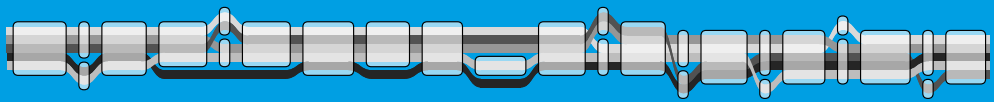


Characterising the CYP21A2 gene with Parakit

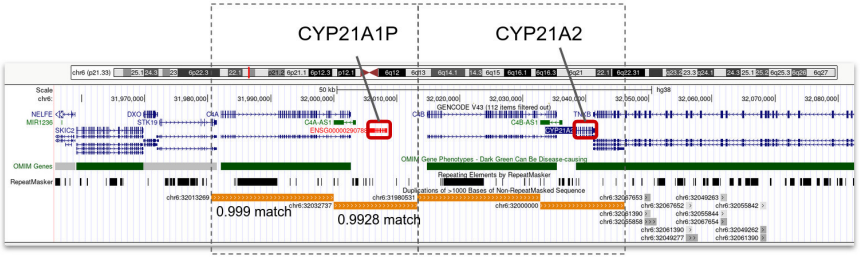
Jean Monlong

International Genome Graph Symposium
02/07/2024

Inserm
La science pour la santé
From science to health

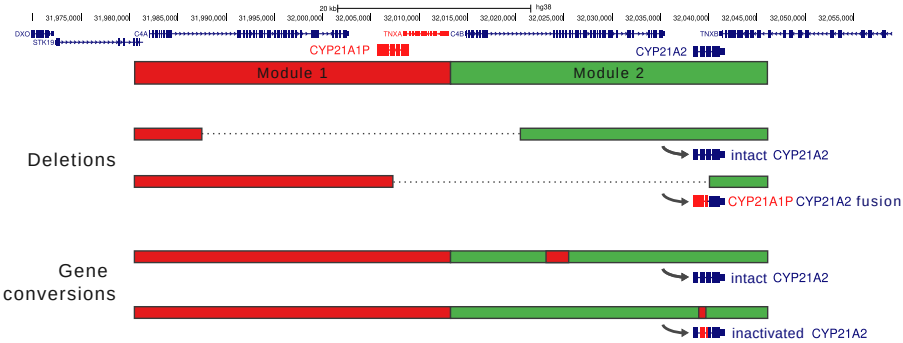
The RCCX module in the HLA region

- ∼30 Kbp genetic module containing the CYP21A2 gene or its paralog CYP21A1P.
- Reference and most individuals have bi-modular alleles.



Pathogenic SNVs/indels or deletions

Most common causes of CYP21A2 disruption and congenital adrenal hyperplasia (autosomal recessive).



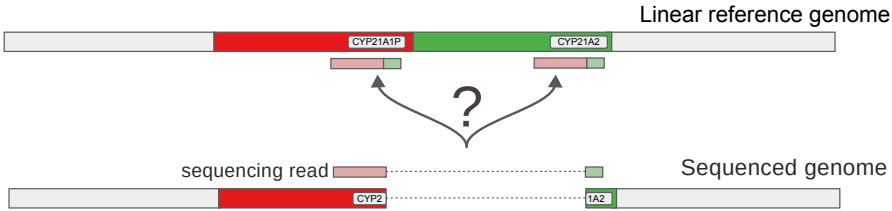
Challenges

- Clinical tests use combination of PCR amplification, Sanger sequencing, probe amplification (MLPA)
- Low resolution or low confidence.

Challenges

- Clinical tests use combination of PCR amplification, Sanger sequencing, probe amplification (MLPA)
- Low resolution or low confidence.

- Multi-mapping confuses variant caller with short reads.



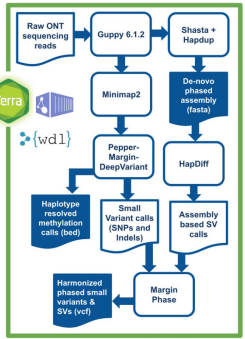
Cost-efficient Nanopore pipeline

- Only one flow-cell of Nanopore
  - ~30X coverage
  - 30 Kbp read N50
  - ~99% accurate

Cost-efficient Nanopore pipeline

- Only one flow-cell of Nanopore
  - ~30X coverage
  - 30 Kbp read N50
  - ~99% accurate
- **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)



Kolmogorov, Billingsley, et al. Nature Methods 2023

Cost-efficient Nanopore sequencing of rare disease patients

- **Four patients** suffering from congenital adrenal hyperplasia.
  (+ two parents)
- All sequenced with ONT ∼30X coverage, 30 Kbp N50, ∼99% accurate.

Cost-efficient Nanopore sequencing of rare disease patients

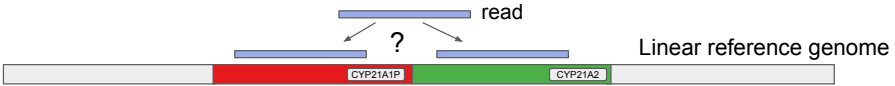
- **Four patients** suffering from congenital adrenal hyperplasia.
  (+ two parents)
- All sequenced with ONT ∼30X coverage, 30 Kbp N50, ∼99% accurate.

- NAPu identified some deletions and pathogenic SNVs.
- Some **missing a second hit**, others with **unreliable phasing**.

Parakit: paralog toolkit using collapsed pangenomes

Goal

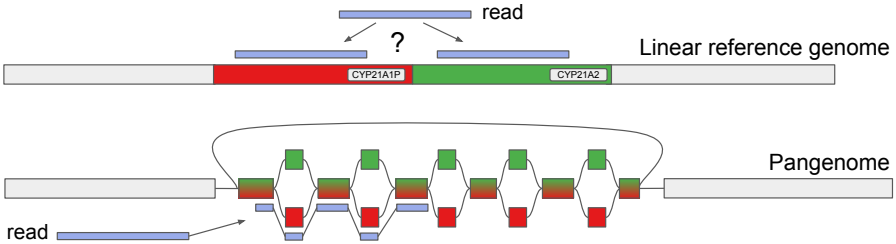
Address multi-mapping confusion by mapping to a **collapsed pangenome** and by analyzing the alignment profile.



Parakit: paralog toolkit using collapsed pangenomes

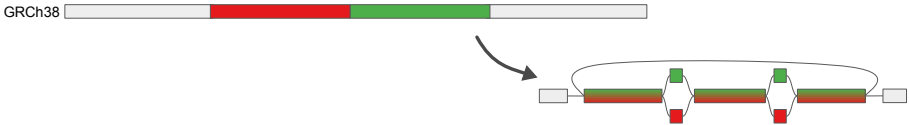
Goal

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



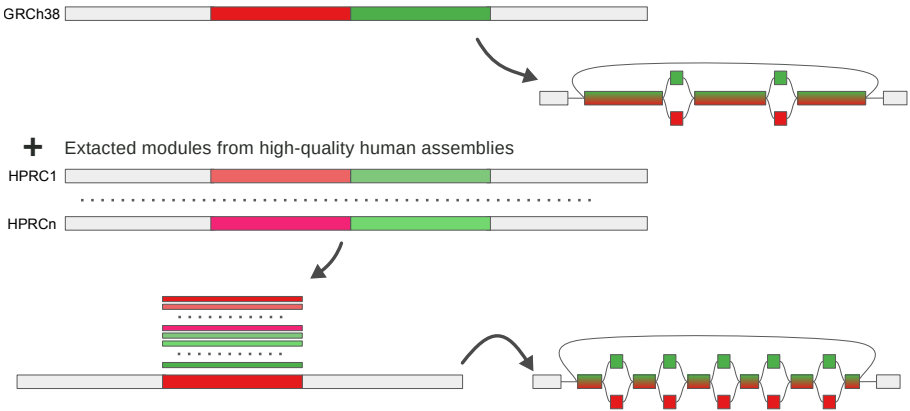
RCCX pangenome construction

Minigraph-Cactus with ugly tricks to force module collapsing.



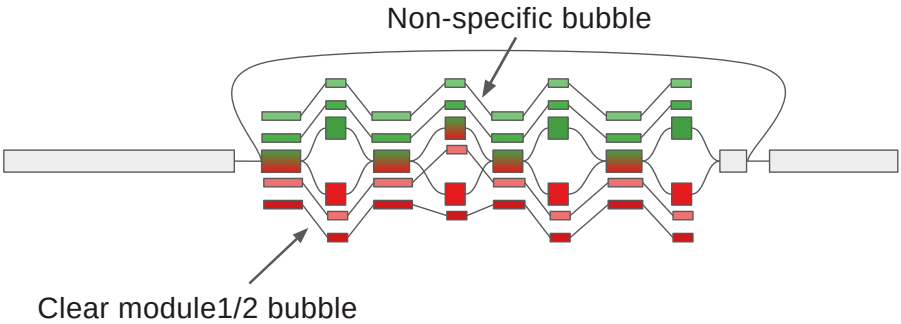
RCCX pangenome construction

Minigraph-Cactus with ugly tricks to force module collapsing.



Pangenome coloring

Identify informative nodes, i.e. specific to module 1 or 2.



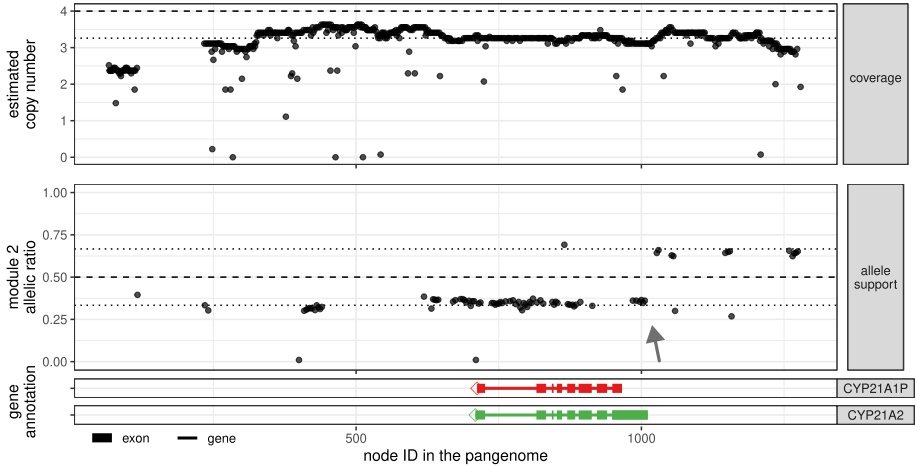
Reanalyzing sequencing reads with this pangenome

Local reads extracted and aligned to the pangenome (GraphAligner).

Then:

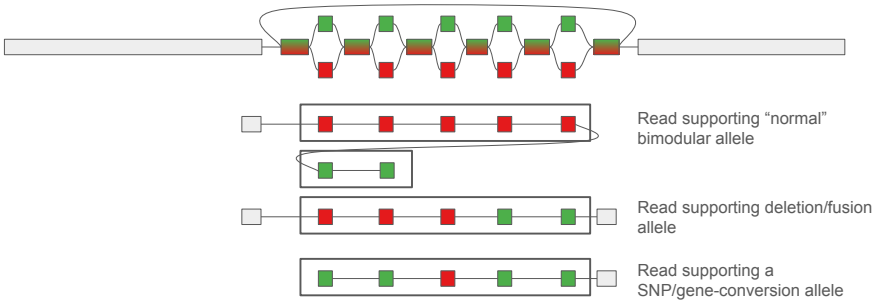
1. Compute **read coverage** along the module
2. Compute **allele support** on module-specific bubbles
3. Find **reads supporting pathogenic variants**
4. Predict **diplotype**

Coverage and allele support on module-specific bubbles



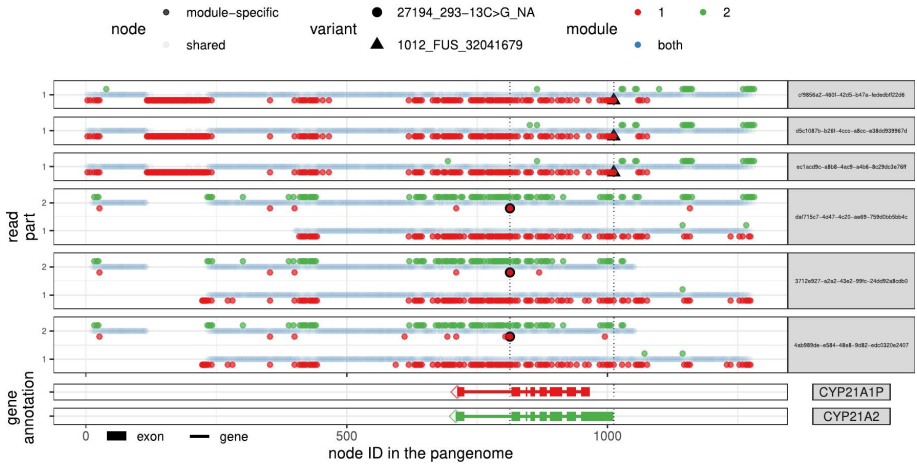
Variant calling from reads supporting a pathogenic signature

Reads represented by path through pangenome, esp. informative nodes.



Sliding window approach to find module switches or isolated nodes.

Read supporting fusion or known pathogenic variants



SNV is a known pathogenic ClinVar variant.

Diplotype prediction

1. Enumerate candidate haplotypes.

   Not too many but not necessarily two.

2. Select most likely pair based on read alignment and coverage.

Diplotype prediction

1. Enumerate candidate haplotypes.

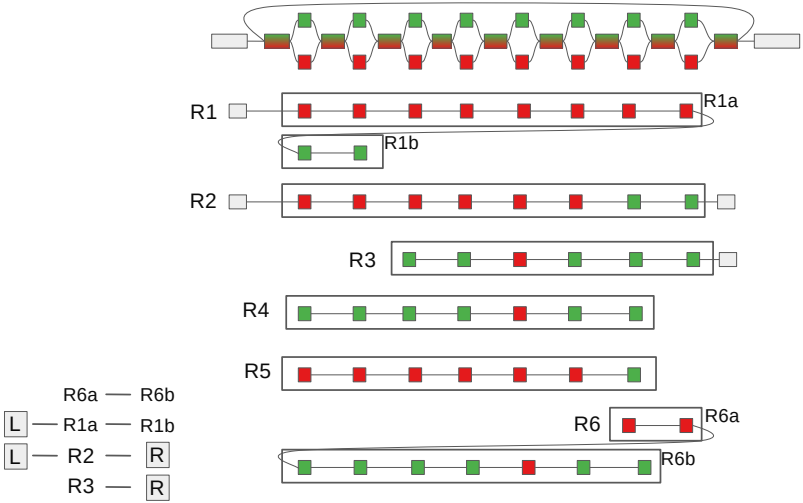
   Not too many but not necessarily two.

2. Select most likely pair based on read alignment and coverage.

Work-in-progress. Currently using a read clustering/consensus approach.

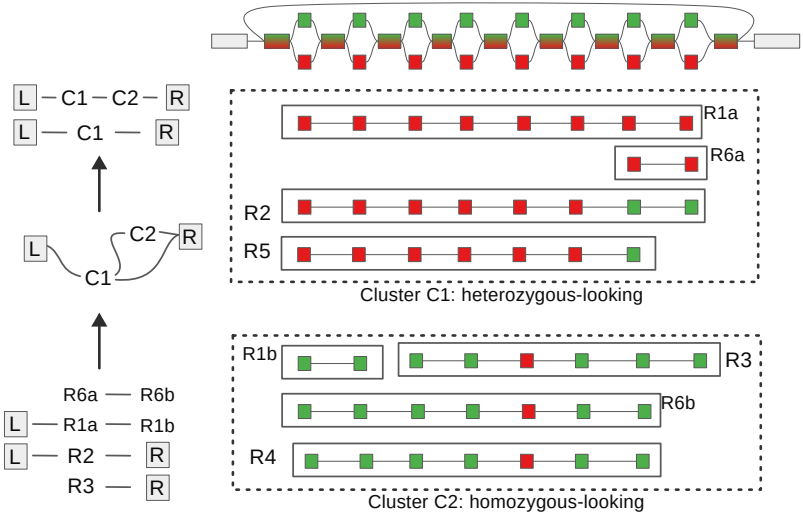
Cluster and consensus of module alleles

Clustering/consensus based on the pangenome profile, module-specific nodes only.



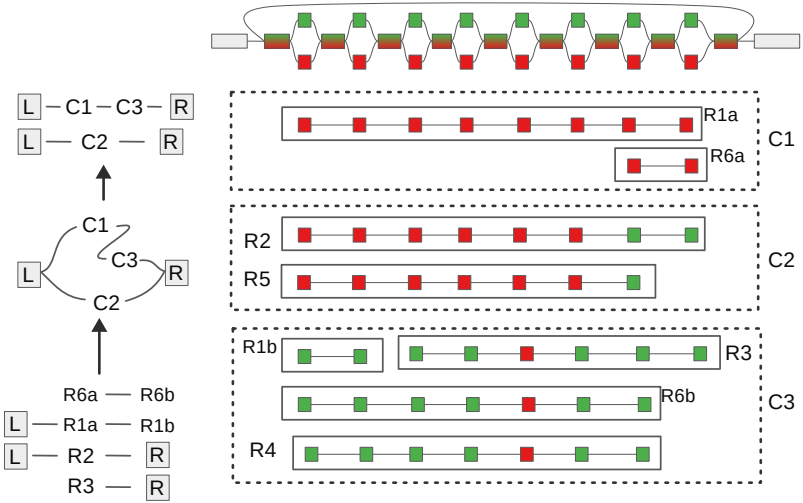
Cluster and consensus of module alleles

Clustering/consensus based on the pangenome profile, module-specific nodes only.



Cluster and consensus of module alleles

Clustering/consensus based on the pangenome profile, module-specific nodes only.



Rank candidate diplotypes

Read alignment

Is there a good place to map the reads, esp. long ones?

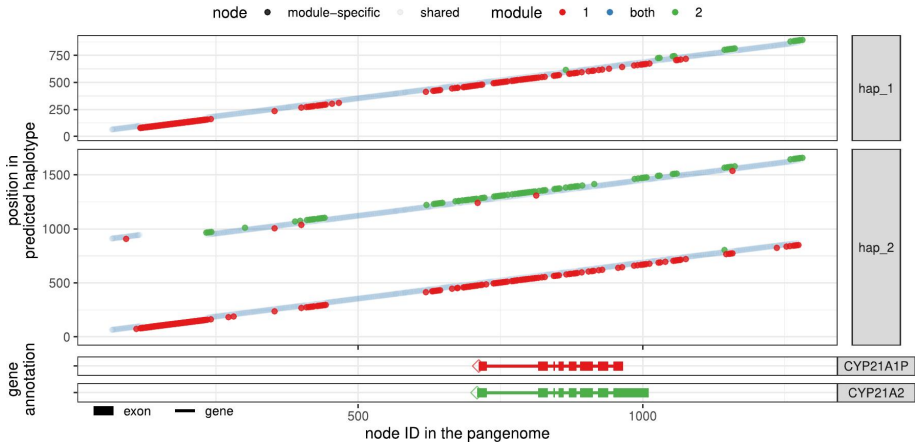
$\Rightarrow$ average alignment score, weighting reads by their length.

Read coverage

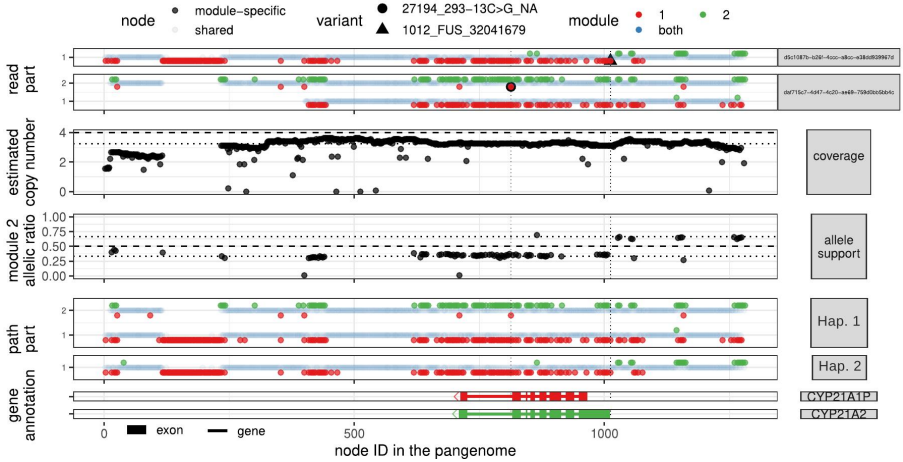
Are the diplotype copy numbers consistent with the read coverage?

$\Rightarrow$ Pearson correlation between node coverage in diplotype and reads.

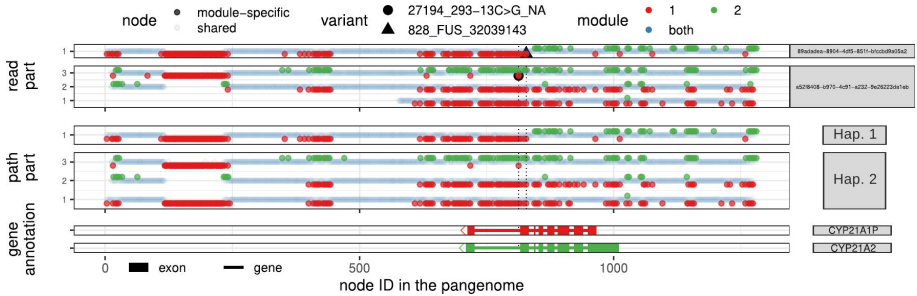
Predicted diplotype



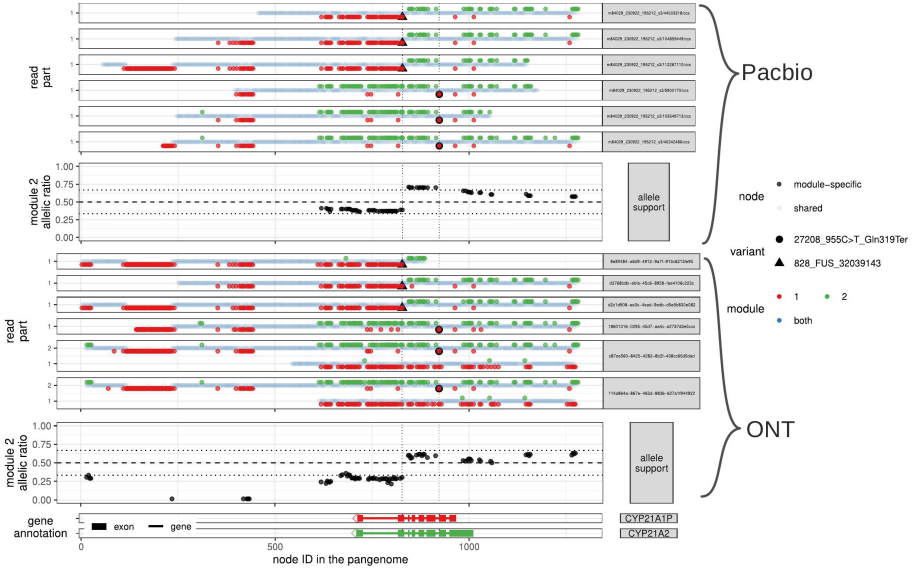
Summary figure



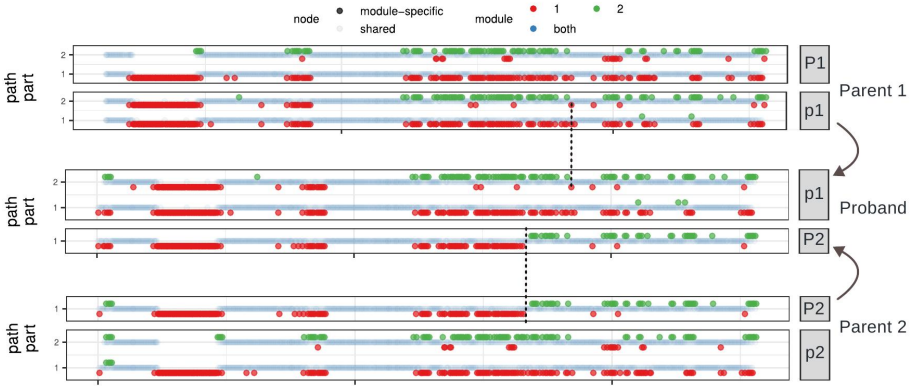
Trimodular alleles also detected



Comparable results with Pacbio HiFi reads



Consistent results with parents



Parakit

- Toolkit to characterize long-reads mapping to the RCCX region hosting CYP21A2.
- Visualize coverage, allele balance, variant-supporting reads, predicted diplotypes.
- Found **compound-heterozygous pathogenic variants** in 4/4 rare disease patients.

`https://github.com/jmonlong/parakit`

Parakit

- Toolkit to characterize long-reads mapping to the RCCX region hosting CYP21A2.
- Visualize coverage, allele balance, variant-supporting reads, predicted diplotypes.
- Found **compound-heterozygous pathogenic variants** in 4/4 rare disease patients.

`https://github.com/jmonlong/parakit`

Next

- Add new features (annotate contigs, better diplotype inference).
- *Manuscript in preparation*.
- Automate construction for other regions.

Acknowledgments

Univ. California, Santa Cruz

- **Benedict Paten**
- **Shloka Negi**
- Karen Miga
- Brandy McNulty
- Ivo Violich

Children's National Hospital

- **Emmanuèle Delot**
- Hayk Barseghyan
- Seth Berger
- Eric Vilain

Chan Zuckerberg Initiative

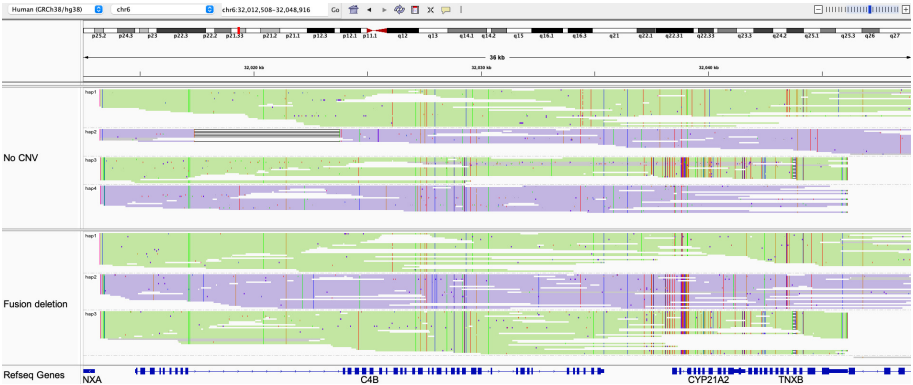
UC SANTA CRUZ Genomics Institute

HUMAN PANGENOME

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

Paraphase, a solution for high-fidelity Pacbio long reads

All reads aligned to the CYP21A2 module, then phased them into haplotypes.



`https://github.com/PacificBiosciences/paraphase`

Chen et al. AJHG 2023

RCCX gene annotation of the HPRC haplotypes