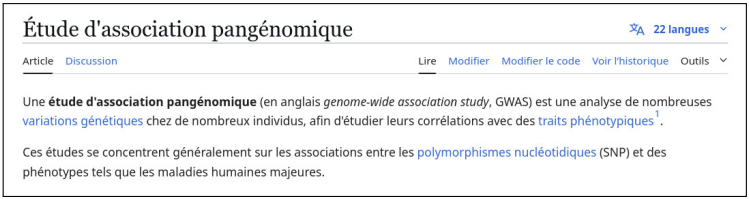


- High functional impact
- Involved in rare and common diseases, and cancers.
- **Hard to detect**

Pangenomics?

Pangenomics?

Not “Genome-wide” in French.



Étude d'association pangénomique

22 langues

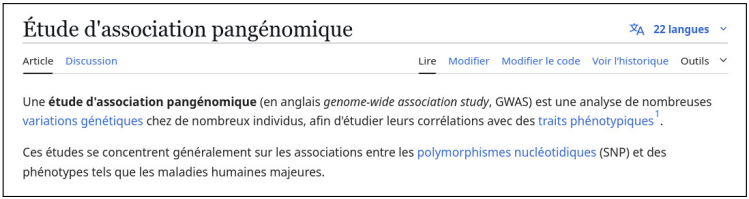
Article Discussion Lire Modifier Modifier le code Voir l'historique Outils

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?

Not “Genome-wide” in French.


Étude d'association pangénomique

22 langues

Article Discussion Lire Modifier Modifier le code Voir l'historique Outils

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].

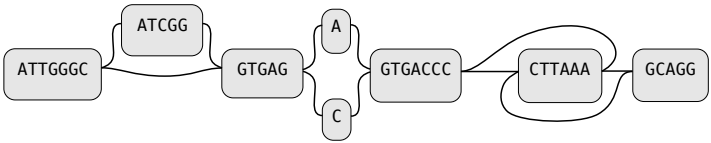
Ces études se concentrent généralement sur les associations entre les polymorphismes nucléotidiques (SNP) et des phénotypes tels que les maladies humaines majeures.

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

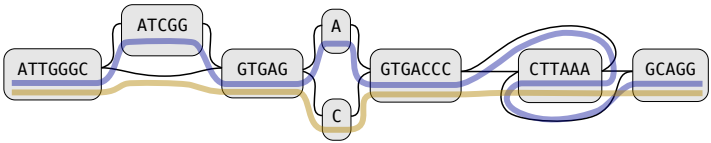


Pangenomes represent genetic diversity succinctly

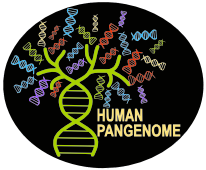
A pangenome represents a **collection of genomes** and the genetic variants among them.

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

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

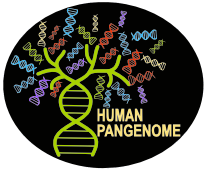


Building a Human pangenome reference

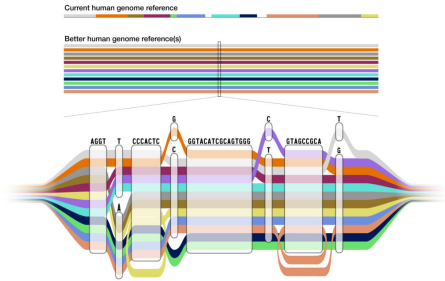


- Human Pangenome Reference Consortium (**HPRC**)
- Latest sequencing technologies for **350 diverse individuals**

Building a Human pangenome reference

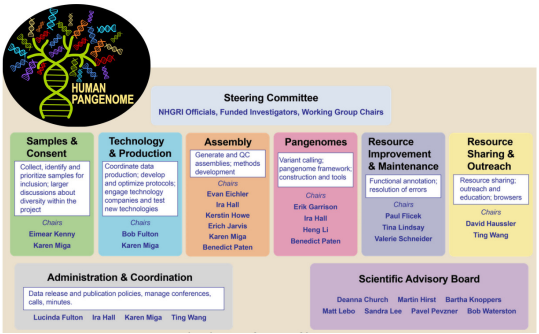


- Human Pangenome Reference Consortium (**HPRC**)
- Latest sequencing technologies for **350 diverse individuals**
- Generate **telomere-to-telomere phased assemblies**
- Pangenome containing a comprehensive catalog of (structural) variants



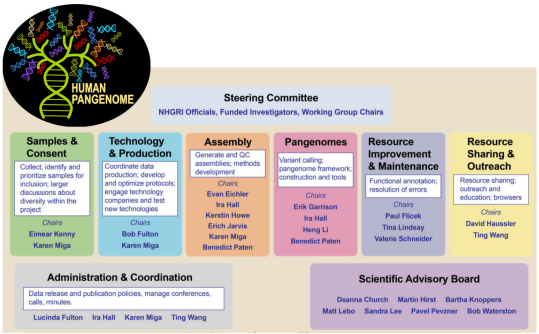
Hickey*, Monlong*, et al. Nat. Biotechnol. 2023

Building a Human pangenome reference



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

Building a Human pangenome reference

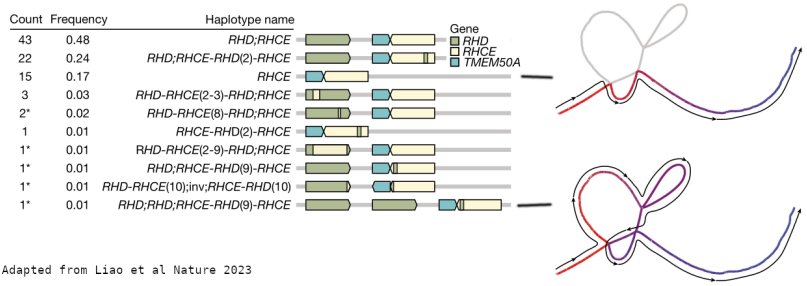


Liao*, Asri*, Ebler*, et al. Nature 2023

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

(Ongoing) Release 2: **466 haplotypes** of near-T2T quality.

Complex structural variants in the HPRC pangenome

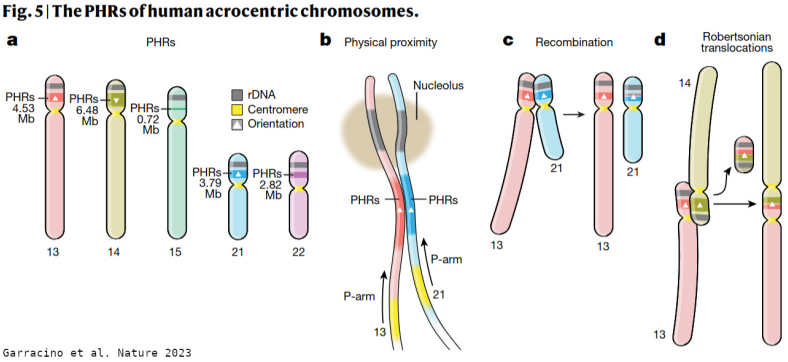


Adapted from Liao et al Nature 2023

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

Recombination between heterologous human acrocentric chromosomes

"we show that acrocentric chromosomes contain pseudo-homologous regions (PHRs) indicative of recombination between non-homologous sequences. "



Guarracino, et al. Nature 2023

The variation and evolution of complete human centromeres

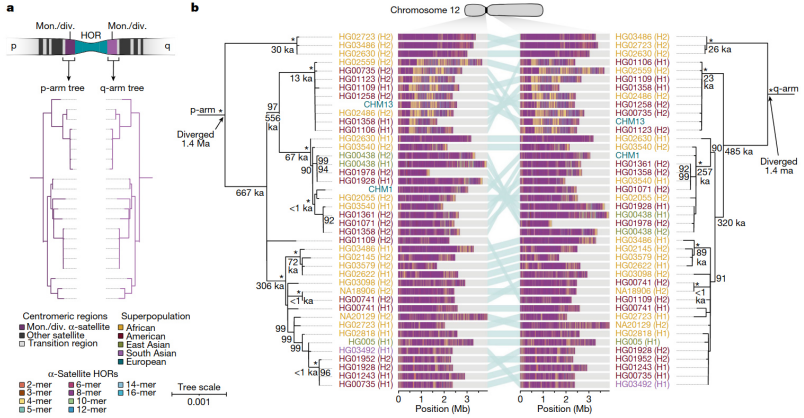


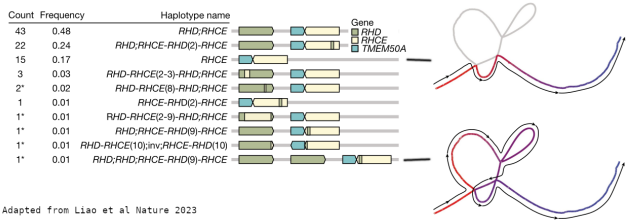
Fig. 7 | Phylogenetic reconstruction of human centromeric haplotypes and the saltatory amplification of new α-satellite HORs.

Logsdon et al. Nature 2024

Logsdon, et al. Nature 2024

Using the human pangenome resources

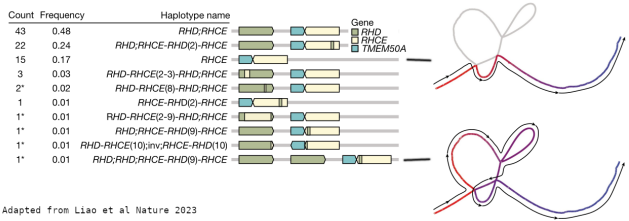
- Explore **high-quality assembled haplotypes** across diverse individuals



Adapted from Liao et al Nature 2023

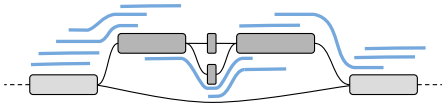
Using the human pangenome resources

- Explore **high-quality assembled haplotypes** across diverse individuals



Adapted from Liao et al Nature 2023

- **As a reference** to study new samples with sequencing data.
  1. Genotype SVs with short-read data
  2. Characterize complex SVs with long-read data

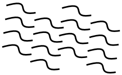


A pangenome reference to genotype SVs with short-read data

Genome sequencing

DNA

DNA fragments



Sequencing machines

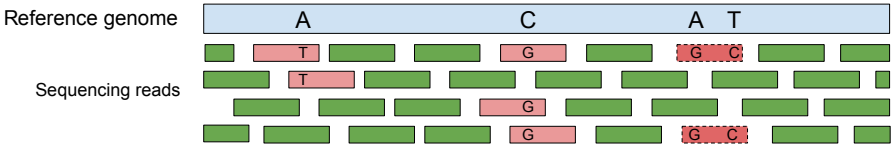
File (~100-300 Gb)

```
@ERR903030.219 HWI-D00574:82:C6L01ANXX:3:1101:3953:1913/1
AGCTCTTATTATTTTGAAATATGTCCCATCAATACCTAATTTATTGAGAGTTTTTAGCATGAAGGGTTGTTGAAT
+
<<A0?FGGGGGGGGGGGEGGGGGGGGGGGGGGGGG:CFCGDFGG>FG1F@0:FGGGGGF@>FG1=F@FFGGGGFGGC
@ERR903030.220 HWI-D00574:82:C6L01ANXX:3:1101:3863:1914/1
ACCATGAAAGACAGGTGTTAGAATCAGTACAAGAAGCAACAGGGAGCCATTGCATTTTGAGCATTTGTTGTATCA
+
<ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGG
@ERR903030.221 HWI-D00574:82:C6L01ANXX:3:1101:3906:1914/1
GATGGGGTTTCACATTGGCCAGGCTGGTCTCAAACTCCTAACCTCAAGGGATCCACCCACCTCGGCCTCCCAAAG
+
<<AAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGG>
@ERR903030.222 HWI-D00574:82:C6L01ANXX:3:1101:3833:1922/1
TACTGAAGAATCAGTGTCAGTTTTGTTAGTTGTGATAATGACATTCTGCTAAGCTAAAGTATAGAGGGGAGAAAT
+
=<BBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGFGFGEGEGGGGFGGG
@ERR903030.223 HWI-D00574:82:C6L01ANXX:3:1101:3942:1927/1
AAAAAATGAACTAAAAATGCATTAAAGACCAAATGTAATACCTAAAAATGTAAAACTTTTAGAAGGAAACATAGG
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@ERR903030.224 HWI-D00574:82:C6L01ANXX:3:1101:3920:1929/1
ATTCCATTCGATTCCATTCGATGATTCCATTGGATTCCATTCGATGATGATTCCATTCAAATCCATTCGATAATT
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGC>FFGGGEF@EDFGCEGFGGGGGGGGEGGGGCFB9EG
...
...
```

Sequencing reads

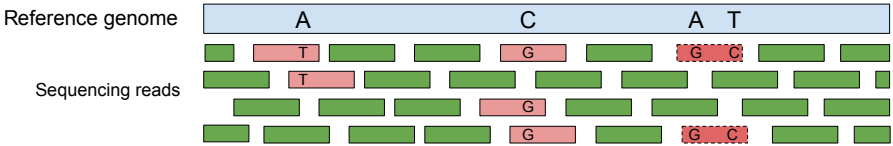
- **Short:** 150-250 bp (current tech)
- **Long:** 10,000s-100,000s bp (new tech. $$$)

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.

Aligning reads to a reference genome

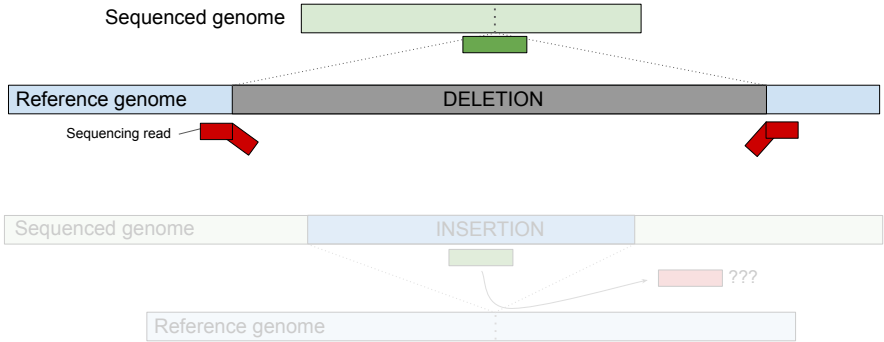


Assuming the reads are correctly placed, small variants are identified as recurrent differences between reads and the reference genome.

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

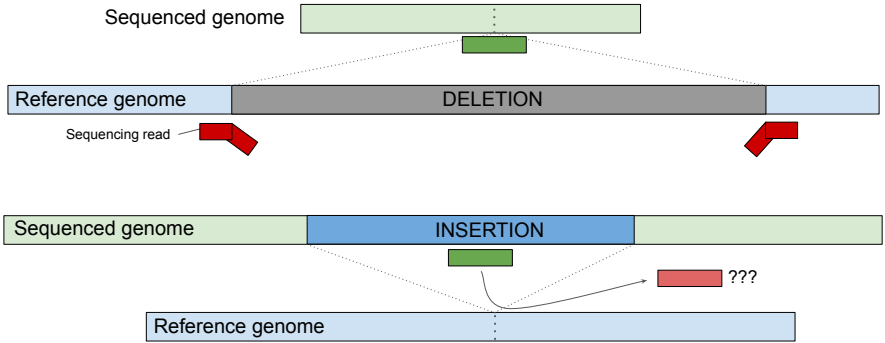
The challenges of structural variant detection

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

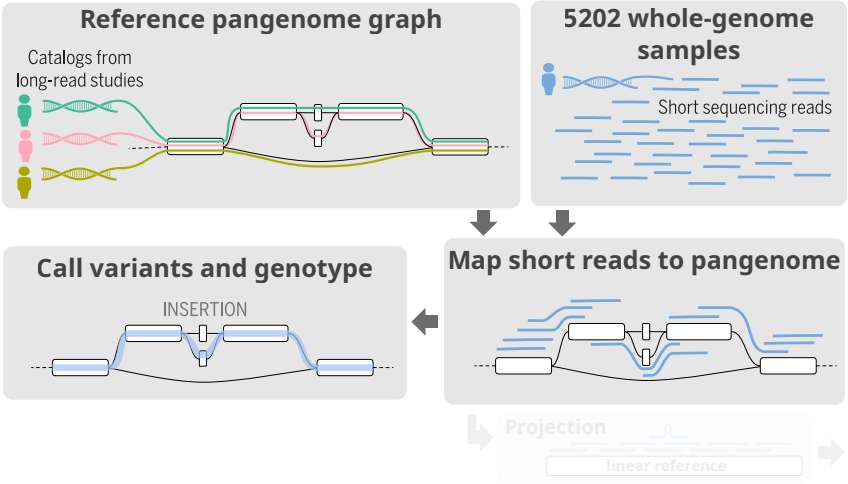


The challenges of structural variant detection

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

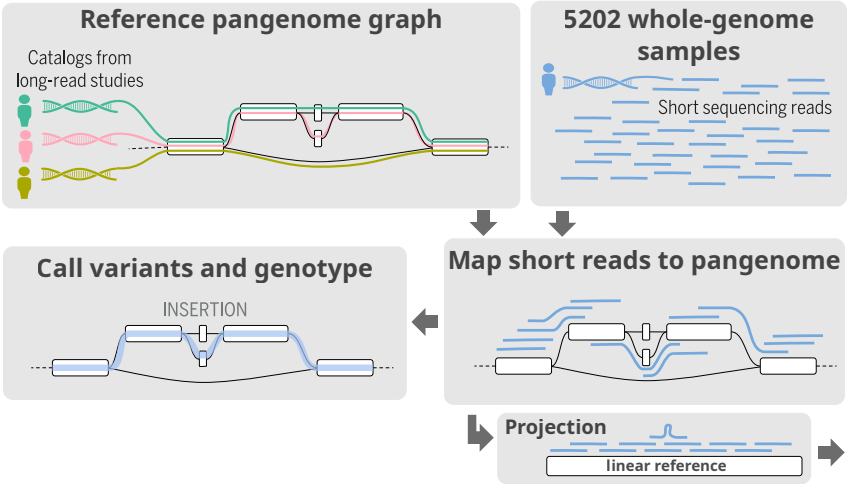


Short-read mapping and structural variant genotyping



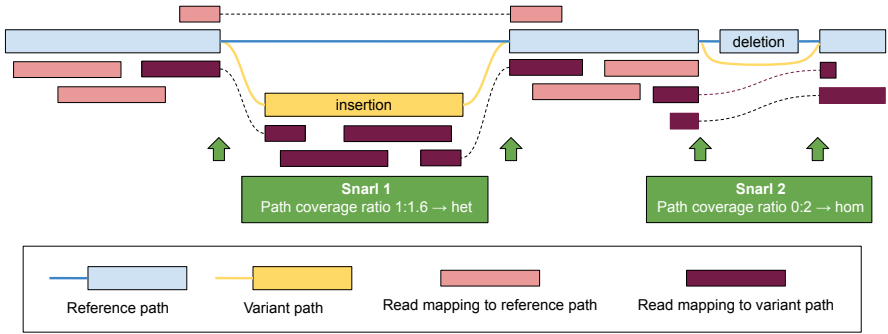
Siren*, Monlong*, Chang*, Novak*, Eizenga*, et al. Science 2021

Short-read mapping and structural variant genotyping



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

Genotyping structural variation from pangenomic mapping

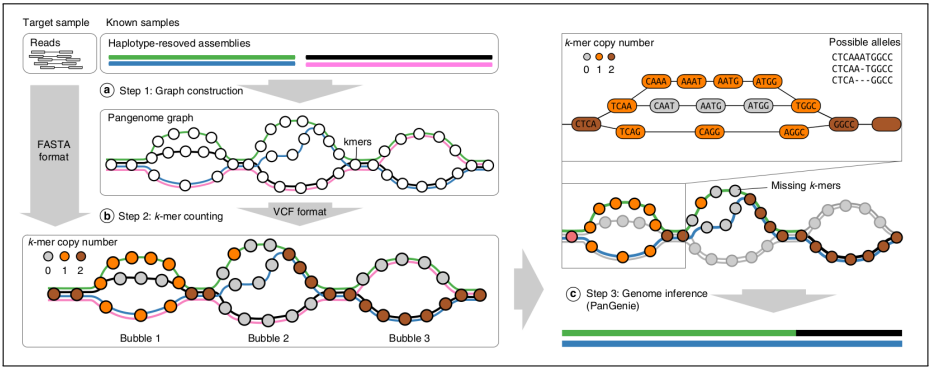


`https://github.com/vgteam/vg`

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

Genotyping structural variation from phased variants

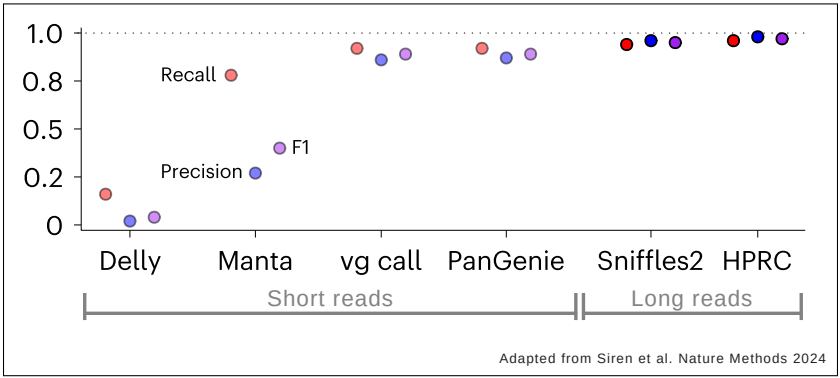
PanGenie uses k-mer and haplotype information to genotype SVs.



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

Ebler et al. Nature Genetics 2022

Structural variant genotyping performance

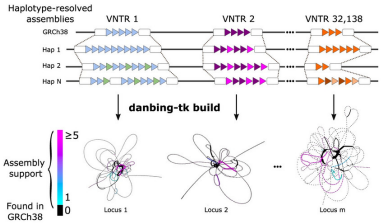


*vg call and PanGenie using the "personalized pangenome" approach (Sirén et al. Nature Methods 2024).

Other specialized methods

Variable Number Tandem Repeats (VNTR) characterization

danbing-tk genotypes them or predict their length from short read data.

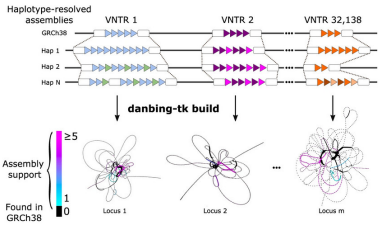


Lu et al. Nature communications 2021

Other specialized methods

Variable Number Tandem Repeats (VNTR) characterization

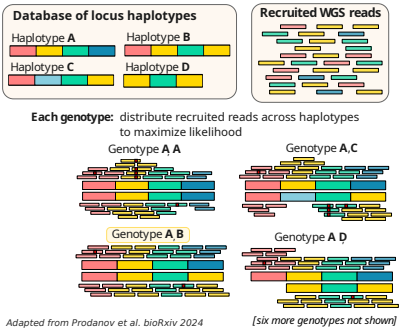
danbing-tk genotypes them or predict their length from short read data.



Lu et al. Nature communications 2021

Targeted genotyping of complex polymorphic genes

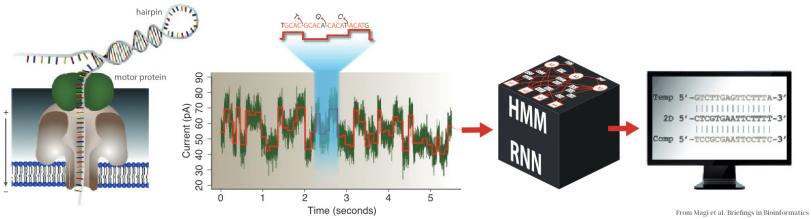
Locityper finds the best pair of known haplotypes from short reads.



Prodanov et al. bioRxiv 2024

A pangenome reference to characterize complex SVs with long-read data

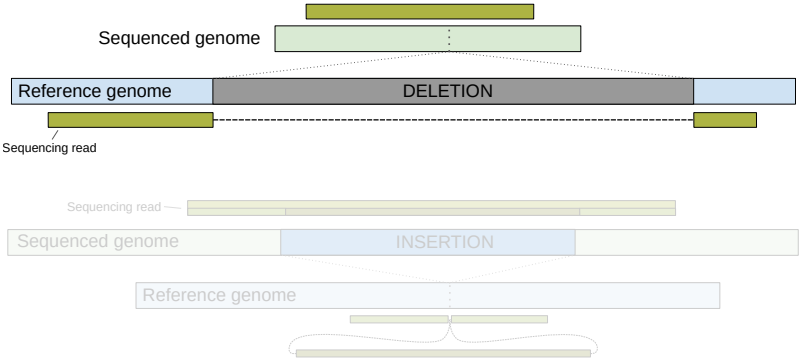
Long-read sequencing with Oxford Nanopore Technologies



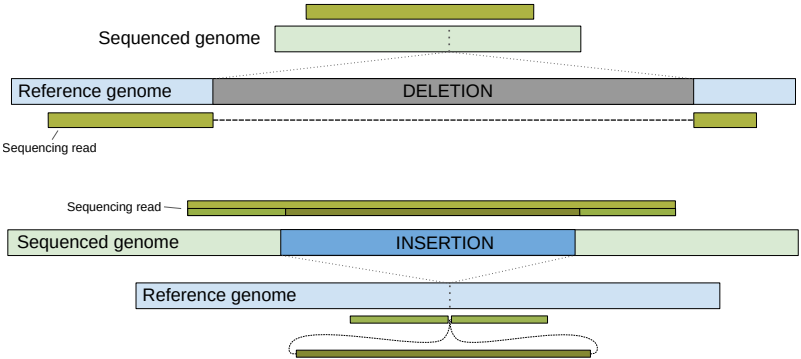
As the DNA (or RNA) fragment passes through the pore, the current changes and is decoded to predict nucleotides.

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

Longer reads improve structural variant detection



Longer reads improve structural variant detection



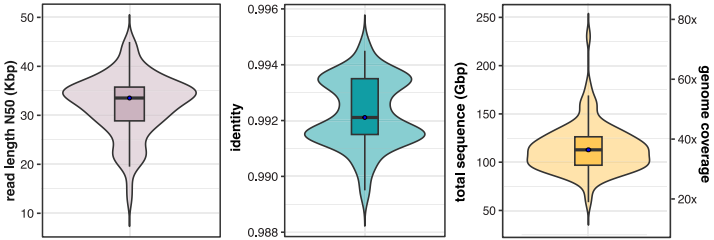
Application to a cohort of rare disease patients

Chan Zuckerberg Initiative

Children's National

GREGoR consortium

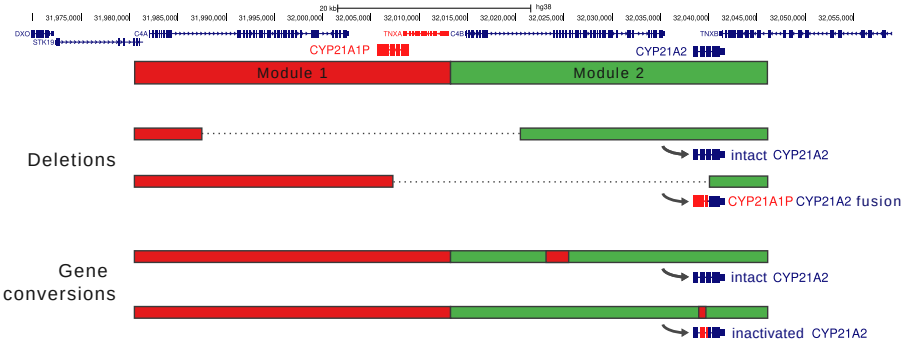
42 probands and 56 unaffected family members, sequenced with one-flowcell of ONT long-read sequencing (R10).



Negi et al. AJHG 2025

Challenging RCCX modules in the HLA region

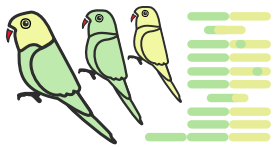
- Tandem-duplication of ∼30 Kbp genetic *module* (99% similar).
- CYP21A1P pseudogene and **CYP21A2 gene**.
- Variants cause congenital adrenal hyperplasia (recessive).

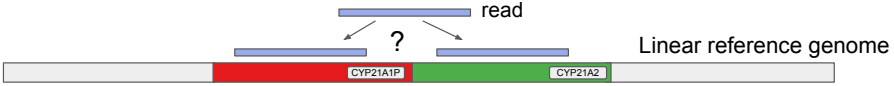


Parakit: paralog toolkit using collapsed pangenomes

Goal

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



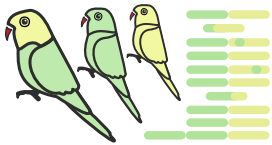


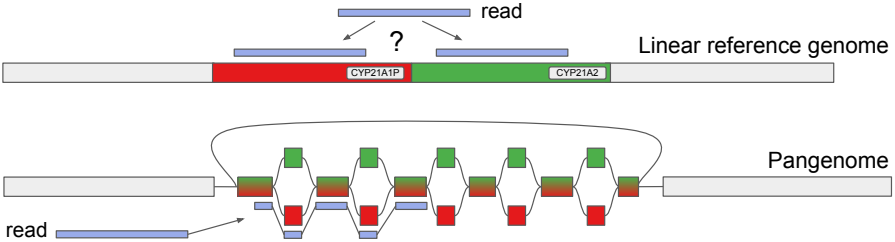
`https://github.com/jmonlong/parakit` Monlong et al. medRxiv 2025

Parakit: paralog toolkit using collapsed pangenomes

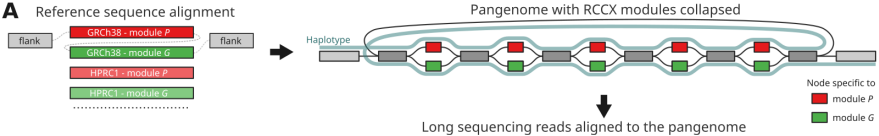
Goal

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



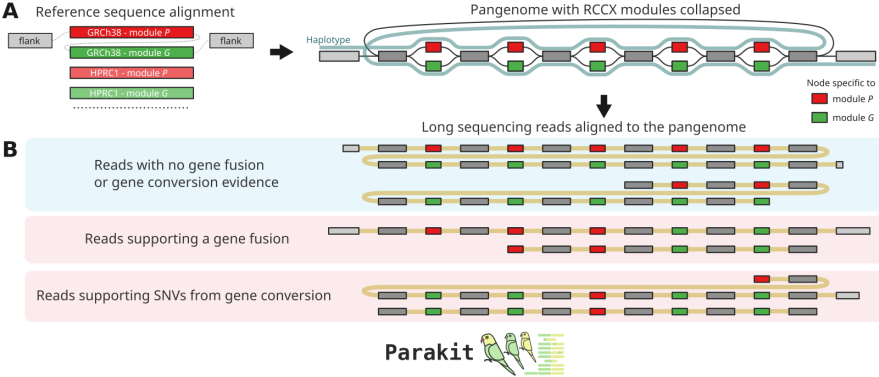


`https://github.com/jmonlong/parakit` Monlong et al. medRxiv 2025

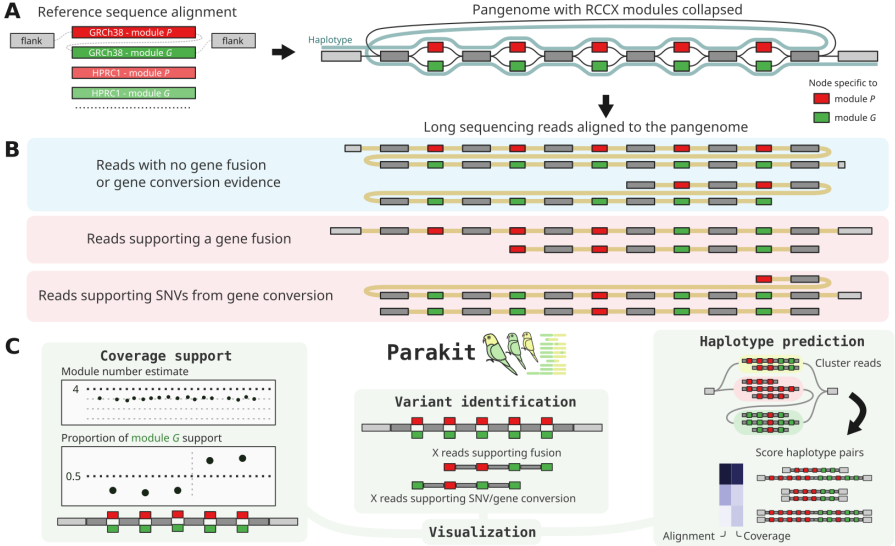


Parakit

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

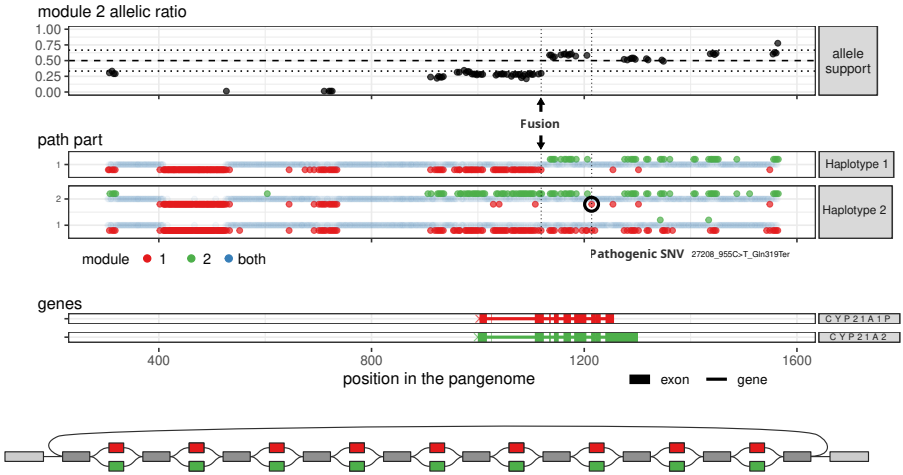


`https://github.com/jmonlong/parakit` Monlong et al. medRxiv 2025



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

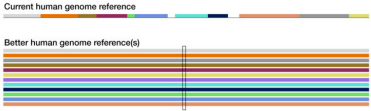
Example: patients with a gene fusion and pathogenic SNV

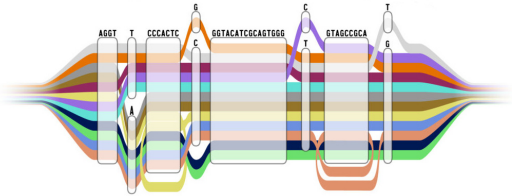


https://github.com/jmonlong/parakit Monlong et al. medRxiv 2025

The human pangenome

→ Complex structural variants at unprecedented resolution





→ Augmented (pan)genomic reference to:

- Genotype SVs from **short-read sequencing** data.
- Characterize complex SVs with **long-read sequencing**.

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

Univ. California, Irvine

- **Emmanuèle Délot**
- Eric Vilain

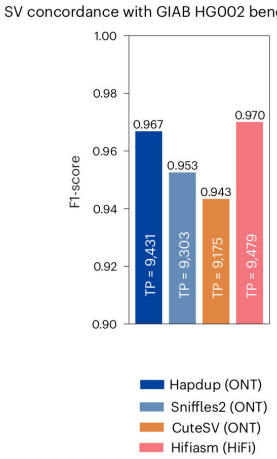
Children's National Research Institute

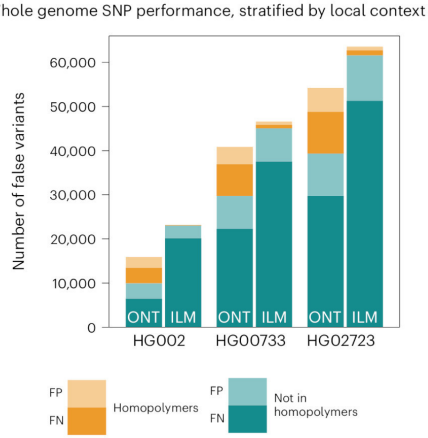
- Seth Berger
- Paolo Canigiula

INRAE

- Xian Chang
- Matthias Zytnicki

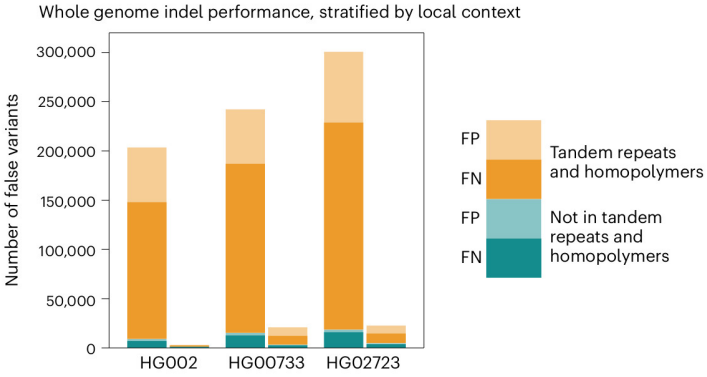
Better calls for both small and structural variants...





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

Small variants found by long-reads only

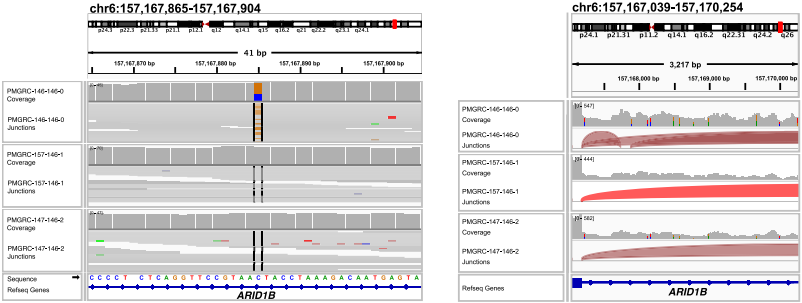
Missense mutation in *KRT86* disease gene (monilethrix) invisible with short reads.

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).



Negi et al. AJHG 2025

Trimodular alleles also detected