Expanding our understanding of human genetic variation through long-read sequencing of 1000 Genomes Project samples

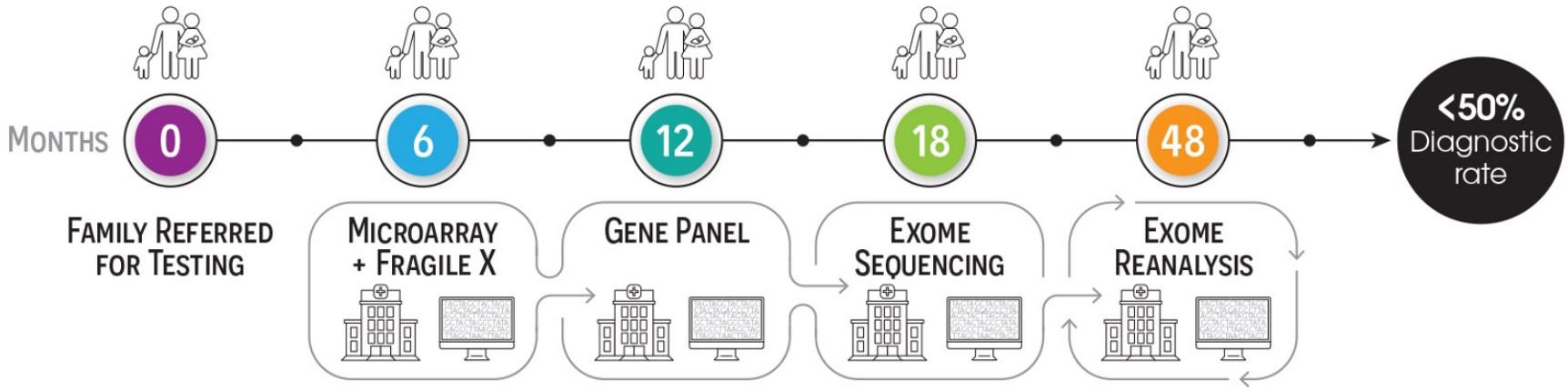
Danny E. Miller, MD, PhD
Assistant Professor
Department of Pediatrics, Division of Genetic Medicine
Department of Laboratory Medicine & Pathology

The University of Washington and Seattle Children's Hospital

ASHG
November 2, 2023

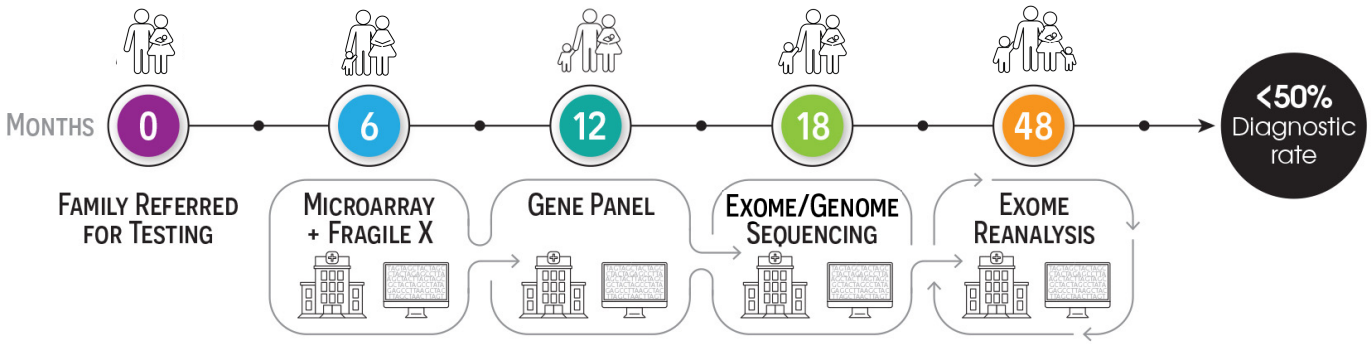
dm1@uw.edu | http://millerlaboratory.com | @danrdanny

Clinical genetic testing often occurs in a stepwise fashion, involving multiple tests and clinic visits

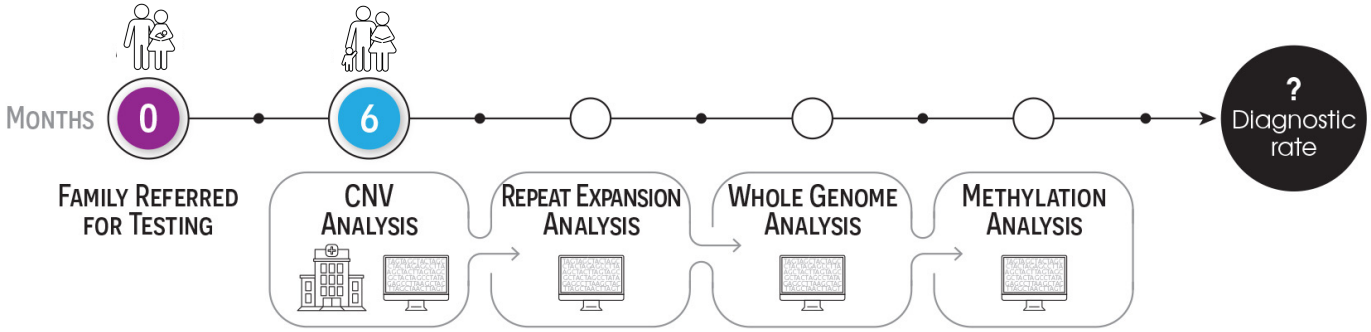


Frésard *et al*. 2018 and Clark *et al*. 2018; Gibson *et al*., Submitted

New technologies, such as long-read sequencing will increase the diagnostic rate

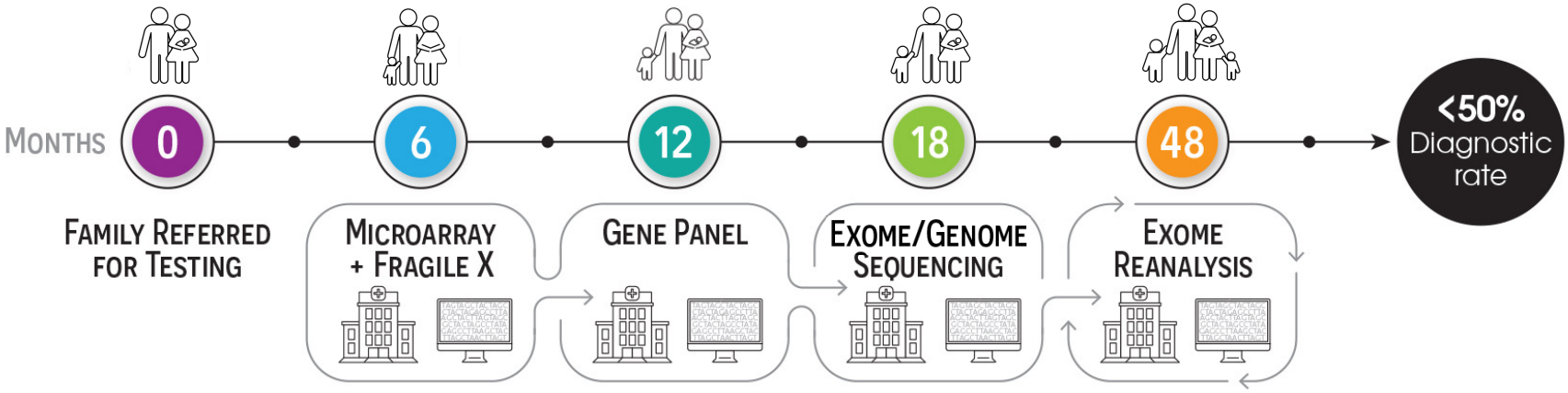


Long-read sequencing as a single test



Frésard *et al*. 2018 and Clark *et al*. 2018; Gibson *et al*., Submitted

A traditional genetic workup is diagnostic in less than 50% of cases



Incomplete gene-phenotype relationships

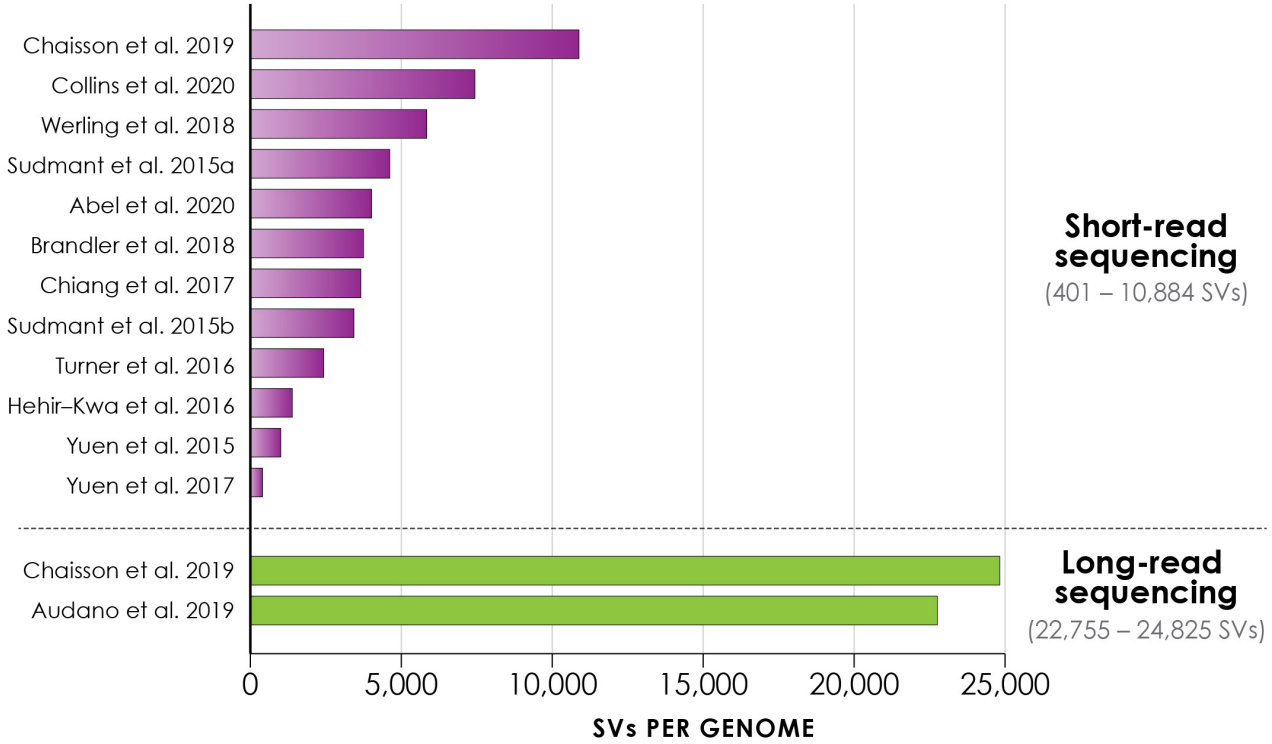
- We do not know the function of all genes

Variants that are difficult to detect or interpret

- Many genes are difficult to sequence
- Structural variants can be difficult to identify
- Predicting the impact of a variant is difficult

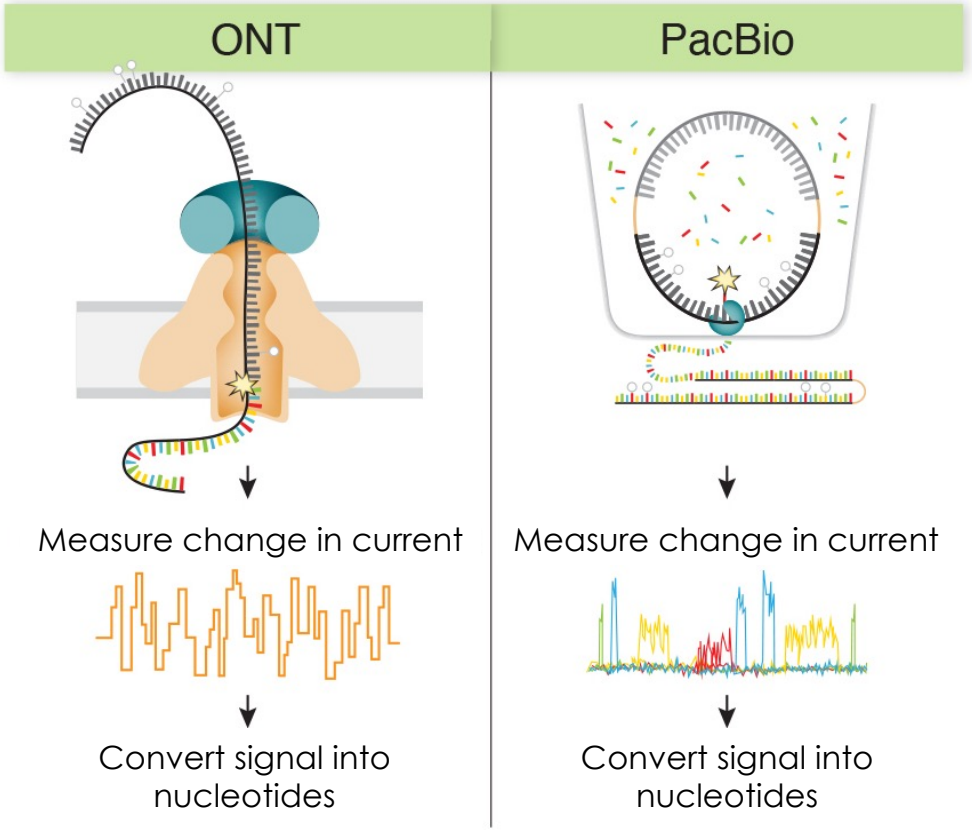
Frésard *et al*. 2018 and Clark *et al*. 2018; Gibson *et al*., Submitted

Short-read sequencing detects less than half the SVs seen by long-read sequencing

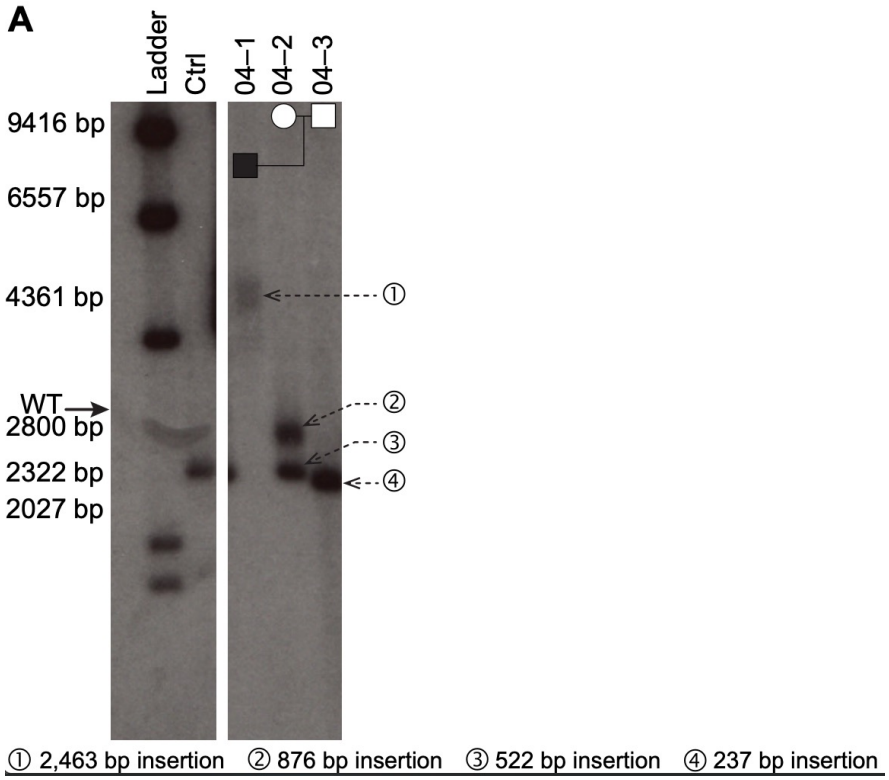


Adapted from Zhao *et al*. 2020

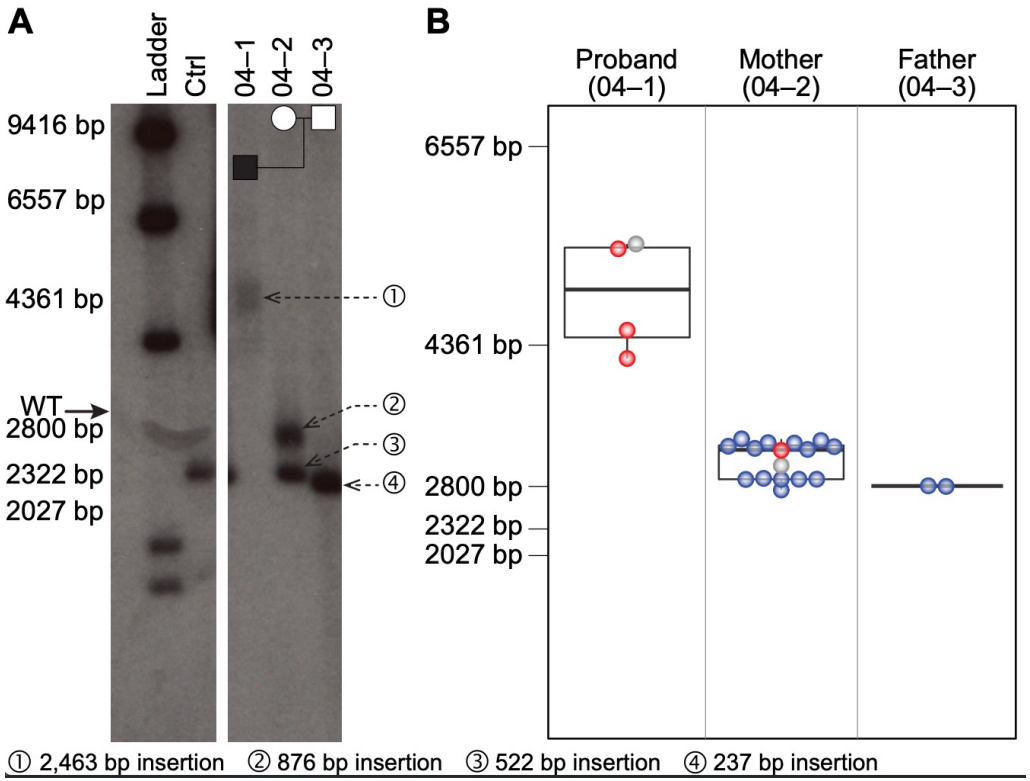
Long-read sequencing technologies



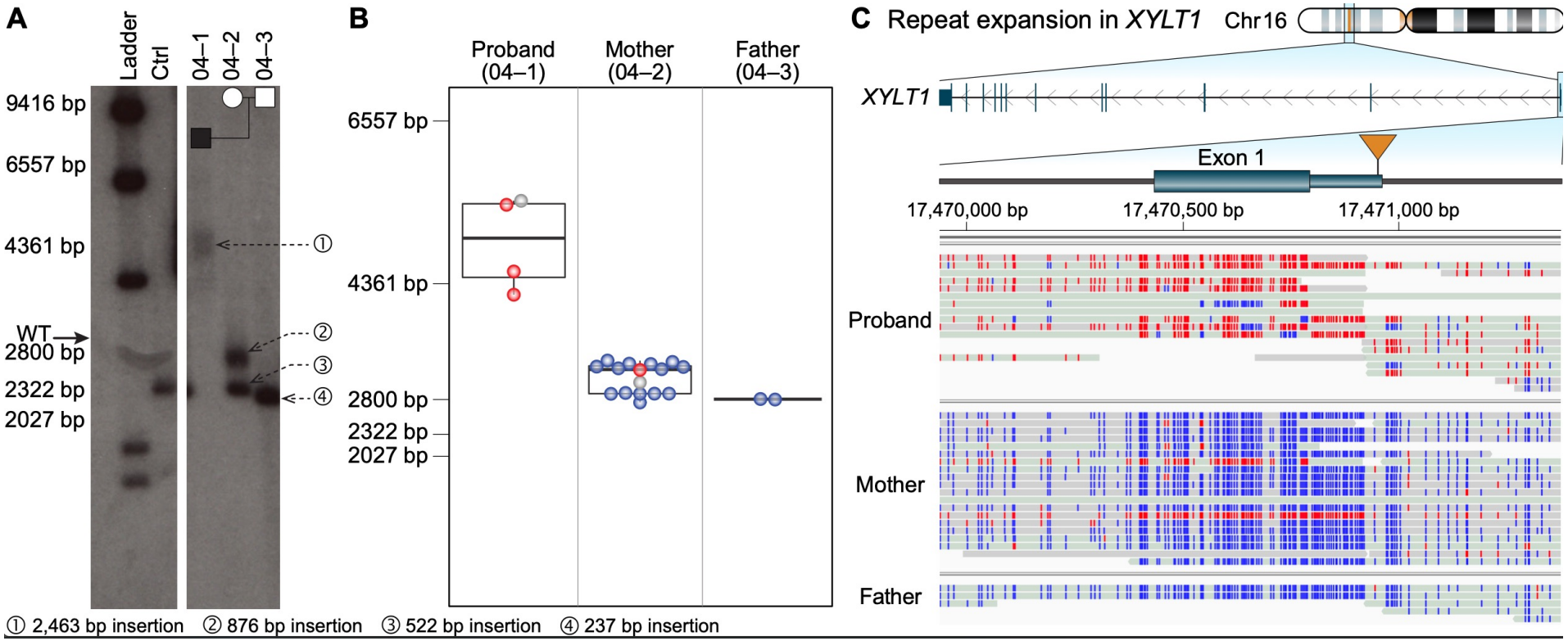
LRS provides detailed information about SVs



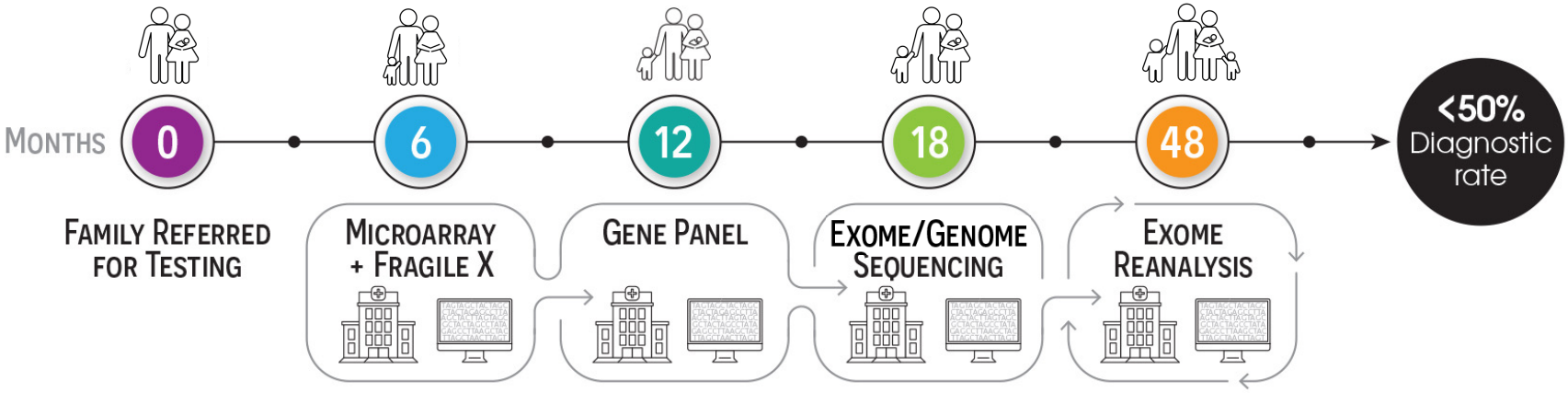
LRS provides detailed information about SVs, including methylation data



LRS provides detailed information about SVs, including methylation data



A traditional genetic workup is diagnostic in less than 50% of cases



Incomplete gene-phenotype relationships

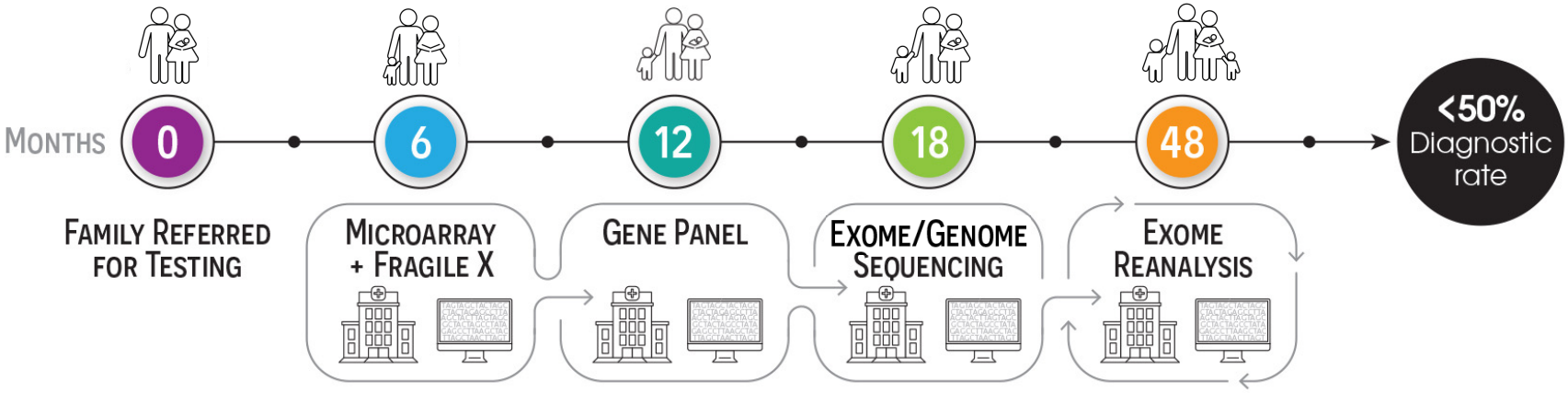
- We do not know the function of all genes

Variants that are difficult to detect or interpret

- Many genes are difficult to sequence
- Structural variants can be difficult to identify
- Predicting the impact of a variant is difficult

Frésard *et al*. 2018 and Clark *et al*. 2018; Gibson *et al*., Submitted

A traditional genetic workup is diagnostic in less than 50% of cases



Incomplete gene-phenotype relationships

- We do not know the function of all genes

Variants that are difficult to detect or interpret

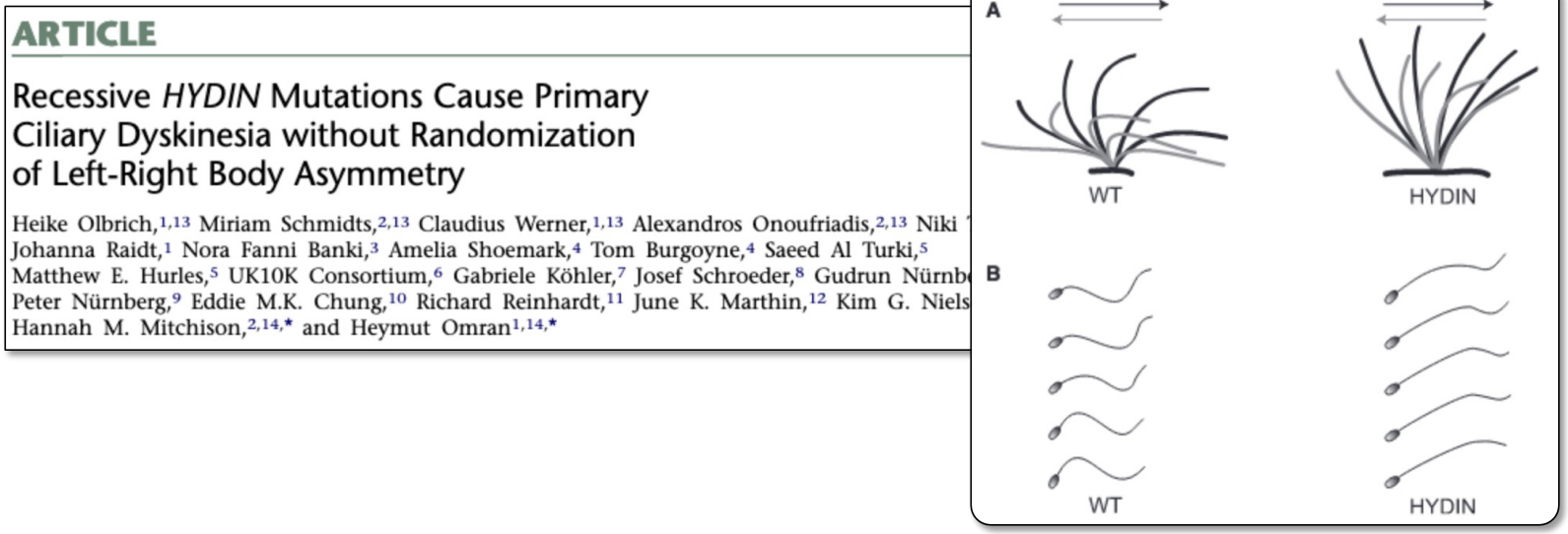
- **Many genes are difficult to sequence**
- Structural variants can be difficult to identify
- Predicting the impact of a variant is difficult

Frésard *et al*. 2018 and Clark *et al*. 2018; Gibson *et al*., Submitted

CASE 1

LRS can be used to identify variants in complex regions of the genome

- Newborn with respiratory failure at birth requiring ECMO
- Duo exome sequencing revealed a likely pathogenic 2-bp deletion in *HYDIN*



ARTICLE

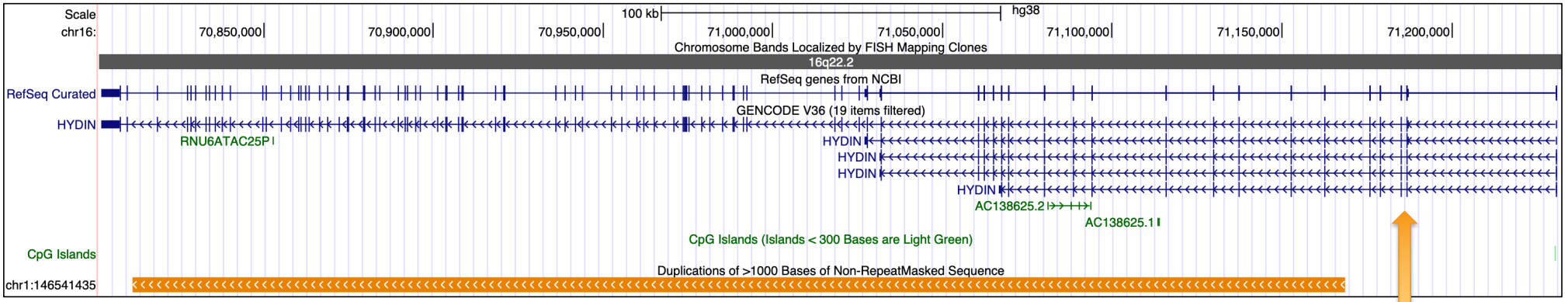
Recessive *HYDIN* Mutations Cause Primary Ciliary Dyskinesia without Randomization of Left-Right Body Asymmetry

Heike Olbrich,[1,13] Miriam Schmidts,[2,13] Claudius Werner,[1,13] Alexandros Onoufriadis,[2,13] Niki
Johanna Raidt,[1] Nora Fanni Banki,[3] Amelia Shoemark,[4] Tom Burgoyne,[4] Saeed Al Turki,[5]
Matthew E. Hurles,[5] UK10K Consortium,[6] Gabriele Köhler,[7] Josef Schroeder,[8] Gudrun Nürnb
Peter Nürnberg,[9] Eddie M.K. Chung,[10] Richard Reinhardt,[11] June K. Marthin,[12] Kim G. Niels
Hannah M. Mitchison,[2,14,*] and Heymut Omran[1,14,*]

Unpublished data

CASE 1

HYDIN is a 400-kb gene containing a 380-kb segmental duplication

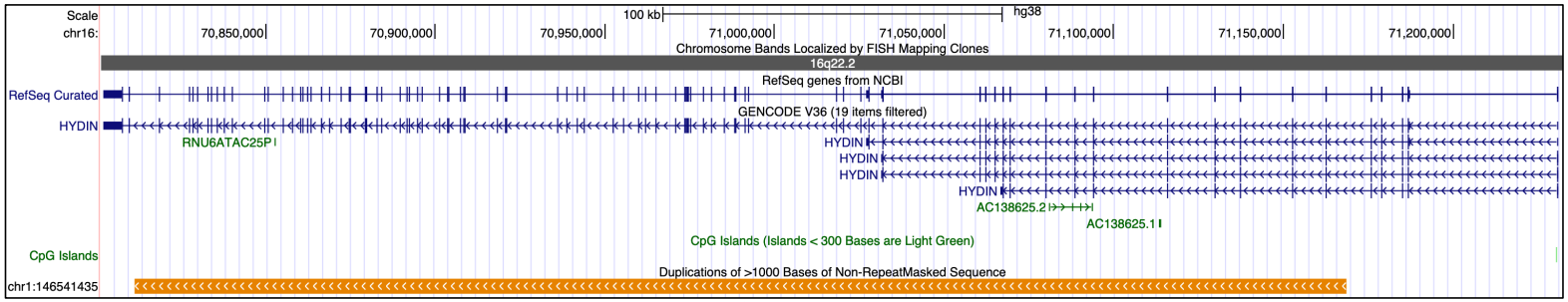


Segmental duplication

2-bp deletion

CASE 1

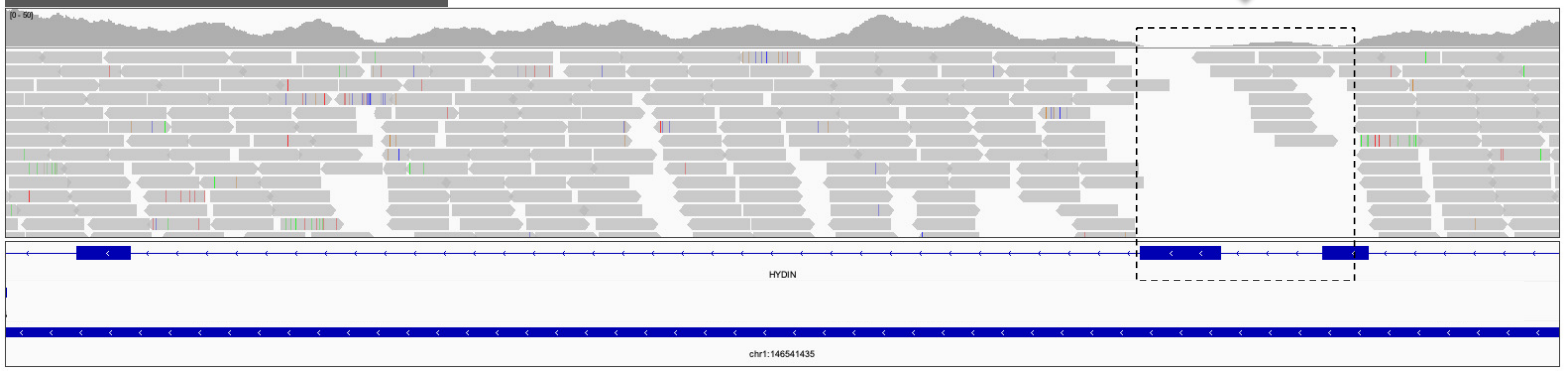
Short reads do not align well within *HYDIN*



Segmental duplication

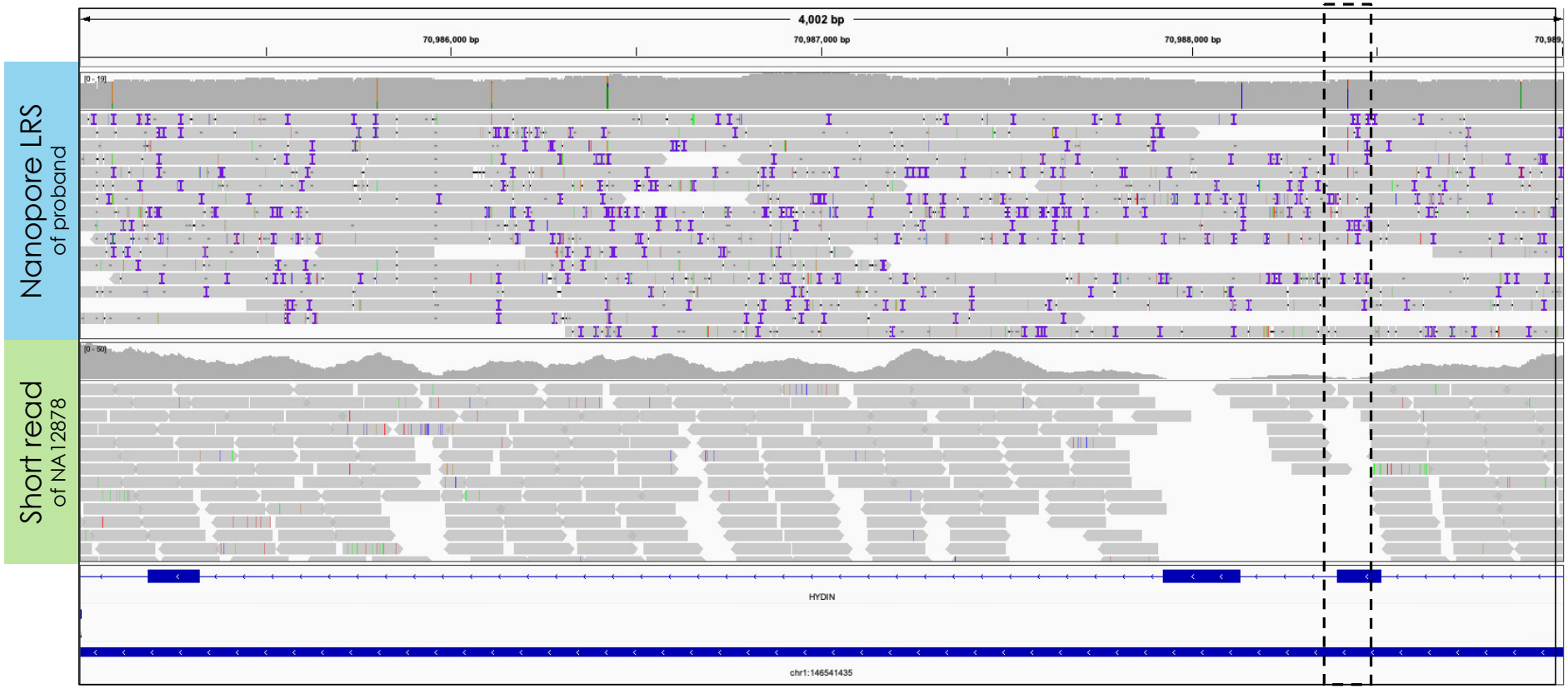
Low coverage

Short-read sequencing
of NA12878



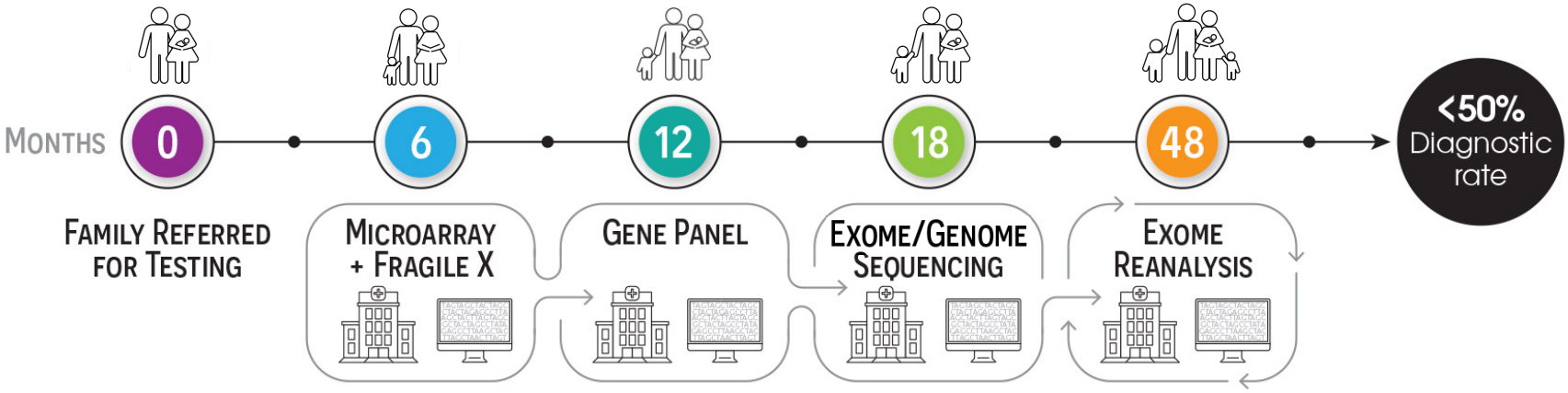
CASE 1

LRS gives even coverage across *HYDIN* and identifies SNVs difficult to detect with short reads



Unpublished data; R9.4 flow cell, superior model, all variants shown

A traditional genetic workup is diagnostic in less than 50% of cases



Incomplete gene-phenotype relationships

- We do not know the function of all genes

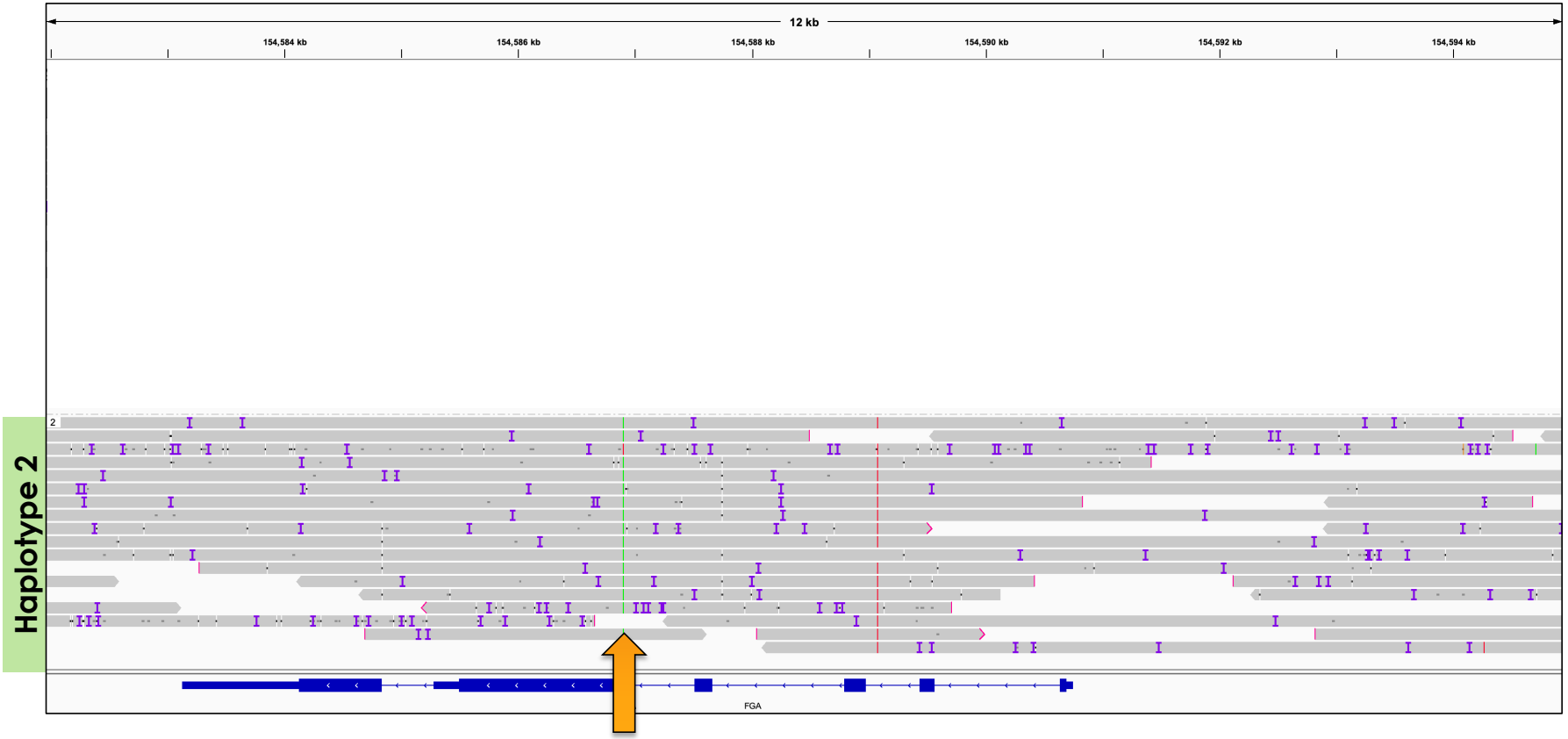
Variants that are difficult to detect or interpret

- Many genes are difficult to sequence
- **Structural variants can be difficult to identify**
- Predicting the impact of a variant is difficult

Frésard *et al*. 2018 and Clark *et al*. 2018; Gibson *et al*., Submitted

CASE 2

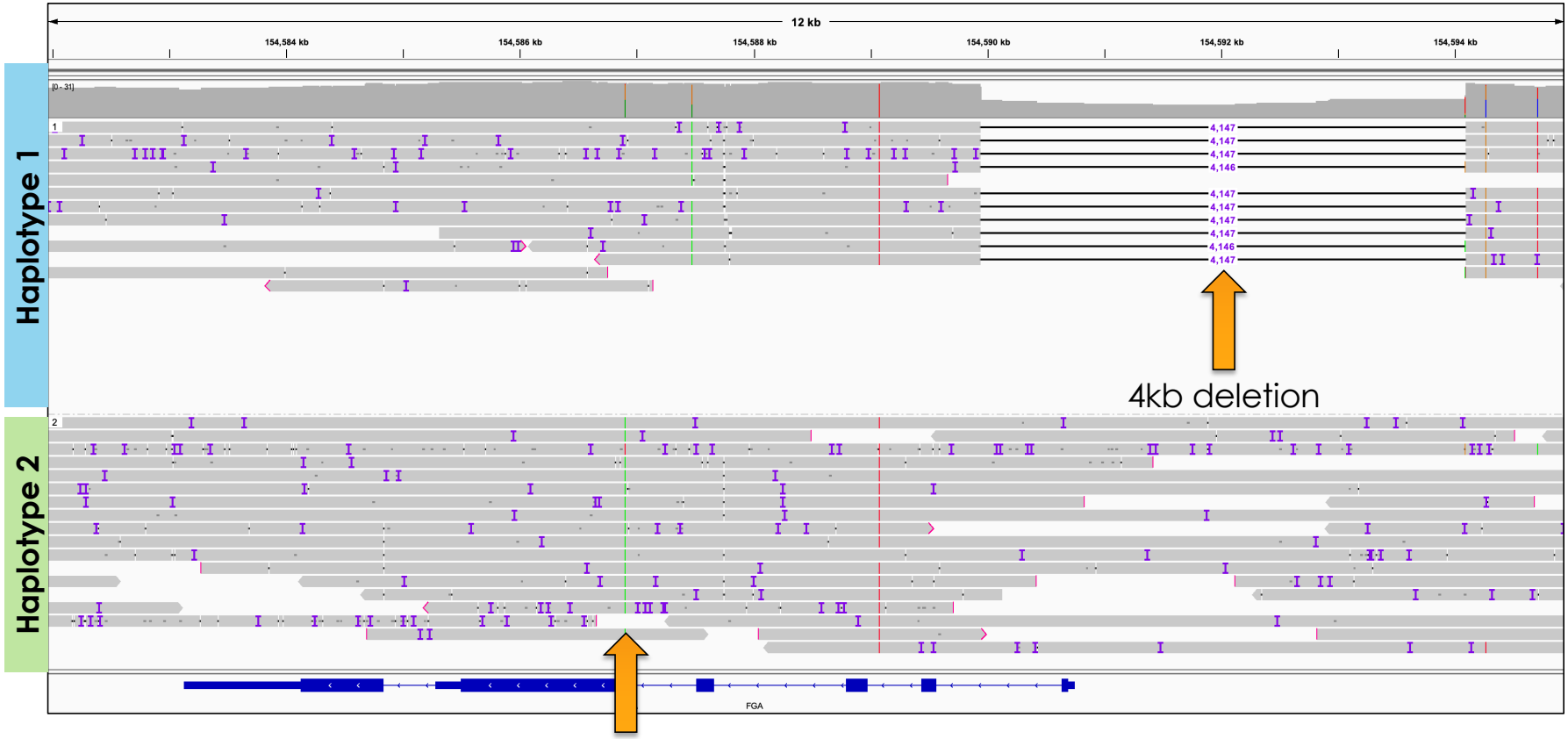
LRS can identify variants missed by prior clinical testing — these are often SVs



Heterozygous for known maternally inherited stop in *FGA* (fibrinogen alpha chain)

CASE 2

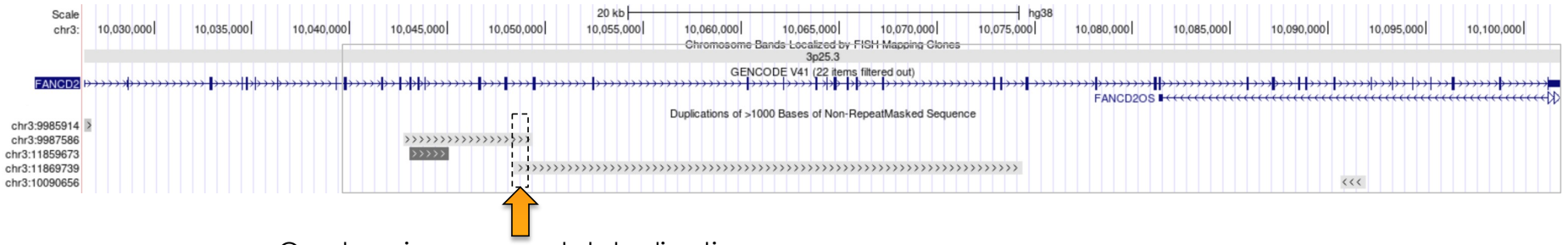
LRS can identify variants missed by prior clinical testing — these are often SVs



Heterozygous for known maternally inherited stop in *FGA* (fibrinogen alpha chain)

CASE 3

LRS can detect variants in regions of the genome difficult to analyze with short reads

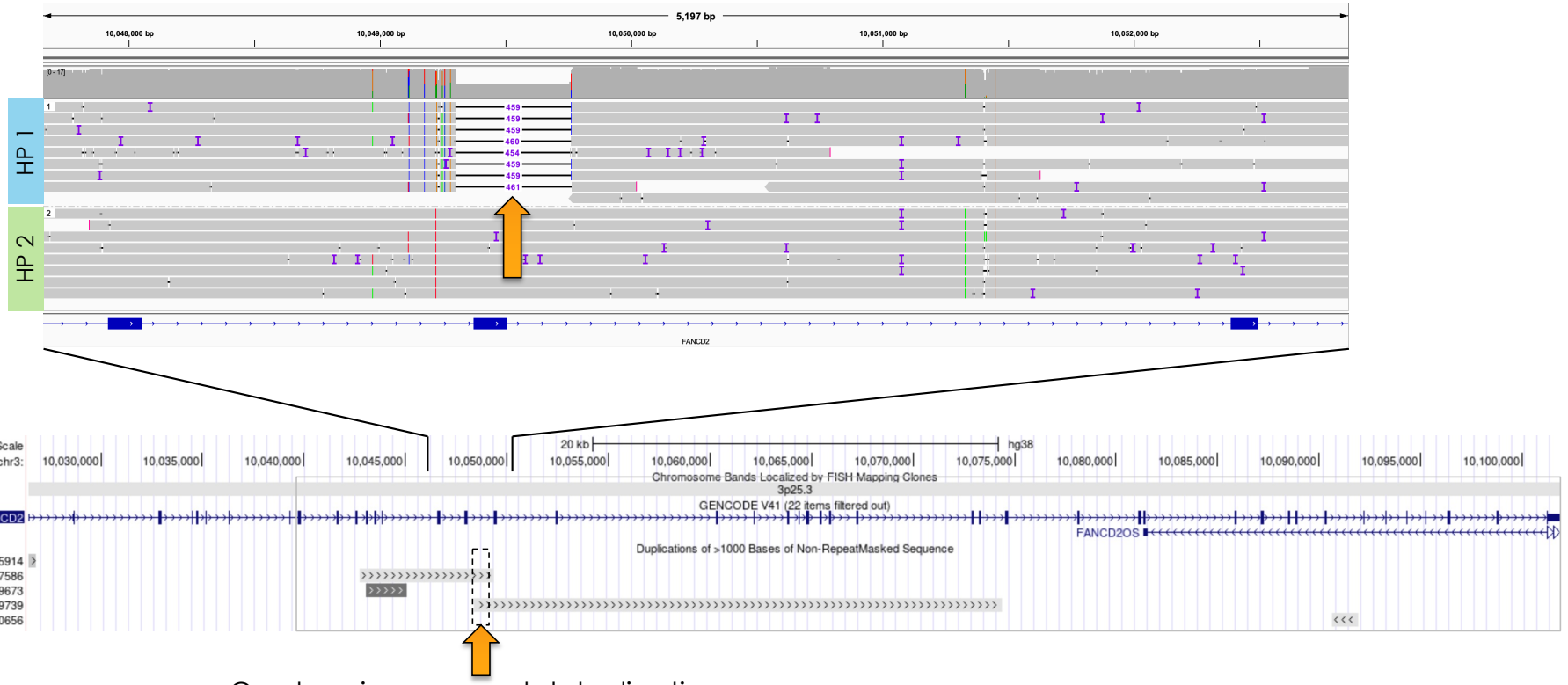


Overlapping segmental duplications

Unpublished data; R9.4.1 flow cell, superior model; indels smaller than 3 bp hidden

CASE 3

LRS can detect variants in regions of the genome difficult to analyze with short reads



Overlapping segmental duplications

Unpublished data; R9.4.1 flow cell, superior model; indels smaller than 3 bp hidden

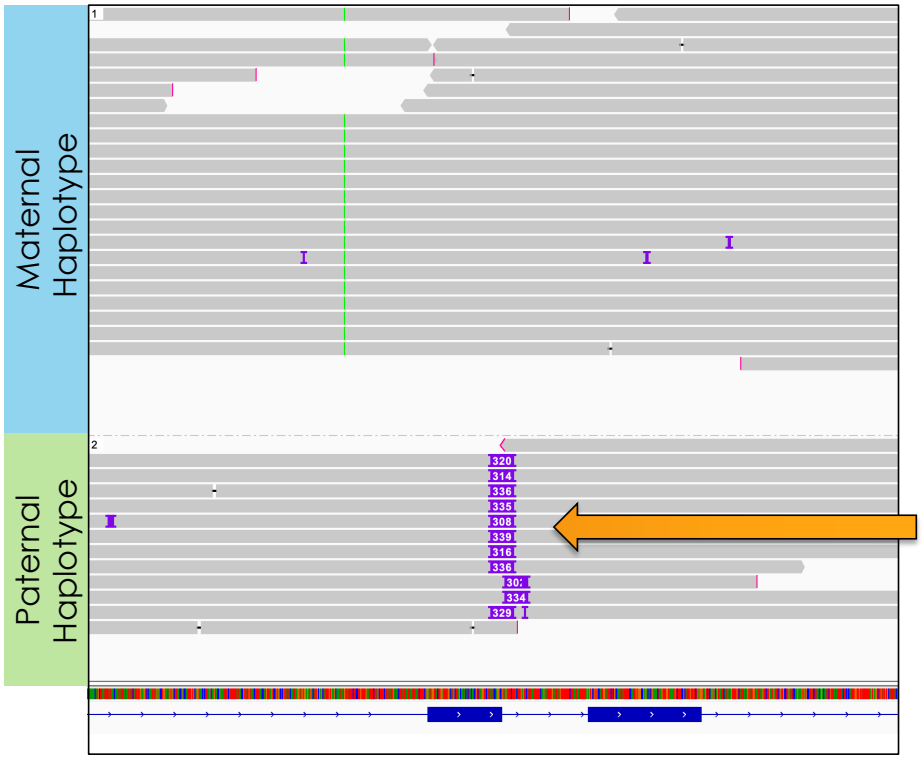
CASE 4

An individual unsolved after clinical testing

- **8-year-old male with suspected glycogen storage disease**
  - Panel identified a single pathogenic variant in *AGL*, but no 2nd variant
- **Family returns for re-evaluation and additional testing**
  - SNP array: negative
  - Exon-level array: negative
- **Research-based short-read WGS**
  - SV identified in *AGL*, thought to be a translocation
  - Clinical optical genome mapping: **negative**
- **Research long-read sequencing**

CASE 4

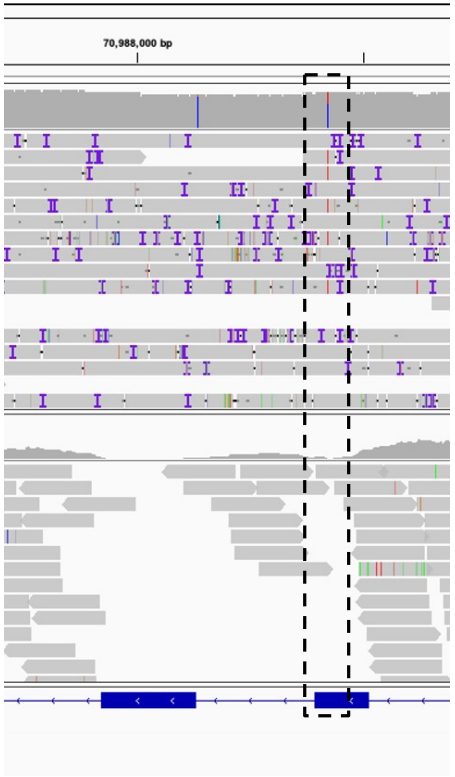
An individual unsolved after clinical testing



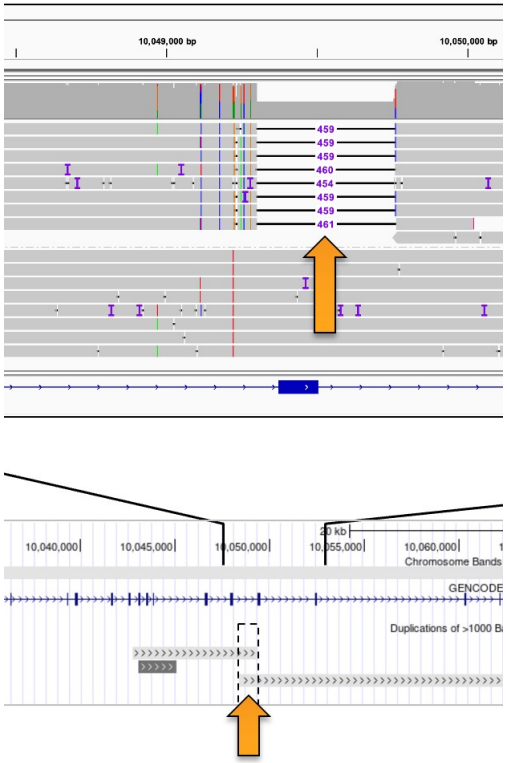
Alu insertion likely to disrupt splicing

Existing databases cannot be used to determine the allele frequency of these variants

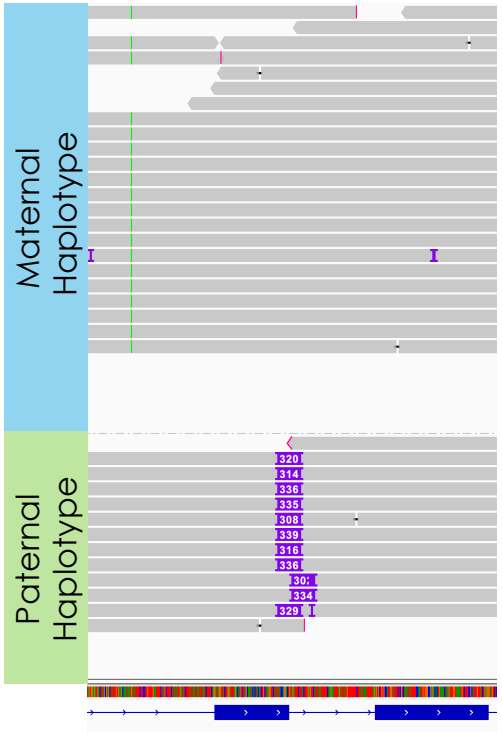
Case 1: Ciliopathy



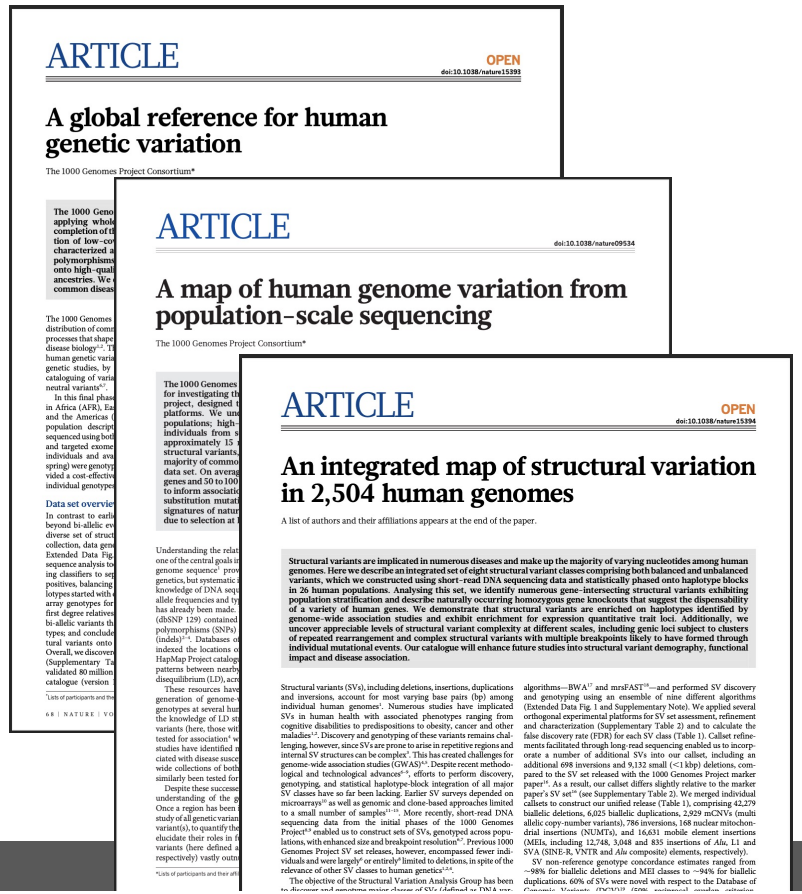
Case 3: Fanconi anemia



Case 4: Glycogen Storage Disease



The 1000 Genomes Project characterized patterns of human genetic variation



ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

ARTICLE

OPEN

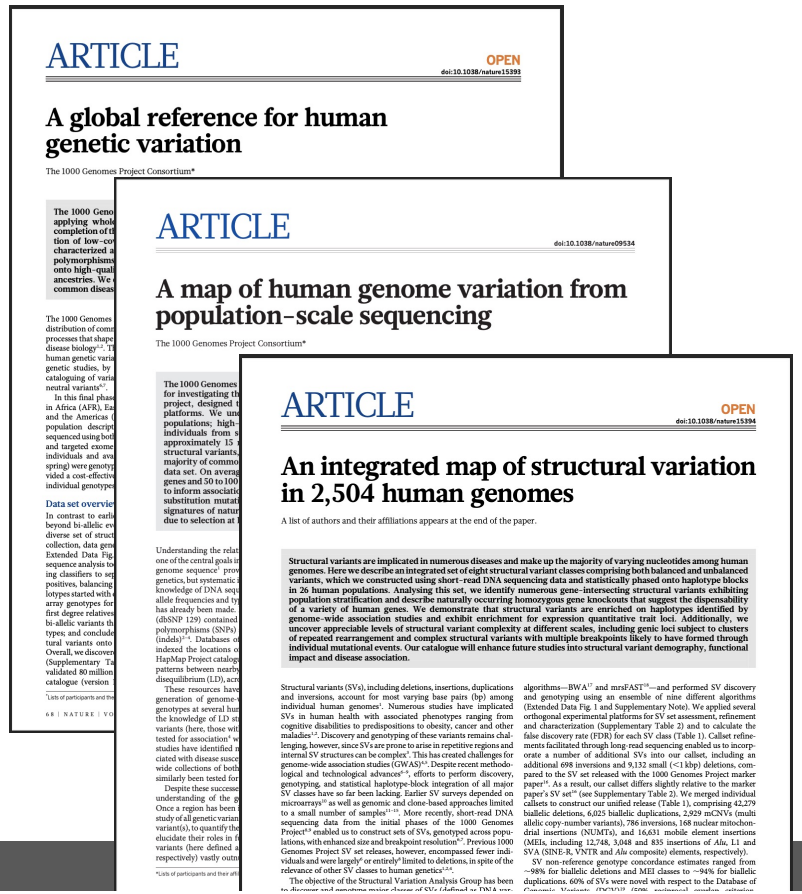
doi:10.1038/nature15394

An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.

Structural variants are implicated in numerous diseases and make up the majority of varying nucleotides among human genomes. Here we describe an integrated set of eight structural variant classes comprising both balanced and unbalanced variants, which we constructed using short-read DNA sequencing data and statistically phased onto haplotype blocks in 26 human populations. Analysing this set, we identify numerous gene–intersecting structural variants exhibiting population stratification and describe naturally occurring homozygous gene knockouts that suggest the dispensability of a variety of human genes. We demonstrate that structural variants are enriched on haplotypes identified by genome-wide association studies and exhibit enrichment for expression quantitative trait loci. Additionally, we uncover appreciable levels of structural variant complexity at different scales, including genic loci subject to clusters of repeated rearrangement and complex structural variants with multiple breakpoints likely to have formed through individual mutational events. Our catalogue will enhance future studies into structural variant demography, functional impact and disease association.

LRS of 1000 Genomes Project samples to characterize previously inaccessible patterns of human variation


ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

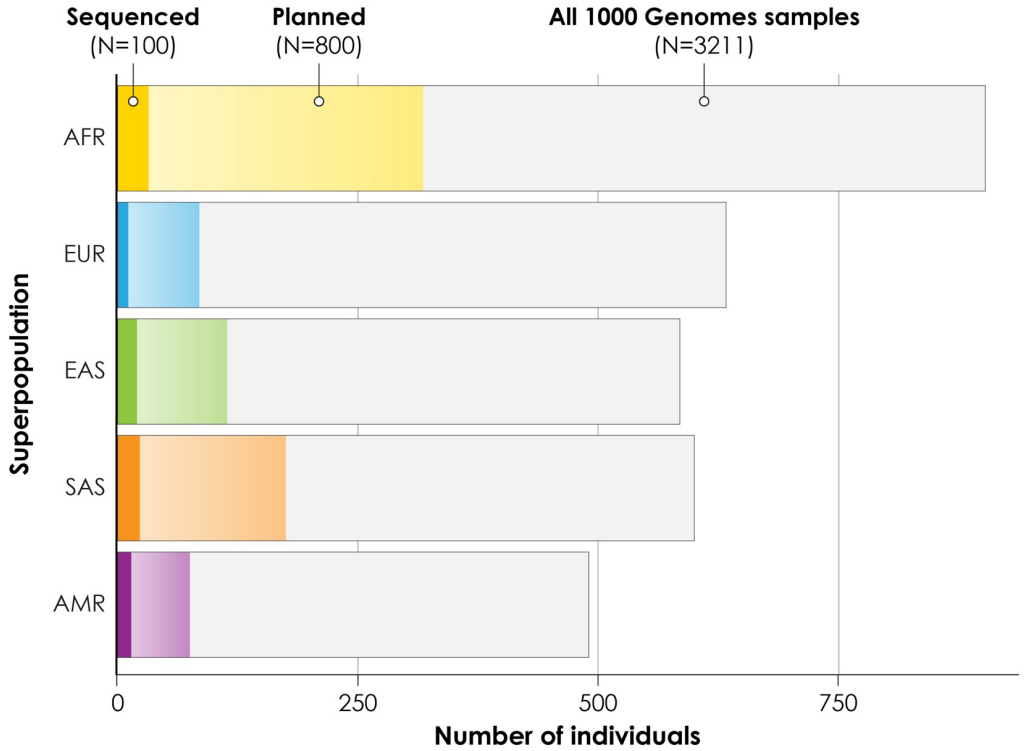
ARTICLE

OPEN

doi:10.1038/nature15394

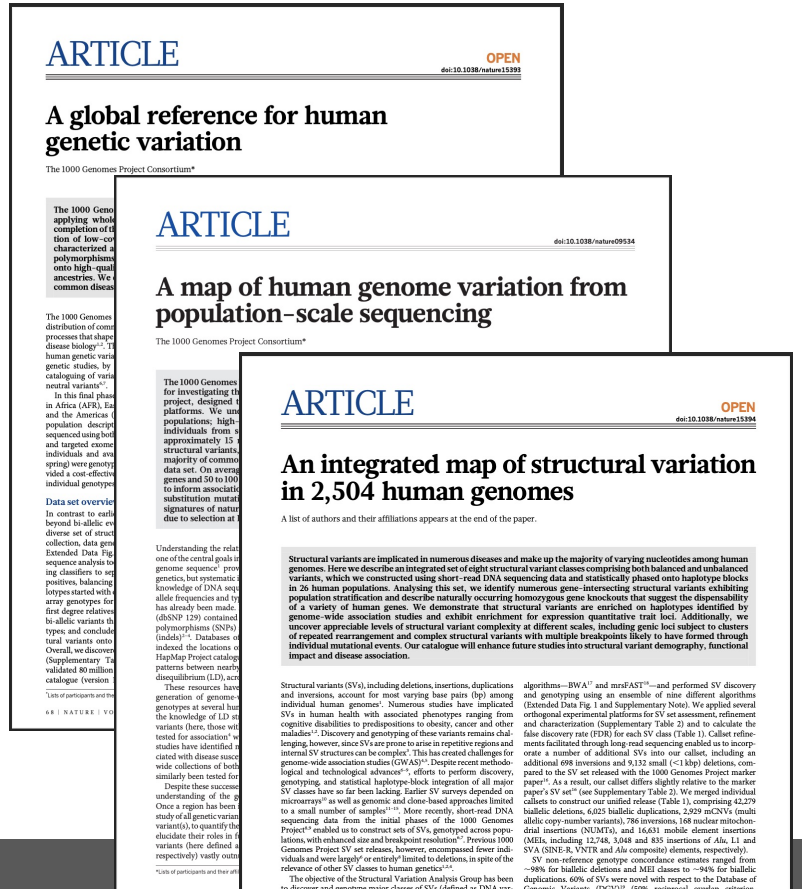
An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.



J. (Gus) Gustafson; 1000 Genomes ONT Sequencing Consortium

LRS of 1000 Genomes Project samples to characterize previously inaccessible patterns of human variation



ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

ARTICLE

OPEN

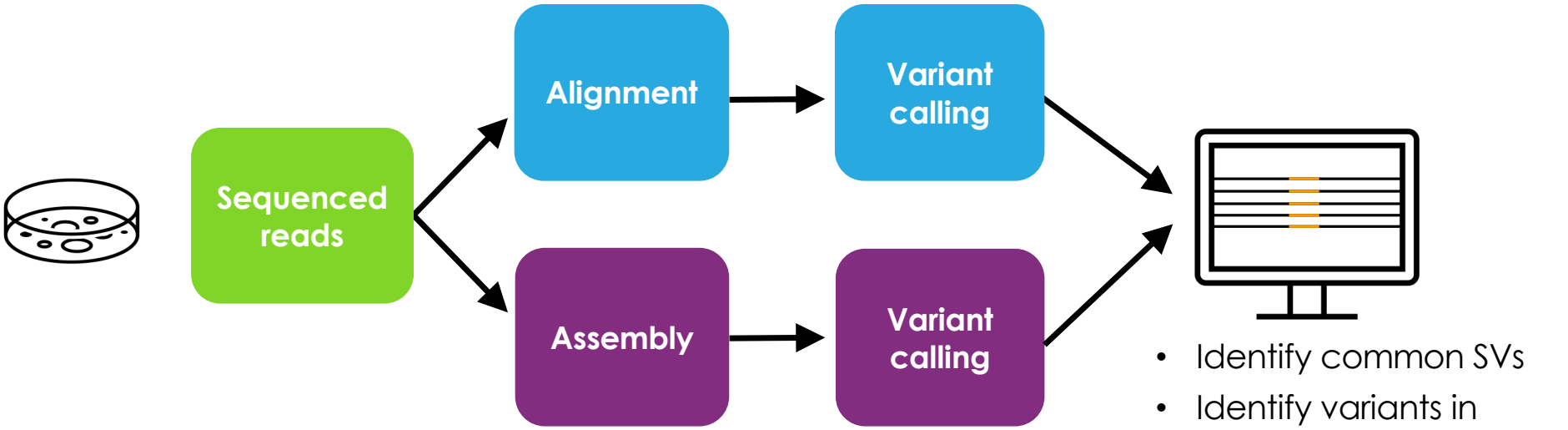
doi:10.1038/nature15394

An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.

- Catalog of SVs for filtering and variant prioritization
- Expand our understanding of variation in difficult to analyze regions of the genome
- Evaluate variation in regions associated with interesting signals in existing data

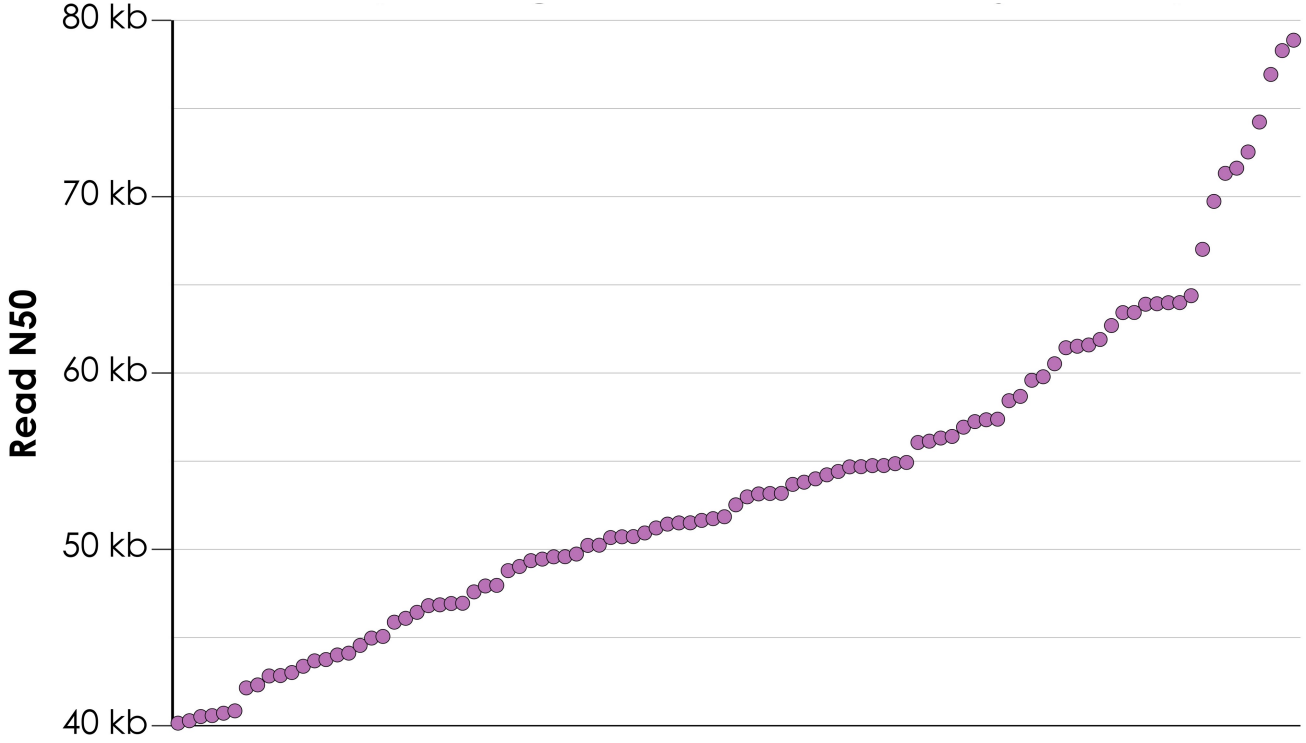
LRS data is analyzed using both alignment-based and assembly-based approaches



Data available at https://s3.amazonaws.com/1000g-ont/index.html

Kolmogorov *et al.*, 2023; 1000 Genomes ONT Sequencing Consortium

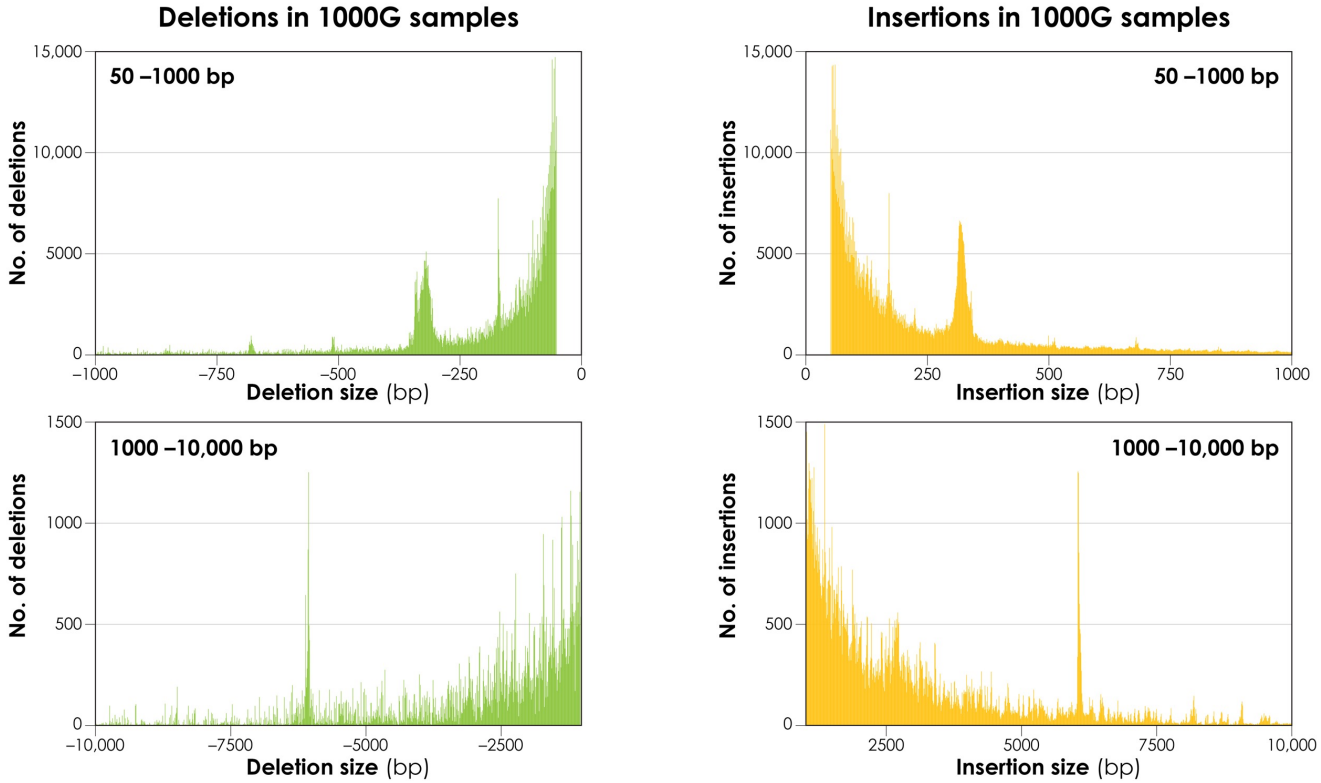
The first 100 samples have read N50s >40 kb with >30x coverage



Data available at https://s3.amazonaws.com/1000g-ont/index.html

J. (Gus) Gustafson; 1000 Genomes ONT Sequencing Consortium

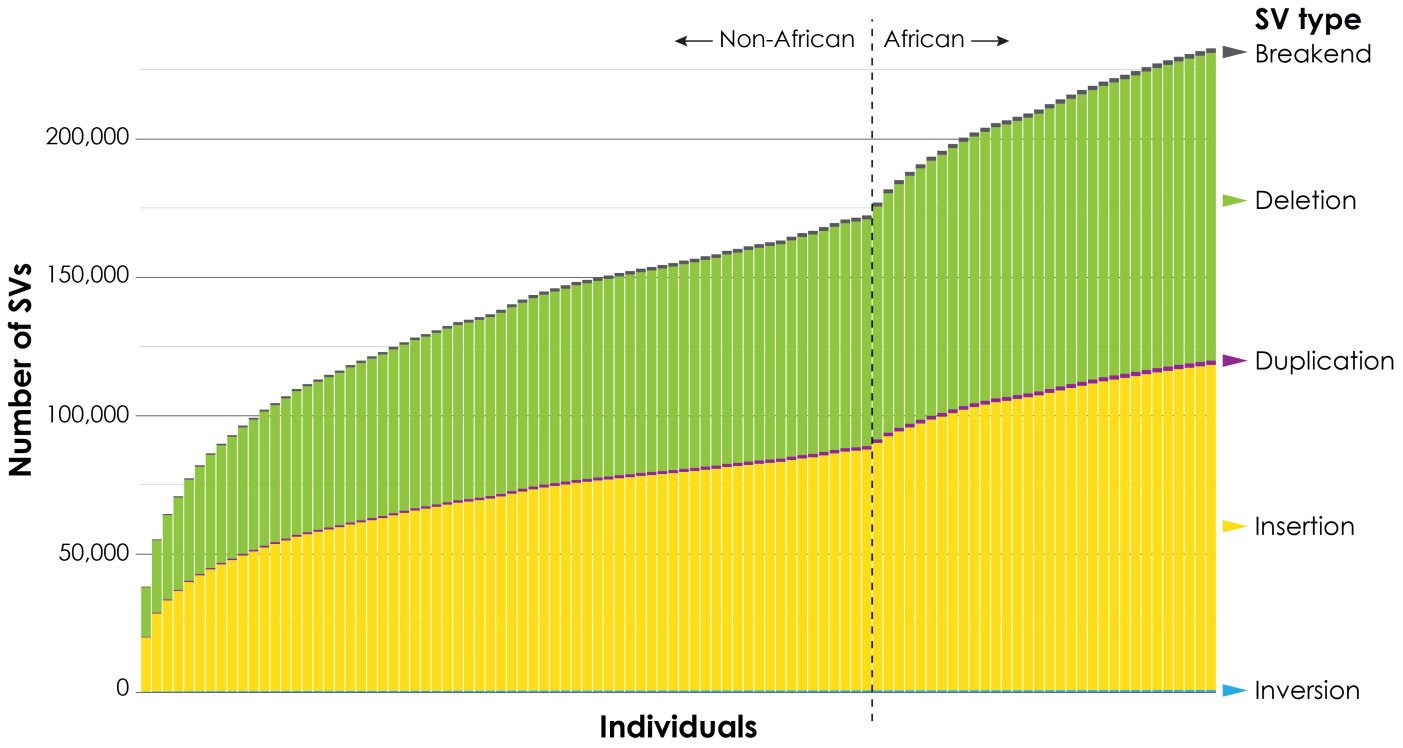
The distribution of insertions and deletions in the first 100 samples follows expected patterns



Data available at https://s3.amazonaws.com/1000g-ont/index.html

J. (Gus) Gustafson; 1000 Genomes ONT Sequencing Consortium

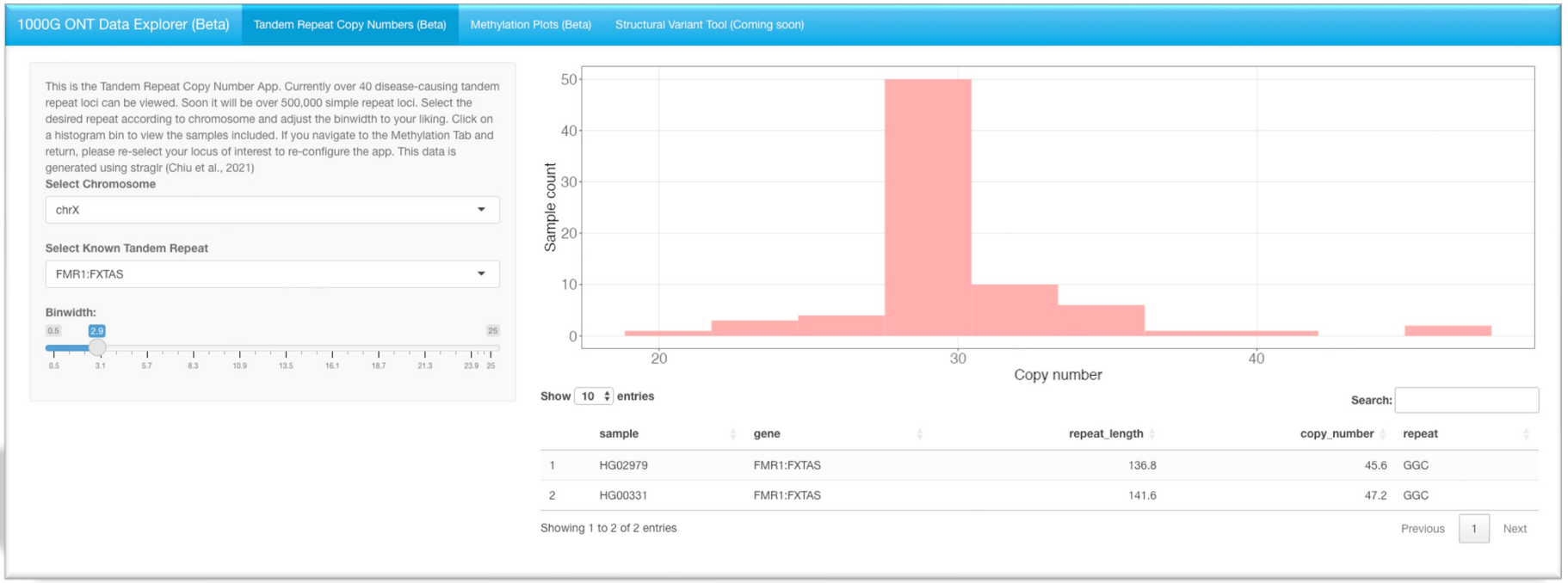
The number of novel SVs increase as more individuals are added



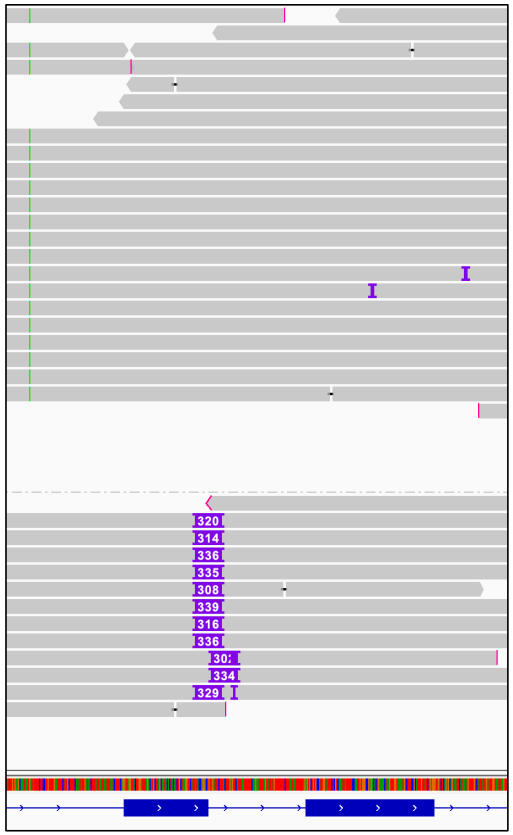
Data available at https://s3.amazonaws.com/1000g-ont/index.html

J. (Gus) Gustafson; 1000 Genomes ONT Sequencing Consortium

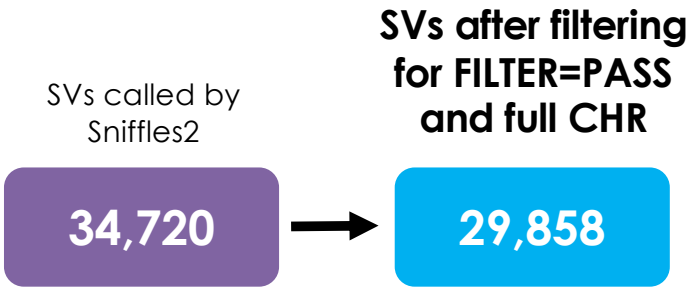
Applications for SV filtering and prioritization using this data are under development



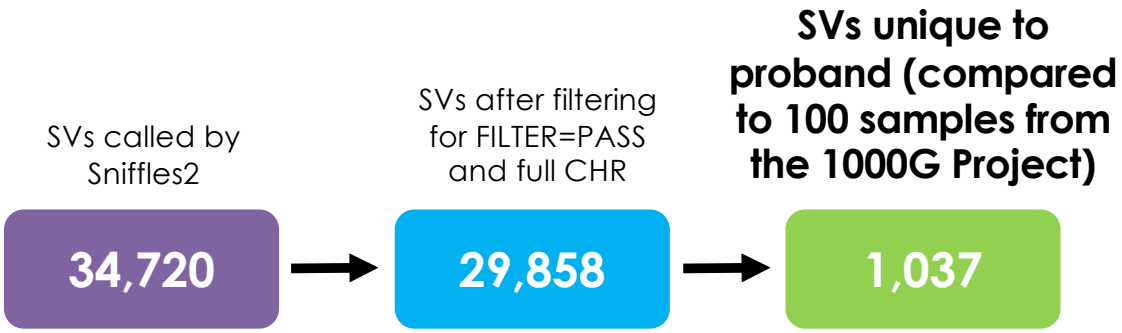
Data from 100 individuals can be used for SV filtering and prioritization



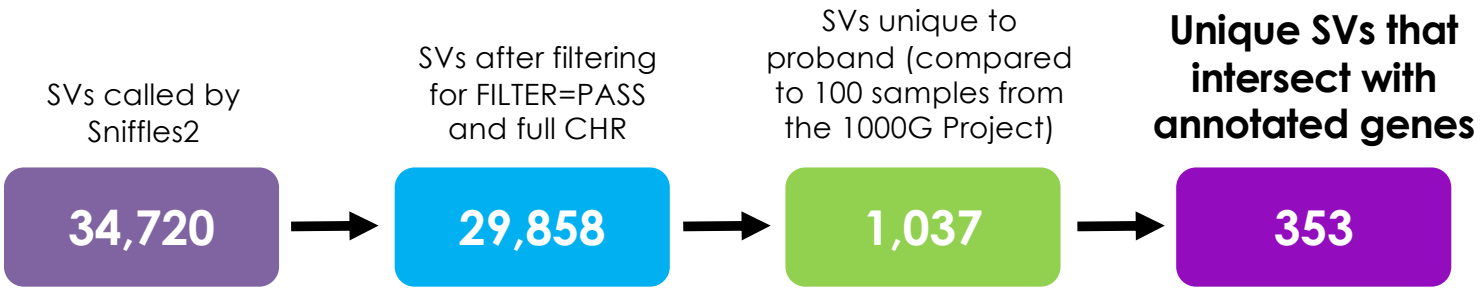
Data from 100 individuals can be used for SV filtering and prioritization



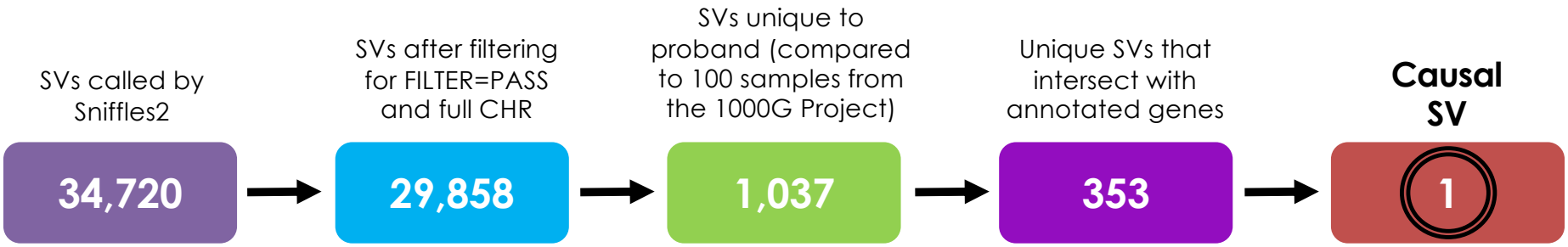
Data from 100 individuals can be used for SV filtering and prioritization



Data from 100 individuals can be used for SV filtering and prioritization



Data from 100 individuals can be used for SV filtering and prioritization



Applications for SV filtering and prioritization using this data are under development

Structural Variant Allele Frequency

Gene_Name

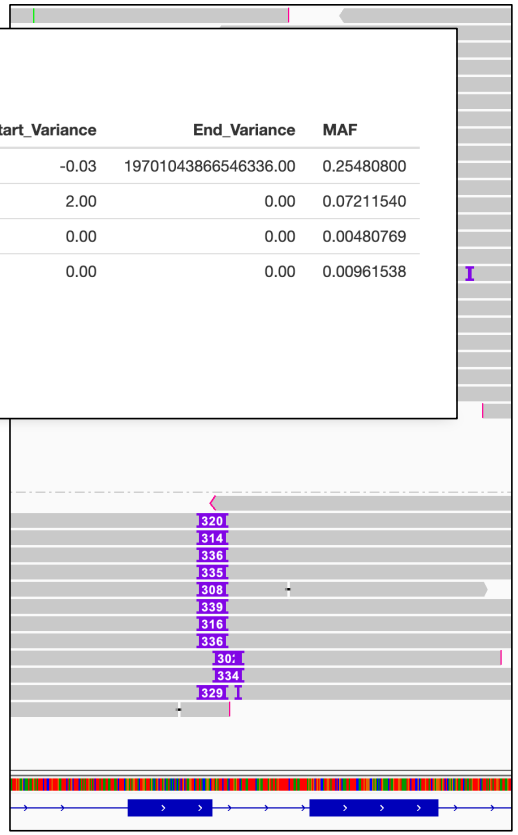
AGL

Filter by:

- ○ position
- ◉ gene

Filter

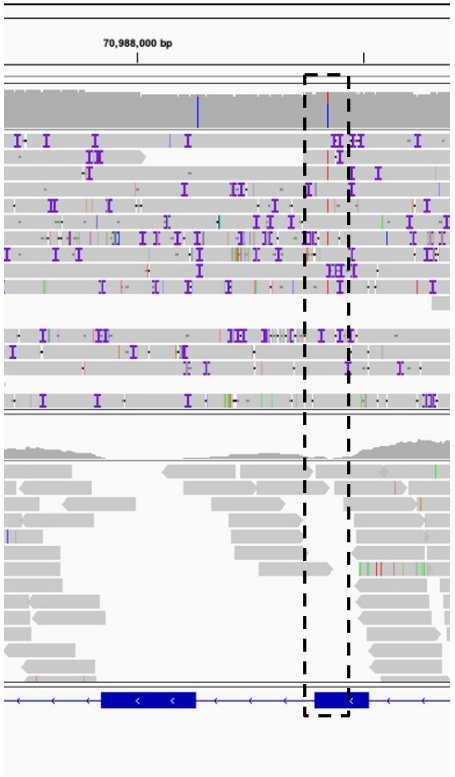
Chromosome	Start_Position	End_Position	Gene_Name	SV_Type	SV_Length	Start_Variance	End_Variance	MAF
chr1	16680842	146052340	AGL	INV	129371498	-0.03	19701043866546336.00	0.25480800
chr1	99856307	99856343	AGL	DEL	-36	2.00	0.00	0.07211540
chr1	99877171	99877487	AGL	INS	316	0.00	0.00	0.00480769
chr1	99890379	99890609	AGL	INS	230	0.00	0.00	0.00961538



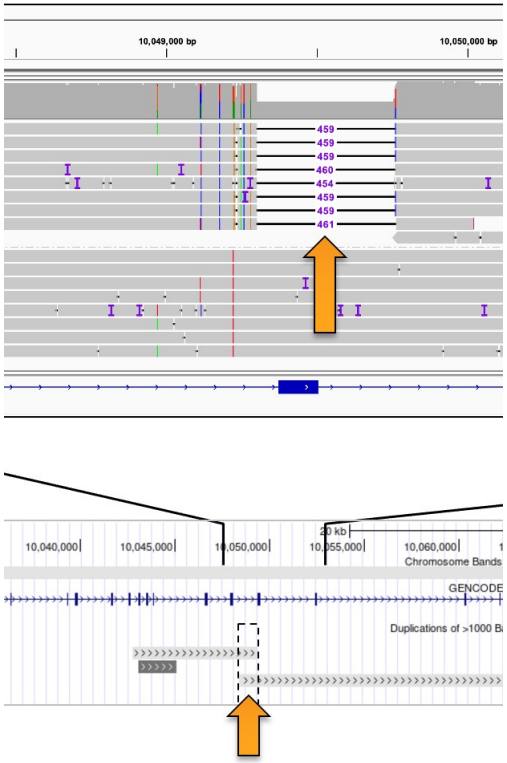
- https://s3.amazonaws.com/1000g-ont/index.html

Data from the 1000G cohort can be used to establish allele frequencies for challenging changes

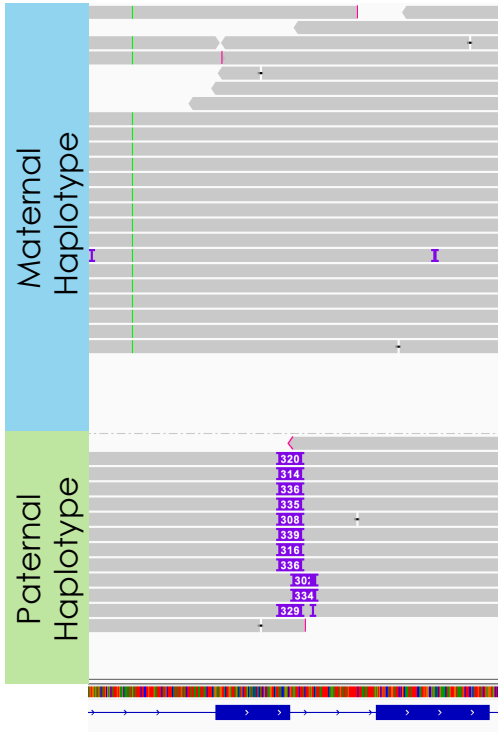
Case 1: Ciliopathy



Case 3: Fanconi anemia



Case 4: Glycogen Storage Disease



Acknowledgements

MILLER LAB

Danny Miller
Zach Anderson
Miranda Galey
Sophia Gibson
Joy Goffena
J. (Gus) Gustafson
Angie Miller
Maisha Sinha
Sophie Storz
Sydney Ward

UNIVERSITY of WASHINGTON

Evan Eichler
Kendra Hoekzema
Jordan Knuth
William Harvey
Brian Shirts

Oxford Nanopore
Androo Markham
Dan Fordham

GREGoR consortium

Sanford Children's
Katie Burns
Laura Davis-Keppen

Northeastern University
Miten Jain

University of Nottingham
Matt Loose

New York Genome Center
Michael Zody
Andre Corvelo
Adrienne Helland
Atit Raval

Cold Spring Harbor Laboratories
Richard McCombie
Cat Reeves
Sarah Goodwin

1000G ONT Sequencing Consortium

Everyone contributing to the 1000G ONT Sequencing Consortium