Long-read sequencing for rare disease diagnostics

Danny E. Miller, MD, PhD
Assistant Professor
Department of Pediatrics, Division of Genetic Medicine
Department of Laboratory Medicine & Pathology
Seattle Children's Hospital and the University of Washington

CSHL Advanced Sequencing Technologies & Bioinformatics Analysis Course
November 17, 2022

dm1@uw.edu | millerlaboratory.com | @danrdanny

We're recruiting!

Data for exercises

- IGV tracks I find useful:
  - hg38: https://tinyurl.com/ycy9wcnh
  - chm13: https://tinyurl.com/mh77yzvn

- Bam files for exercises:
  - https://tinyurl.com/4xvu2adu

Why long-read sequencing?
A traditional genetic workup is nondiagnostic in 50% of cases

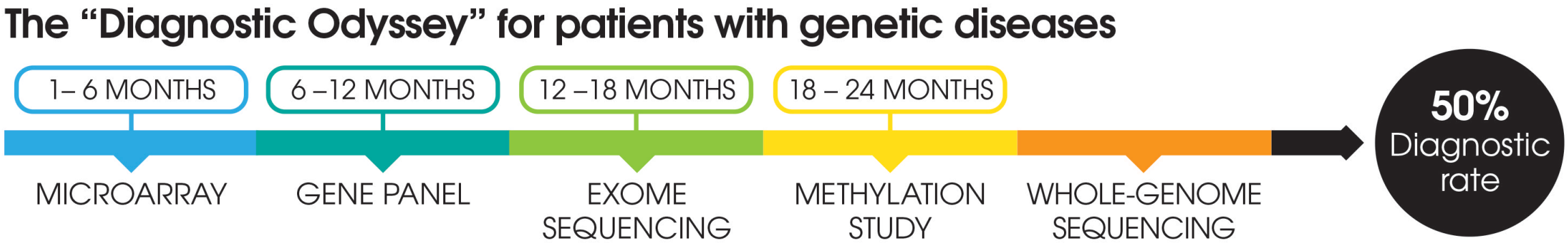
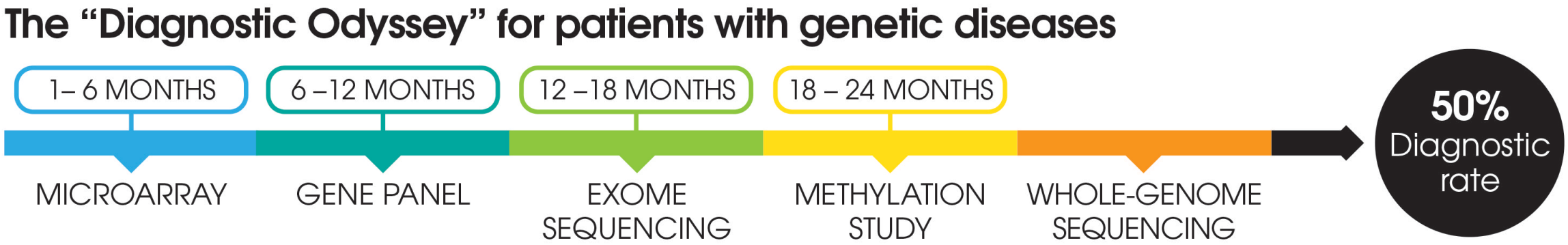


Figure based on data from Frésard *et al*. 2018 and Clark *et al*. 2018

Why long-read sequencing?
A traditional genetic workup is nondiagnostic in 50% of cases



Incomplete gene-phenotype relationships

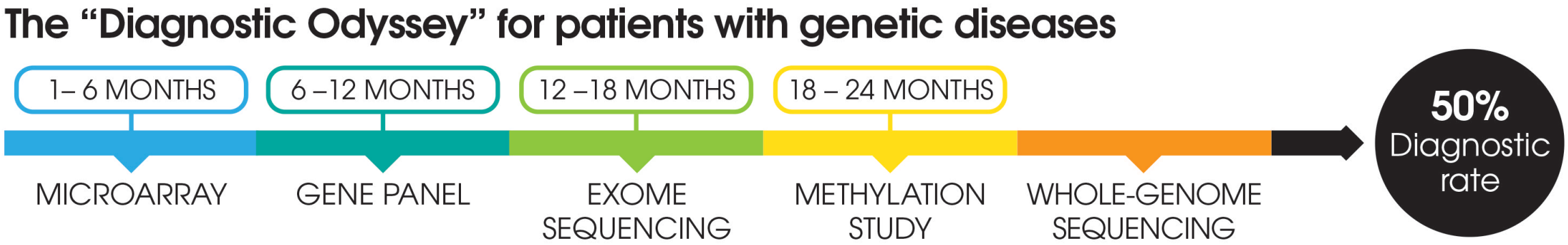
- We do not know the function of all genes

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

Figure based on data from Frésard *et al*. 2018 and Clark *et al*. 2018

Why long-read sequencing?
A traditional genetic workup is nondiagnostic in 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

Figure based on data from Frésard *et al*. 2018 and Clark *et al*. 2018

Targeted LRS can resolve variants in (some) segmental duplications

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

Targeted LRS can resolve segmental dup

- Newborn with respiratory failure at b
- Duo exome sequencing revealed c deletion in *HYDIN*

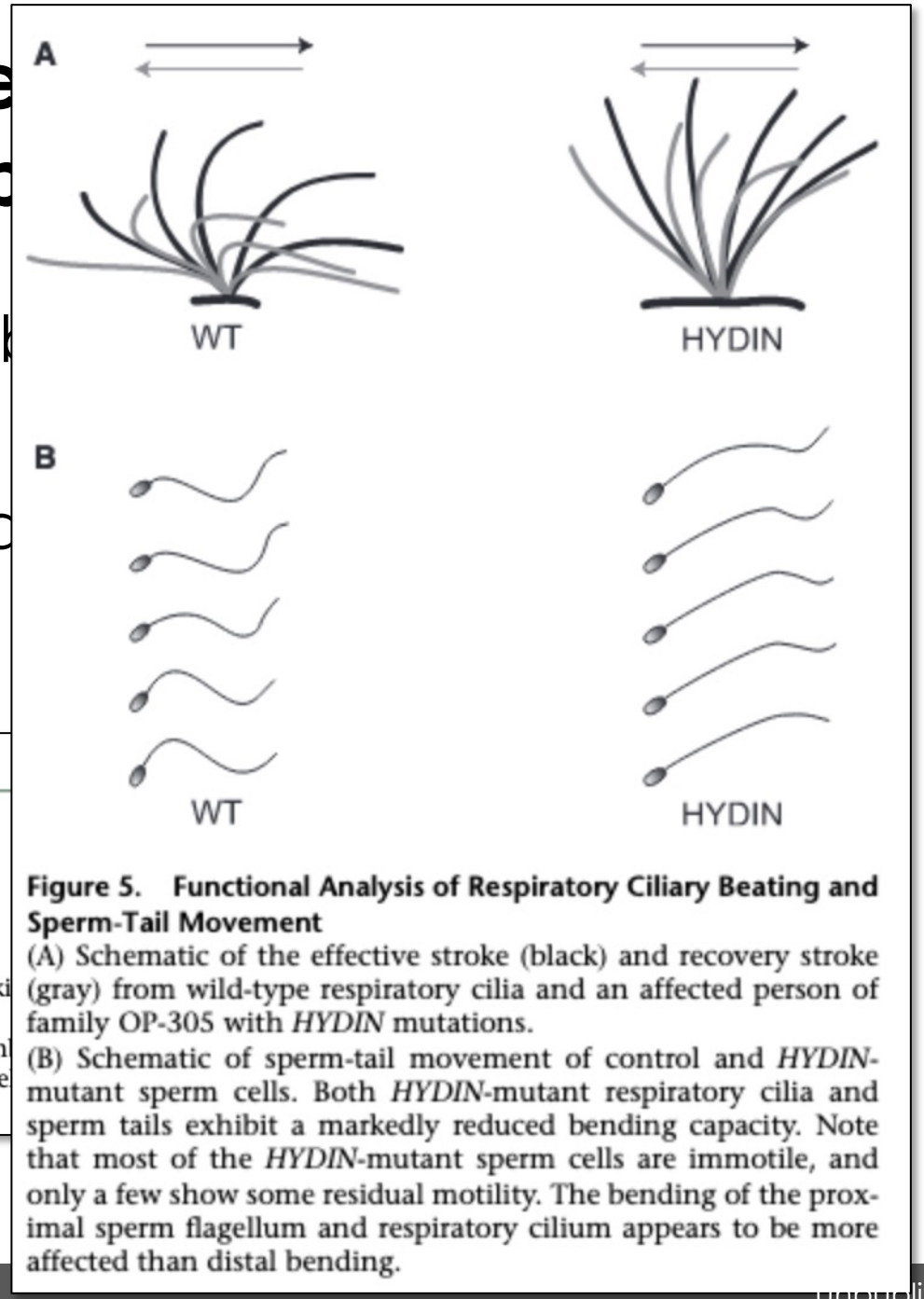


Figure 5. Functional Analysis of Respiratory Ciliary Beating and Sperm-Tail Movement

(A) Schematic of the effective stroke (black) and recovery stroke (gray) from wild-type respiratory cilia and an affected person of family OP-305 with *HYDIN* mutations.

(B) Schematic of sperm-tail movement of control and *HYDIN*-mutant sperm cells. Both *HYDIN*-mutant respiratory cilia and sperm tails exhibit a markedly reduced bending capacity. Note that most of the *HYDIN*-mutant sperm cells are immotile, and only a few show some residual motility. The bending of the proximal sperm flagellum and respiratory cilium appears to be more affected than distal bending.

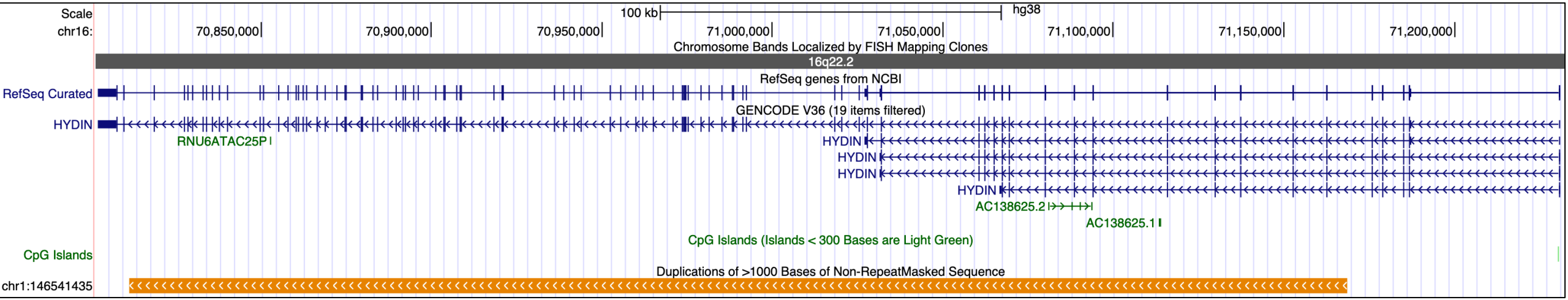
ARTICLE

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 Nürnl Peter Nürnberg,[9] Eddie M.K. Chung,[10] Richard Reinhardt,[11] June K. Marthin,[12] Kim G. Nie Hannah M. Mitchison,[2,14,*] and Heymut Omran[1,14,*]

unpublished data

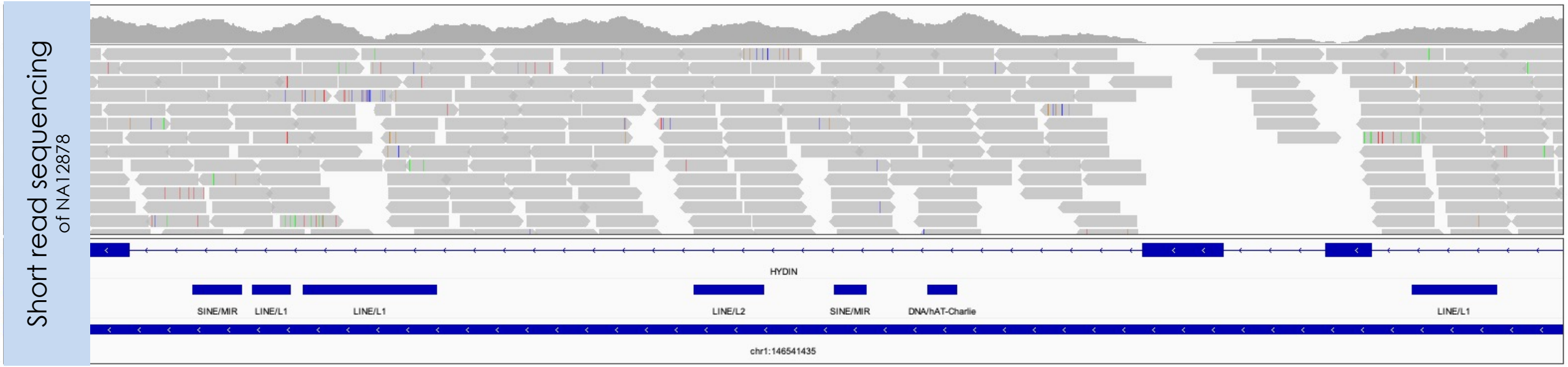
HYDIN is a 400-kb gene containing a 380-kb segmental duplication



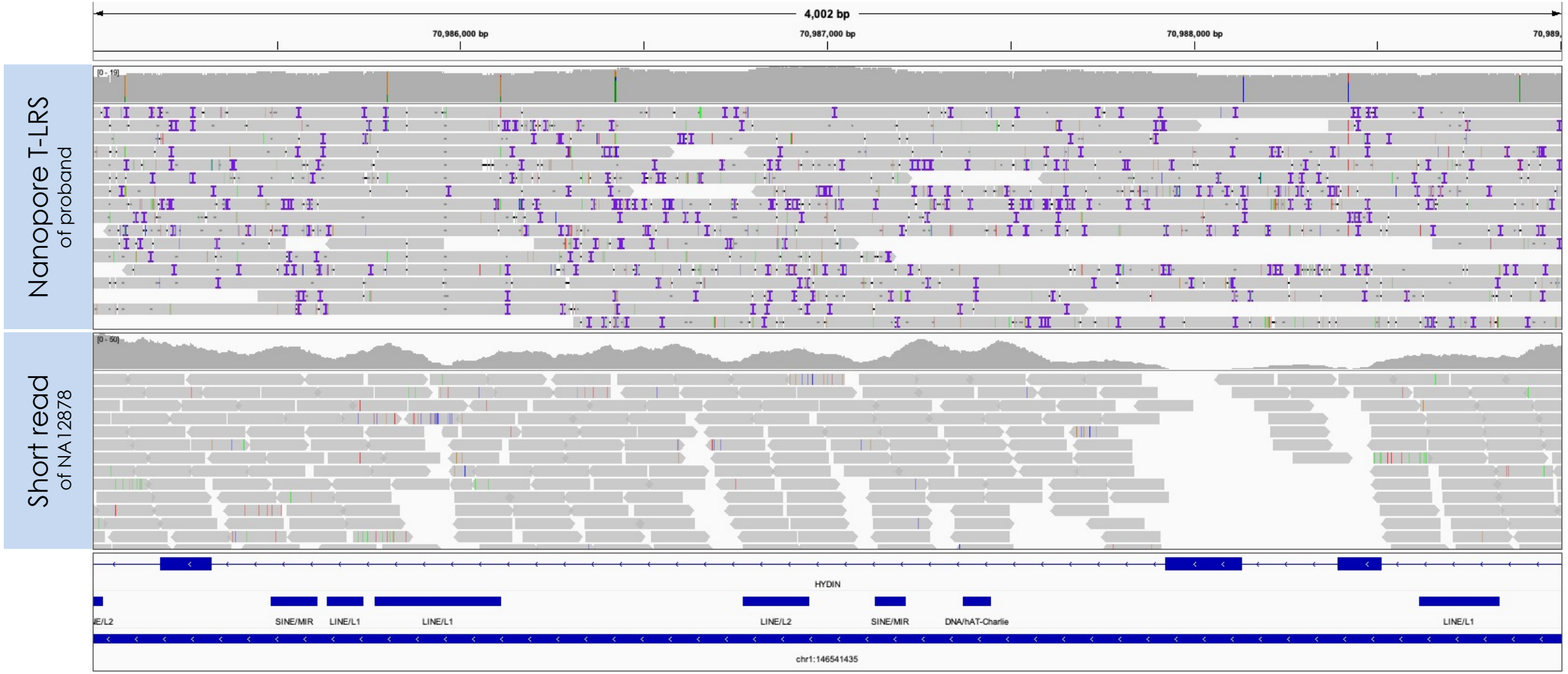
Segmental duplication

Short reads do not align well within *HYDIN*

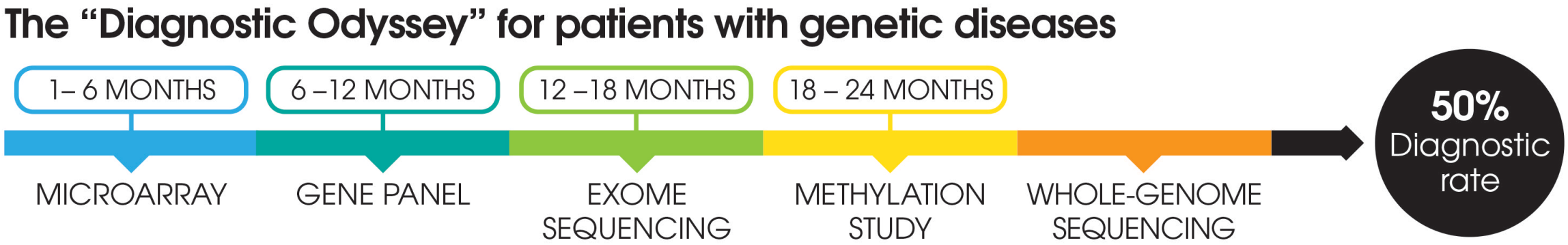
Low coverage



T-LRS yields even coverage across *HYDIN* and did not identify a second pathogenic variant



A traditional genetic workup is nondiagnostic in 50% of cases



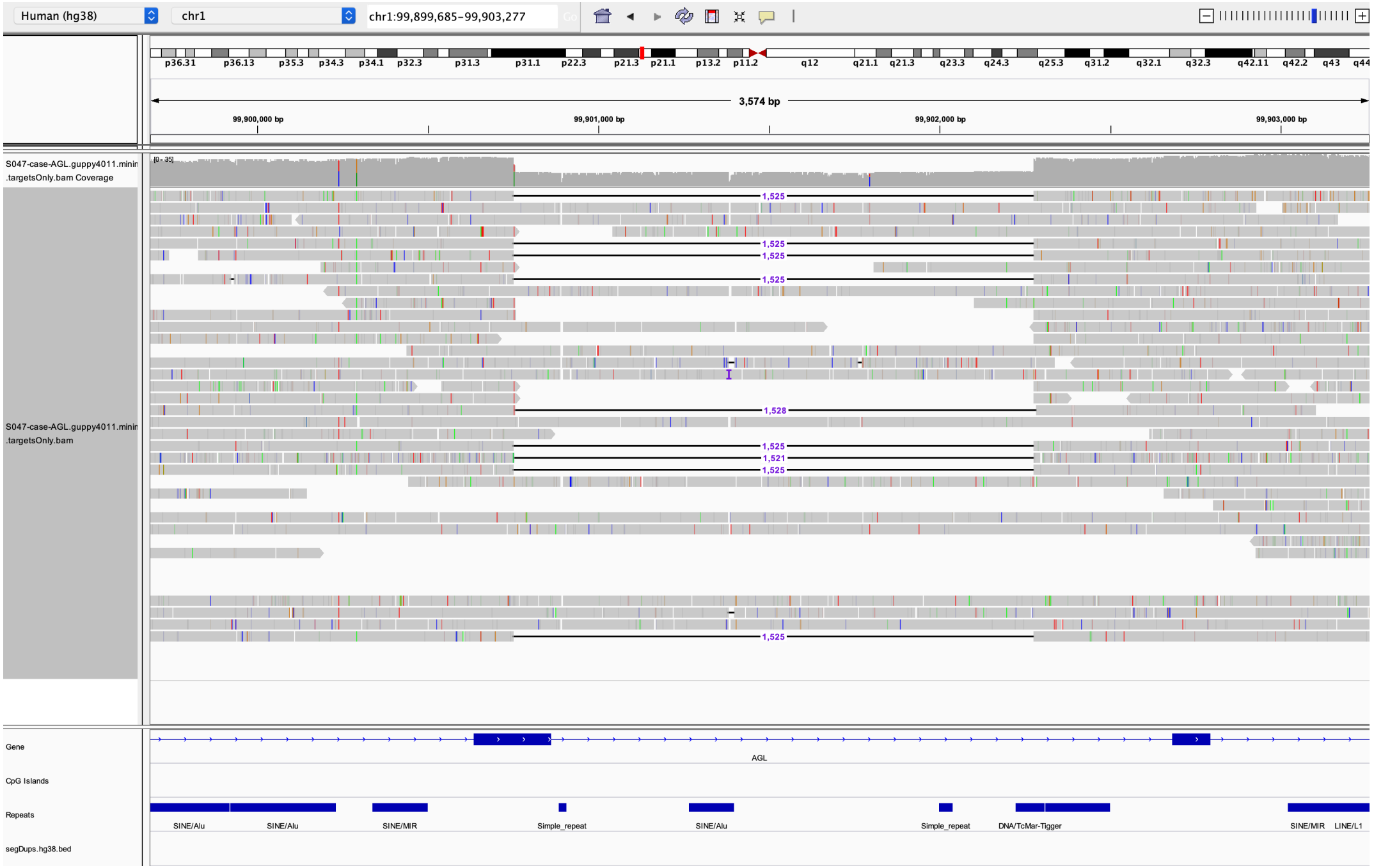
Incomplete gene-phenotype relationships

- We do not know the function of all genes

Variants that are difficult to detect or interpret

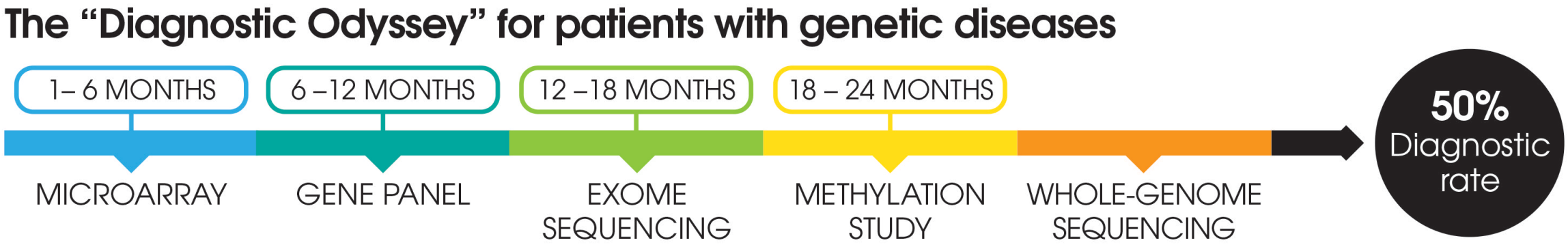
- Many genes are difficult to sequence
- **Structural variants can be difficult to identify**
- Predicting the impact of a variant is difficult

Figure based on data from Frésard *et al*. 2018 and Clark *et al*. 2018



Miller *et al.*, 2021

A traditional genetic workup is nondiagnostic in 50% of cases



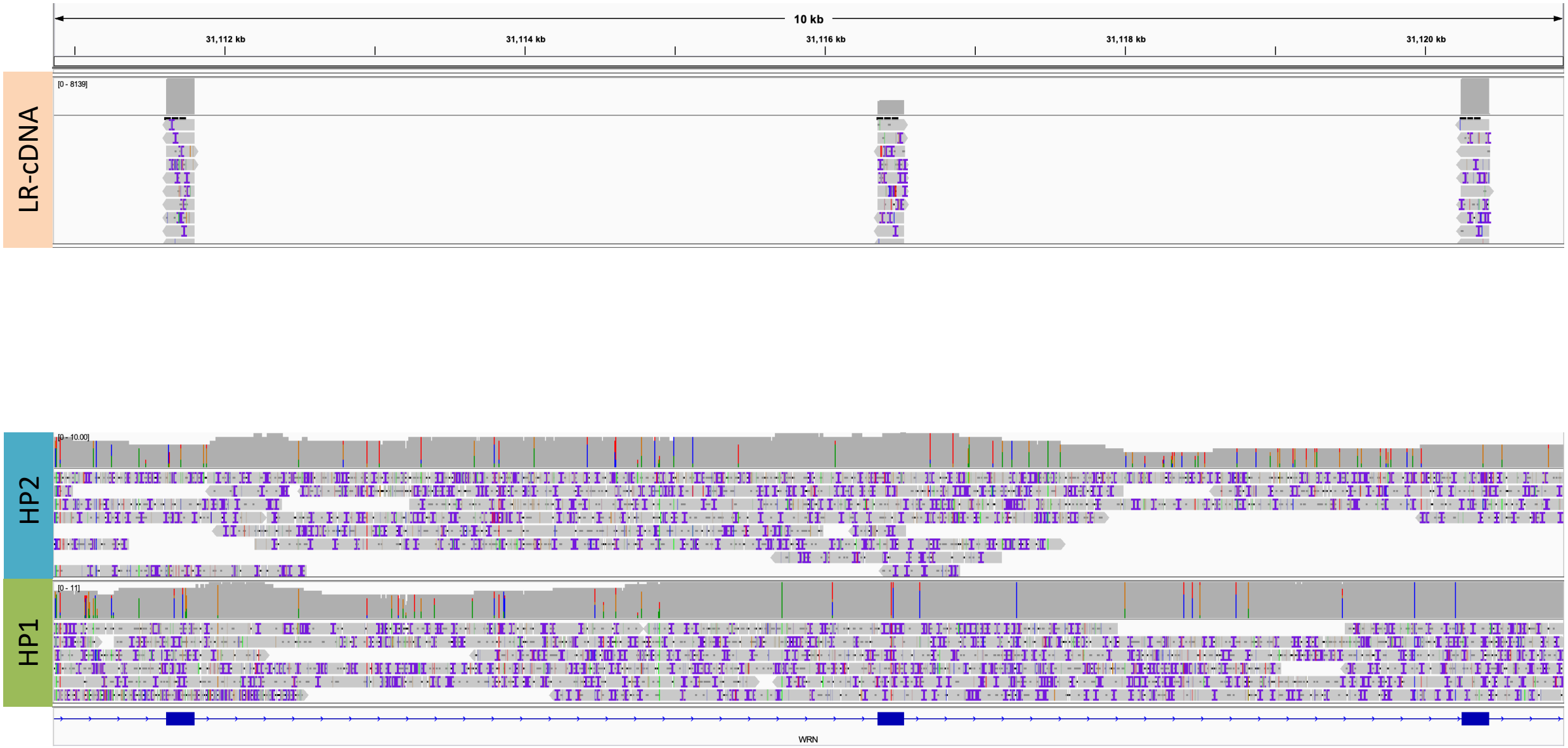
Incomplete gene-phenotype relationships

- 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**

Figure based on data from Frésard *et al*. 2018 and Clark *et al*. 2018


10 kb
31,112 kb
31,114 kb
31,116 kb
31,118 kb
31,120 kb
LR-cDNA
[0 - 8139]
HP2
[0 - 10.00]
HP1
[0 - 11]
WRN

Long-read sequencing will shorten the time to diagnosis and increase the rate of diagnosis

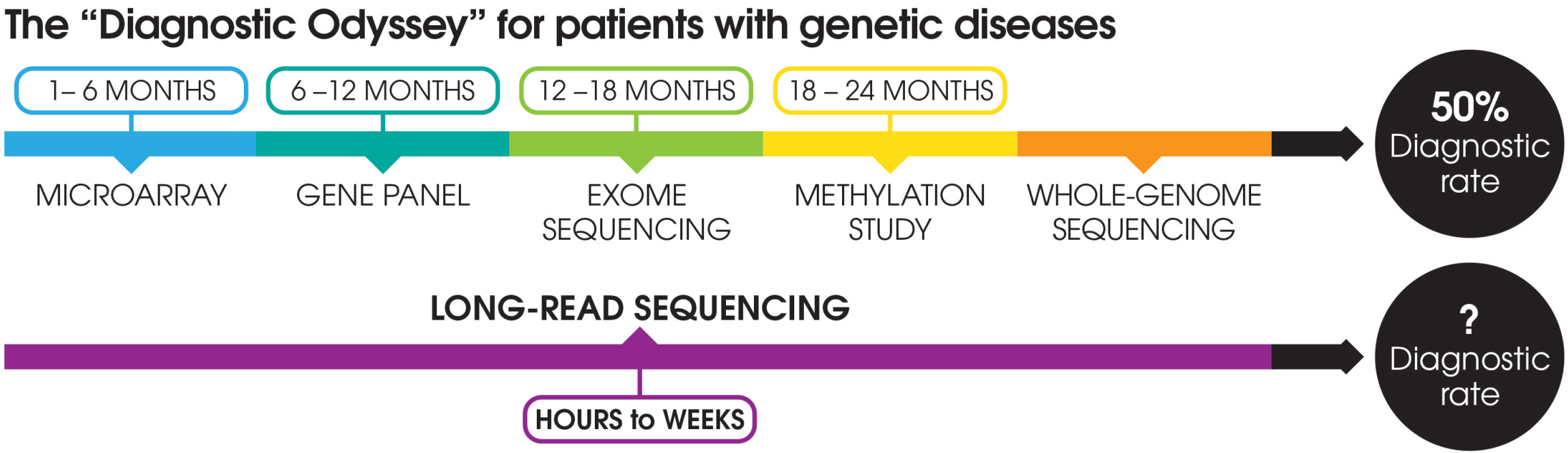
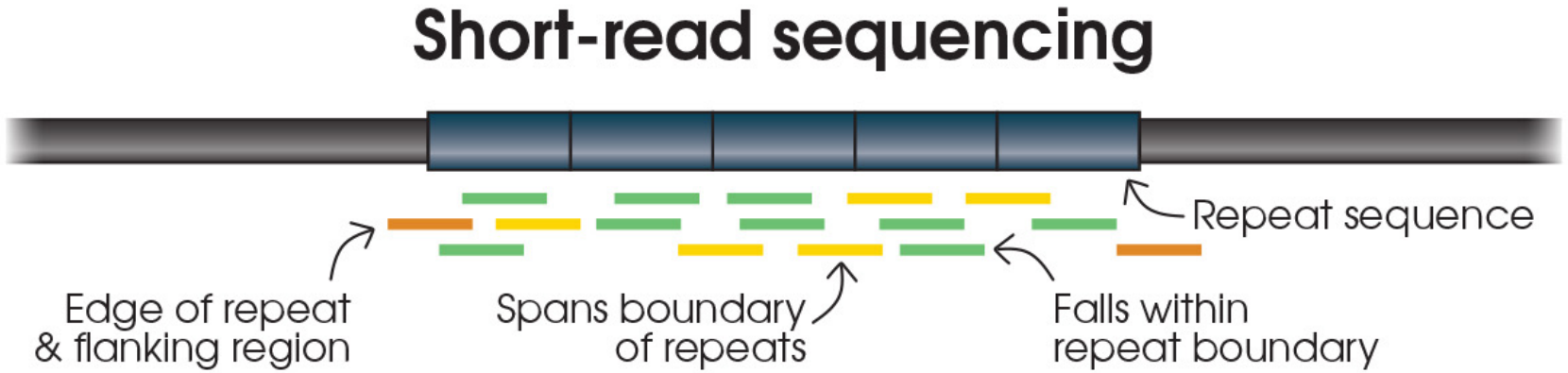


Figure based on data from Frésard *et al*. 2018 and Clark *et al*. 2018

LONG-READ SEQUENCING

technology

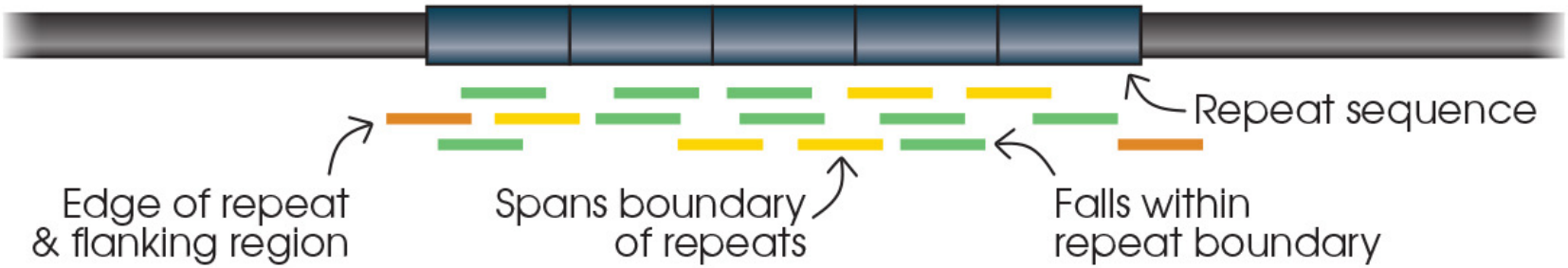
Long-read sequencing simplifies analysis of complex genomic regions



- Read length: **100 – 300 bp**
- Read accuracy: **>99%**
- Cost to sequence a human to 30x coverage: <**$500**

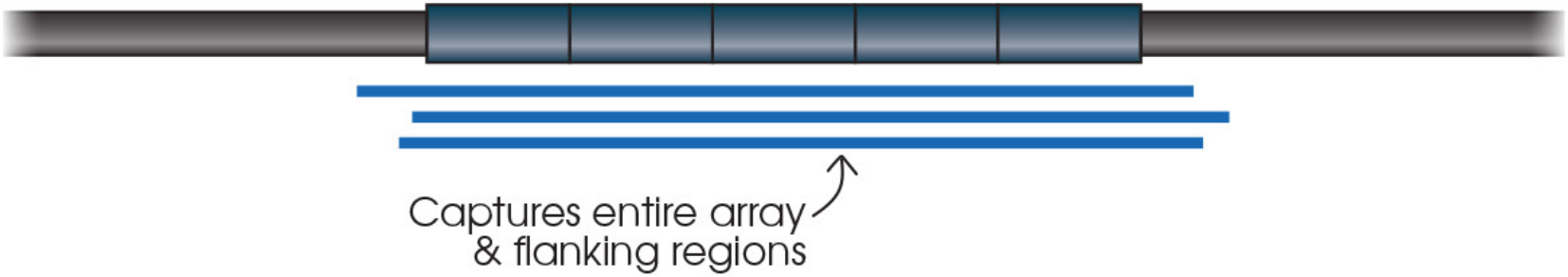
Long-read sequencing simplifies analysis of complex genomic regions

Short-read sequencing



- Read length: **100 – 300 bp**
- Read accuracy: **>99%**
- Cost to sequence a human to 30x coverage: <**$500**

Long-read sequencing



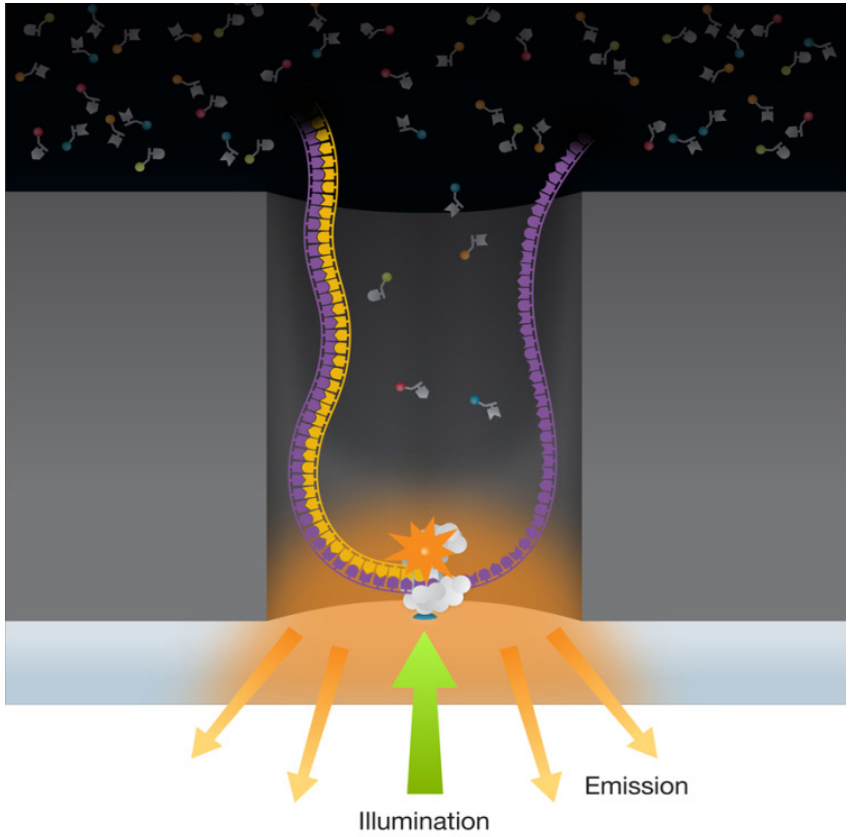
- Read length: **1 kb to >2 Mb**
- Read accuracy: **90–99%**
- Cost to sequence a human to 30x coverage: **$500 – $3k**
- **Information-rich reads**

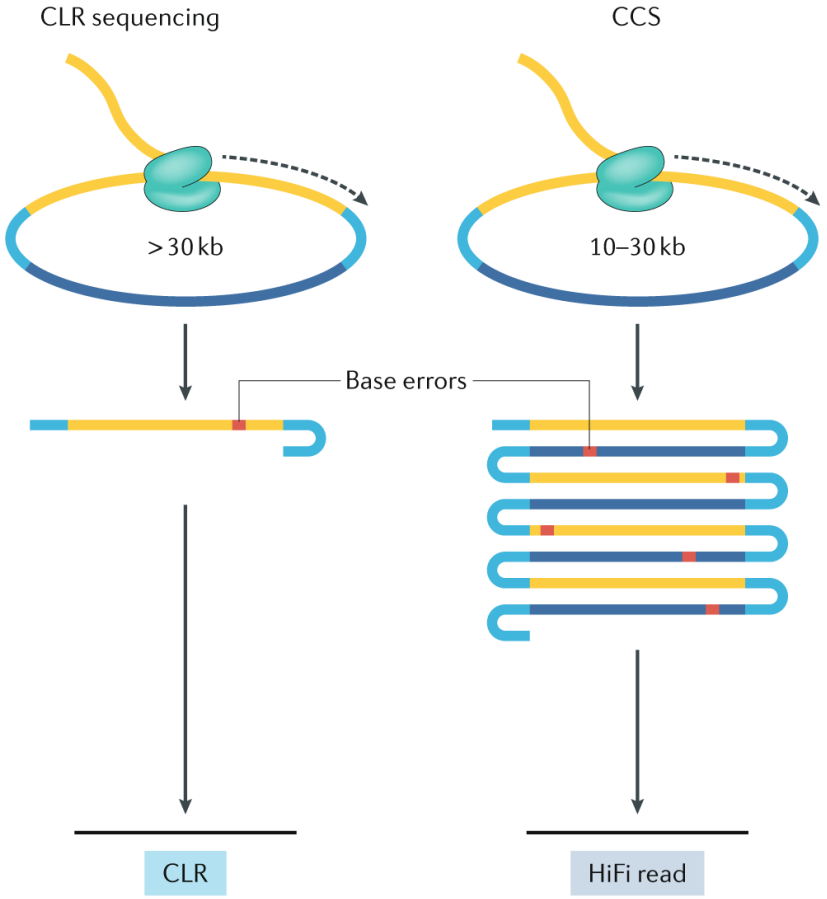
Two commercial long-read sequencing technologies are available

Pacific Biosciences

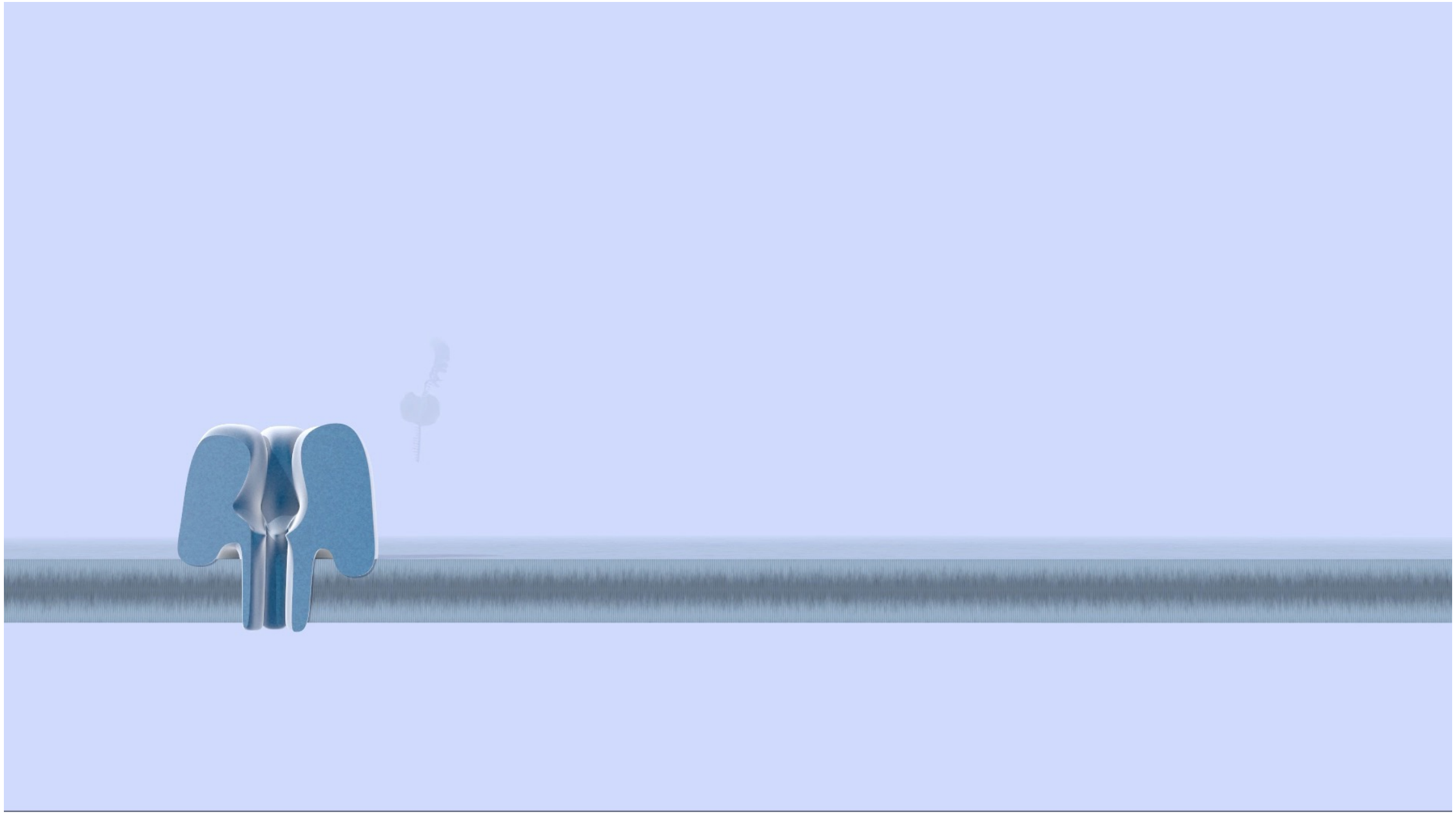
Oxford Nanopore

PacBio sequencing can produce extremely high-quality reads



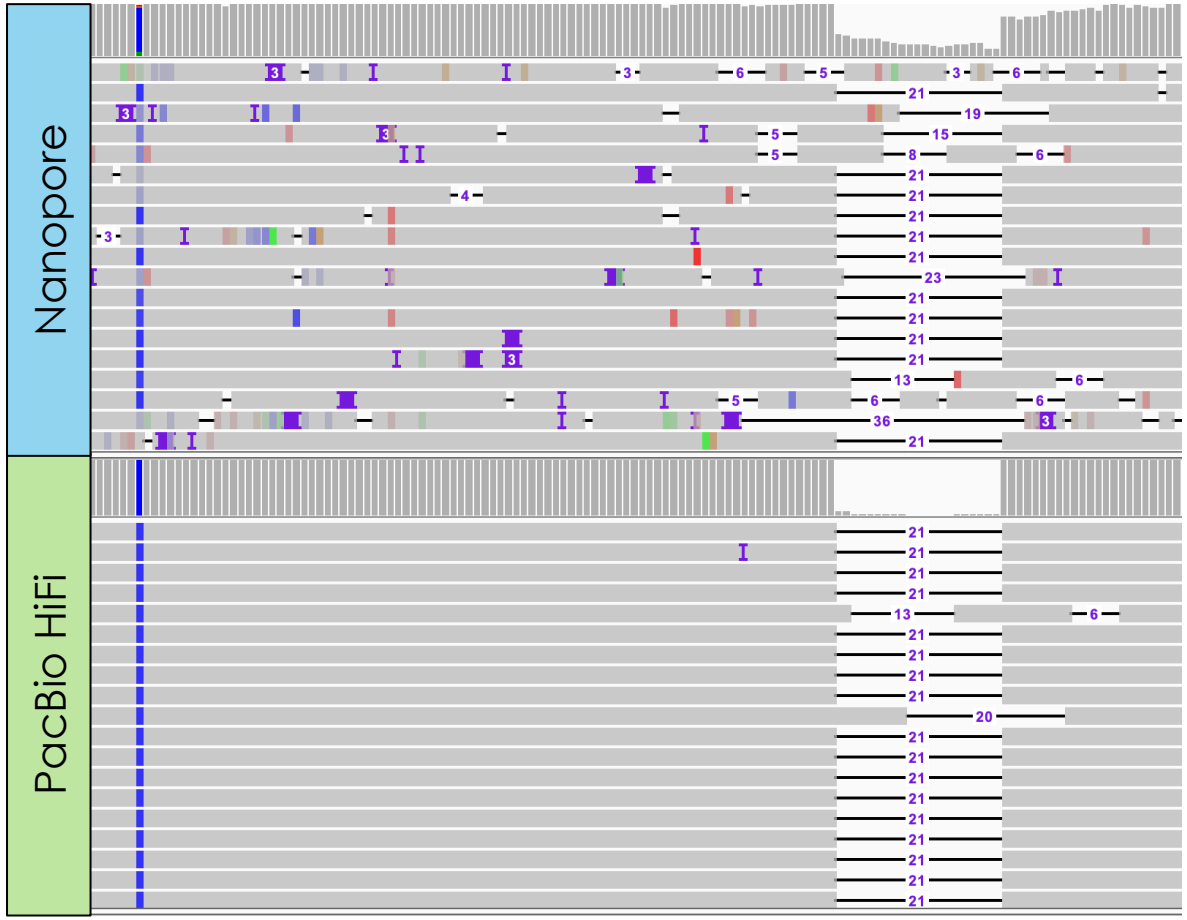


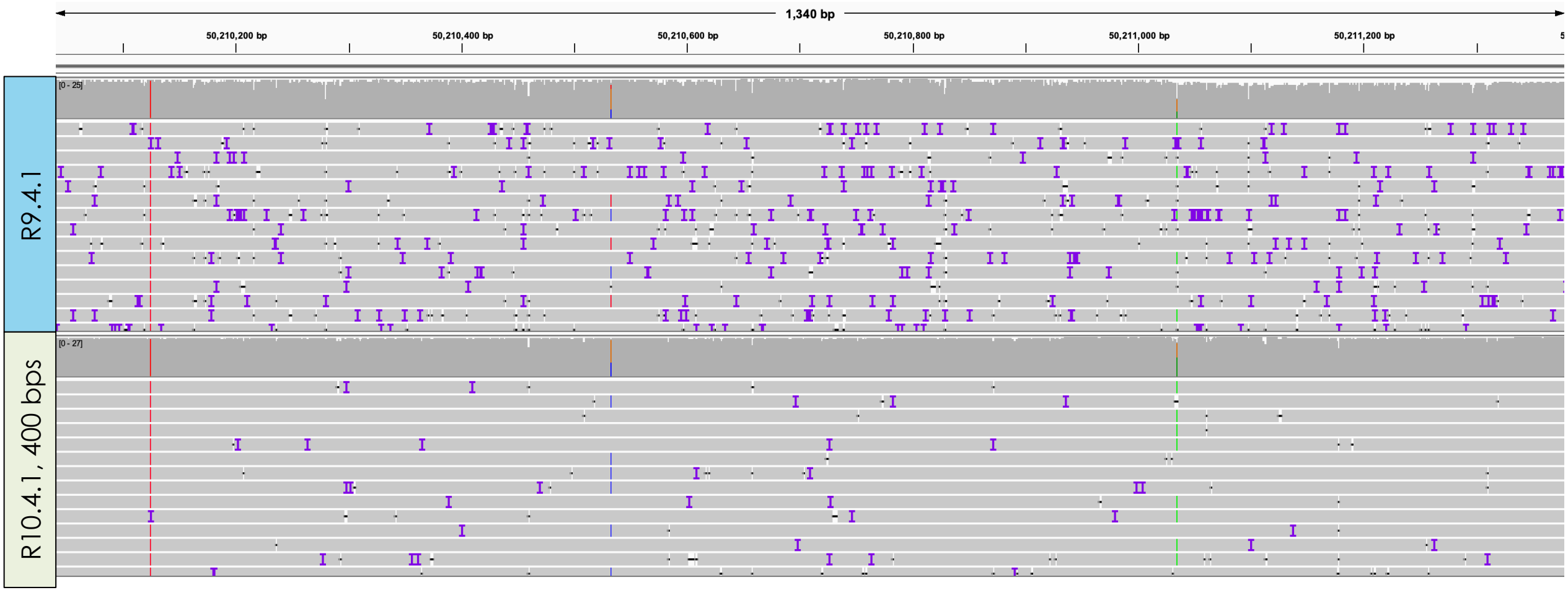
Source: Pacific Biosciences; Logsdon, Vollger, & Eichler, *Nature Reviews Genetics*, 2021



Source: Oxford Nanopore

Although Nanopore data are noisier, essential information is preserved




1,340 bp
50,210,200 bp
50,210,400 bp
50,210,600 bp
50,210,800 bp
50,211,000 bp
50,211,200 bp
[0 - 25]
R9.4.1
[0 - 27]
R10.4.1, 400 bps

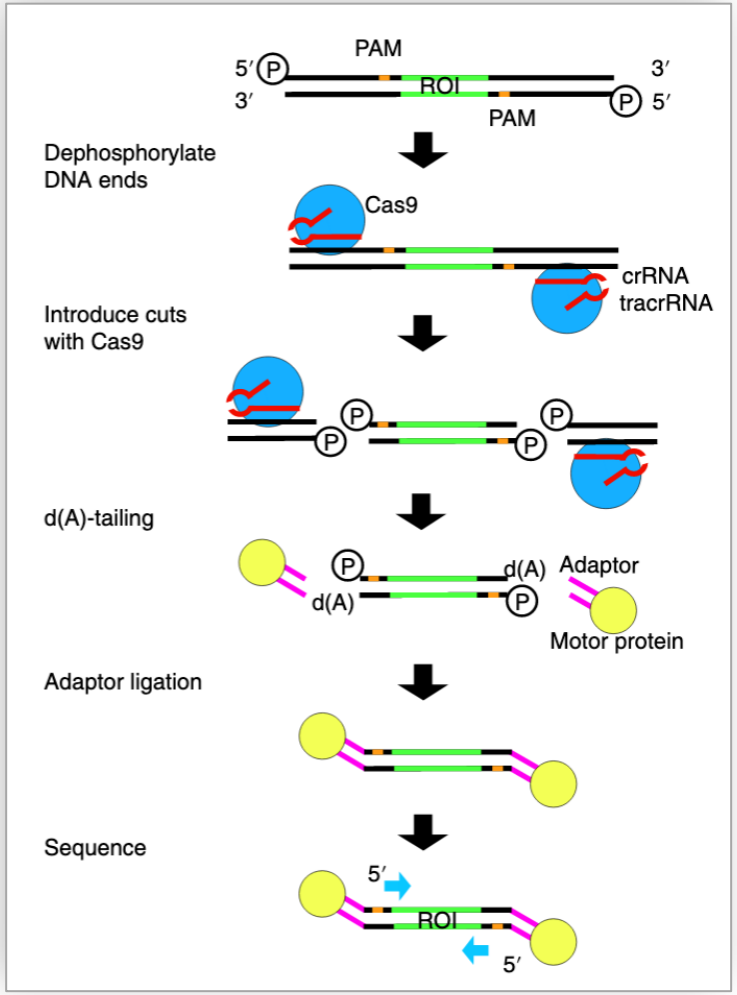
Exercise: compare ONT and PacBio data

- **HIFI.COL1A1.bam**: 25x coverage HiFi dataset. Not phased, no methylation.

- Compare this to **HG00382.R10.COL1A1.bam,** R10 flow cell, sup model, no methylation, not phased.

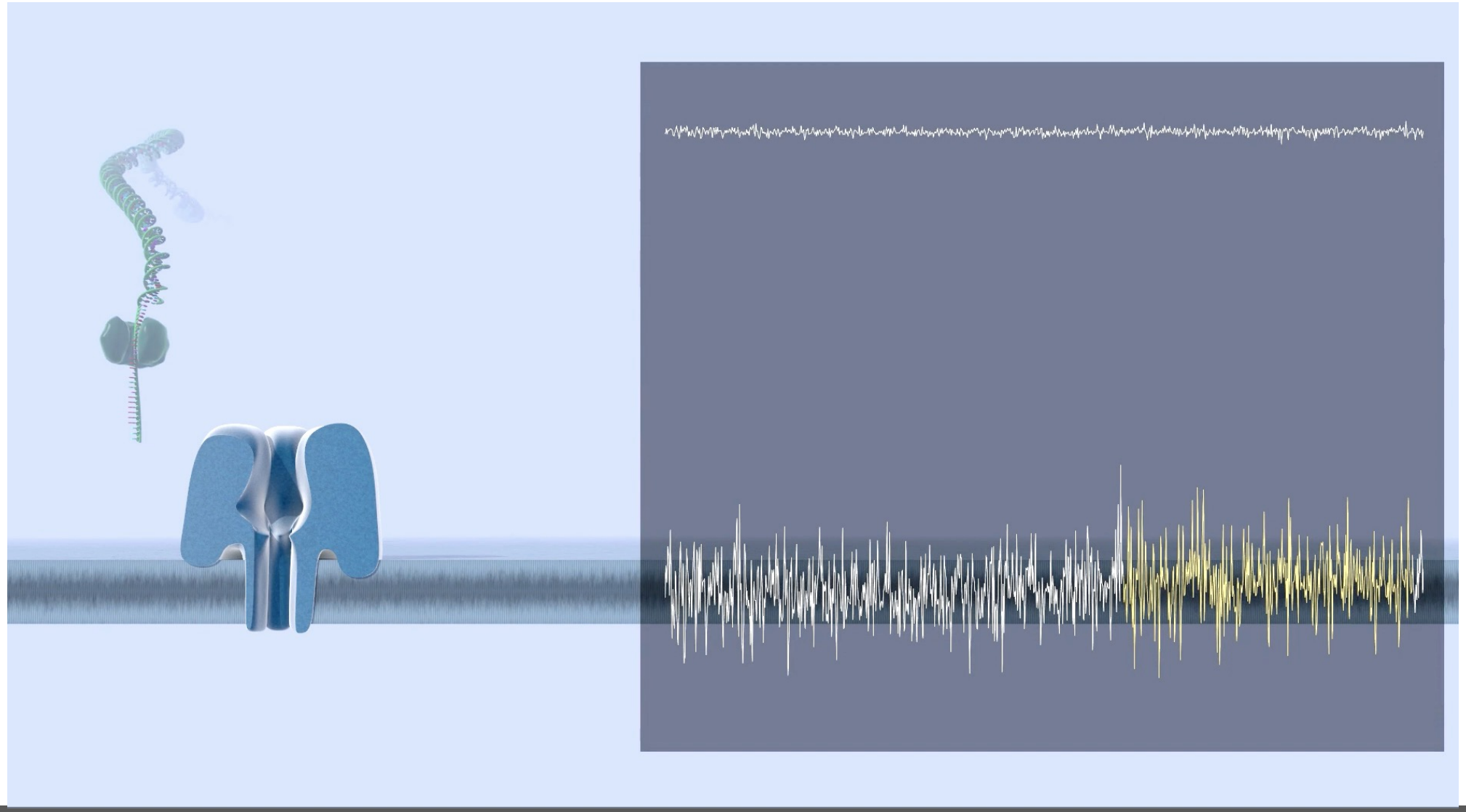
Targeting specific regions for LRS decreases costs

- **PCR amplify** target regions
  - Limited fragment sizes (<15 kb?)
  - Inconsistent success with amplification
  - Lose epigenetic information
- **Cas9**
  - Reagents expensive
  - Limited to <200 kb fragments
  - Challenging to target multiple regions
- **Adaptive sampling (ReadFish)**
  - Computational selection of regions of interest



Gilpatrick *et al.* 2019

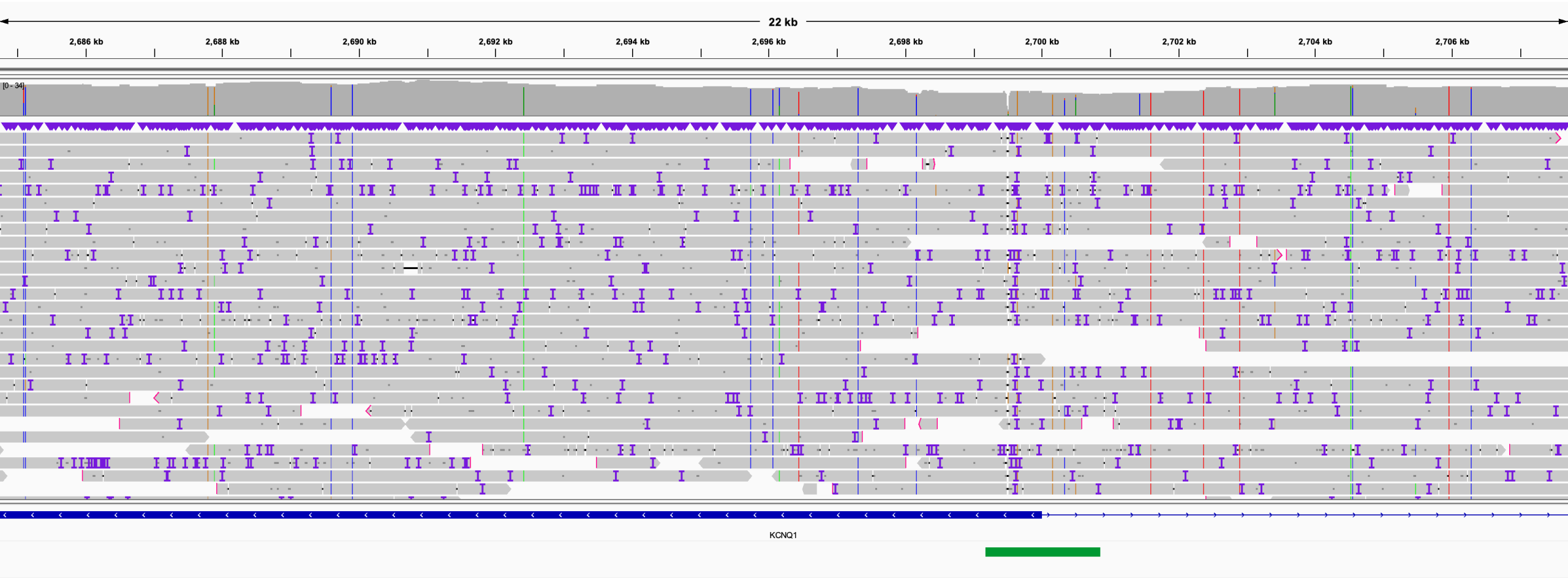
Adaptive sampling: computational targeting of specific regions of the genome for sequencing



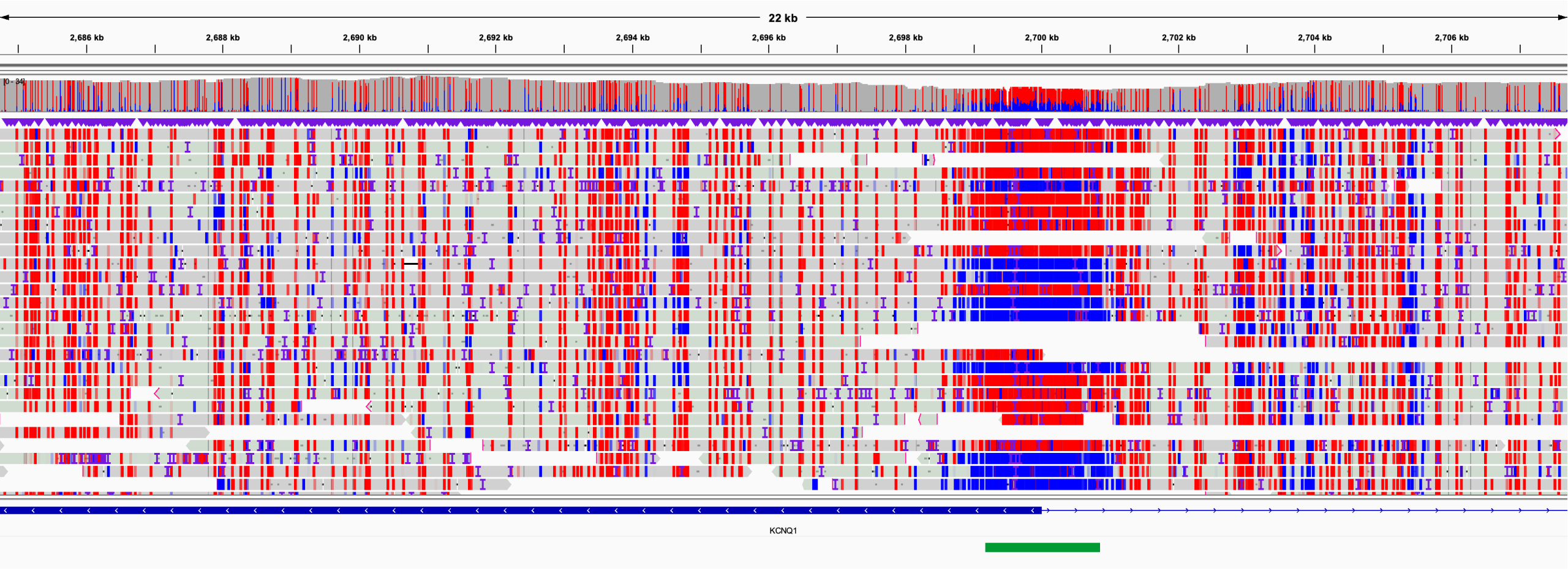
Exercise: variants, phasing, methylation

- Open **BWS.region.bam**, R9, superior model, with methylation, phased
- Go to *KCNQ1*

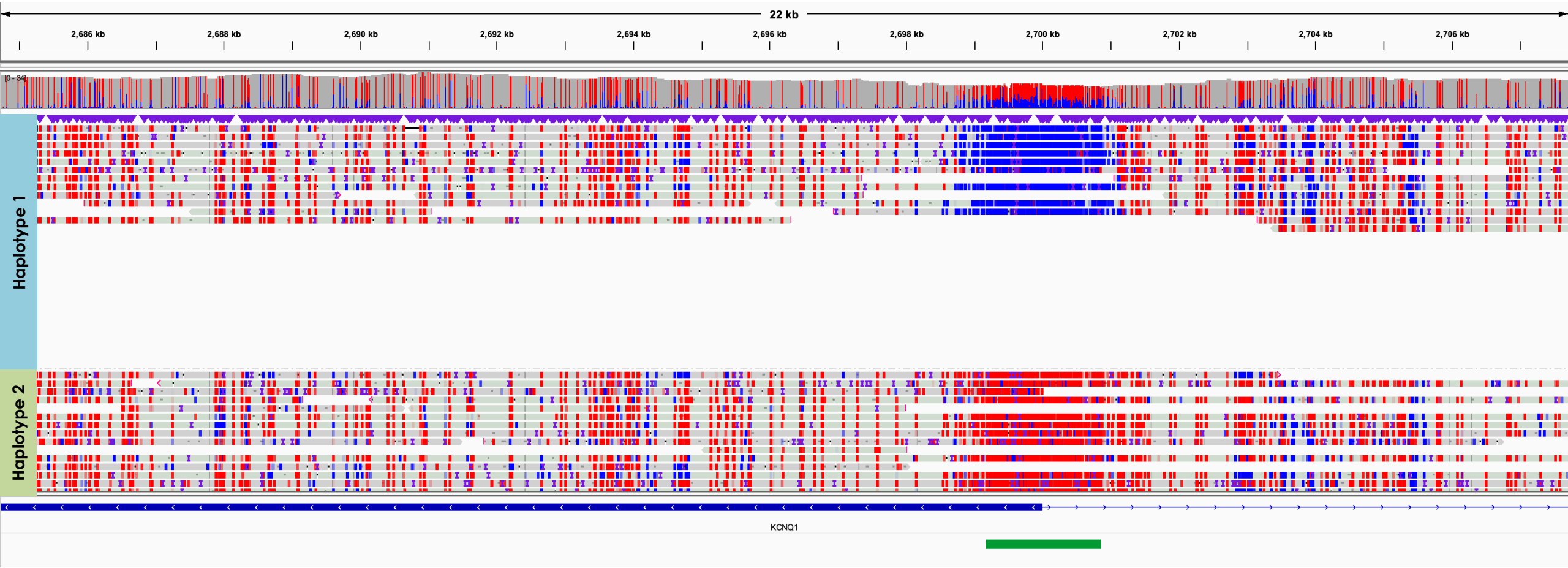
The advantage of LRS: SNVs, indels, SVs, and methylation can be called and phased from a single data source



The advantage of LRS: SNVs, indels, SVs, and methylation can be called and phased from a single data source



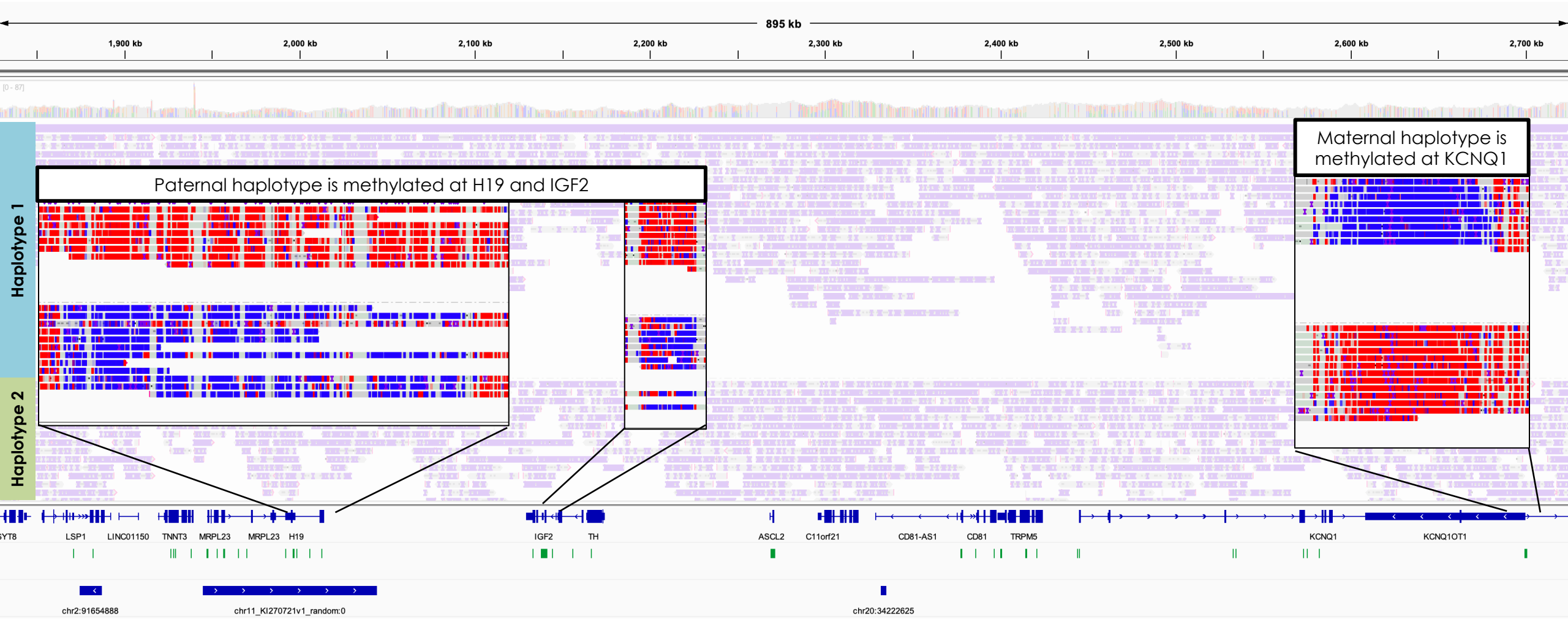
The advantage of LRS: SNVs, indels, SVs, and methylation can be called and phased from a single data source



Large segments of the genome can be phased accurately

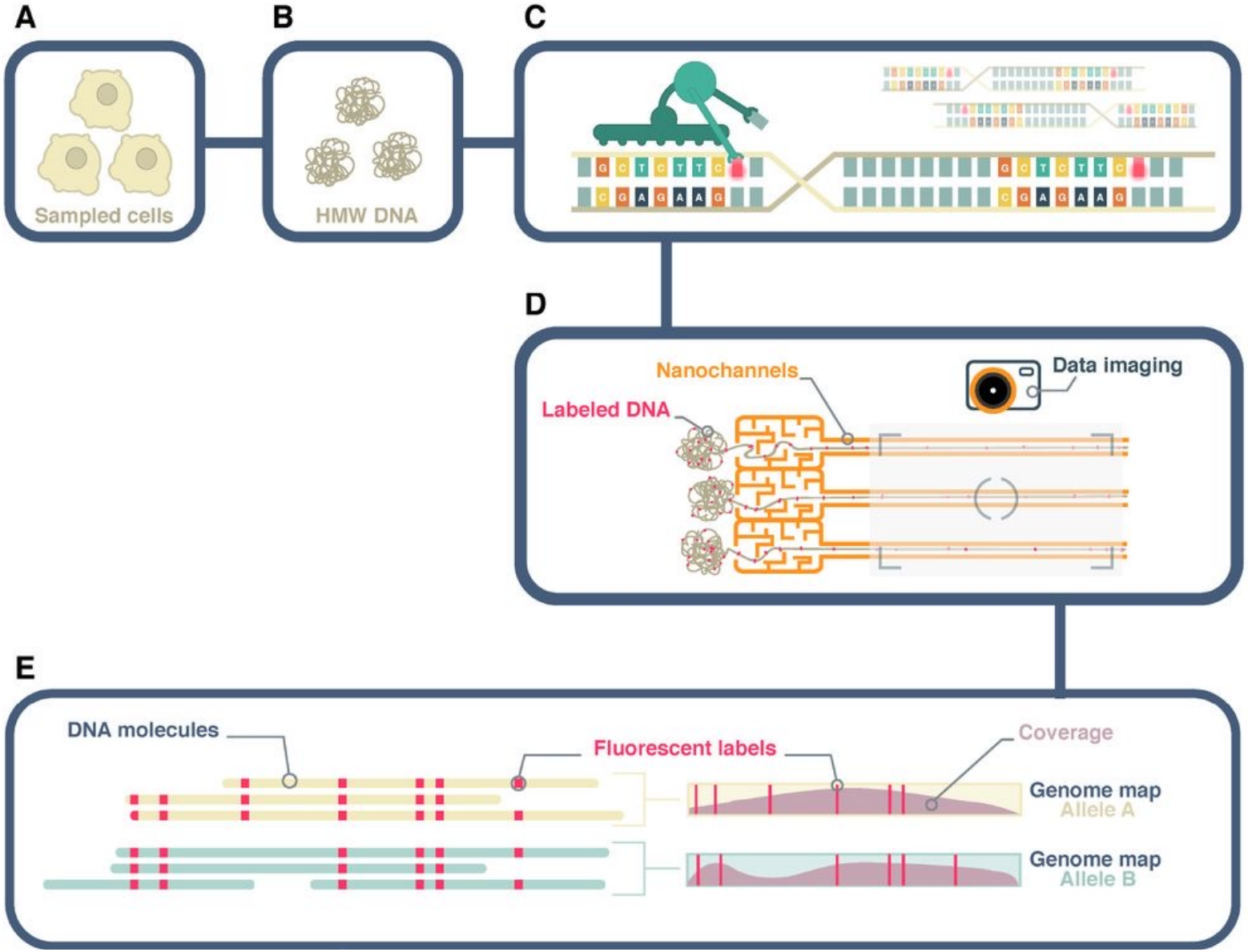
This sample had an average read length of 15kb

Large segments of the genome can be phased accurately

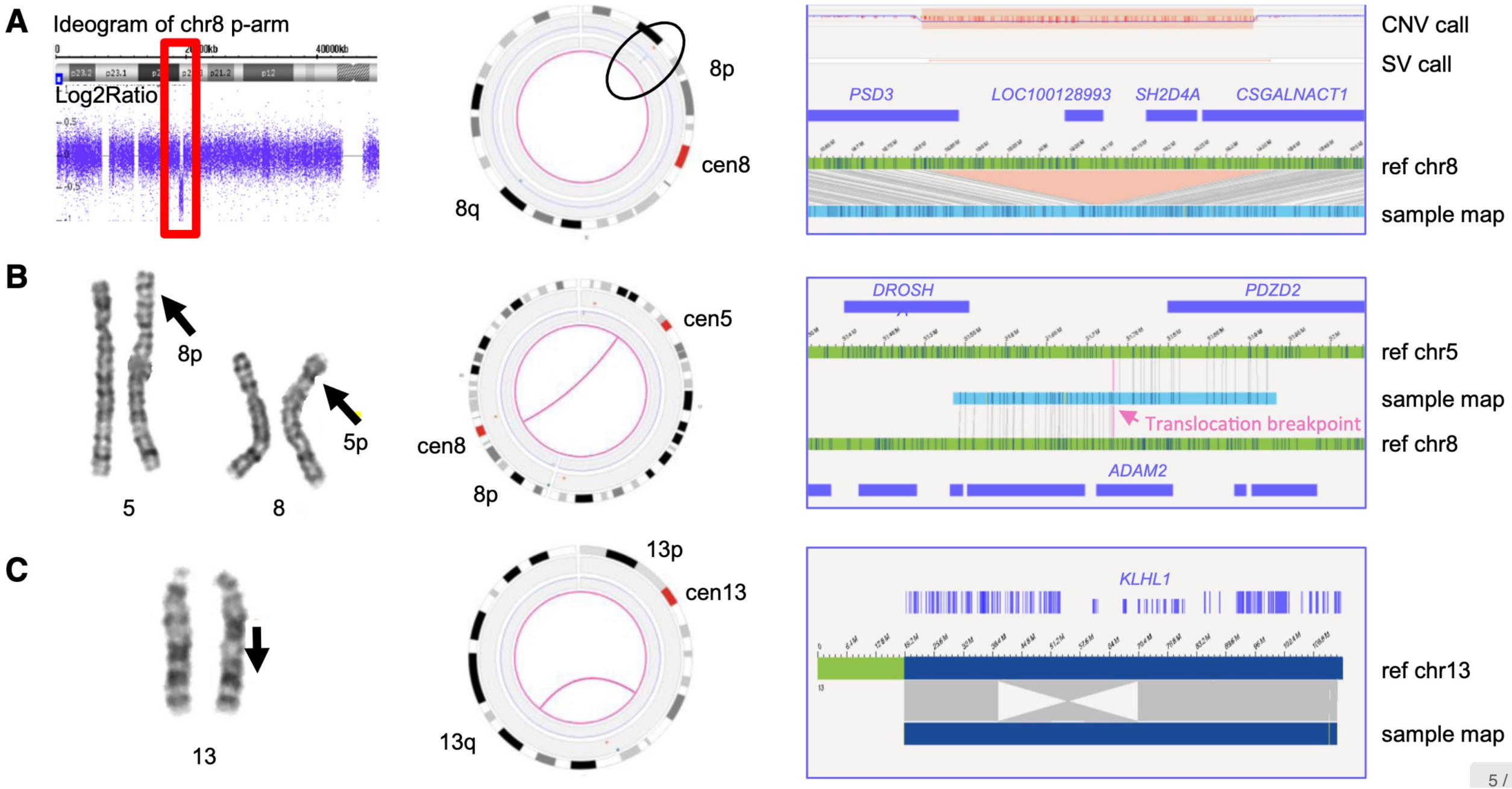


This sample had an average read length of 15kb

Large SVs can be identified by optical genome mapping



Chan *et al.*, 2018


A
Ideogram of chr8 p-arm
Log2Ratio
8p
cen8
8q
CNV call
SV call
PSD3
LOC100128993
SH2D4A
CSGALNACT1
ref chr8
sample map
B
8p
5p
5
8
cen5
cen8
8p
DROSH
PDZD2
ref chr5
sample map
Translocation breakpoint
ref chr8
ADAM2
C
13
13p
cen13
13q
KLHL1
ref chr13
sample map

How quickly can LRS be used to make a diagnosis?

LONG-READ SEQUENCING

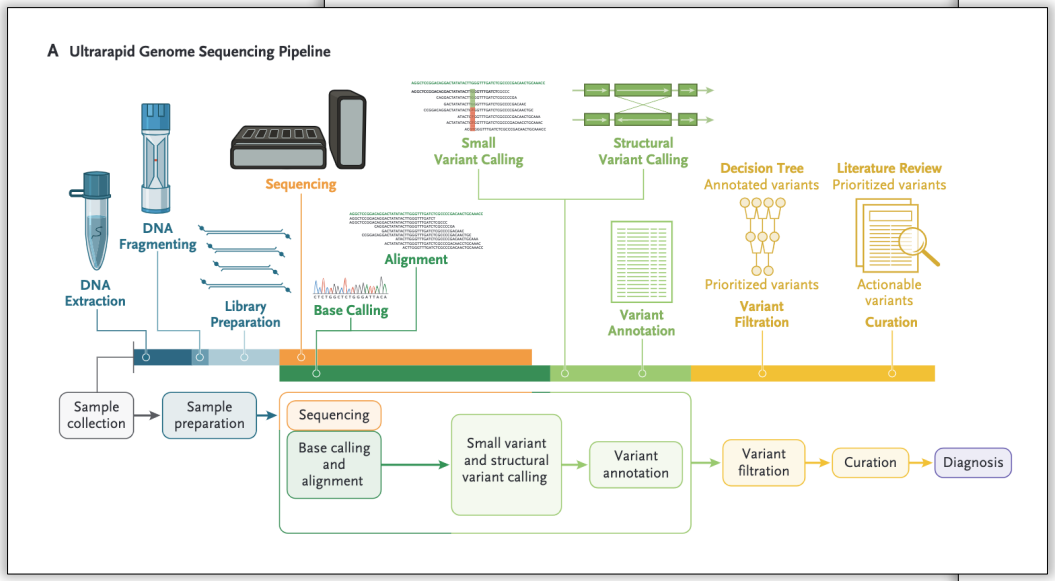
HOURS to WEEKS

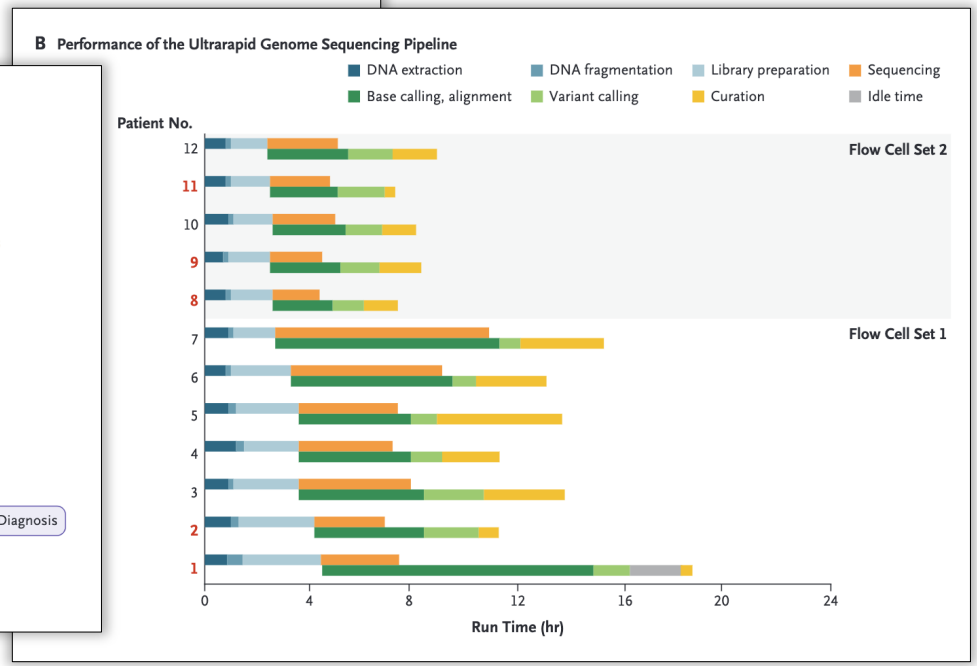
? Diagnostic rate

The NEW ENGLAND JOURNAL of MEDICINE

CORRESPONDENCE

Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting

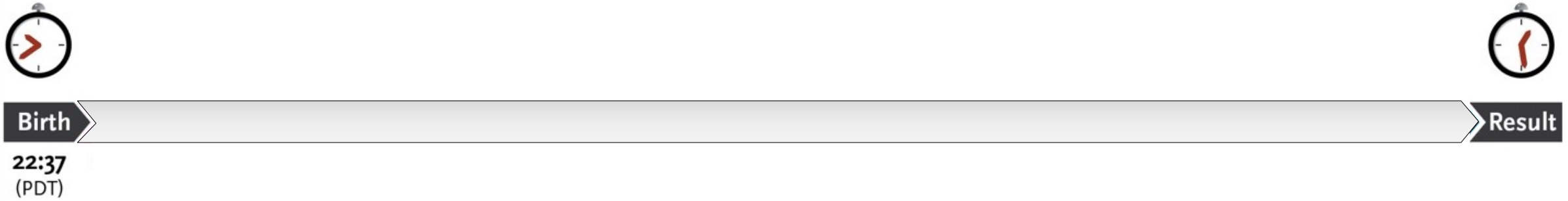




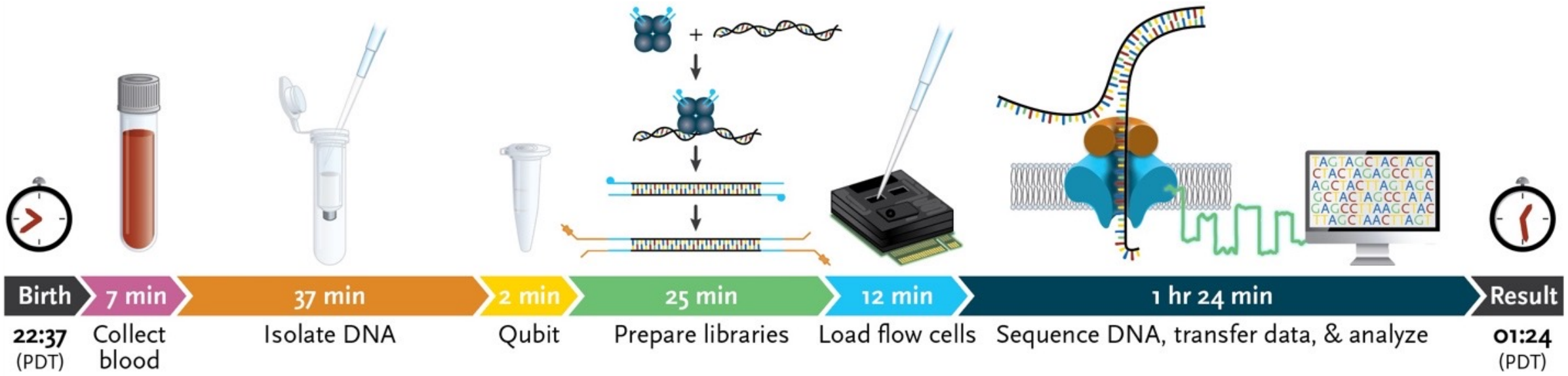
Gorzynski *et al*. 2022

Rapid assessment of genetic risk using WG-LRS

- Newborn at risk of an inherited genetic disorder
- Family wanted to know, did he inherit one or both causative variants?



Within 3 hours of birth, determined the newborn did not inherit either causative variant



2 hours 47 minutes

LONG-READ SEQUENCING

to resolve challenging clinical cases

LRS allows for comprehensive evaluation of clinical cases, including methylation

- Cases with **structural variants**
  - Multiple deletions or duplications on one or more chromosomes
  - Translocations in which the precise position of the breakpoints is unknown

- **Missing variant** cases
  - Single variant is found in a gene associated with a recessive disease
  - No variants identified for an X-linked or dominant disorder

- **Phasing of known variants**

- Cases with suspected **imprinting disorders**

LRS allows for comprehensive evaluation of clinical cases, including methylation

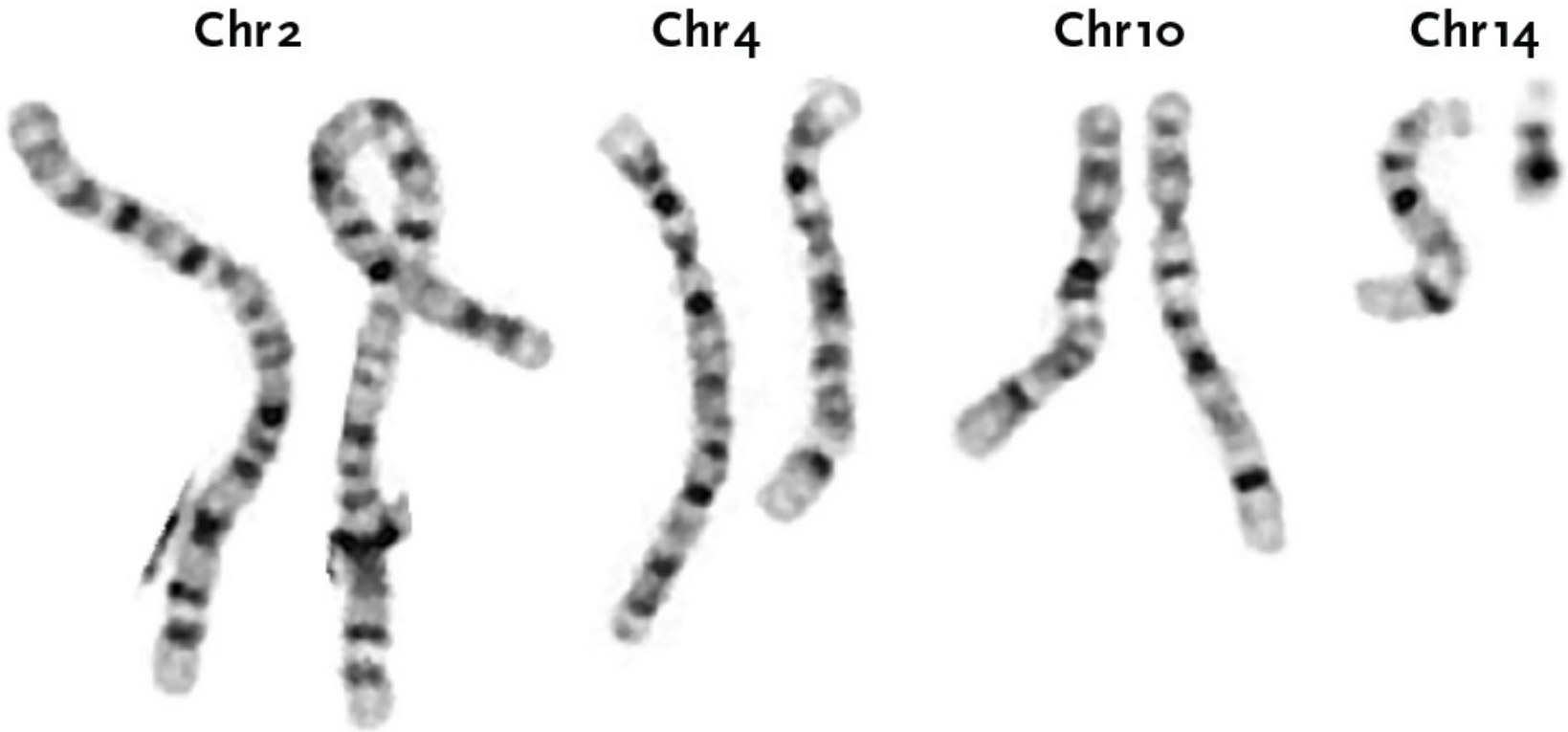
- Cases with **structural variants**
  - Multiple deletions or duplications on one or more chromosomes
  - Translocations in which the precise position of the breakpoints is unknown

- **Missing variant** cases
  - Single variant is found in a gene associated with a recessive disease
  - No variants identified for an X-linked or dominant disorder

- **Phasing of known variants**

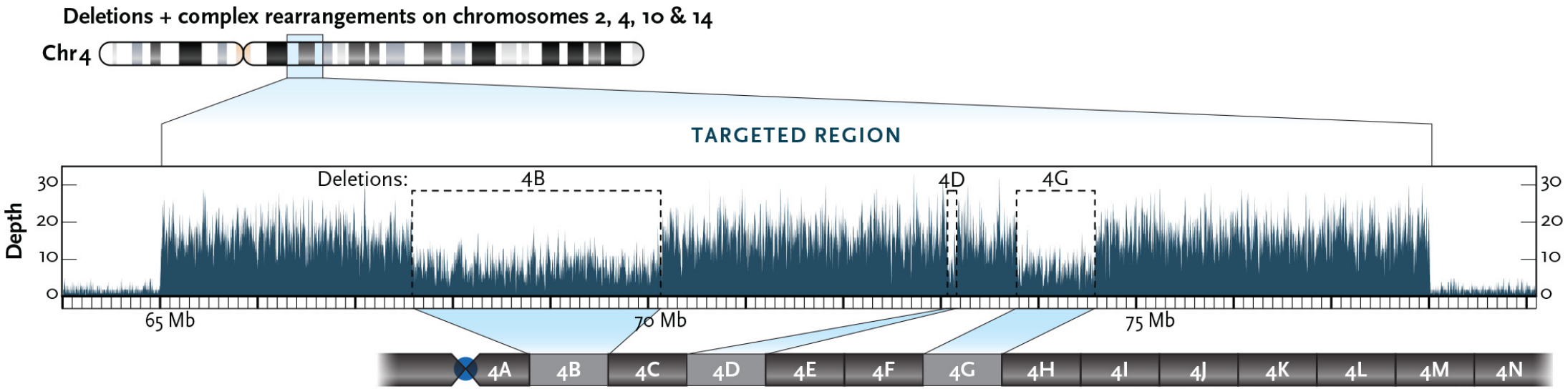
- Cases with suspected **imprinting disorders**

LRS can resolve events with multiple copy number changes and translocations

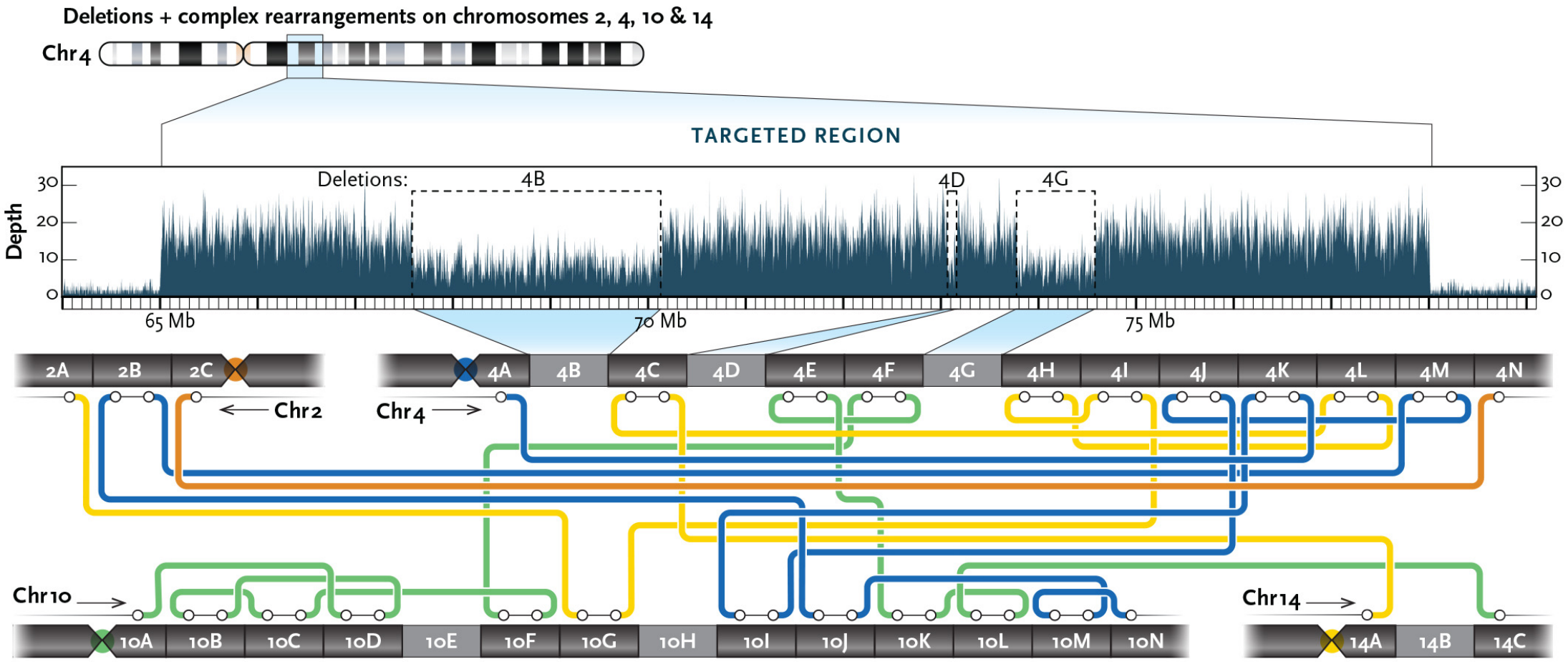


Miller *et al.* 2021, *AJHG*

LRS can resolve events with multiple copy number changes and translocations



LRS identified multiple rearrangements between and within the four chromosomes



Exercise: complex SV

- Open **F8.multipleInversions.bam**: T-LRS, R9, Guppy 5.0.7, sup model (but older), no methylation, no phasing

- This male has Hemophilia A, why?

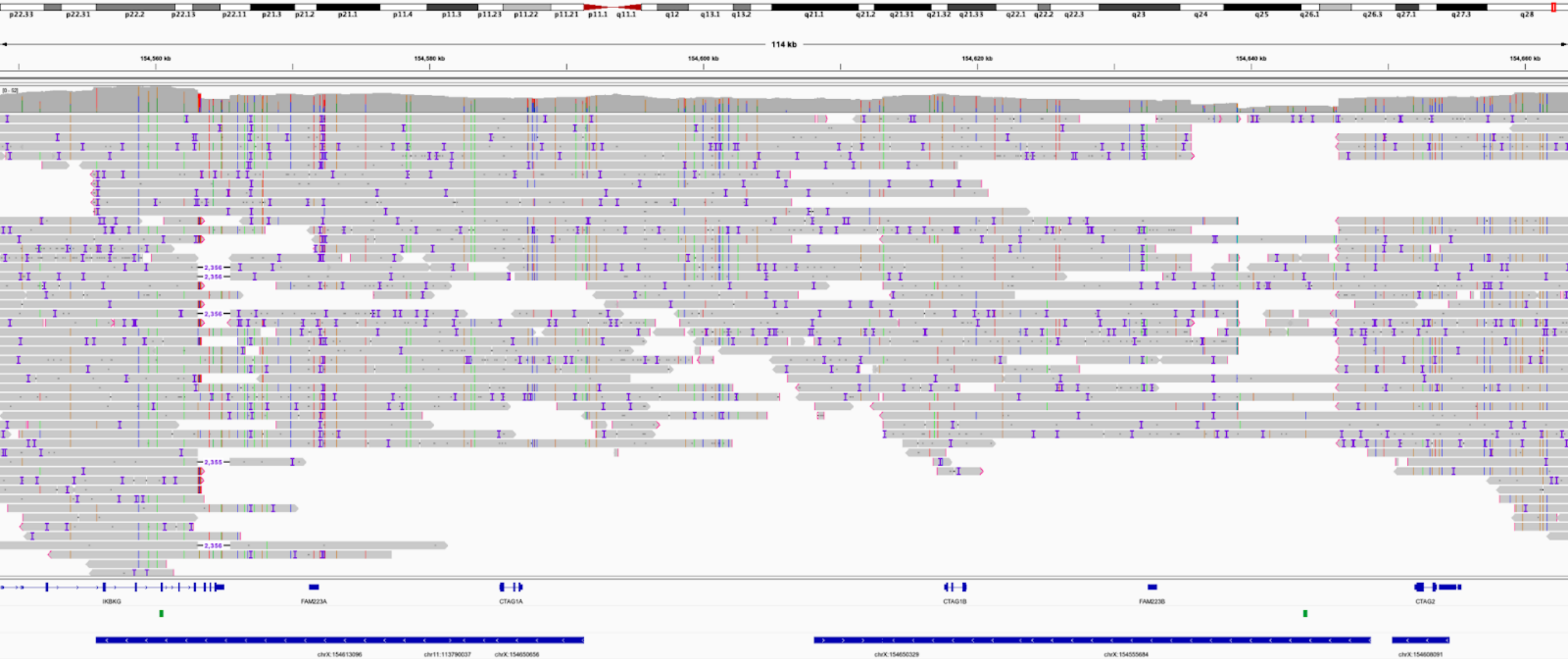
Exercise: clarifying a simple duplication

- Open **CSH.AdvSeq.Examples.bam**
  - Look at *CTNND2*, do you see the aberration?
  - What is the structure of the event?

Exercise: difficulty calling a repeat expansion

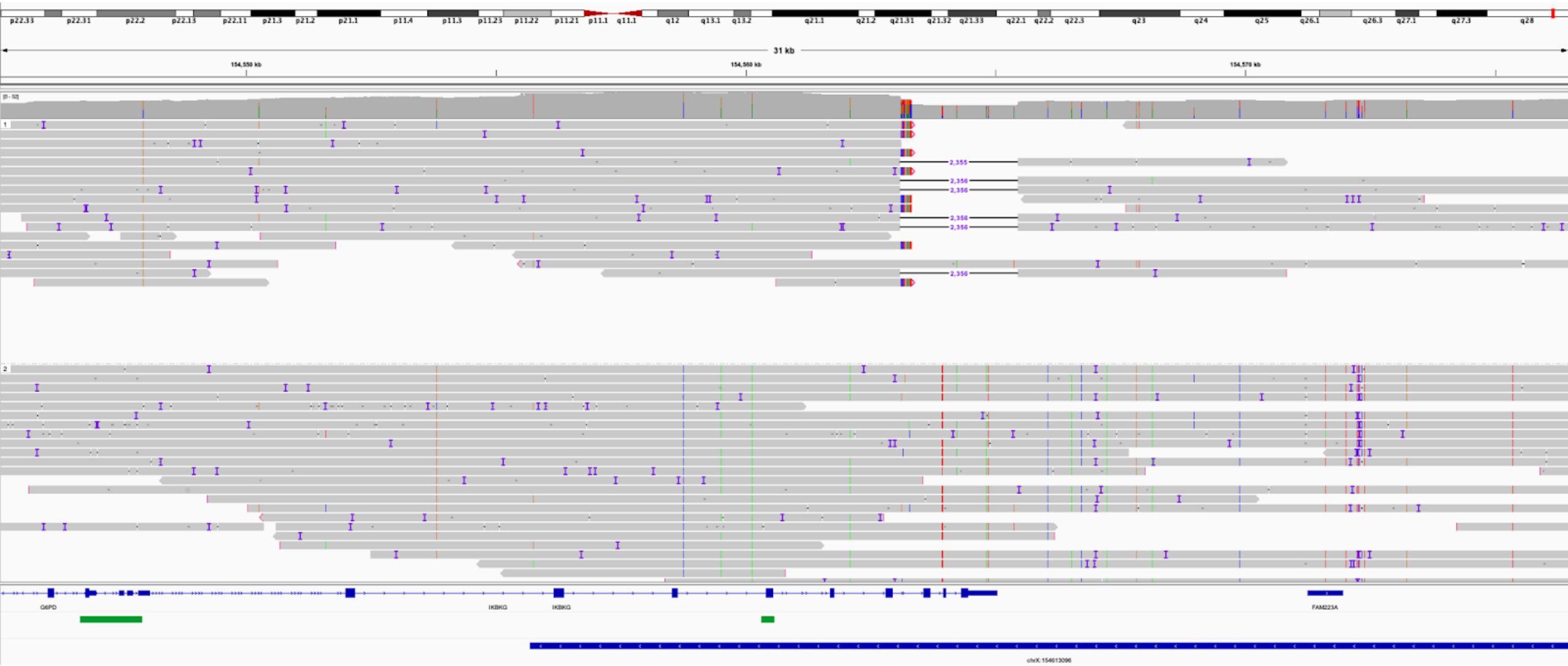
- Open **CSH.AdvSeq.Examples.bam**
  - Look at *FXN*, the pathogenic expansion is near chr9:69,037,300.

Whole-genome LRS can be used to resolve SVs in complex regions



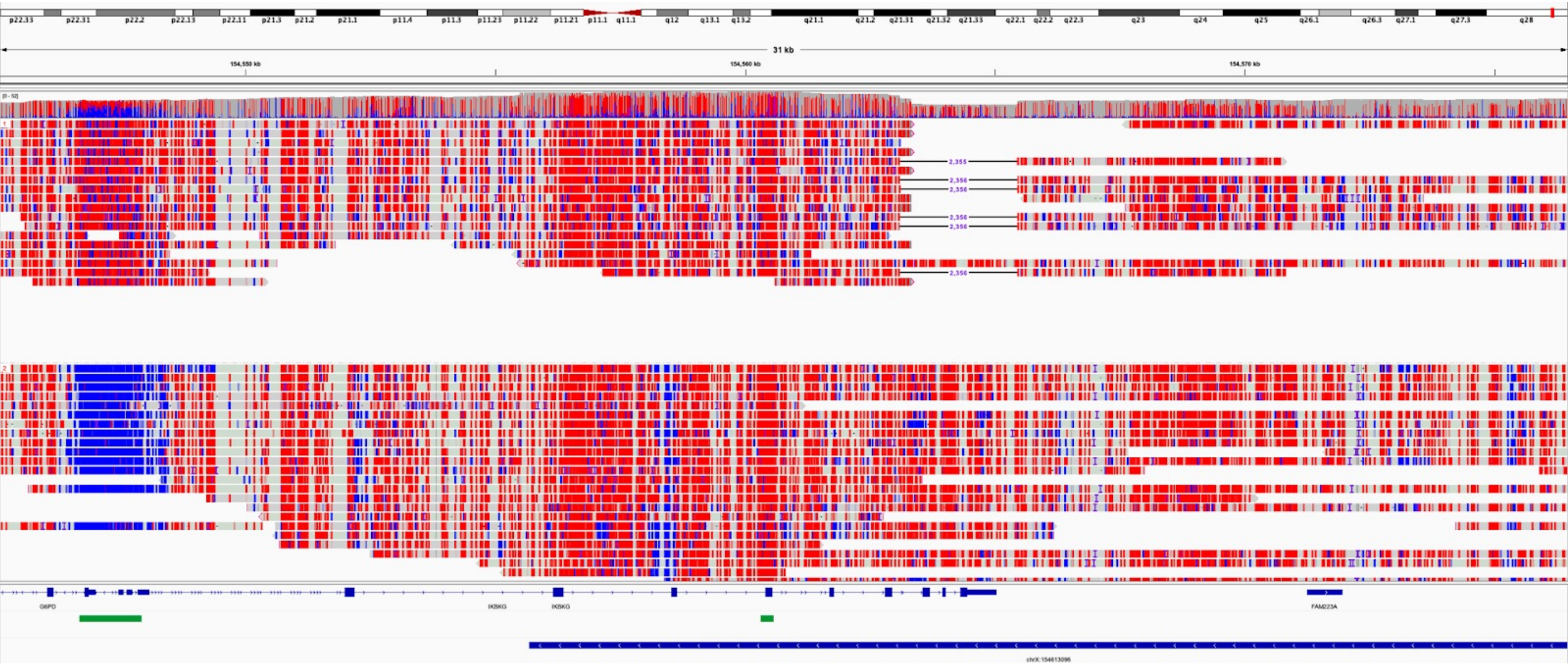
Danny Miller, Monica Wojick; unpublished

Whole-genome LRS can be used to resolve SVs in complex regions



Danny Miller, Monica Wojick; unpublished

Skewed X-inactivation supports pathogenicity of the SV



Danny Miller, Monica Wojick; unpublished

Exercise: challenging genomic region

- Open **CSH.AdvSeq.Examples.bam**
  - Look at *OPN1LW*, what's so difficult about this region?

LRS allows for comprehensive evaluation of clinical cases, including methylation

- Cases with **structural variants**
  - Multiple deletions or duplications on one or more chromosomes
  - Translocations in which the precise position of the breakpoints is unknown

- **Missing variant** cases
  - Single variant is found in a gene associated with a recessive disease
  - No variants identified for an X-linked or dominant disorder

- **Phasing of known variants**

- Cases with suspected **imprinting disorders**

LRS allows for comprehensive evaluation of clinical cases, including methylation

- Cases with **structural variants**
  - Multiple deletions or duplications on one or more chromosomes
  - Translocations in which the precise position of the breakpoints is unknown

- **Missing variant** cases
  - Single variant is
  - No variants iden

- **Phasing of know**

- Cases with suspe

Easiest way to tell us what we are missing with traditional clinical testing

- Variants in repetitive regions, variants that are difficult to detect, or variants that are difficult to interpret

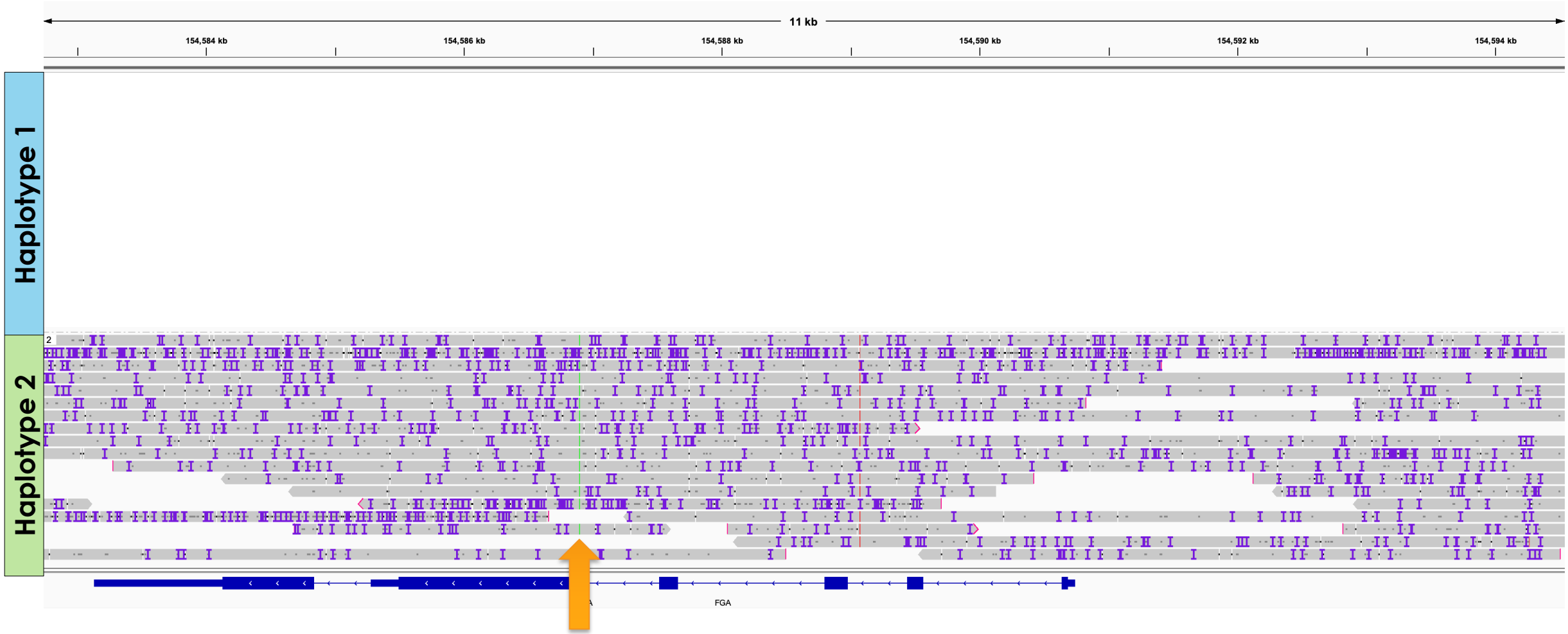
This leads to

- Improved analysis algorithms
- Unique mechanisms of disease
- Identification of variation in complex regions

Exercise: missing variant case

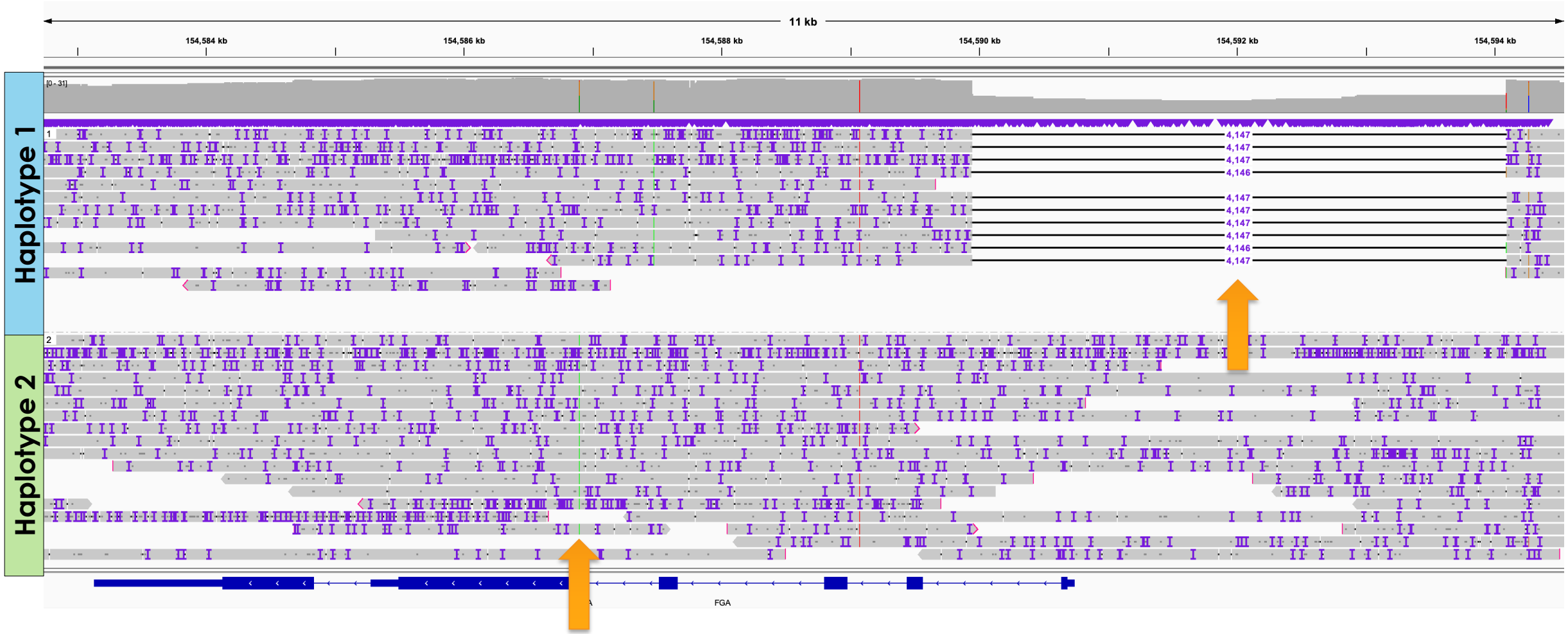
- Open **FGA.ONT.phased.meth.bam**. R9 pore, sup model, phased data, with methylation calls.
  - Individual has hypofibrinogenemia and one inherited pathogenic variant in *FGA*.
  - What's the second variant?

LRS can identify variants missed by prior clinical testing



Known maternally inherited stop in *FGA* (fibrinogen alpha chain)

LRS can identify variants missed by prior clinical testing



Known maternally inherited stop in *FGA* (fibrinogen alpha chain)

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

Acrodermatitis enteropathica with a single known pathogenic variant

- Toddler w/severe eczema, allergies
- Family suspected acrodermatitis enteropathica
  - Congenital form is caused by malabsorption of Zinc in the intestine
  - He improved on Zinc supplementation
- Clinical testing revealed a single nonsense variant in *SLC39A4*

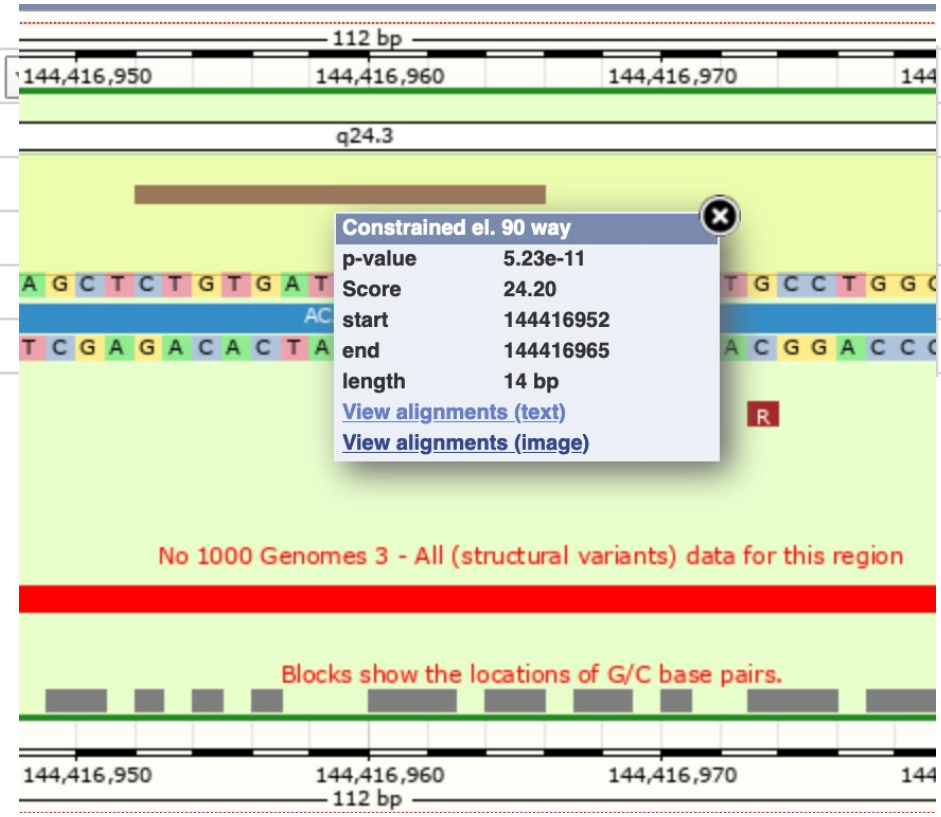
Unpublished data, photo shown with permission of family

LRS identified the known stop and three candidate variants on the other haplotype

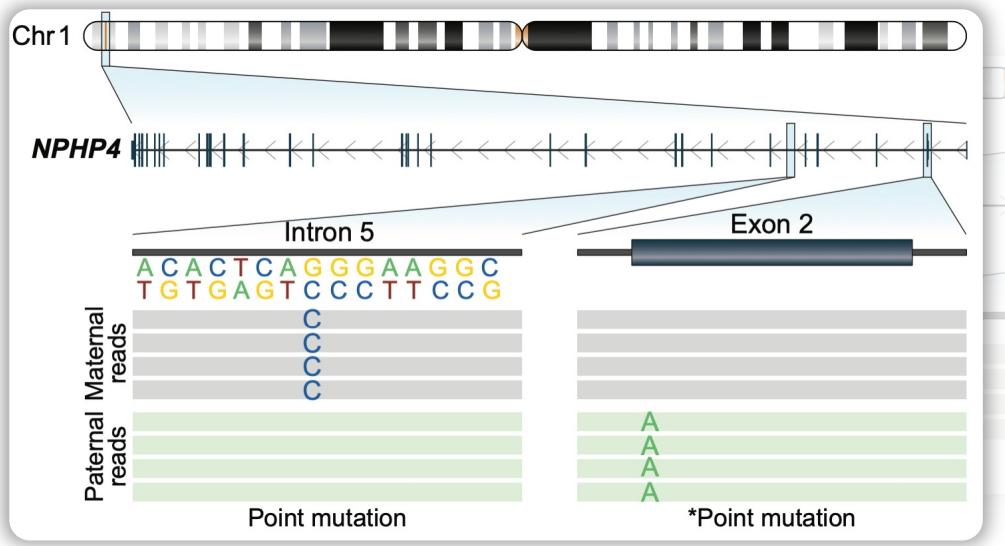
CHROM	POS	REF	ALT	QUAL	TYPE	ZYGOUS	HP	Consequence	IMPACT	SYMBOL	CADD_PHRE
chr8	144414042	C	T	17.32	SNV	'0/1	1|0	stop_gained	HIGH	SLC39A4	38
chr8	144413722	T	C	16.78	SNV	'0/1	0|1	intron_variant	MODIFIER	SLC39A4	7.361
chr8	144414753	G	A	16.81	SNV	'0/1	0|1	synonymous_variant	LOW	SLC39A4	0.405
chr8	144416958	T	C	21.04	SNV	'0/1	0|1	upstream_gene_variant	MODIFIER	SLC39A4	18.22

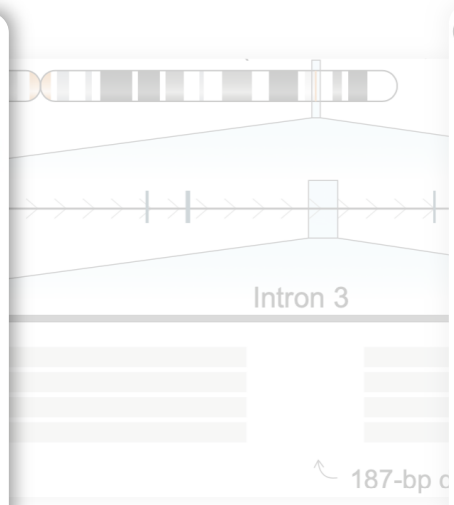
The second variant sits within a highly conserved transcription factor binding site only active in GI cells

CHROM	POS	REF	ALT	QUAL	TYPE	ZYGOUS	HP	CADD_PHRE
chr8	144414042	C	T	17.32	SNV	'0/1	1|0	38
chr8	144413722	T	C	16.78	SNV	'0/1	0|1	7.361
chr8	144414753	G	A	16.81	SNV	'0/1	0|1	0.405
chr8	144416958	T	C	21.04	SNV	'0/1	0|1	18.22



LRS identifies variants not detected by standard clinical testing



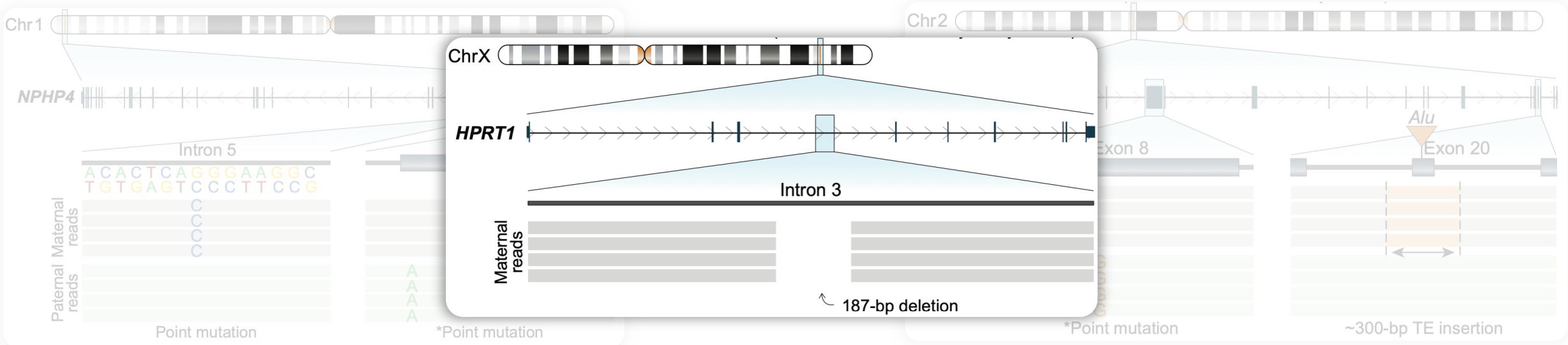


Intronic variant
Disrupts splicing

Inversion
Bisects the gene

Transposable element insertion
Disrupts gene function

LRS identifies variants not detected by standard clinical testing

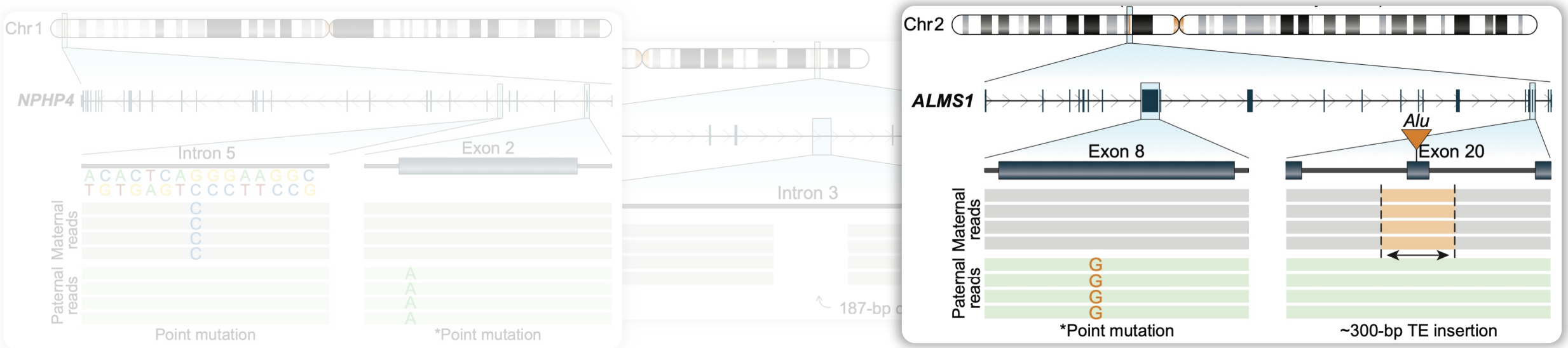


Intronic variant
Disrupts splicing

Inversion
Bisects the gene

Transposable element insertion
Disrupts gene function

LRS identifies variants not detected by standard clinical testing



Intronic variant
Disrupts splicing

Inversion
Bisects the gene

Transposable element insertion
Disrupts gene function

Our experience with missing variants

- Two thirds of missing variant cases could be solved with LRS

	Submitted	Solved	Not Solved
Autosomal recessive	38	22 (58%)	16
Autosomal dominant	2	2	--
X-linked	8	7	1*
Total	48	31 (65%)	17

- In the unsolved cases we can focus on:
  - Rare intronic variants (evaluate with RNA-seq)
  - Variants in regulatory regions
  - Variants in other genes that may modify the phenotype

*Not solved by LRS, solved by another approach; summary data is unpublished

Our experience with missing variants

- Two thirds of missing variant cases could be solved with LRS

	Submitted	Solved	Not Solved
Autosomal recessive	38	22 (58%)	16
Autosomal dominant	2	2	--
X-linked	8	7	1*
Total	48	31 (65%)	17

- In the unsolved cases we can focus on:
  - Rare intronic variants (evaluate with RNA-seq)
  - Variants in regulatory regions
  - Variants in other genes that may modify the phenotype

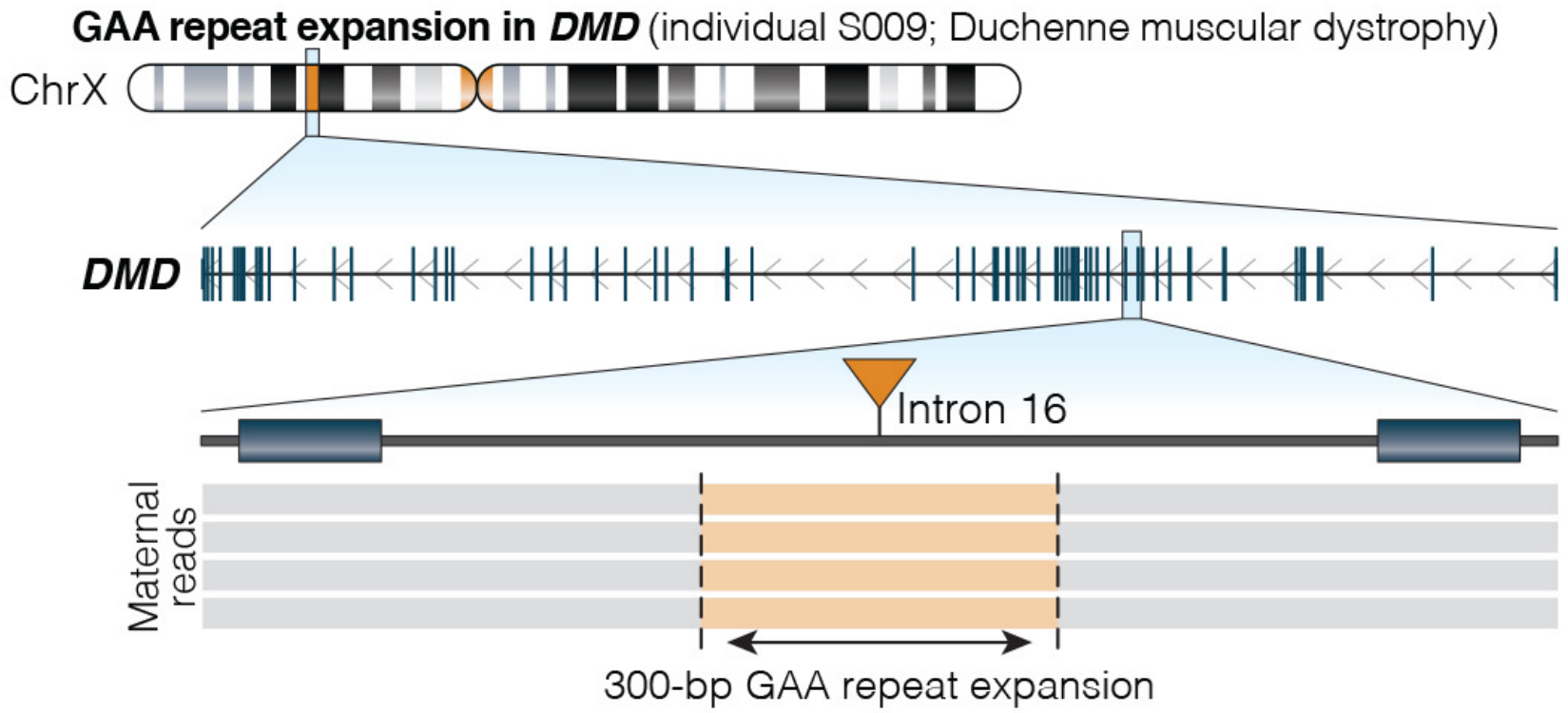
- In any clinical case with a compelling variant or phenotype **LRS is probably the next best test**
  - Please send us your missing variant cases!

*Not solved by LRS, solved by another approach; summary data is unpublished

ONT has difficulty with homopolymers

How can this impact interpretation?

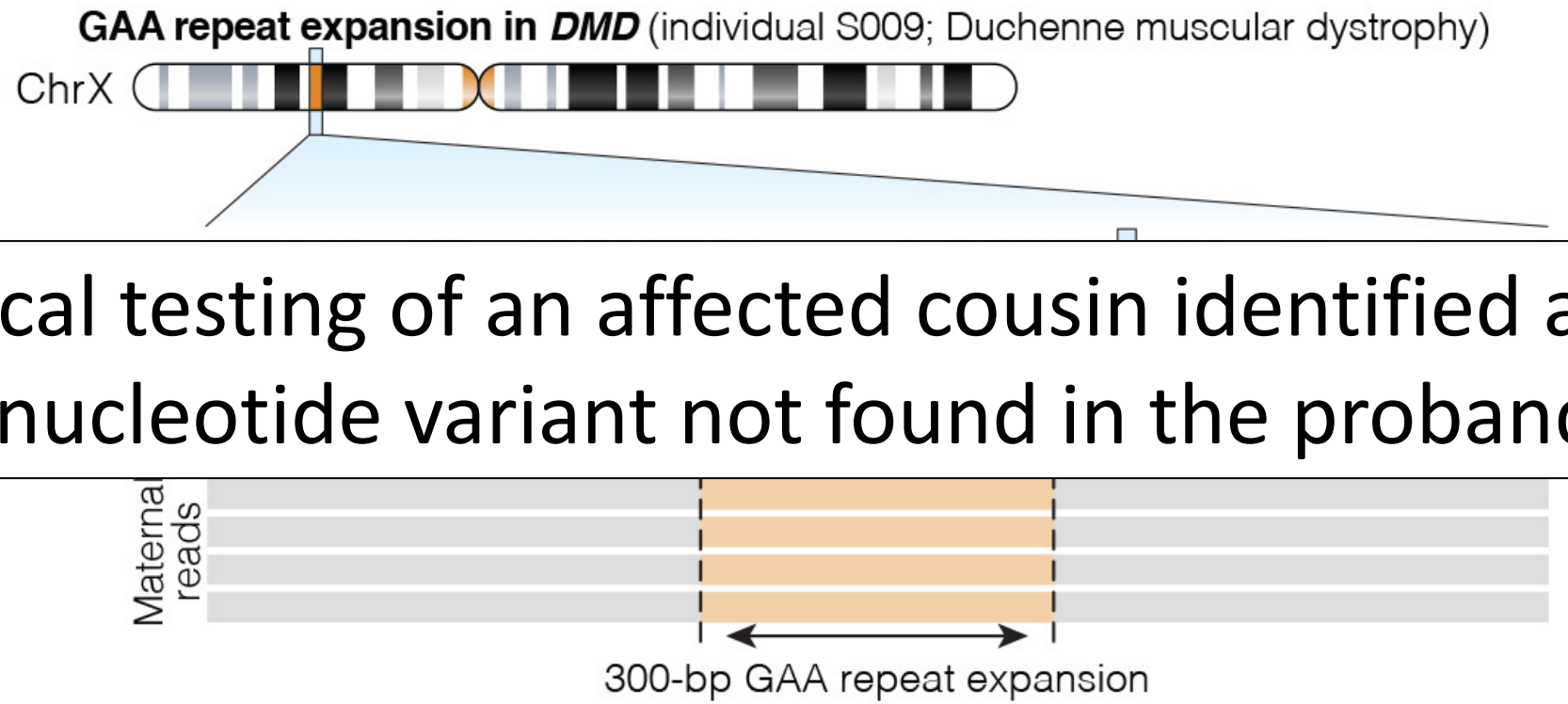
Patient with a family history of Duchenne muscular dystrophy, but no molecular diagnosis



Miller *et al.*, 2021 AJHG

Array: negative | **Sequencing of exons**: negative | No protein seen on **Western** from muscle biopsy

Patient with a family history of Duchenne muscular dystrophy, but no molecular diagnosis

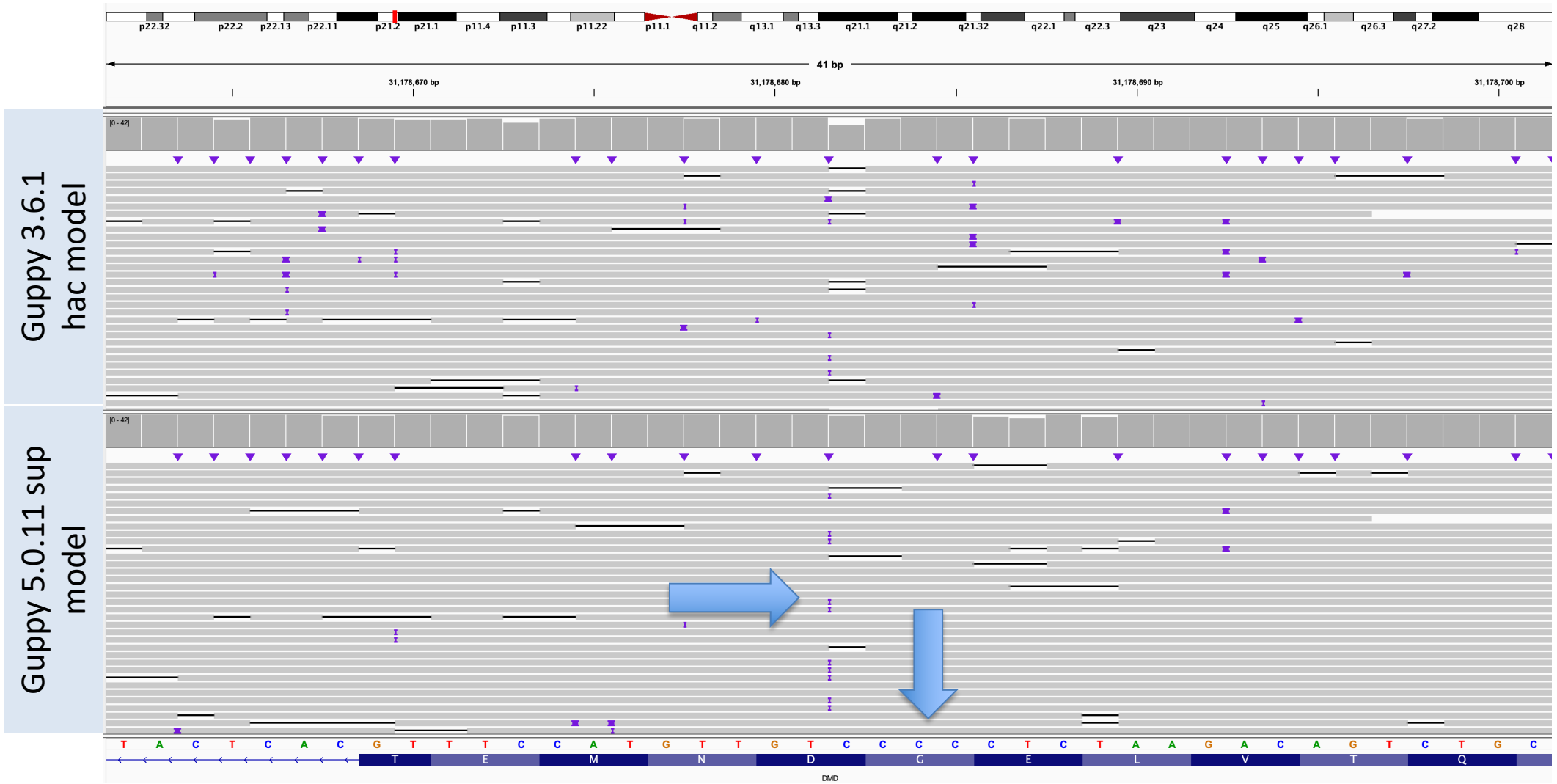


Clinical testing of an affected cousin identified a single nucleotide variant not found in the proband…

Miller *et al.*, 2021 AJHG

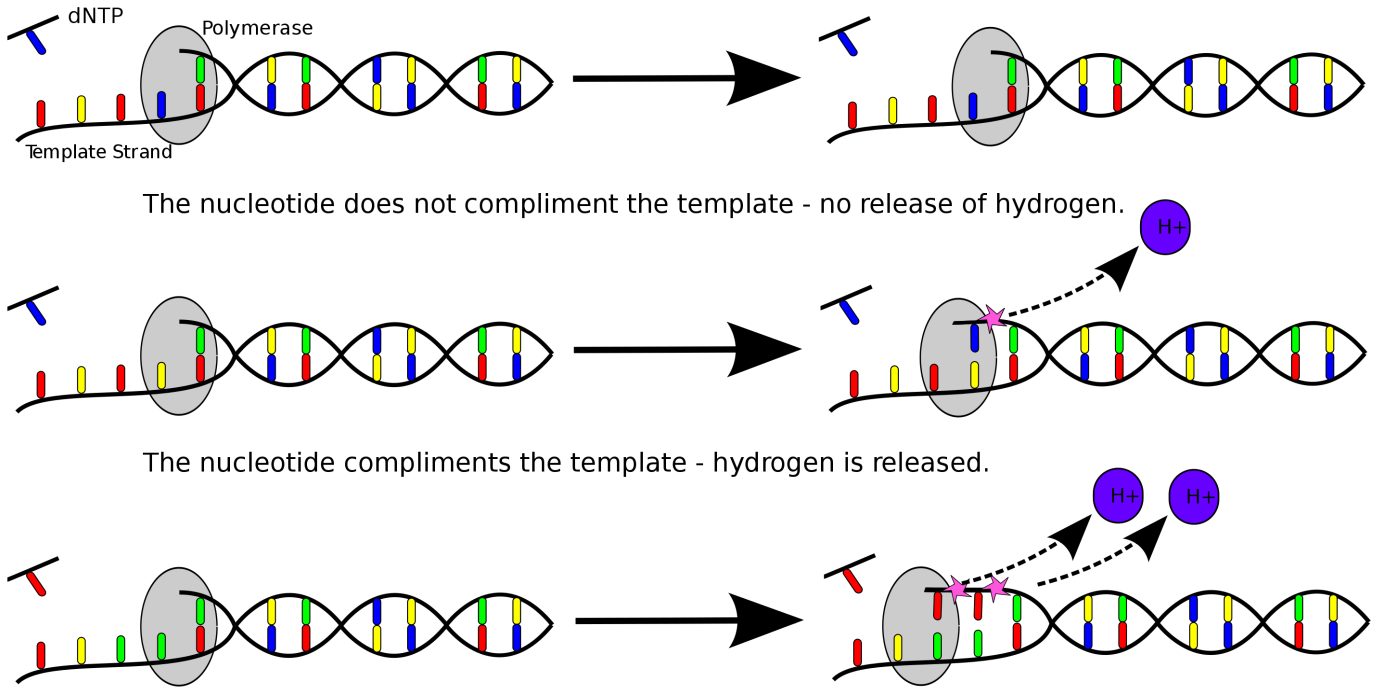
Array: negative | **Sequencing of exons**: negative | No protein seen on **Western** from muscle biopsy

Clinical testing of a relative identified a pathogenic single nucleotide insertion not identified by clinical or research testing



Why did we (and the original clinical lab) not identify a single nucleotide insertion in a homopolymer?

Genomic deoxyribonucleic acid (gDNA) was isolated from the patient's specimen via a standardized DNA isolation kit. All of the coding sequences and the flanking splice junctions of 10 genes (see table below) in the Neuromuscular Disorders (NMD) Panel were sequenced simultaneously by massive parallel (Next-Generation Sequencing) via the Ion Torrent instrument. The DNA sequence was aligned

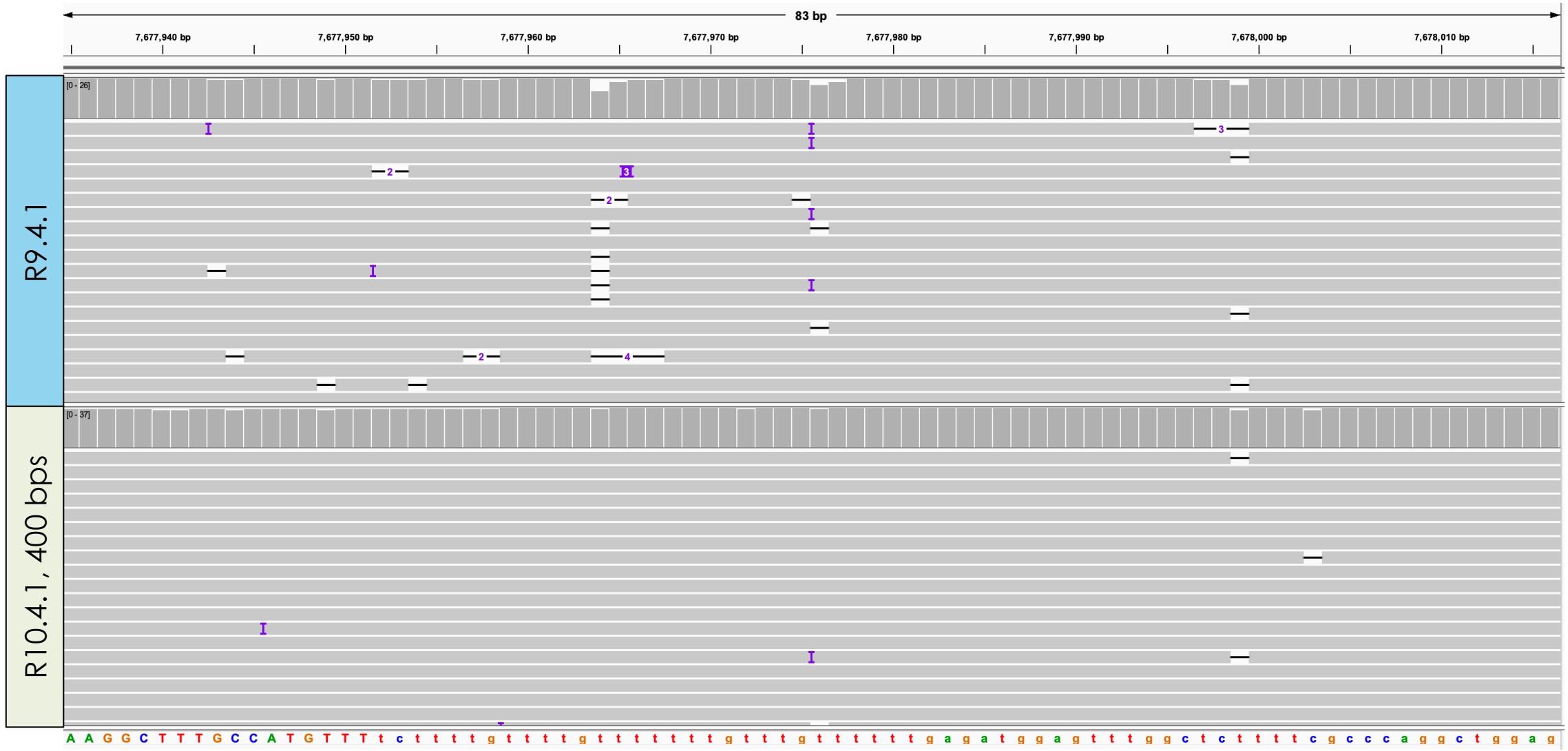


The nucleotide does not compliment the template - no release of hydrogen.

The nucleotide compliments the template - hydrogen is released.

The nucleotide compliments several bases in a row - multiple hydrogen ions are released.

Improved homopolymer calling with R10.4.1



Base called using Guppy 6.3.2; superior model

LRS to detect differences in methylation

LRS allows for comprehensive evaluation of clinical cases, including methylation

- Cases with **structural variants**
  - Multiple deletions or duplications on one or more chromosomes
  - Translocations in which the precise position of the breakpoints is unknown

- **Missing variant** cases
  - Single variant is found in a gene associated with a recessive disease
  - No variants identified for an X-linked or dominant disorder

- **Phasing of known variants**

- Cases with suspected **imprinting disorders**

Genome-wide detection of differences in methylation

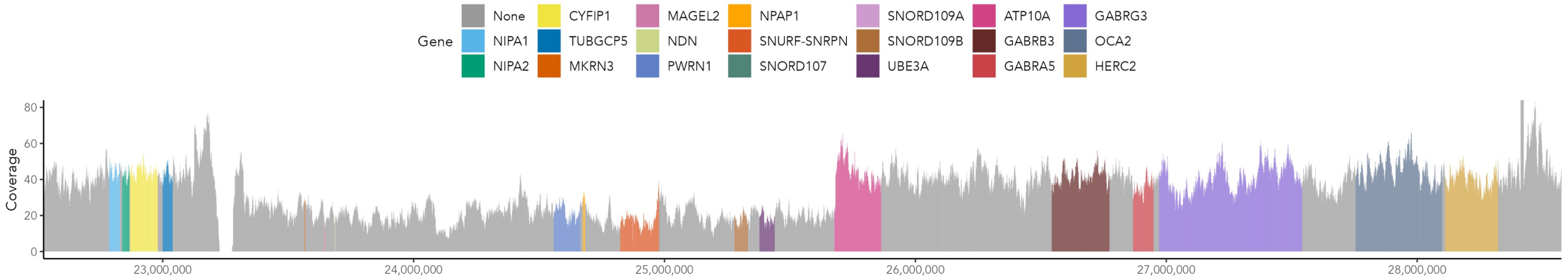
- More than 31,000 CpG islands in the human genome
- Well-known imprinting disorders

Exercise: imprinting disorder example

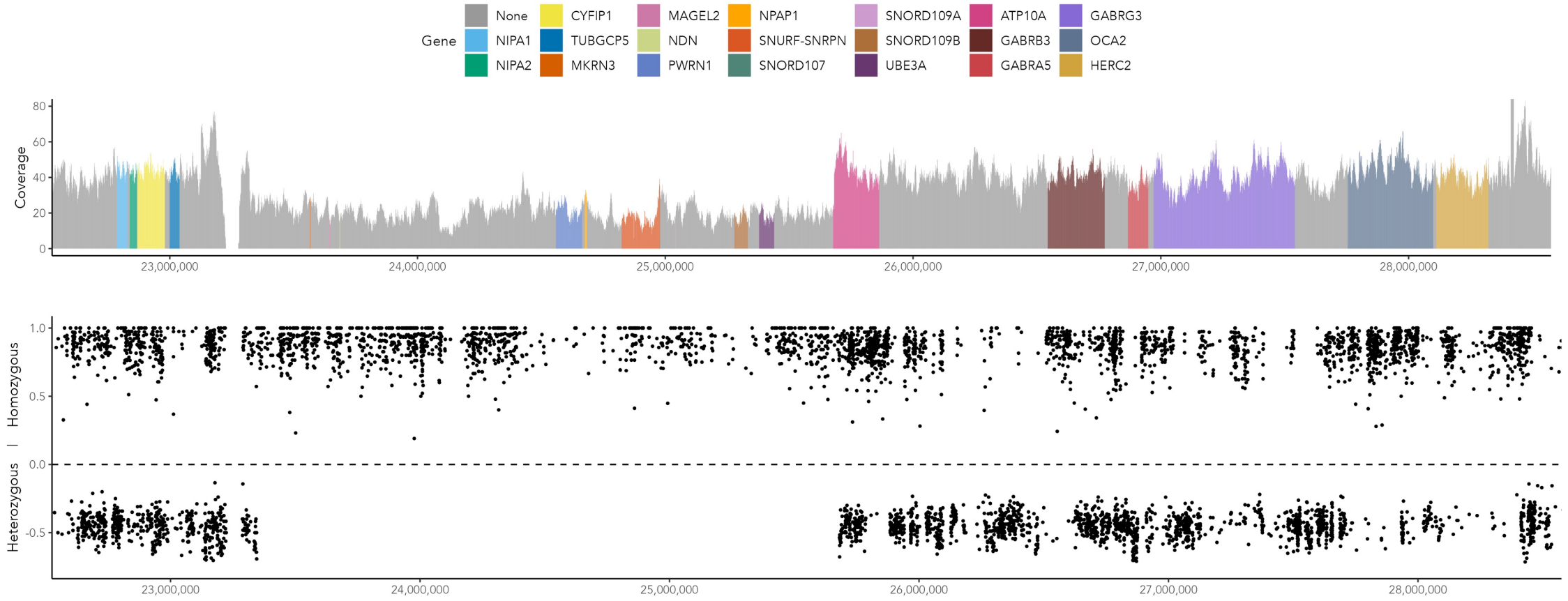
- Open **CSH.AdvSeq.Examples.bam**
  - This file contains data from different experiments and individuals, all ONT.
  - Go to SNURF and look at methylation, what do you see at the CpG?

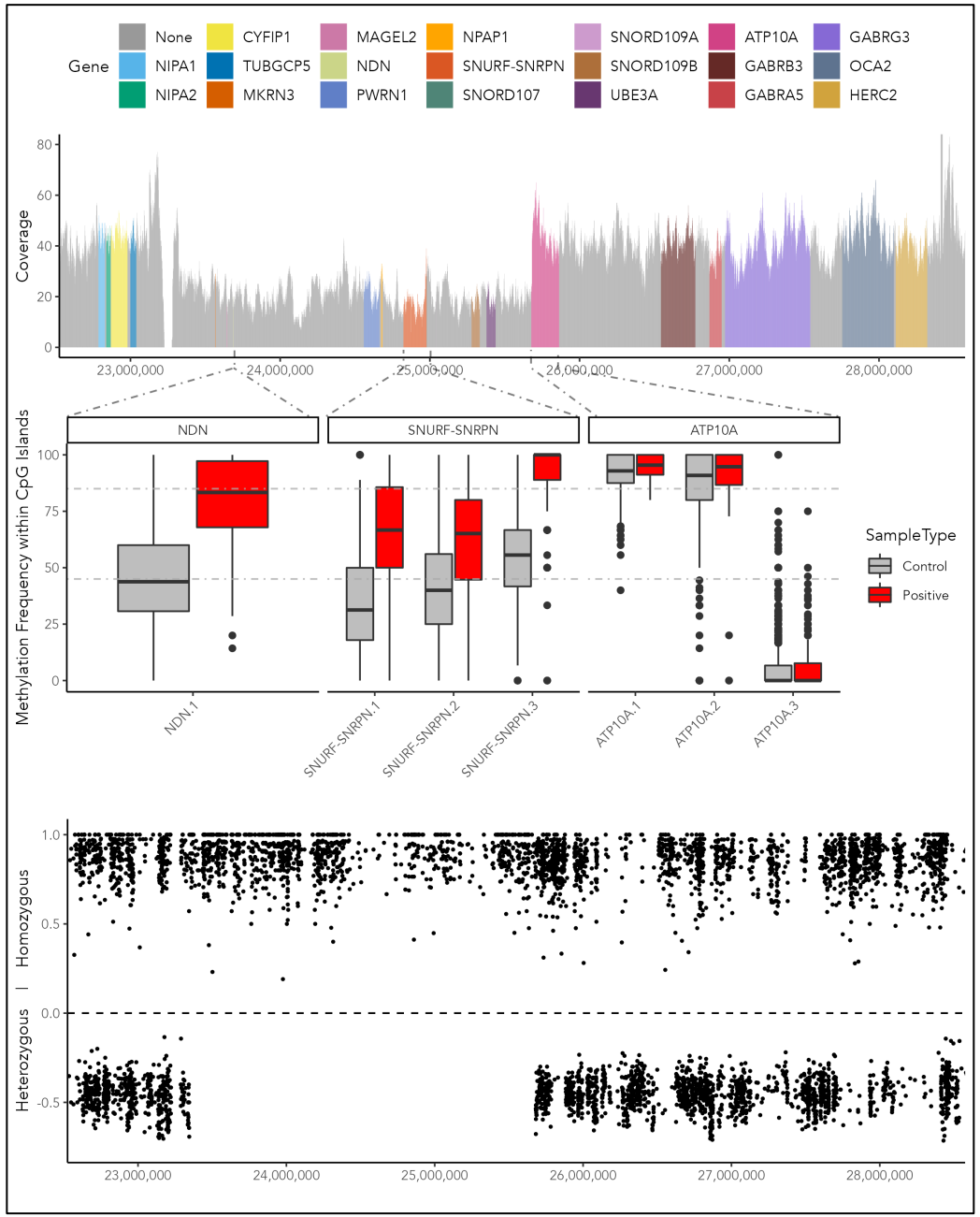
- Open **PWS.ONT.control.bam**
  - Compare methylation here to the previous sample.

MeOW identifies differentially methylated regions in a known case of Prader-Willi syndrome



MeOW identifies differentially methylated regions in a known case of Prader-Willi syndrome



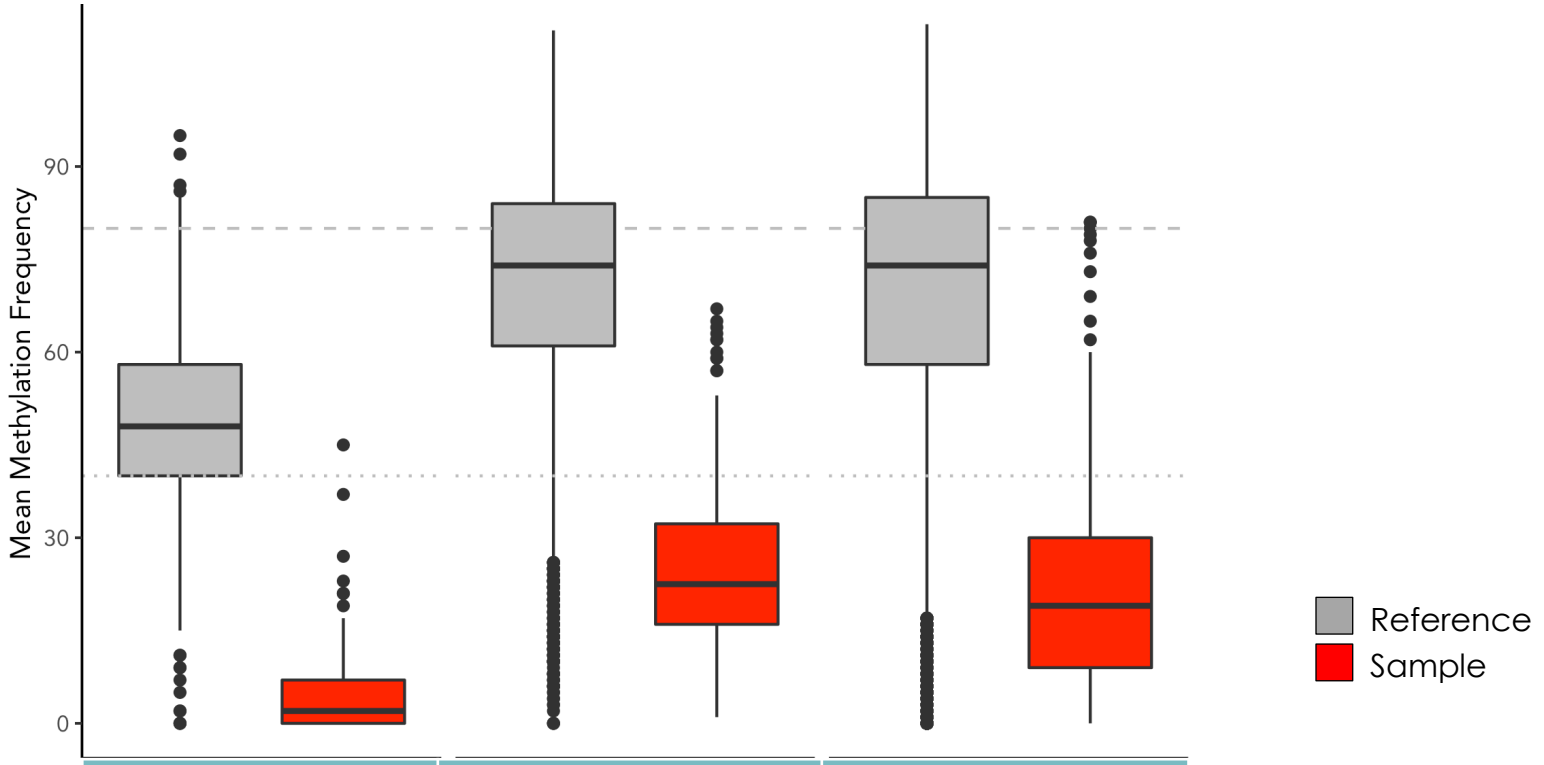


Differences in methylation are seen at select CpG islands

Genome-wide detection of differences in methylation

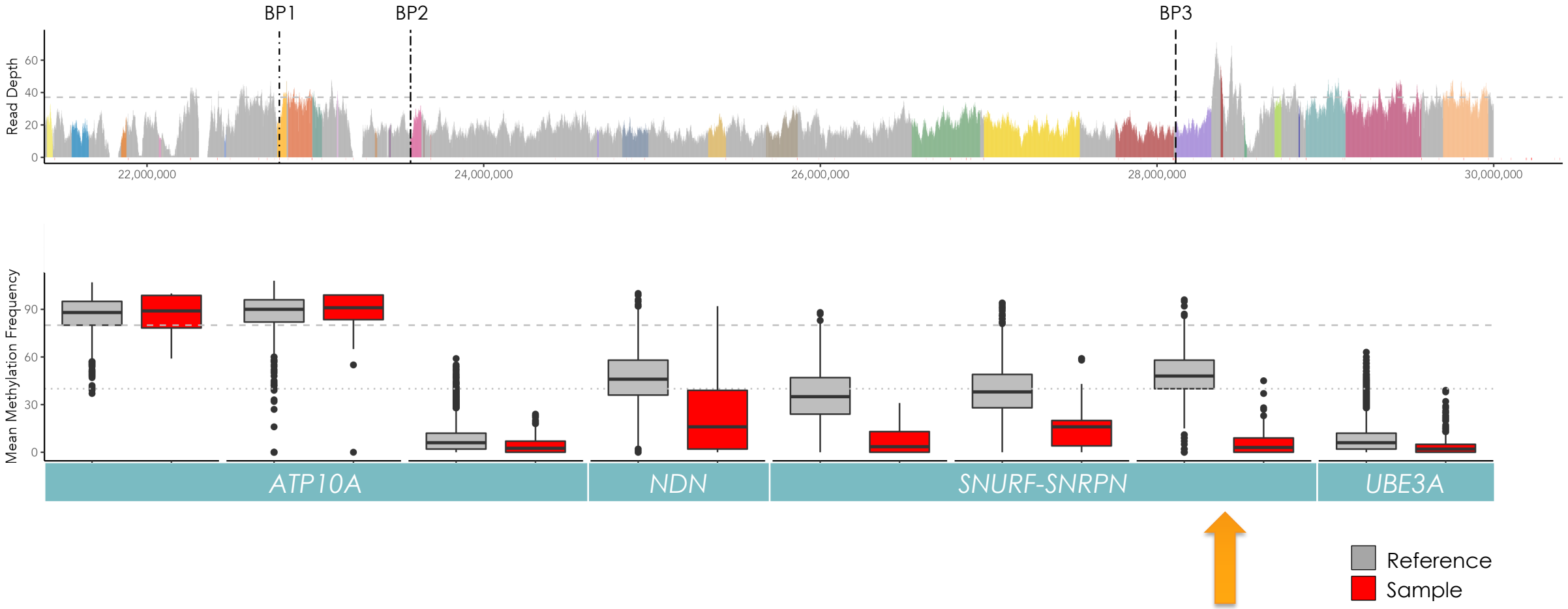
- More than 31,000 CpG islands in the human genome
- Well-known imprinting disorders
- **What if you don't know where to look?**

MeOW identified 3 DMRs in an unknown case

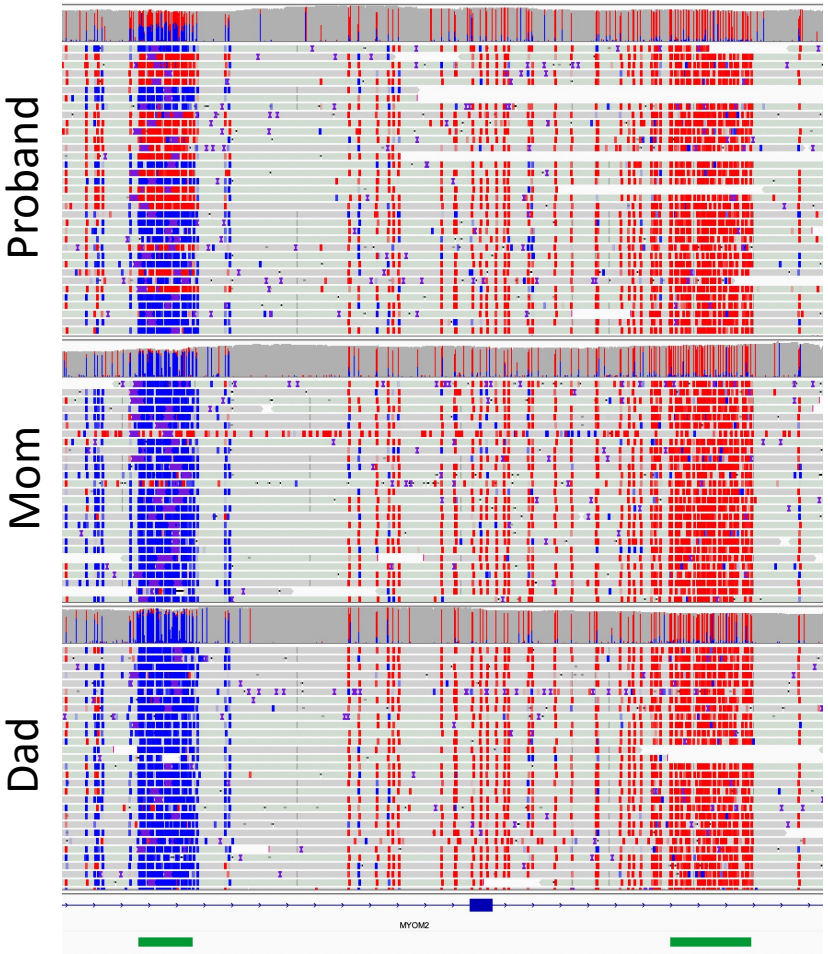


Chromosome	15	19	2
Gene	*SNURF-SNRPN*	*A1BG*	None
T-test P value	1.05×10^{-5}	2.3×10^{-61}	4.4×10^{-89}
Beta test P value	1.49×10^{-5}	0.0097	0.035

Analysis of the full region confirms a deletion that includes the imprinting locus

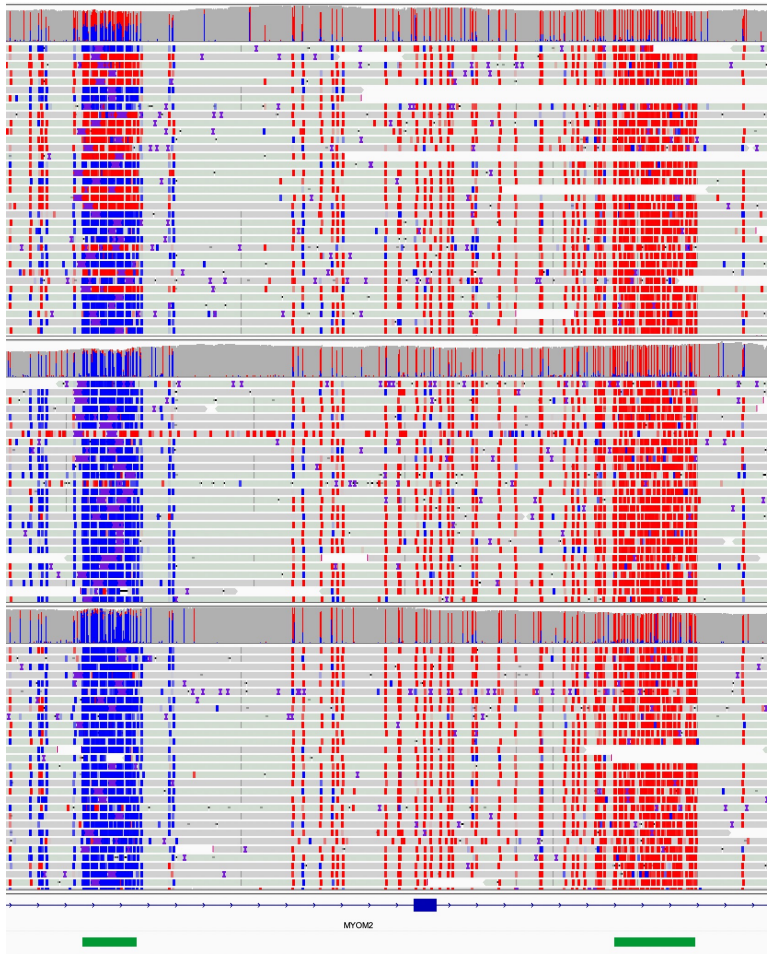


MeOW identifies age-specific DMRs



In the genomes of 8 newborns, MeOW identified DMRs not shared with parents or adult controls.

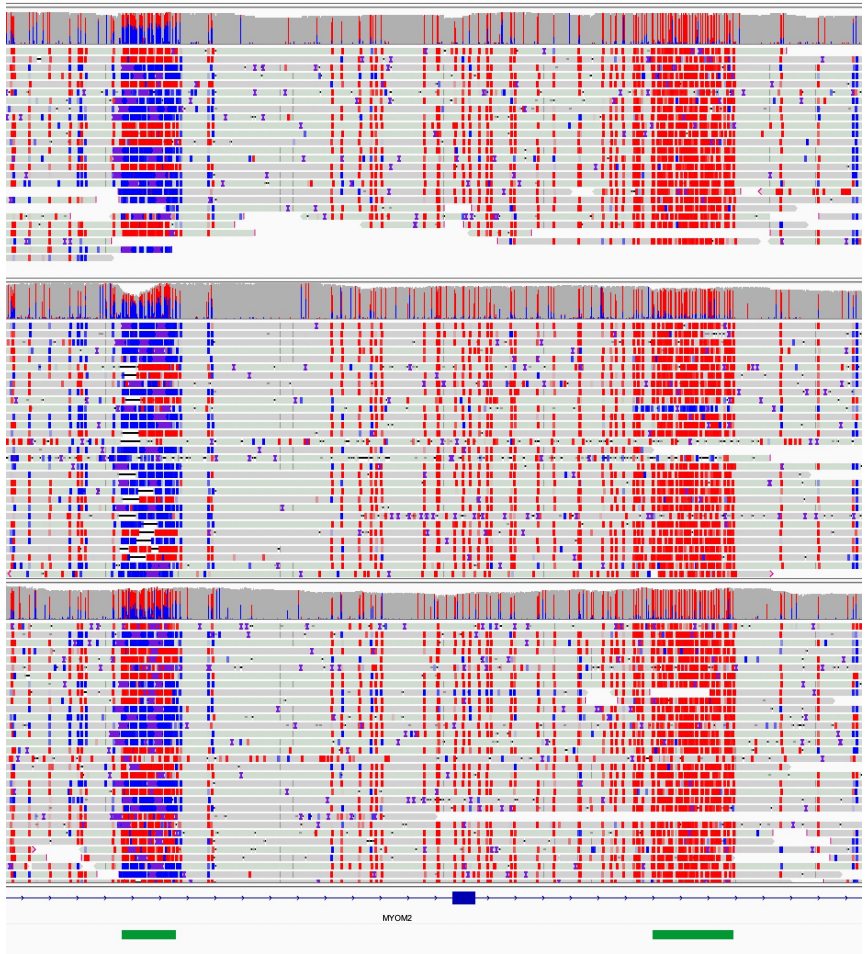
MeOW identifies age-specific DMRs



In the genomes of 8 newborns, MeOW identified DMRs not shared with parents or adult controls.

…out of the 36 identified CpG islands, 9 DMRs were shared between at least three individuals, suggesting an age-related pattern.

Age-matched controls



Exercise: female with X-linked disorder

- Open **HemB-female.ONT.hg38.phased.meth.bam**. R9 pore, sup model, phased data, with methylation calls.
  - Individual has Hemophilia B, why?
  - Go to *FMR1* and look at methylation. What stands out to you?

Exercise: RNA sequencing

- Open **isoSeq.example.bam**.
  - HiFi iso-seq data, only reads from chrX:153,847,386-155,249,315.
  - Find reads with multiple isoforms.
    - GDI1, for example.

Additional Exercises

- Using the Hemophilia B sample, call SVs and identify the pathogenic aberration.
  - Software you need
    - Sniffles: https://github.com/fritzsedlazeck/Sniffles
    - CuteSV: https://github.com/tjiangHIT/cuteSV

- For a challenge
  - Re-align the data to hg38 and preserve methylation tags.
    - Software: samtools, minimap2
  - Phase your re-aligned data
    - Software: Clair3, LongPhase
  - Re-align the data to CHM13. Compare alignment in challenging regions to see if CHM13 improves your alignment.
    - Software: samtools, minimap2

If you want to see some direct RNA data

- https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md

Thank you!