

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

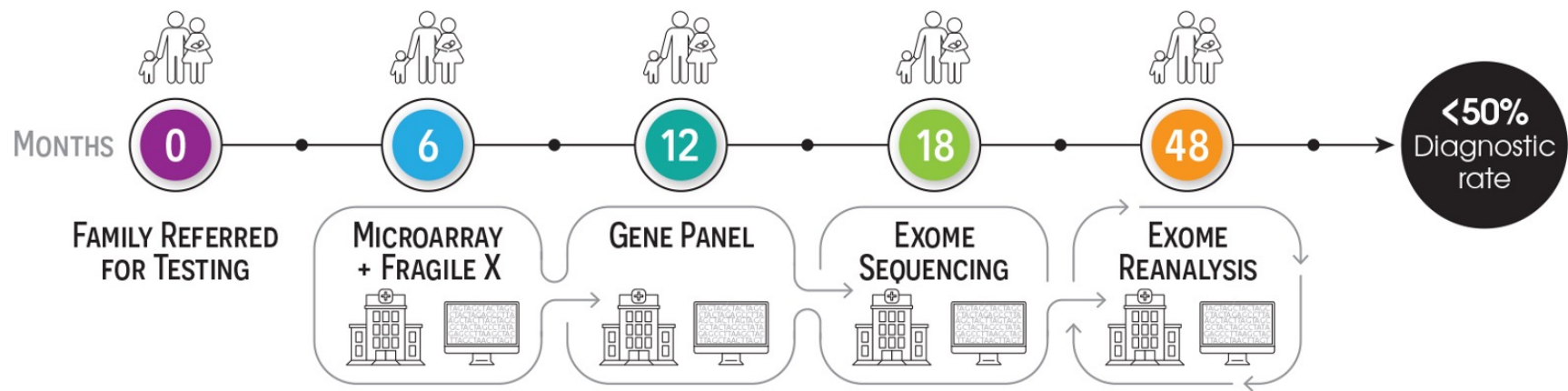
The University of Washington and Seattle Children's Hospital

ASHG

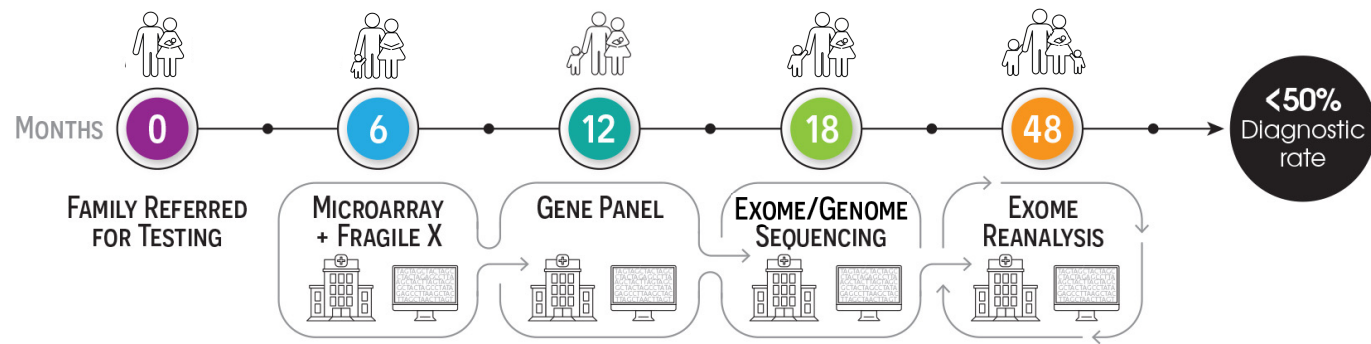
November 2, 2023

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

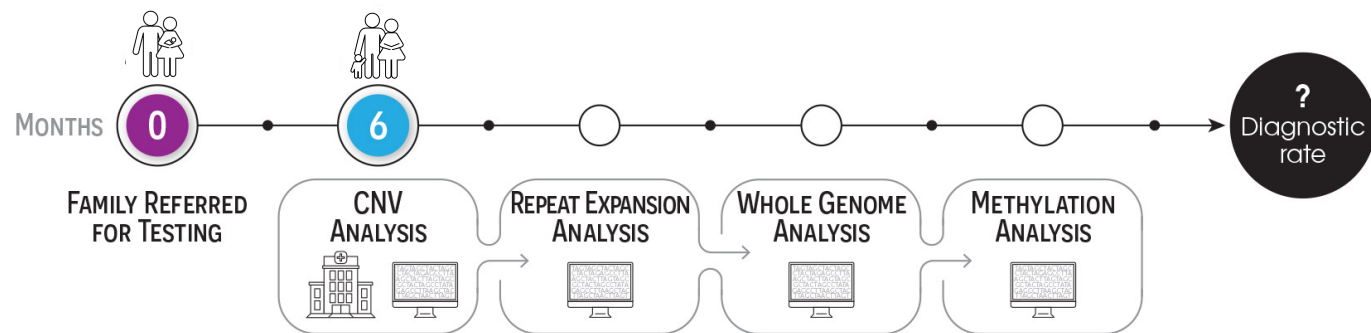
Clinical genetic testing often occurs in a stepwise fashion, involving multiple tests and clinic visits



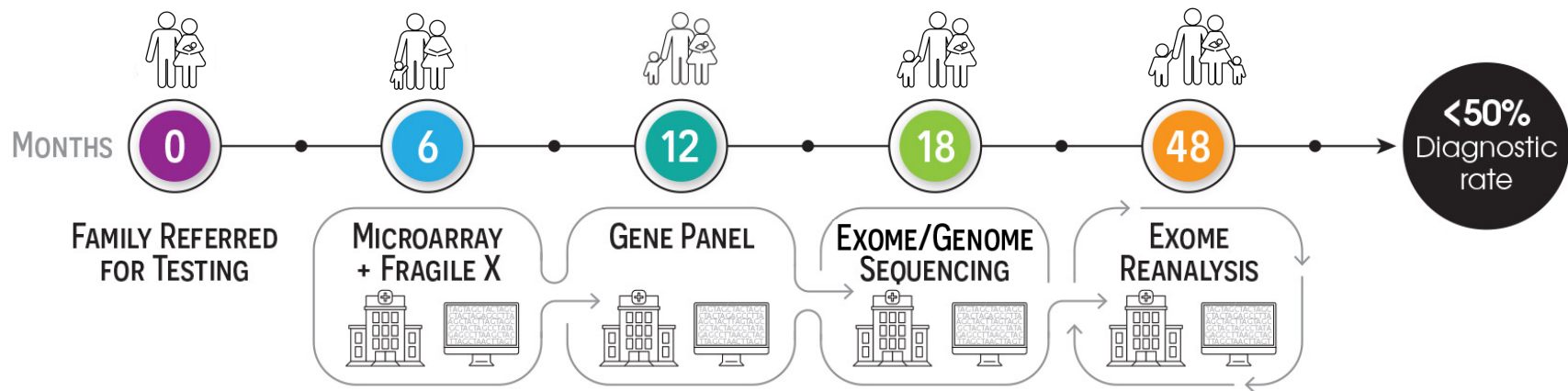
New technologies, such as long-read sequencing will increase the diagnostic rate



Long-read sequencing as a single test



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



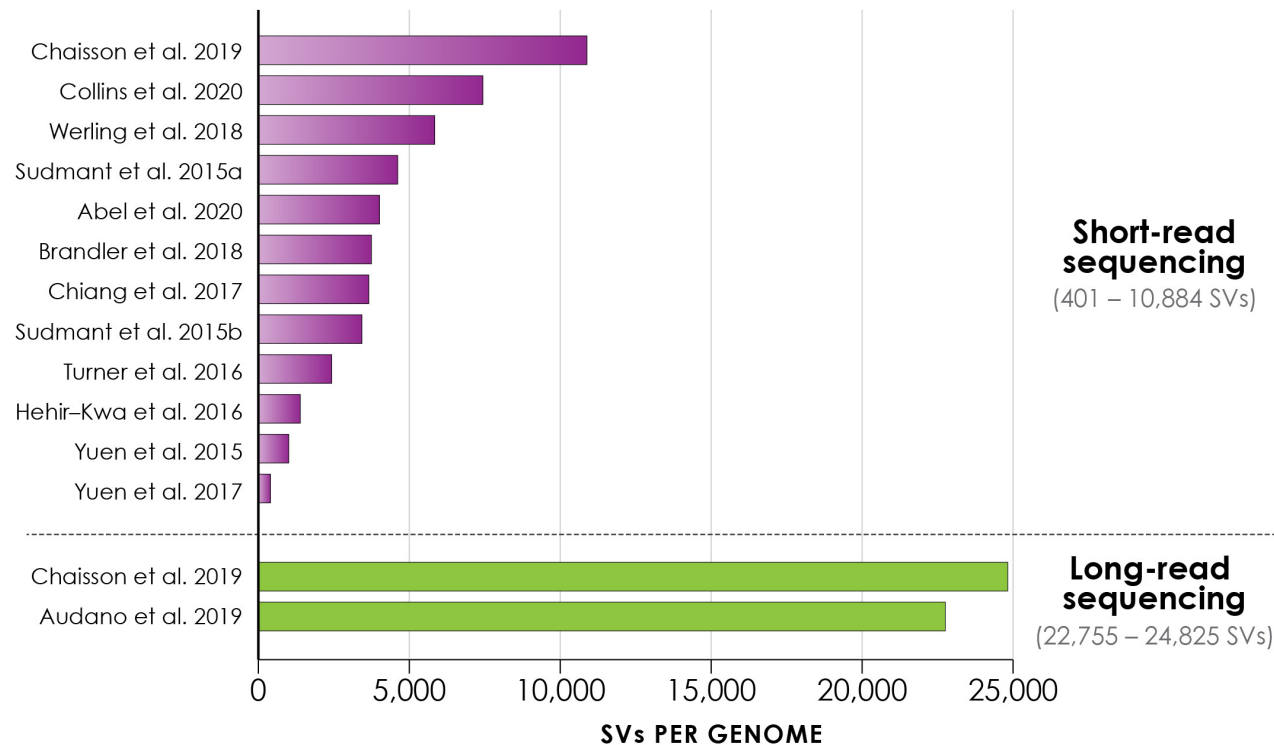
Incomplete gene-phenotype relationships

- We do not know the function of all genes

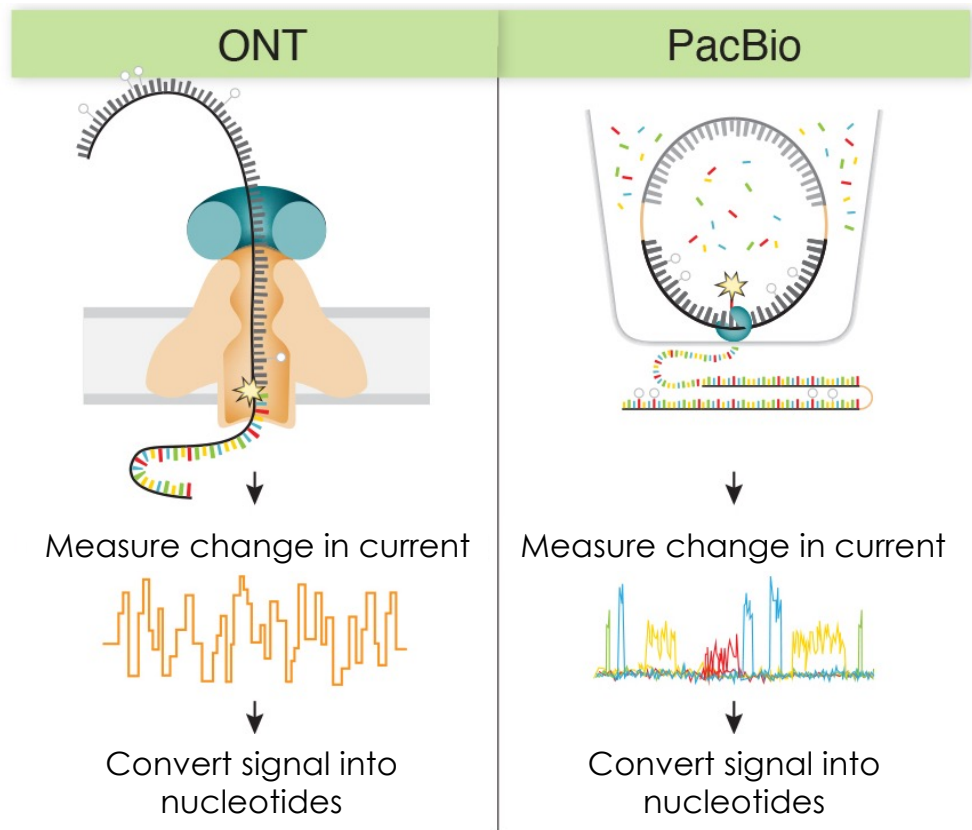
Variants that are difficult to detect or interpret

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

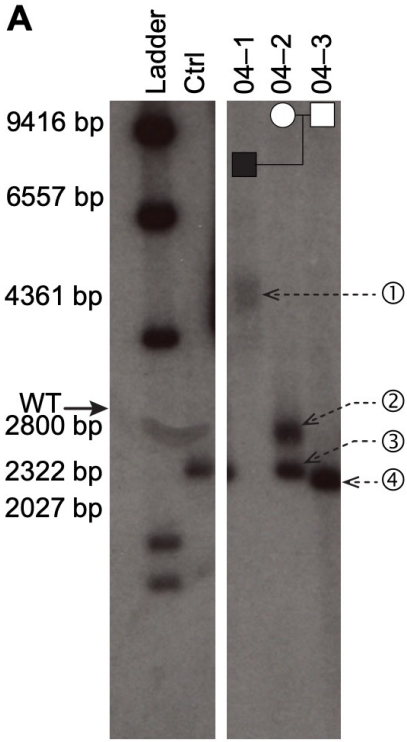
Short-read sequencing detects less than half the SVs seen by long-read sequencing



Long-read sequencing technologies

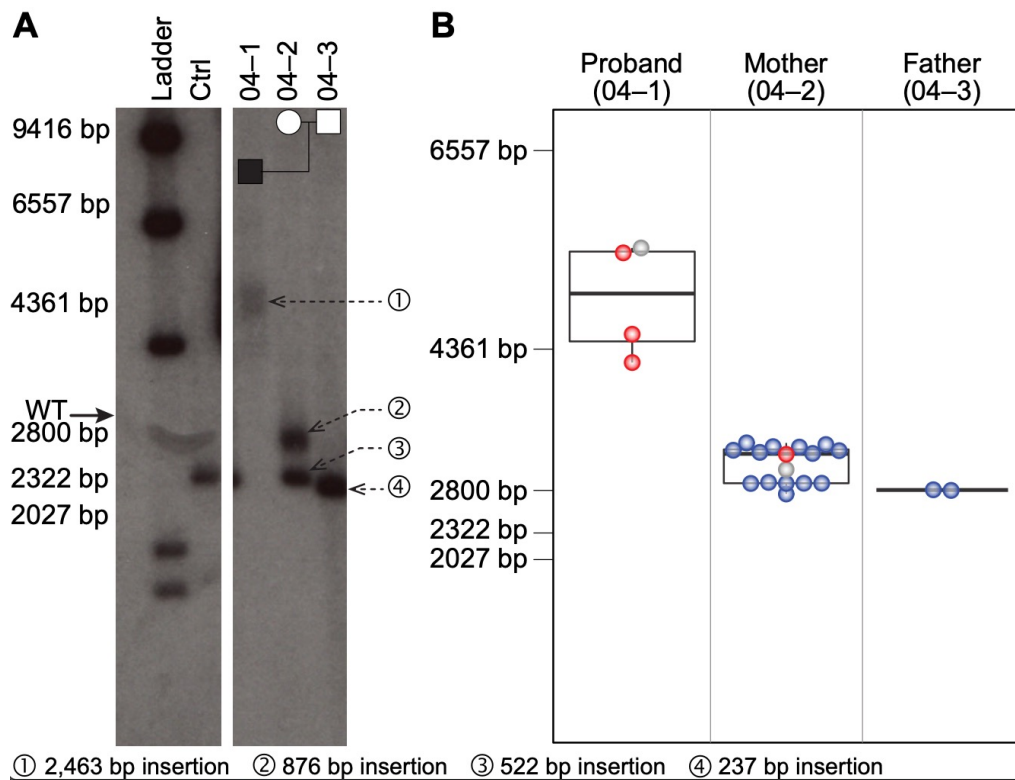


LRS provides detailed information about SVs

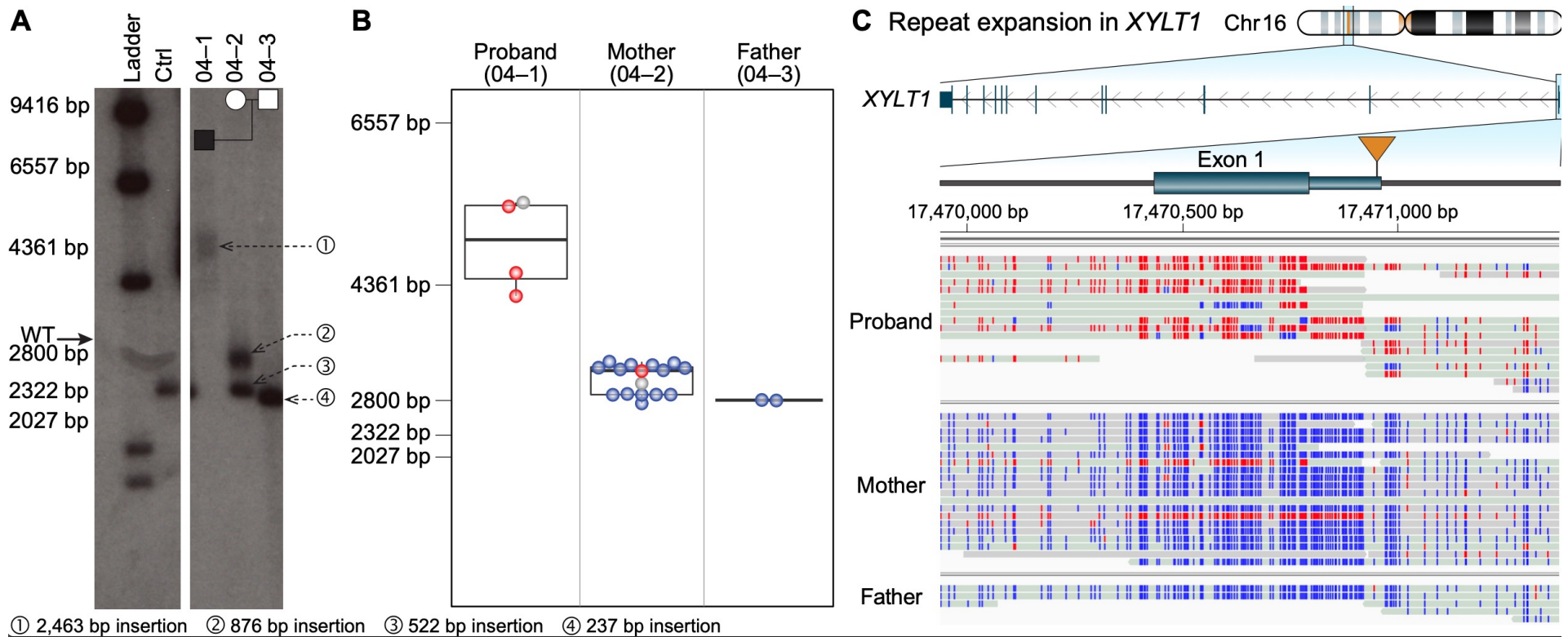


① 2,463 bp insertion ② 876 bp insertion ③ 522 bp insertion ④ 237 bp insertion

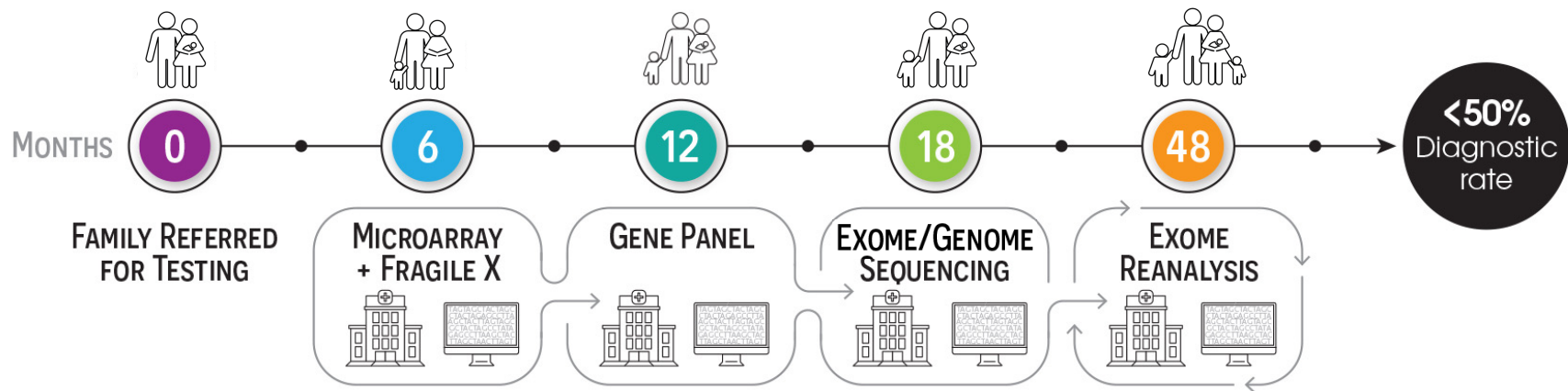
LRS provides detailed information about SVs, including methylation data



LRS provides detailed information about SVs, including methylation data



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



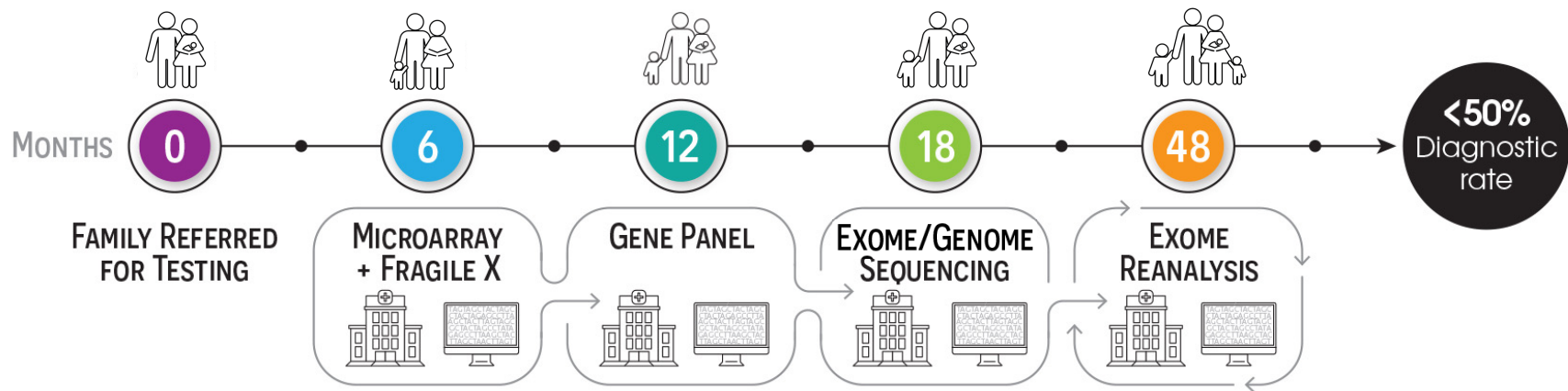
Incomplete gene-phenotype relationships

- 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

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

CASE 1

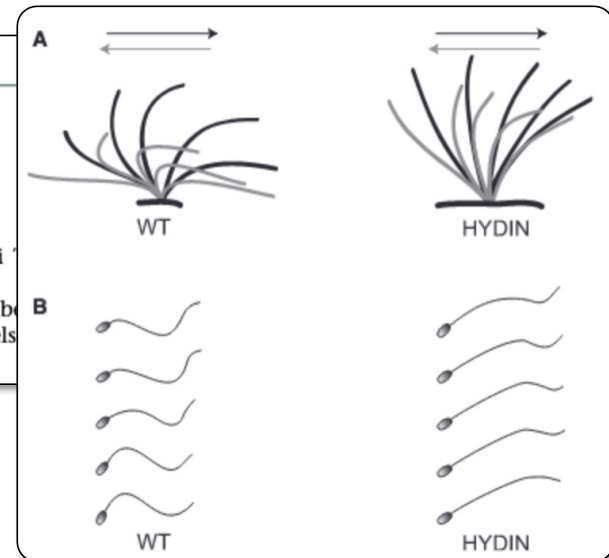
LRS can be used to identify variants in complex regions of the genome

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

ARTICLE

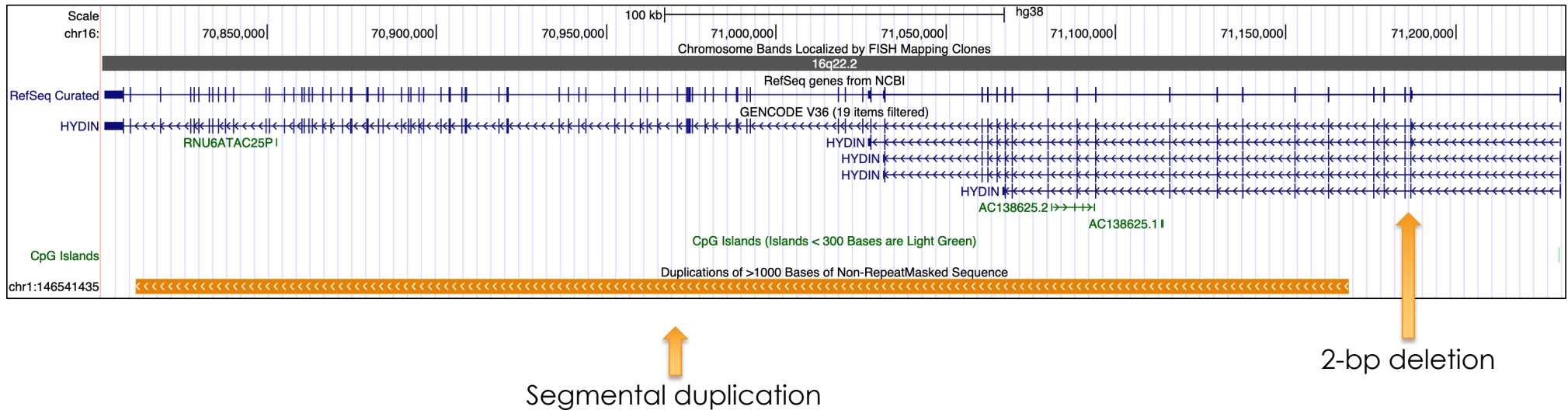
Recessive *HYDIN* Mutations Cause Primary Ciliary Dyskinesia without Randomization of Left-Right Body Asymmetry

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 Nürnberg,⁹ Peter Nürnberg,⁹ Eddie M.K. Chung,¹⁰ Richard Reinhardt,¹¹ June K. Marthin,¹² Kim G. Nielsen,¹³ Hannah M. Mitchison,^{2,14,*} and Heymut Omran^{1,14,*}



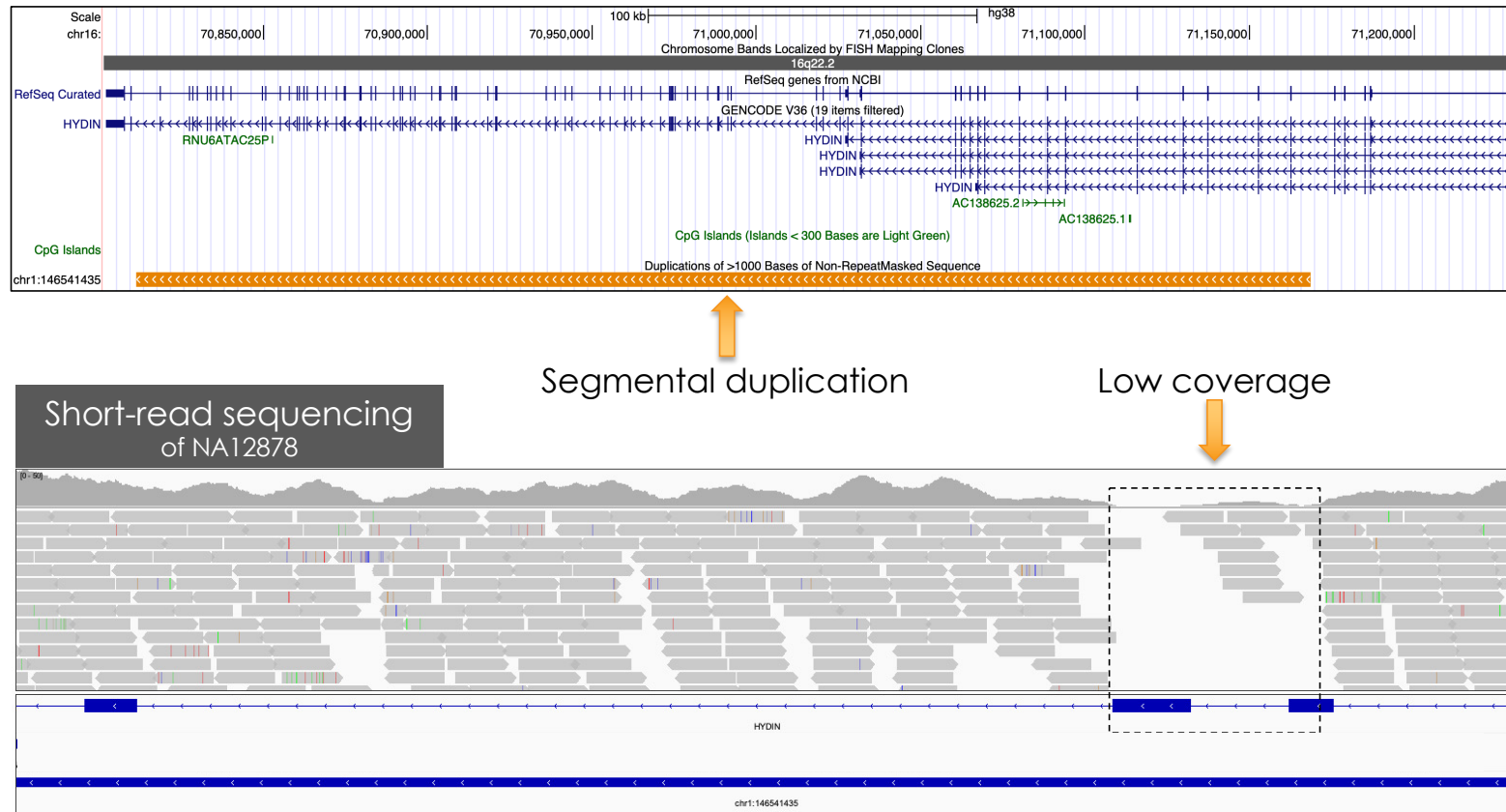
CASE 1

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



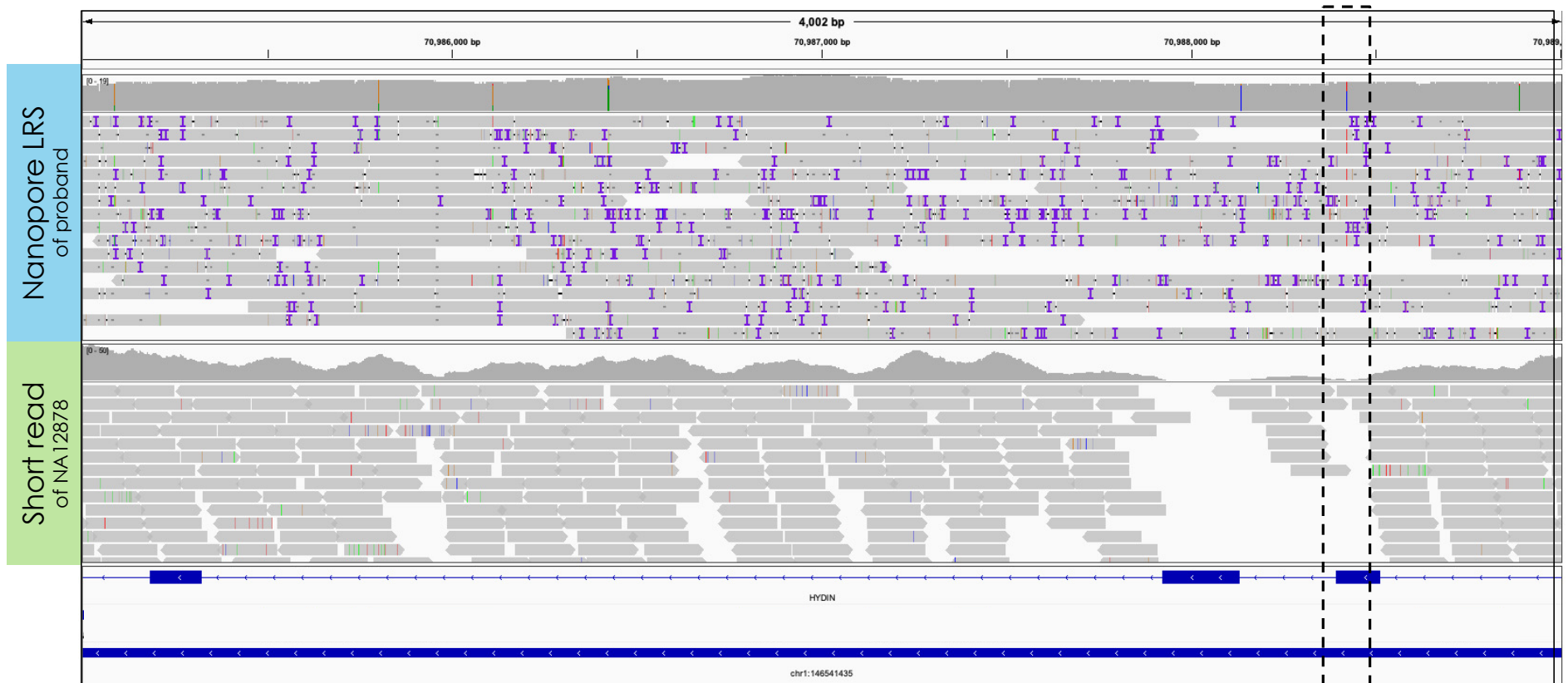
CASE 1

Short reads do not align well within *HYDIN*

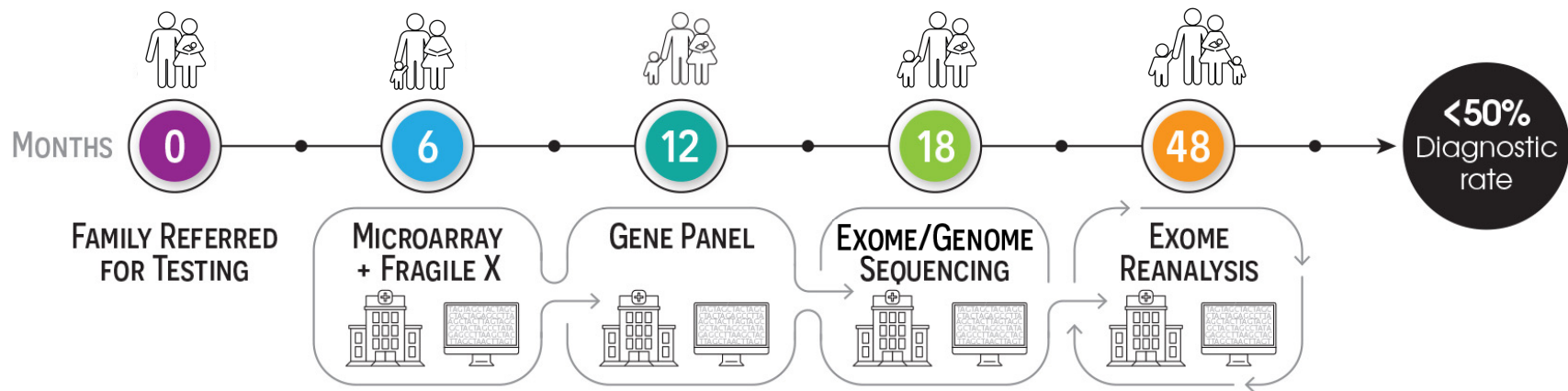


CASE 1

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



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



Incomplete gene-phenotype relationships

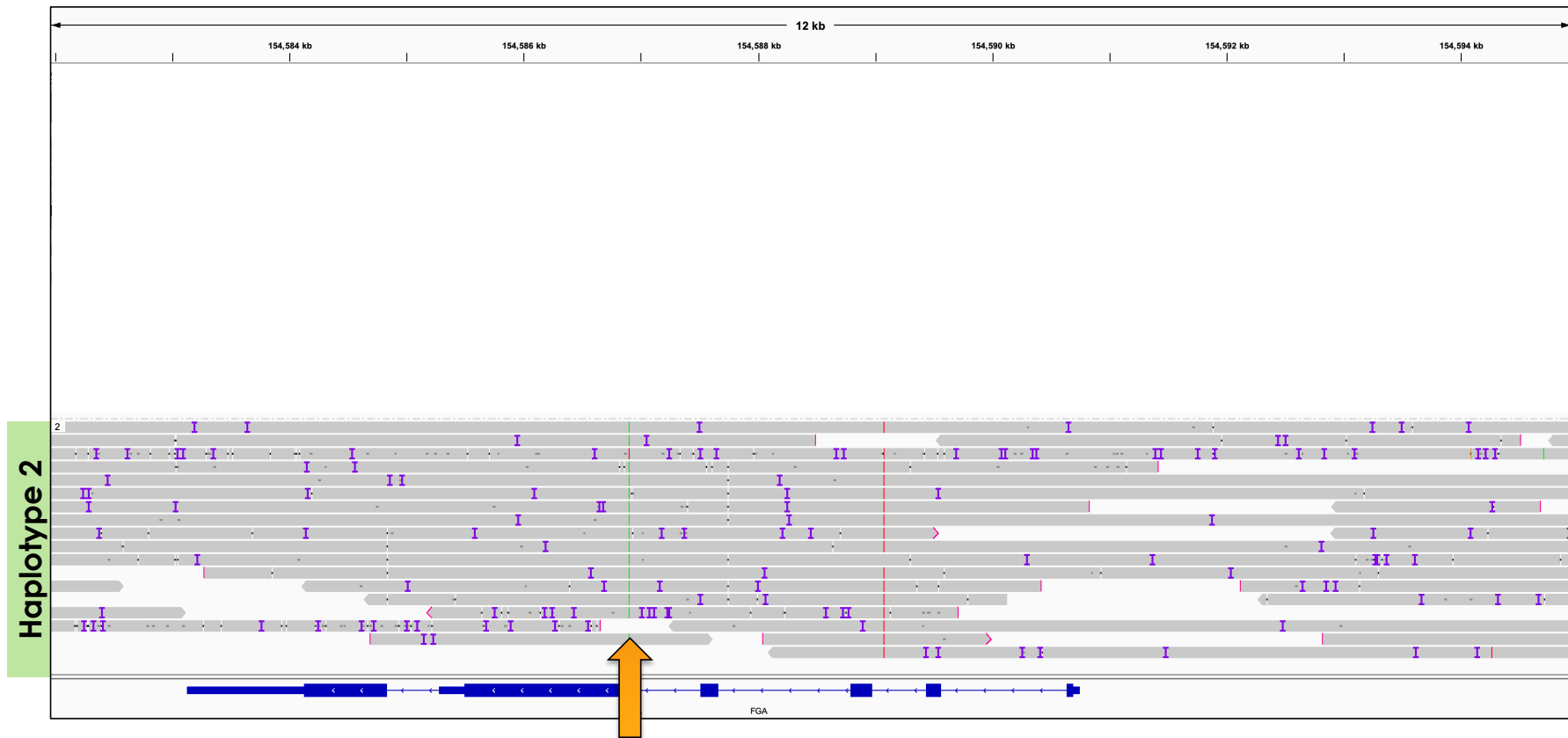
- We do not know the function of all genes

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

CASE 2

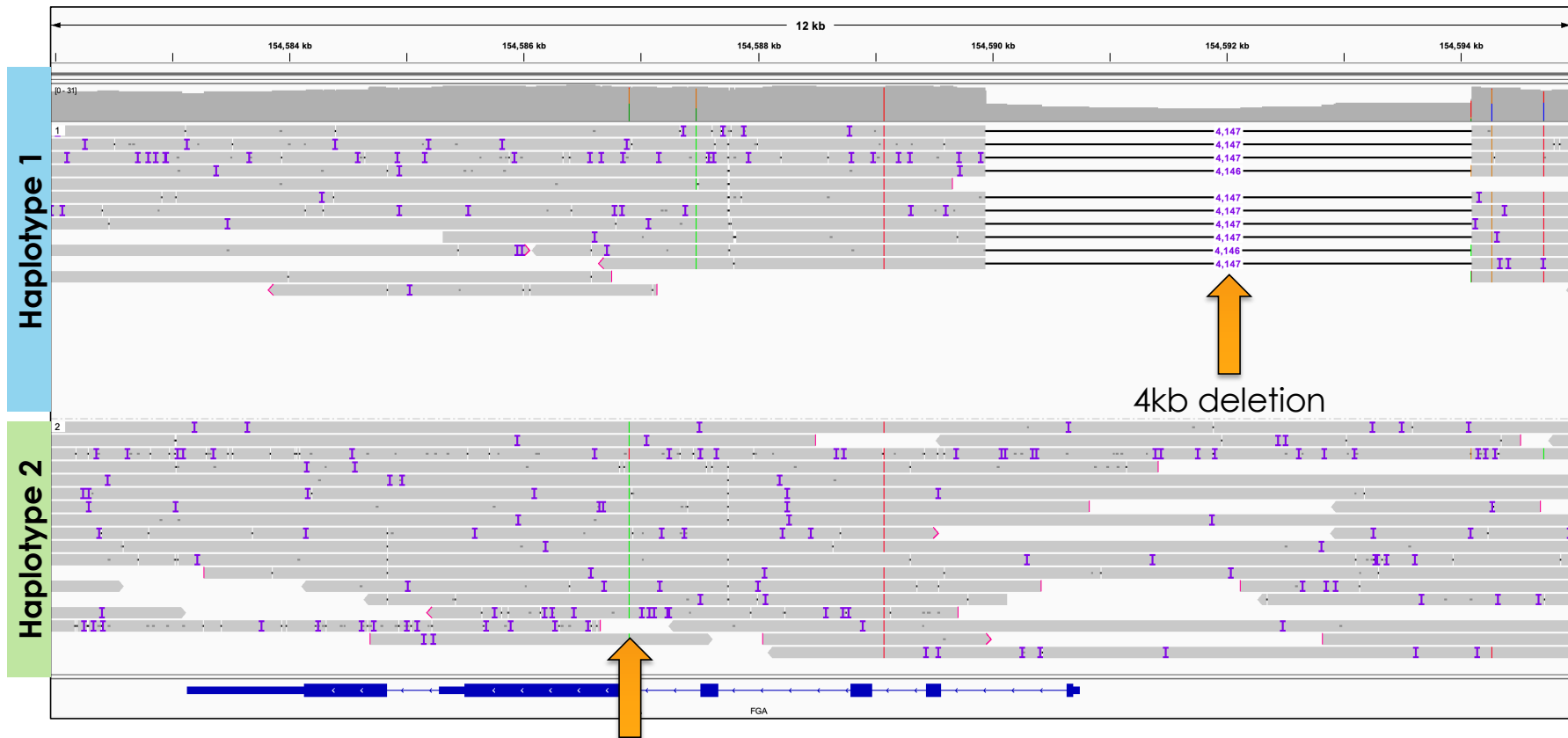
LRS can identify variants missed by prior clinical testing — these are often SVs



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

CASE 2

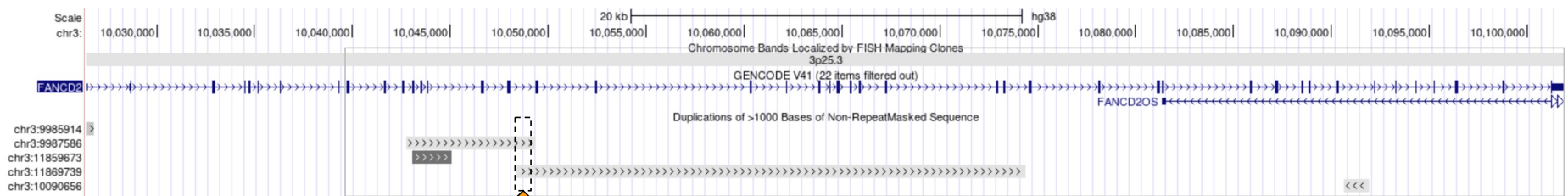
LRS can identify variants missed by prior clinical testing — these are often SVs



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

CASE 3

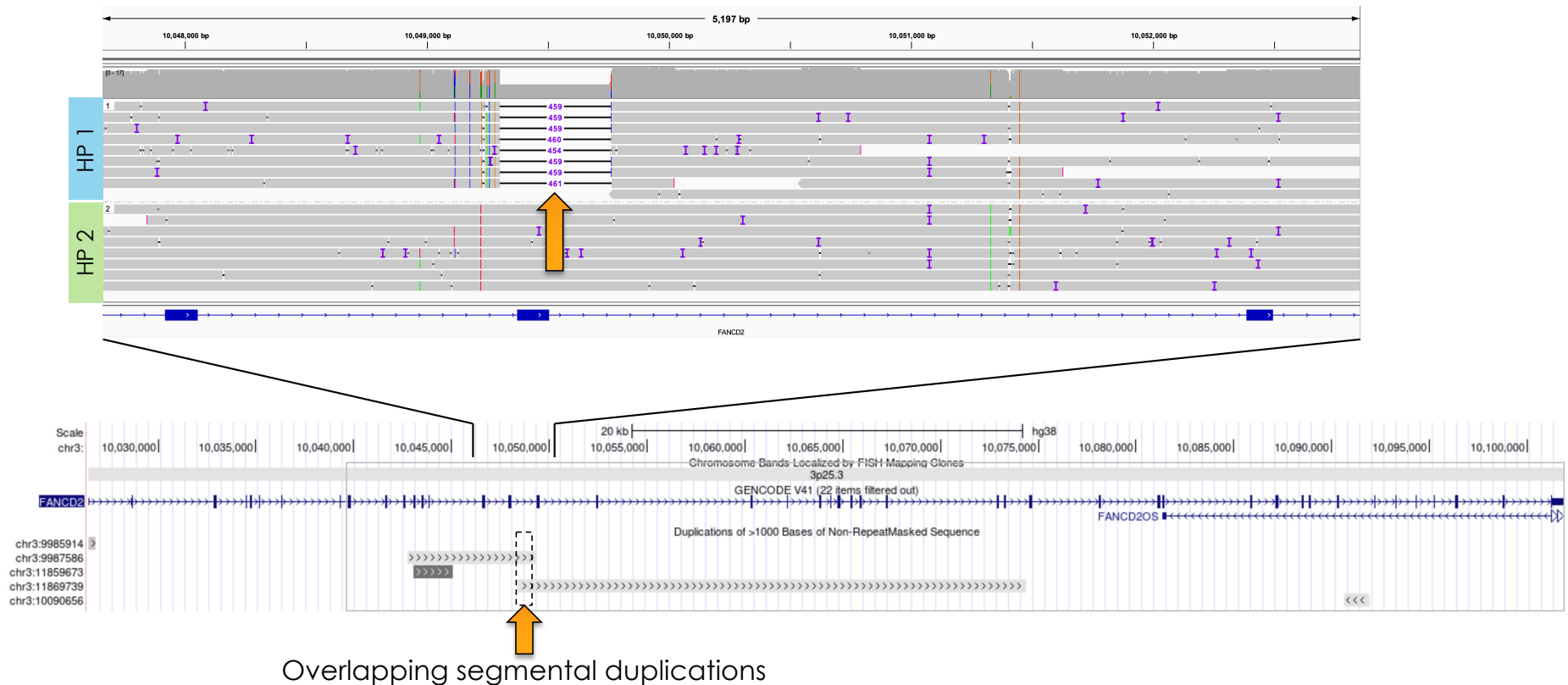
LRS can detect variants in regions of the genome difficult to analyze with short reads



Overlapping segmental duplications

CASE 3

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



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

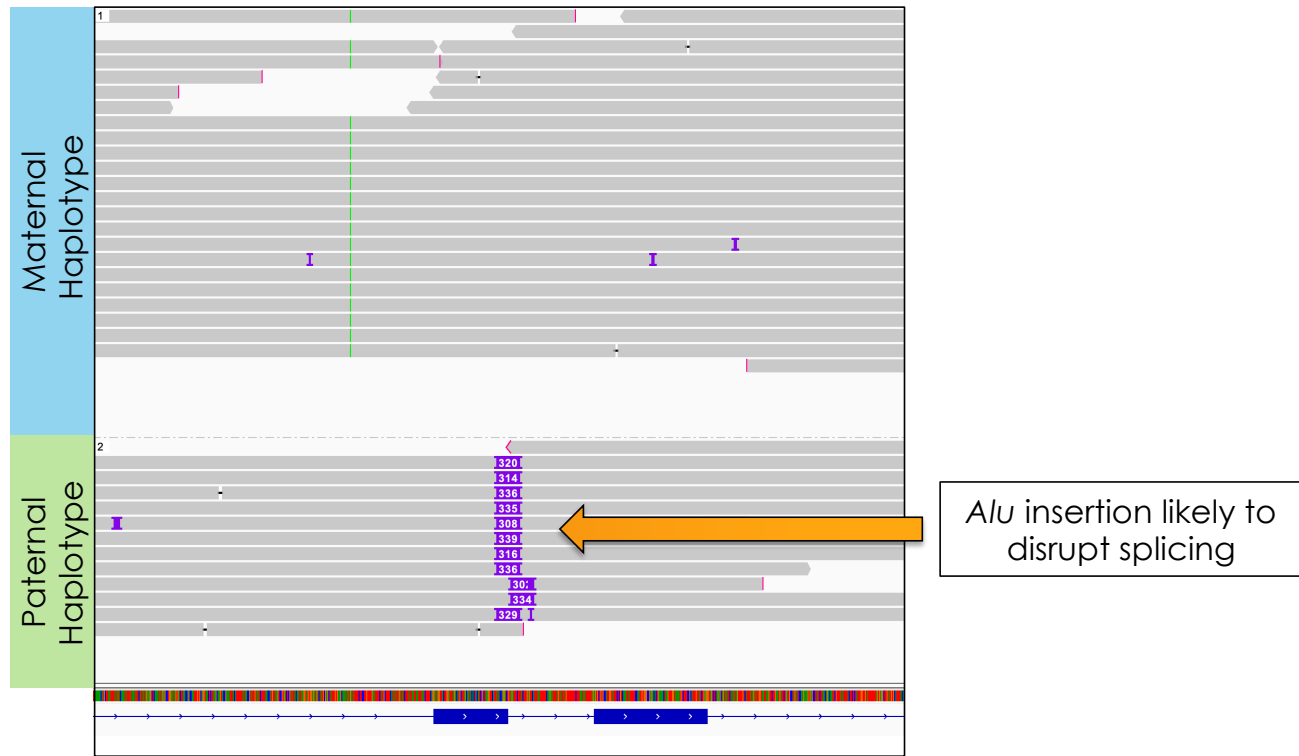
CASE 4

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**

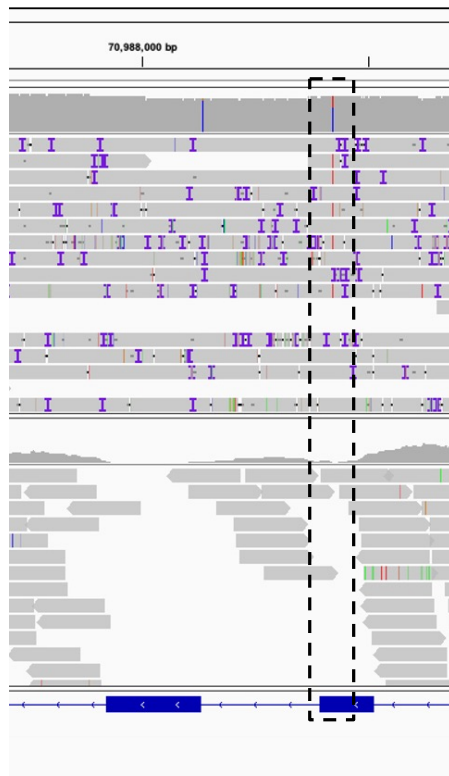
CASE 4

An individual unsolved after clinical testing

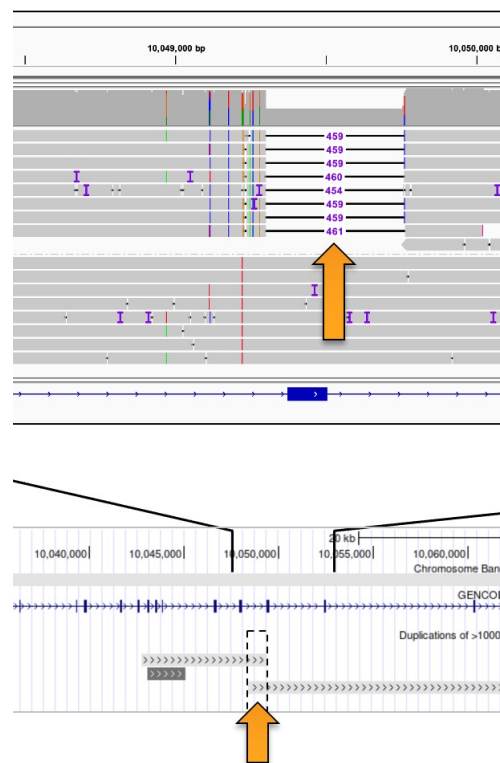


Existing databases cannot be used to determine the allele frequency of these variants

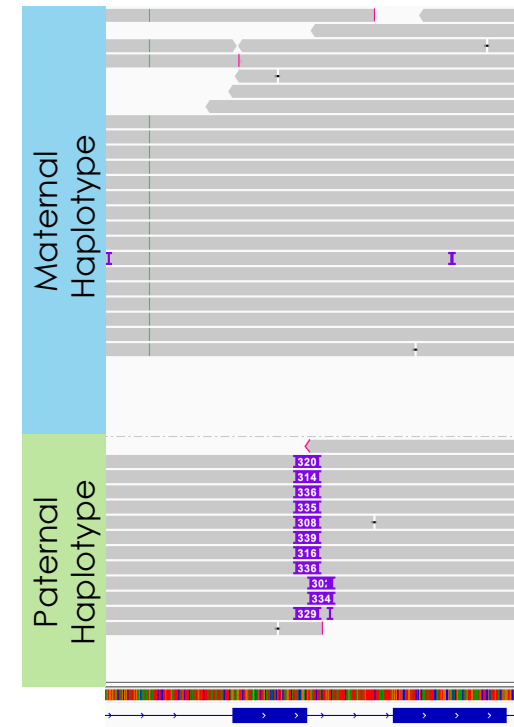
Case 1: Ciliopathy



Case 3: Fanconi anemia



Case 4: Glycogen Storage Disease



The 1000 Genomes Project characterized patterns of human genetic variation

ARTICLE

OPEN
doi:10.1038/nature15339

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

ARTICLE

OPEN
doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

The 1000 Genomes Project Consortium has completed the first phase of the project, which has characterized a polymorphism onto high-quality ancestries. We

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 gene 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.

Structural variants (SVs), including deletions, insertions, duplications and inversions, account for most varying base pairs (bp) among individual human genomes¹. Numerous studies have implicated SVs in human health with associated phenotypes ranging from cognitive disabilities to predispositions to obesity, cancer and other maladies^{2–5}. Discovery and genotyping of these variants remains challenging, however, since SVs are prone to arise in repetitive regions and internal SV structures can be complex⁶. This has created challenges for genome-wide association studies (GWAS)^{7,8}. Despite recent methodological and technological advances^{9–11}, efforts to perform discovery, genotyping and statistical haplotype-block integration of all major SV classes have so far been lacking. Earlier SV surveys depended on microarrays¹² as well as genomic clone-based approaches limited to a small number of samples^{13,14}. More recently, short-read DNA sequencing data from the initial phases of the 1000 Genomes Project¹⁵ enabled us to construct sets of SVs genome-wide across populations, with enhanced size and breakpoint resolution¹⁶. Previous 1000 Genomes Project SV set releases, however, encompassed fewer individuals and were largely or entirely limited to deletions, in spite of the relevance of other SV classes to human genetics^{17,18}.

The objective of the Structural Variation Analysis Group has been to discover and catalogue the classes of SVs that are common to DNA

algorithms—BWA¹⁹ and mntAST²⁰—and performed SV discovery and genotyping using an ensemble of nine different algorithms (Extended Data Fig. 1 and Supplementary Note). We applied several orthogonal experimental platforms for SV set assessment, refinement and characterization (Supplementary Table 2) and to calculate the false discovery rate (FDR) for each SV class (Table 1). Callset refinements facilitated through long-read sequencing enabled us to incorporate a number of additional SVs into our callset, including an additional 698 inversions and 9,132 small (<1 kbp) deletions, compared to the SV set released with the 1000 Genomes Project marker paper¹⁵. As a result, our callset differs slightly relative to the marker paper's SV set¹⁵ (see Supplementary Table 2). We merged individual callsets to construct our unified release (Table 1), comprising 42,279 biallelic deletions, 6,025 biallelic duplications, 2,929 mCNVs (multi-allelic copy-number variants), 786 inversions, 168 nuclear mitochondrial insertions (NMIs), and 16,631 mobile element insertions (MEIs), including 12,748, 3,048 and 835 insertions of Alu, L1 and SVA (SINE-B, VNTR and Alu composite) elements, respectively.

SV non-reference genotype concordance estimates ranged from ~98% for biallelic deletions and MEI classes to ~94% for biallelic duplications. 69% of SVs were noted with respect to the Database of Genetic Variants (DGV)²¹ (see Supplementary Table 2).

The objective of the Structural Variation Analysis Group has been to discover and catalogue the classes of SVs that are common to DNA

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

*List of participants and their affiliations

LRS of 100 Genomes Project samples to characterize previously inaccessible patterns of human variation

ARTICLE
OPEN
doi:10.1038/nature15399

A global reference for human genetic variation
The 1000 Genomes Project Consortium*

The 1000 Genomes Project Consortium has applied whole-genome sequencing to a diverse set of individuals from 26 human populations. We have characterized a polymorphism onto high-quality ancestries. We

ARTICLE
doi:10.1038/nature15394

A map of human genome variation from population-scale sequencing
The 1000 Genomes Project Consortium*

The 1000 Genomes Project Consortium has applied whole-genome sequencing to a diverse set of individuals from 26 human populations. We have characterized a polymorphism onto high-quality ancestries. We

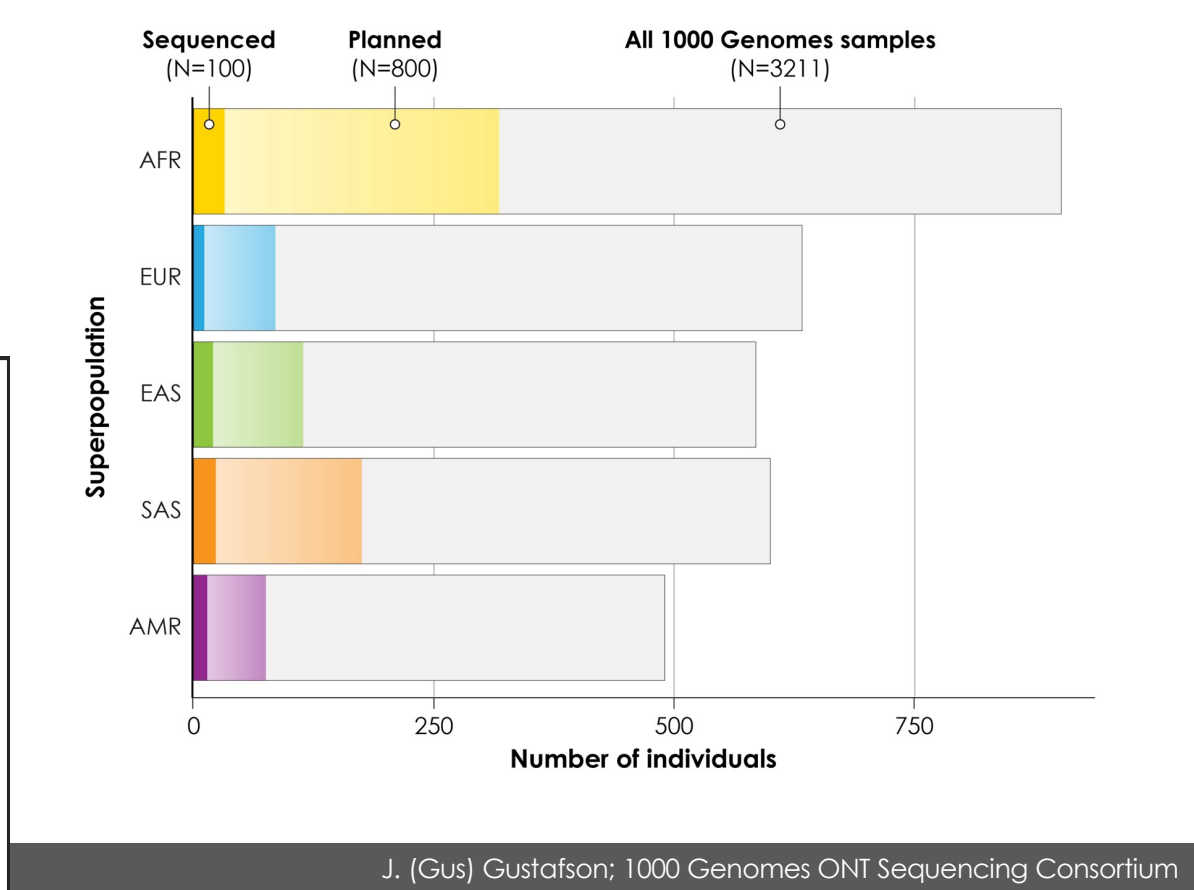
ARTICLE
OPEN
doi:10.1038/nature15394

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 gene 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.

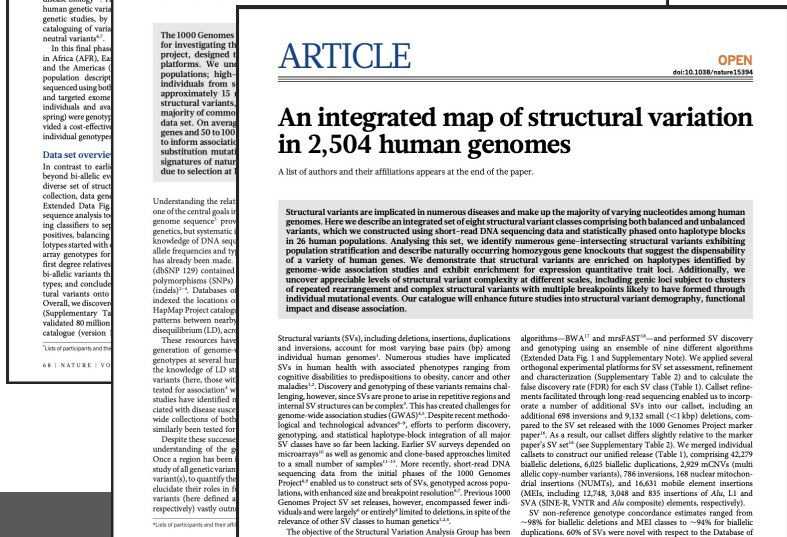
Structural variants (SVs), including deletions, insertions, duplications and inversions, account for most varying base pairs (bp) among individual human genomes. Numerous studies have implicated SVs in human health with associated phenotypes ranging from cognitive disabilities to predispositions to obesity, cancer and other maladies^{1–3}. Discovery and genotyping of these variants remains challenging, however, since SVs are prone to arise in repetitive regions and internal SV structures can be complex⁴. This has created challenges for genome-wide association studies (GWAS)^{5,6}. Despite recent methodological and technological advances^{7–9}, efforts to perform discovery, genotyping and statistical haplotype-block integration of all major SV classes have so far been lacking. Earlier SV surveys depended on microarrays¹⁰ as well as genomic clone-based approaches limited to a small number of samples^{11,12}. More recently, short-read DNA sequencing data from the initial phases of the 1000 Genomes Project¹³ enabled us to construct sets of SVs genotyped across populations, with enhanced size and breakpoint resolution¹⁴. Previous 1000 Genomes Project SV set releases, however, encompassed fewer individuals and were largely or entirely limited to deletions, in spite of the relevance of other SV classes to human genetics¹⁵.

The objective of the Structural Variation Analysis Group has been to discover and catalogue structural classes of SVs defined by DNA sequence. We have applied a combination of BWA¹⁶ and mntAST¹⁷—and performed SV discovery and genotyping using an ensemble of nine different algorithms (Extended Data Fig. 1 and Supplementary Note). We applied several orthogonal experimental platforms for SV set assessment, refinement and characterization (Supplementary Table 2) and to calculate the false discovery rate (FDR) for each SV class (Table 1). Callset refinements facilitated through long-read sequencing enabled us to incorporate a number of additional SVs into our catalog, including an additional 698 inversions and 9,132 small (<1 kbp) deletions, compared to the SV set released with the 1000 Genomes Project marker paper¹³. As a result, our callset differs slightly relative to the marker paper's SV set¹³ (see Supplementary Table 2). We merged individual callsets to construct our unified release (Table 1), comprising 42,279 biallelic deletions, 6,025 biallelic duplications, 2,929 mCNVs (multi-allelic copy-number variants), 786 inversions, 168 nuclear mitochondrial insertions (NMTIs), and 16,651 mobile element insertions (MEIs), including 12,748, 3,048 and 835 insertions of Alu, L1 and SVA (SINE, VNTR and Alu composite) elements, respectively). SV non-reference genotype concordance estimates ranged from ~98% for biallelic deletions and MEI classes to ~94% for biallelic duplications. 60% of SVs were noted with respect to the Database of Genetic Variants (DGV)¹⁸ (see Supplementary Note).

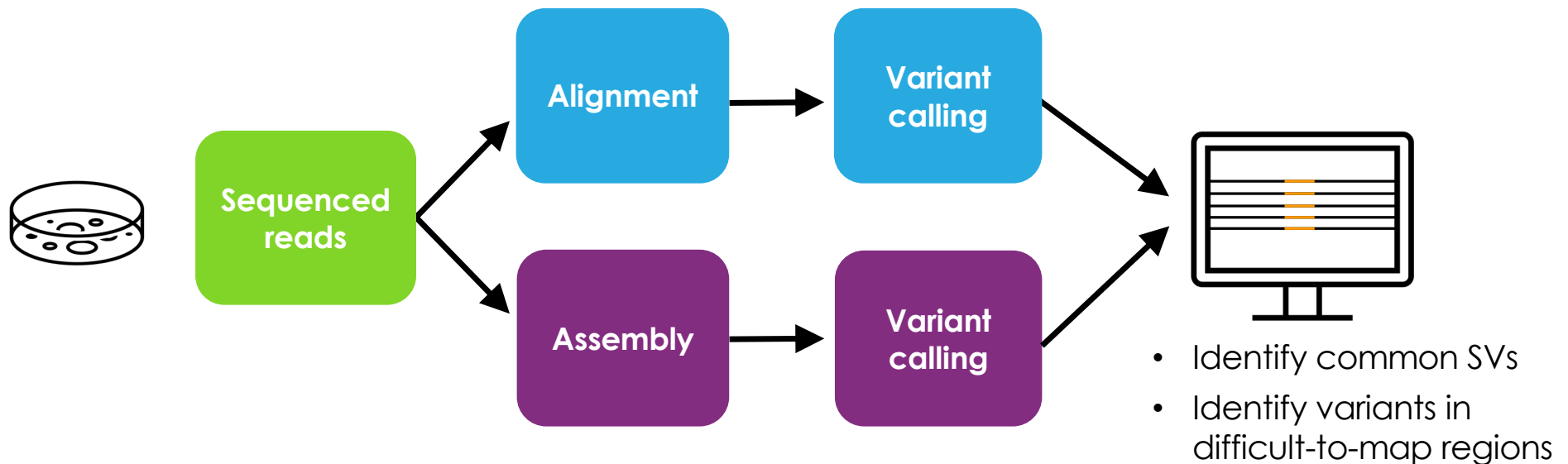


LRS of 100 Genomes Project samples to characterize previously inaccessible patterns of human variation

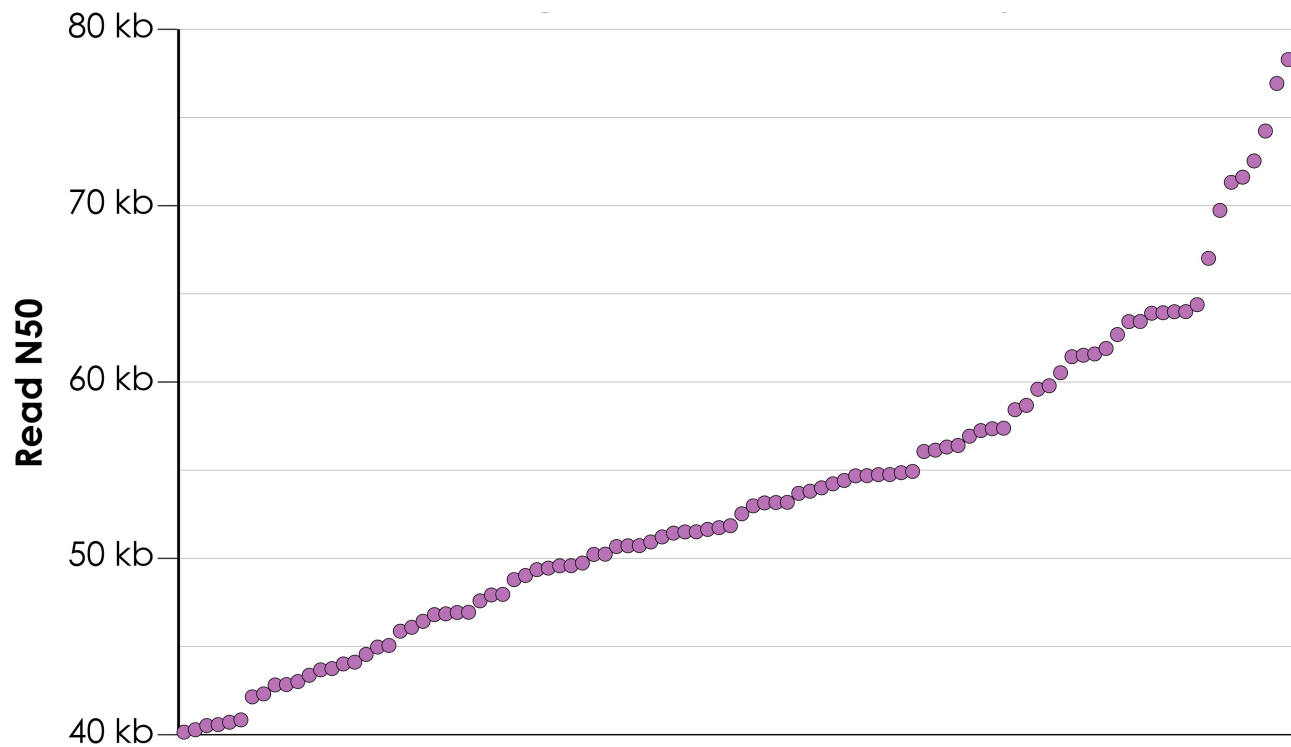
- Catalog of SVs for filtering and variant prioritization
- Expand our understanding of variation in difficult to analyze regions of the genome
- Evaluate variation in regions associated with interesting signals in existing data



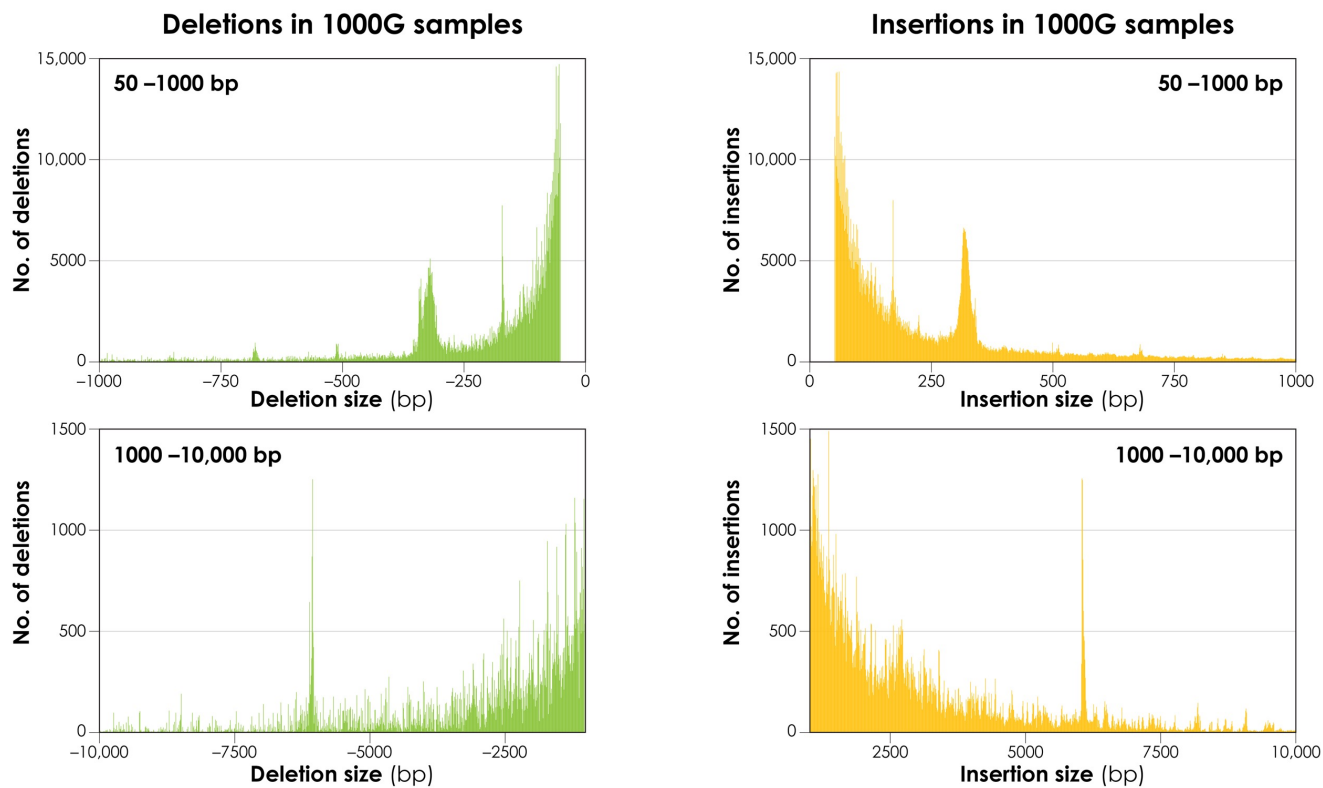
LRS data is analyzed using both alignment-based and assembly-based approaches



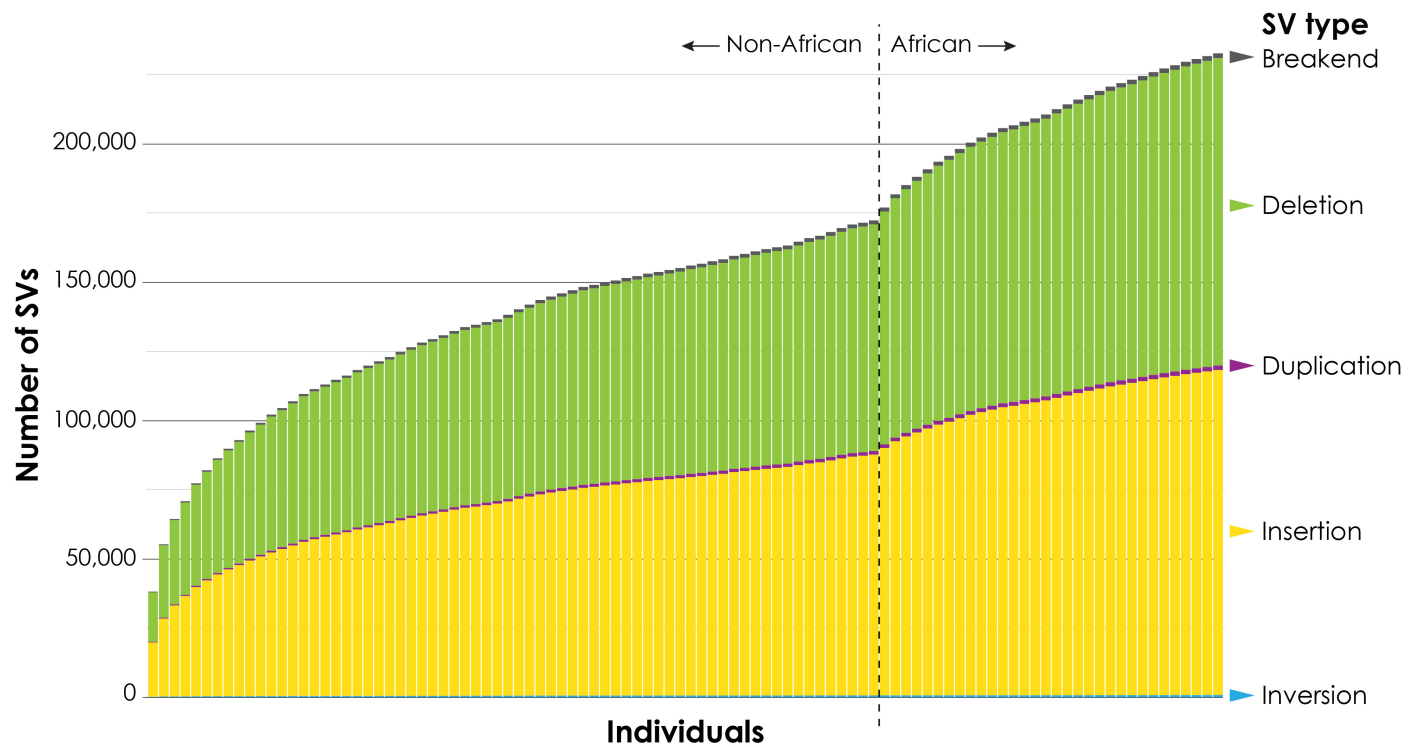
**The first 100 samples have read N50s >40 kb
with >30x coverage**



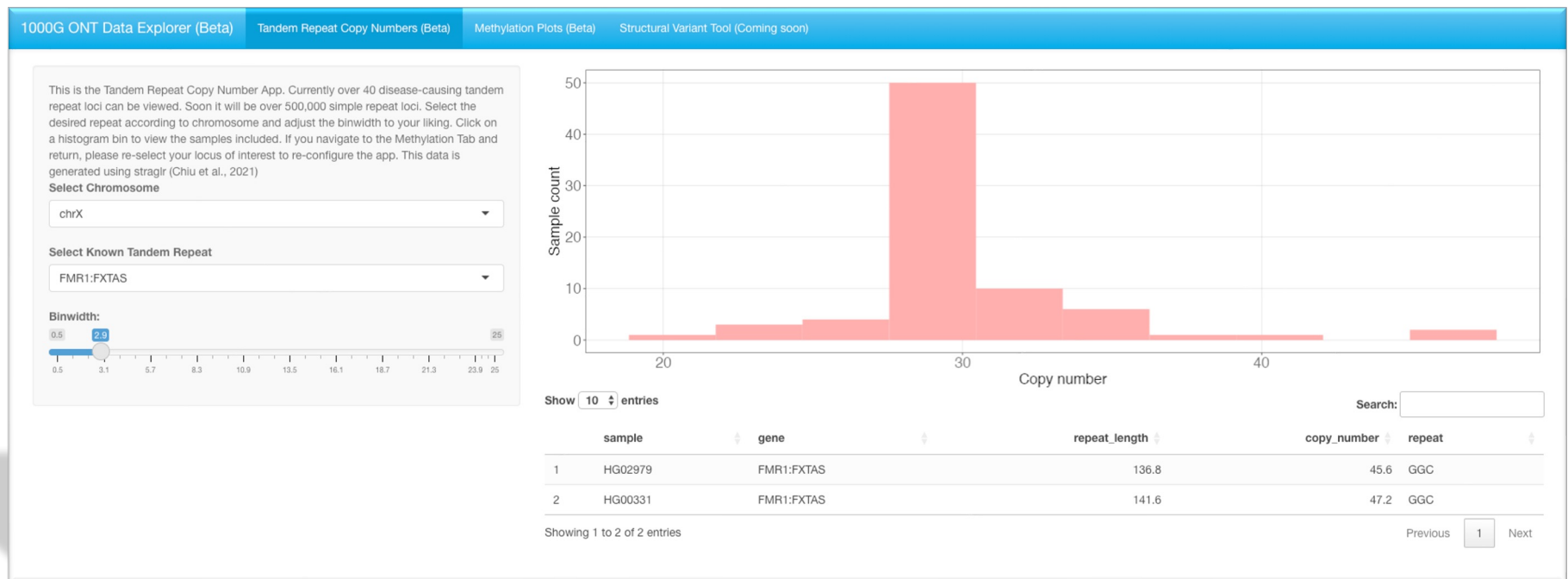
The distribution of insertions and deletions in the first 100 samples follows expected patterns



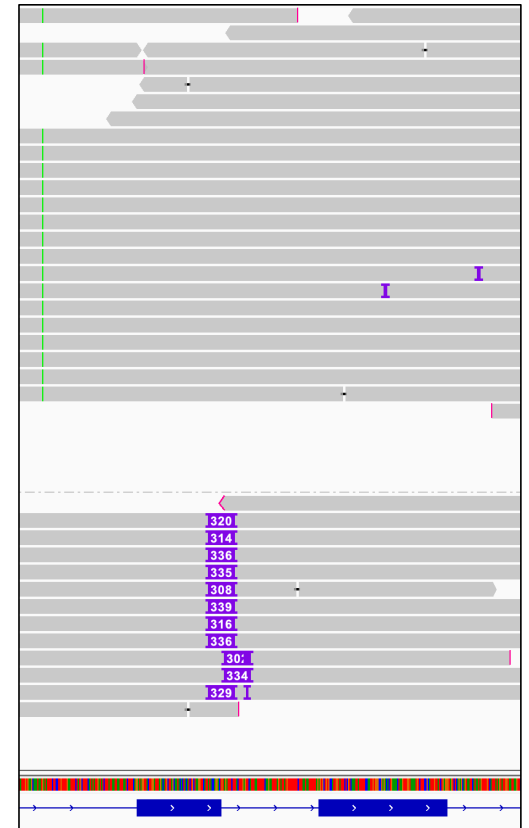
The number of novel SVs increase as more individuals are added



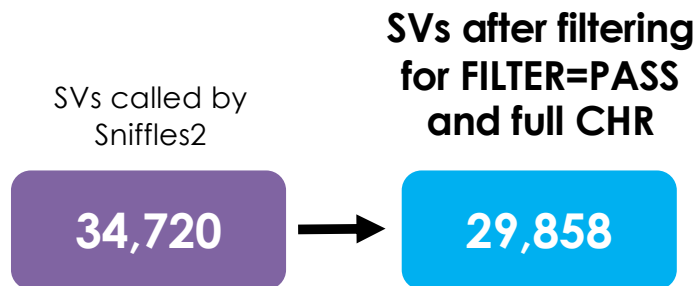
Applications for SV filtering and prioritization using this data are under development



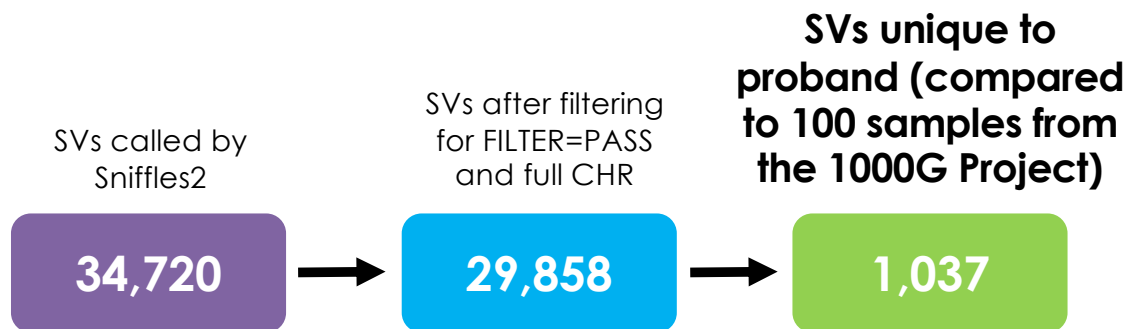
Data from 100 individuals can be used for SV filtering and prioritization



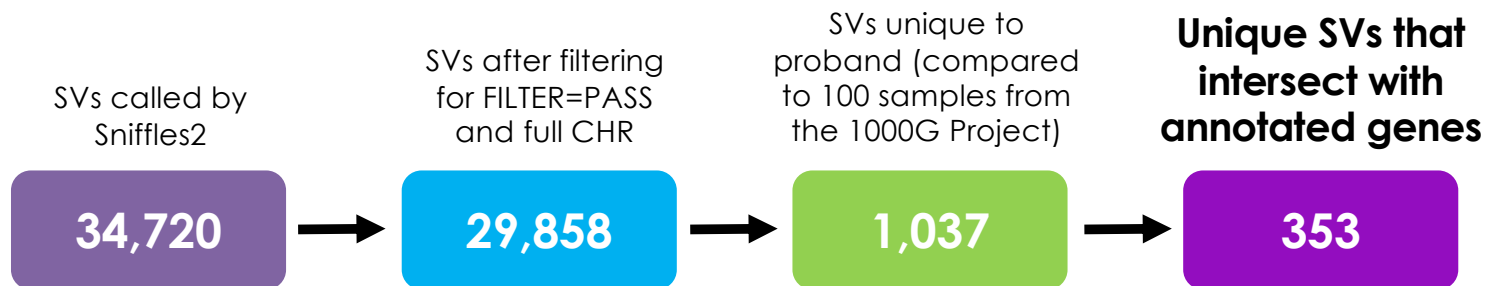
Data from 100 individuals can be used for SV filtering and prioritization



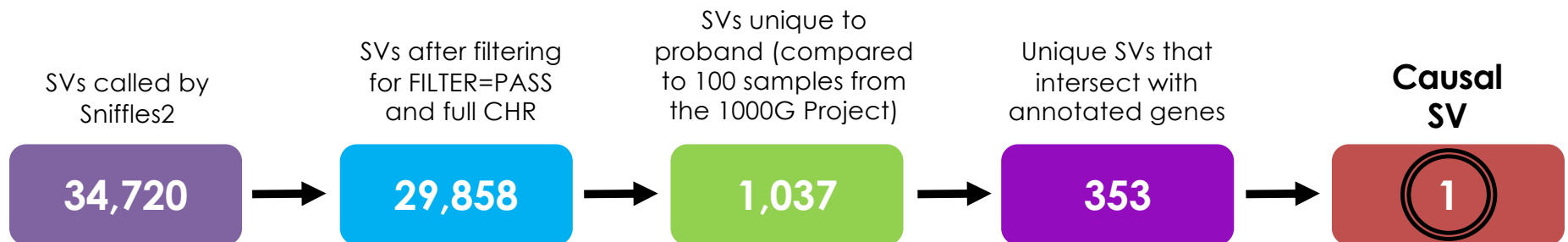
Data from 100 individuals can be used for SV filtering and prioritization



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

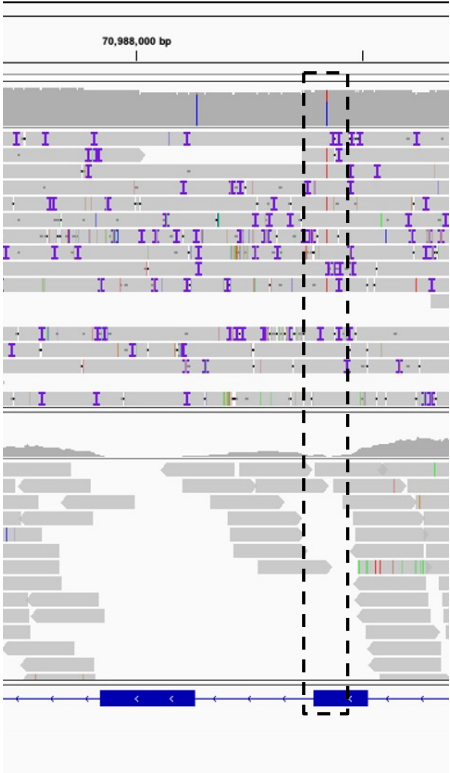
Filter

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

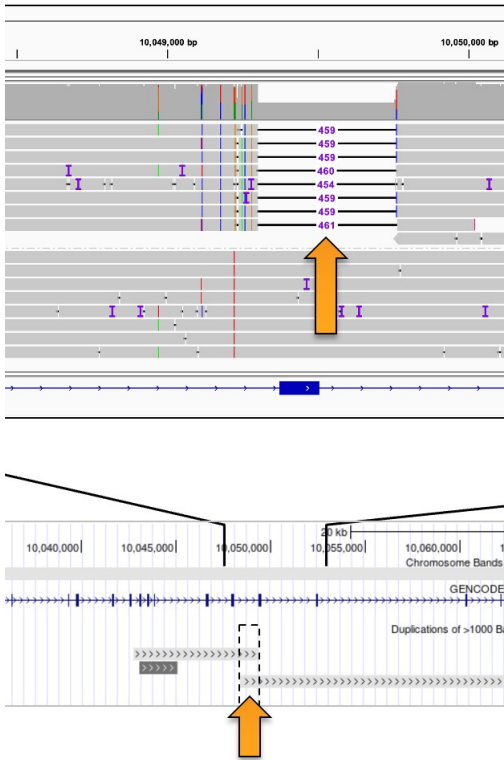
- <https://s3.amazonaws.com/1000g-ont/index.html>

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

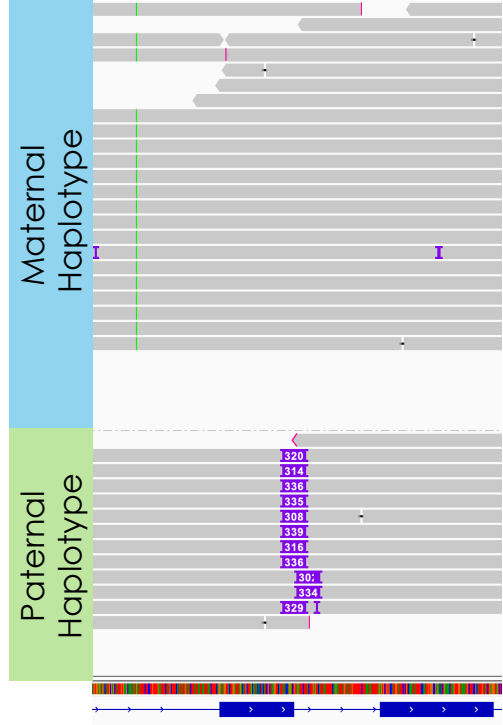
Case 1: Ciliopathy



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



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



The logo for the 1000G ONT Sequencing Consortium is a dark blue circle with a white dashed line around its perimeter. Inside the circle, the text "1000G ONT Sequencing Consortium" is written in a green, sans-serif font, with "1000G ONT" on the top line, "Sequencing" on the second line, and "Consortium" on the third line.

1000G ONT
Sequencing
Consortium

Everyone contributing
to the 1000G ONT
Sequencing Consortium

