Illuminating the Dark Genome: evaluation of technologies to assess medically relevant dark spots

Fritz Sedlazeck
@sedlazeck

Oct.18, 2018

Emerging technologies improve genomes

Technology	Yield/flow cell	Read Length/Range
PacBio Sequel	~10 Gbp	~15 kbp *
ONT MINion	~10-12 Gbp	~30 kbp *
ONT Promethion	~80 Gbp	~30 kbp *
10x Linked Reads	~30x	100bp paired end/ 100kbp

REVIEWS

COMPUTATIONAL TOOLS

Piercing the dark matter: bioinformatics of long-range sequencing and mapping

Fritz J. Sedlazeck[1], Hayan Lee[2], Charlotte A. Darby[3] and Michael C. Schatz[3,4]*

BRIEF REPORT | Genetics inMedicine

Long-read genome sequencing identifies causal structural variation in a Mendelian disease

Jason D. Merker, MD, PhD[1,2], Aaron M. Wenger, PhD[3], Tam Sneddon, DPhil[2],

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Detection of GBA missense mutations and other variants using the Oxford Nanopore MinION

Melissa Leija-Salazar, Fritz J Sedlazeck, Katya Mokretar, Stephen Mullin, Marco Toffoli, Maria Athanasopoulou,

New Results

Multi-platform discovery of haplotype-resolved structural variation in human genomes

Mark J.P. Chaisson, Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J.

New Results

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

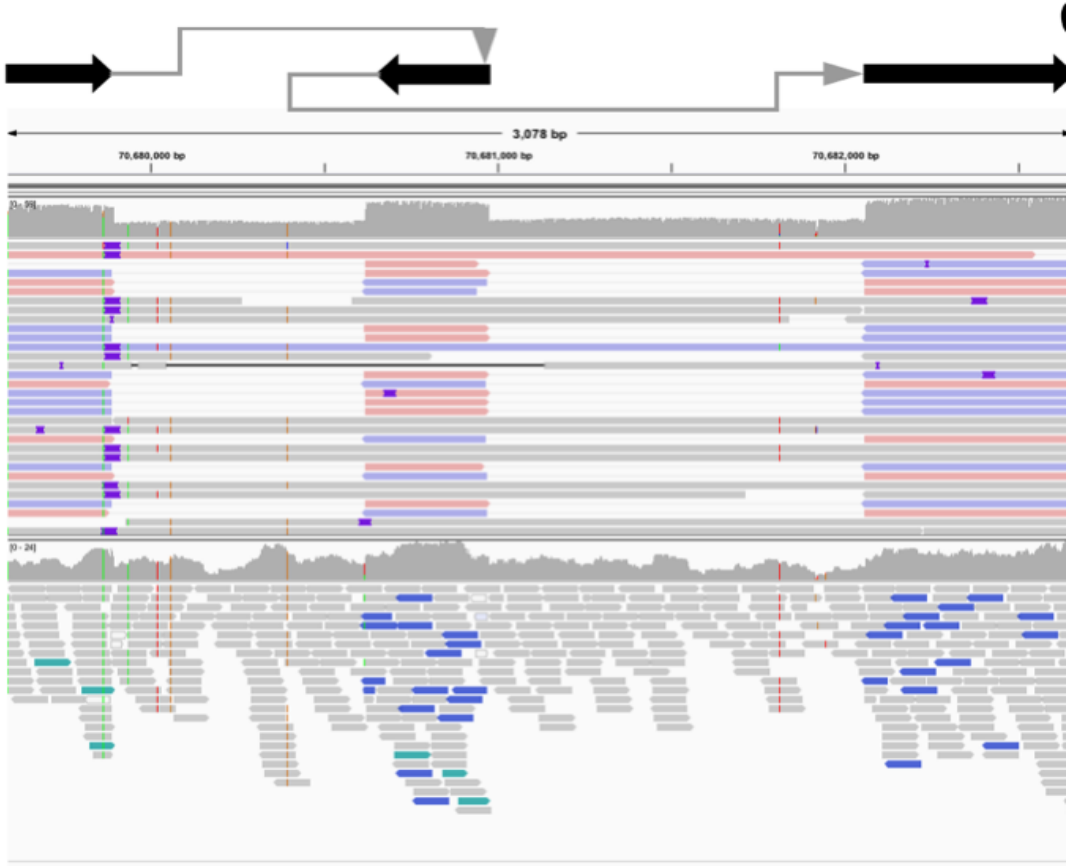
Maria Nattestad, Sara Goodwin, Karen Ng, Timour Baslan, Fritz J. Sedlazeck, Philipp Rescheneder, Tyler Garvin,

Identification of complex SVs with long reads

Long reads:
(+) SVs in repetitive regions
(+) Span SVs
(+) Uniform coverage
(+) Can identify more complex SVs

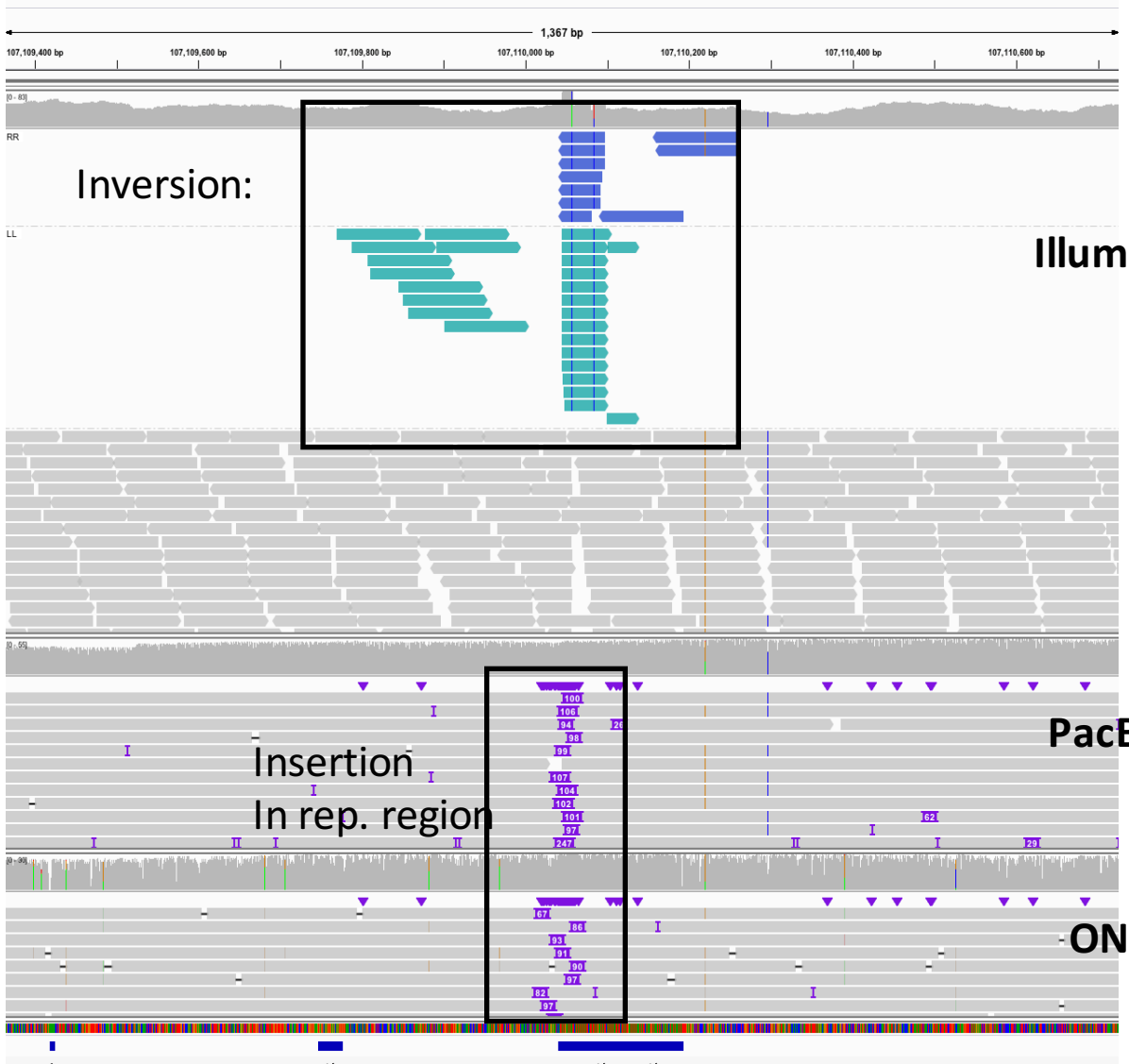
Inversion flanked by deletions:

- Haemophilia A
- Only found over long range PCR! (2007)



Sedlazeck et.al. (2018)

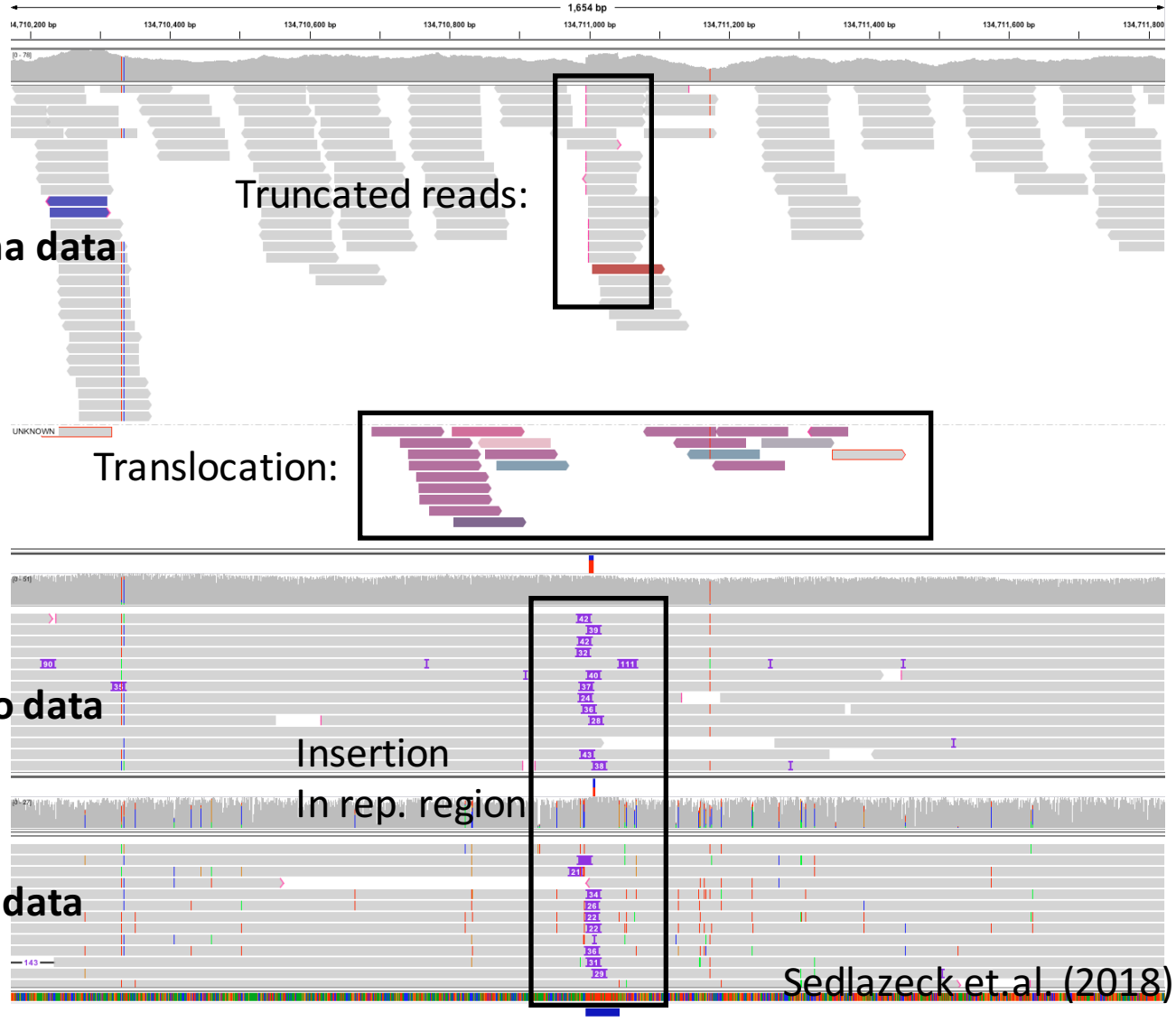
Identification of false positive signals.



Illumina data

PacBio data

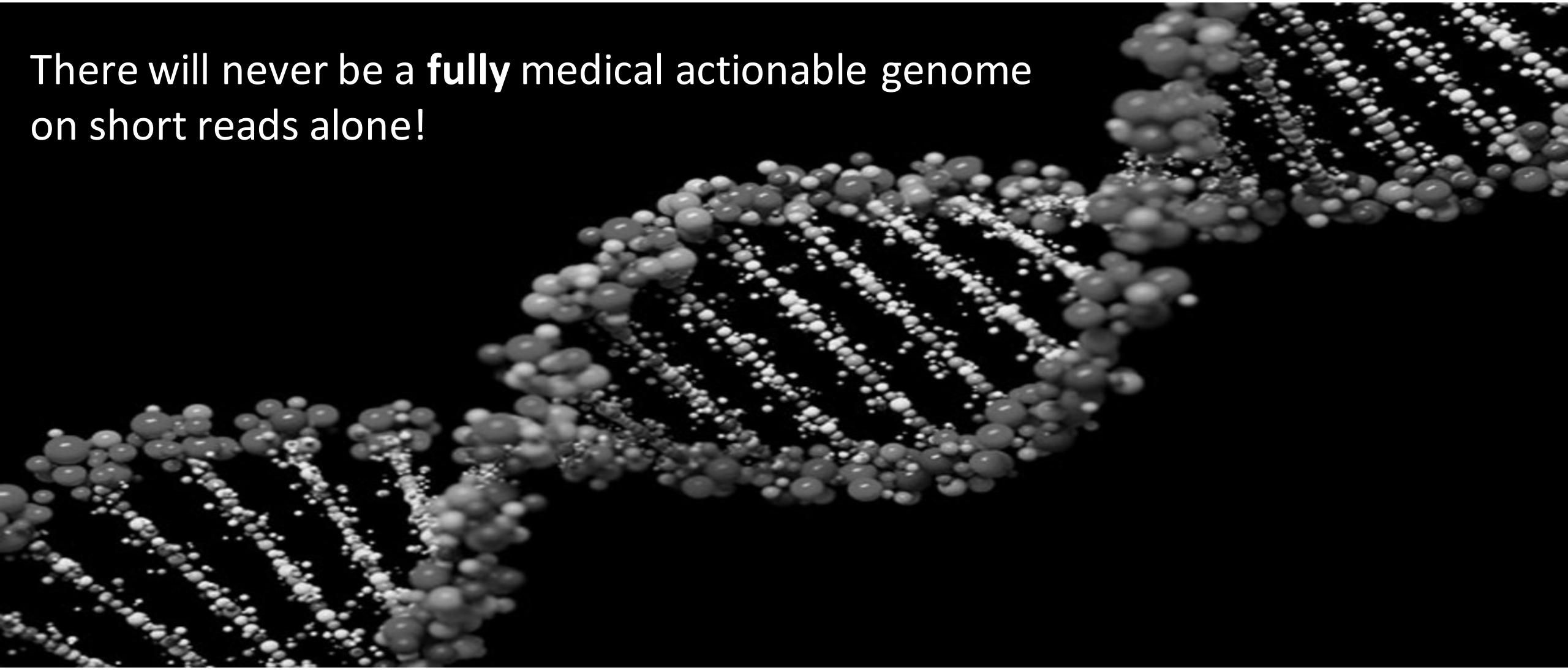
ONT data



Sedlazeck et.al. (2018)

The dark genome

There will never be a **fully** medical actionable genome on short reads alone!



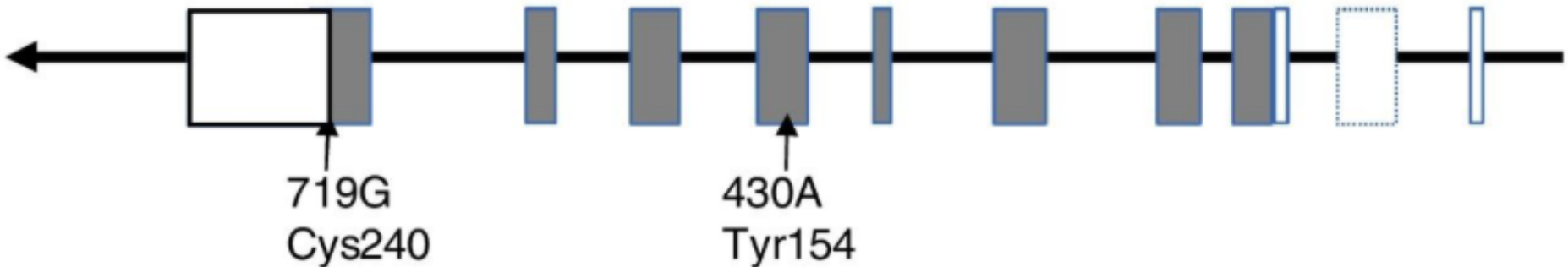
Comprehensive Genomes Pilot: CCDG

Family Study	Illumina, PCR-Free	PacBio	Oxford Nanopore	10X Genomics	RNA-Seq (Read-pairs – M)
Ashkenazi Jewish trio	38x	18x	11x	30x	50
HGSC Control Trio	>100x	19x	14x	56x	68
HapMap CEU Trio	>100x	40x	25x	30x	35

@HGSC
External

TPMT

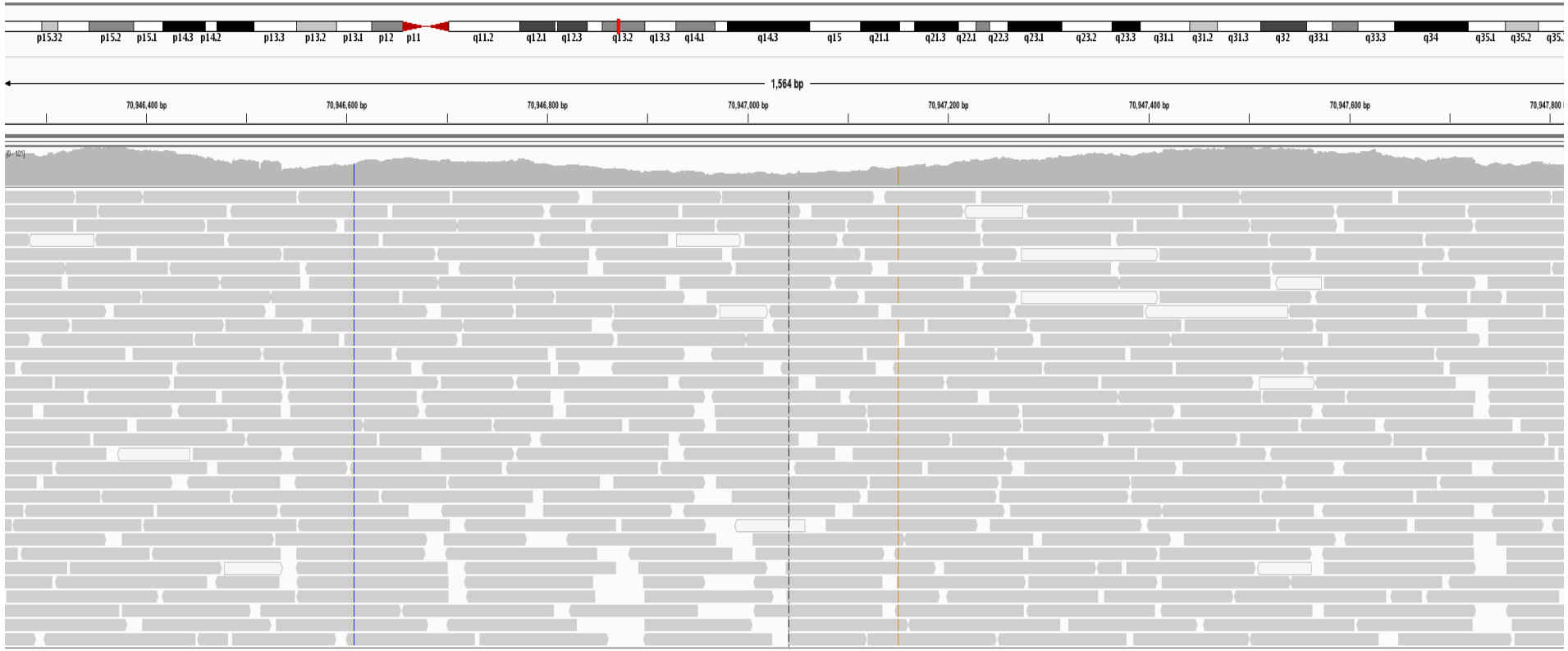
- TPMT gene - encodes the enzyme that metabolizes thiopurine drugs
- 3 drugs: azathioprine, mercaptopurine, thioguanine
  - Chemotherapy for pediatric leukemia
- Two variants, rs1800460 and rs1142345 (>8000 bases apart)
  - cis: TPMT*1/*3A diplotype (intermediate metabolizer)
  - trans: TPMT*3B/*3C diplotype (poor metabolizer)
  - *1/*3A diplotype is more common but less severe than *3B/*3C



IGV viewer

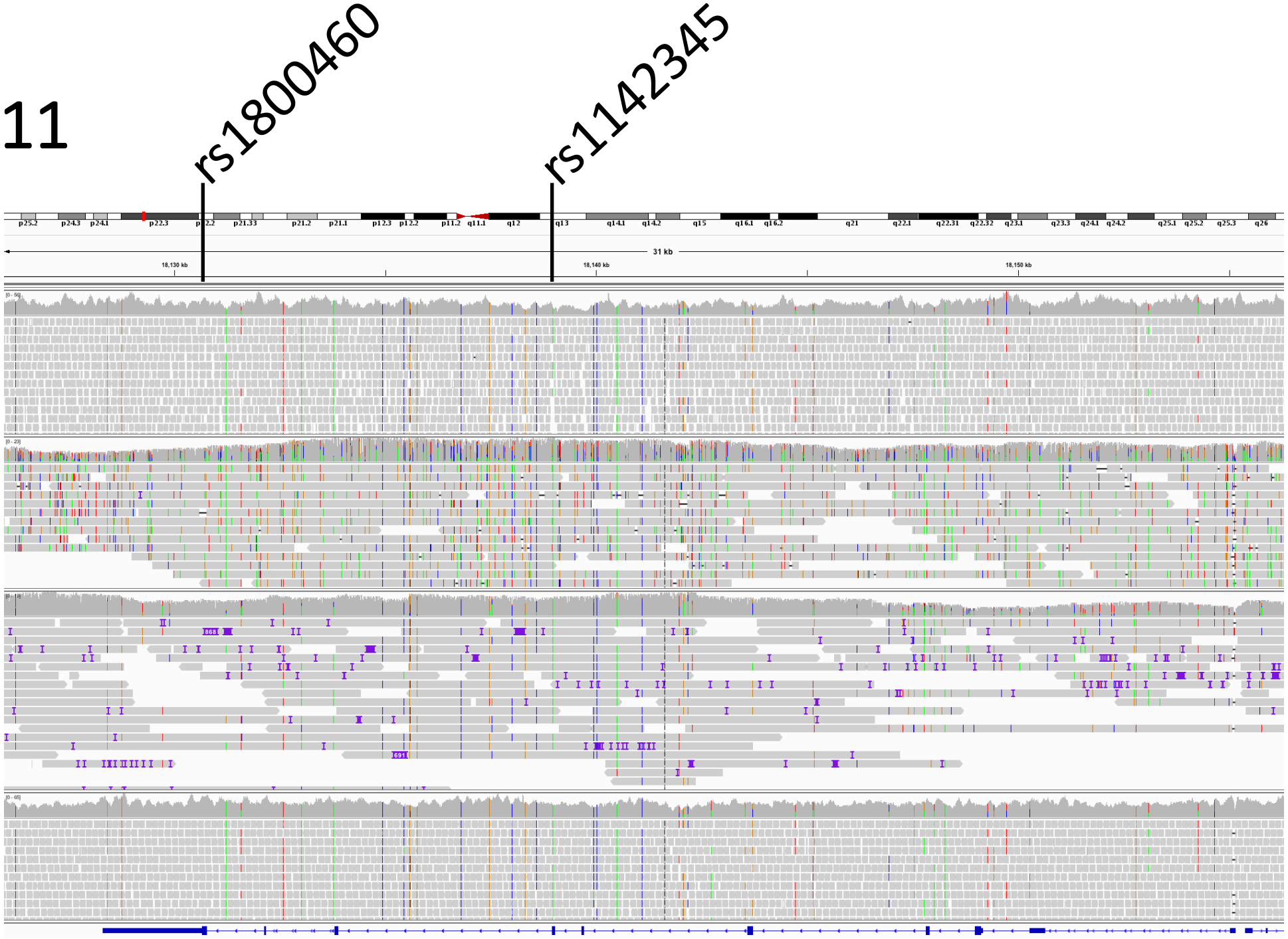
Coordinates
Coverage

Reads

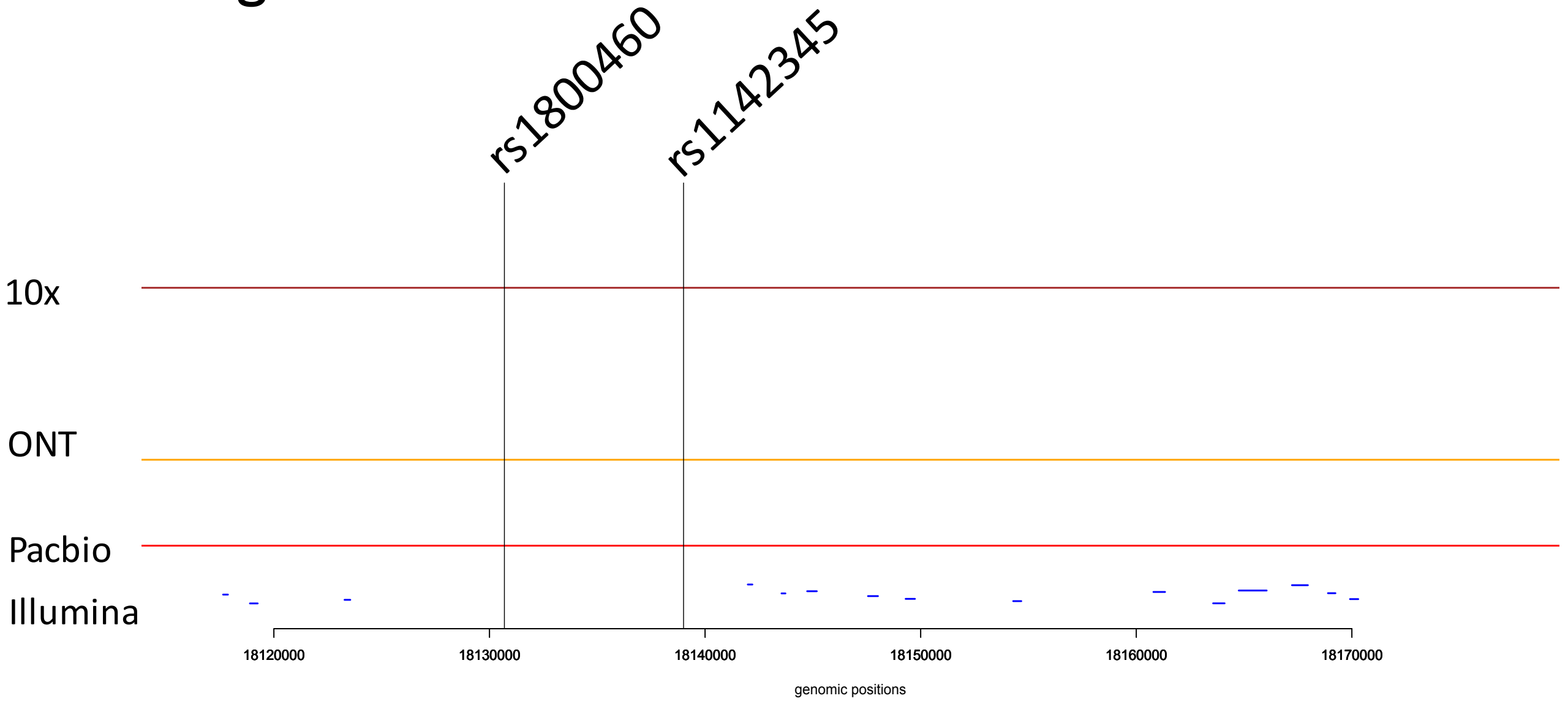


Genome

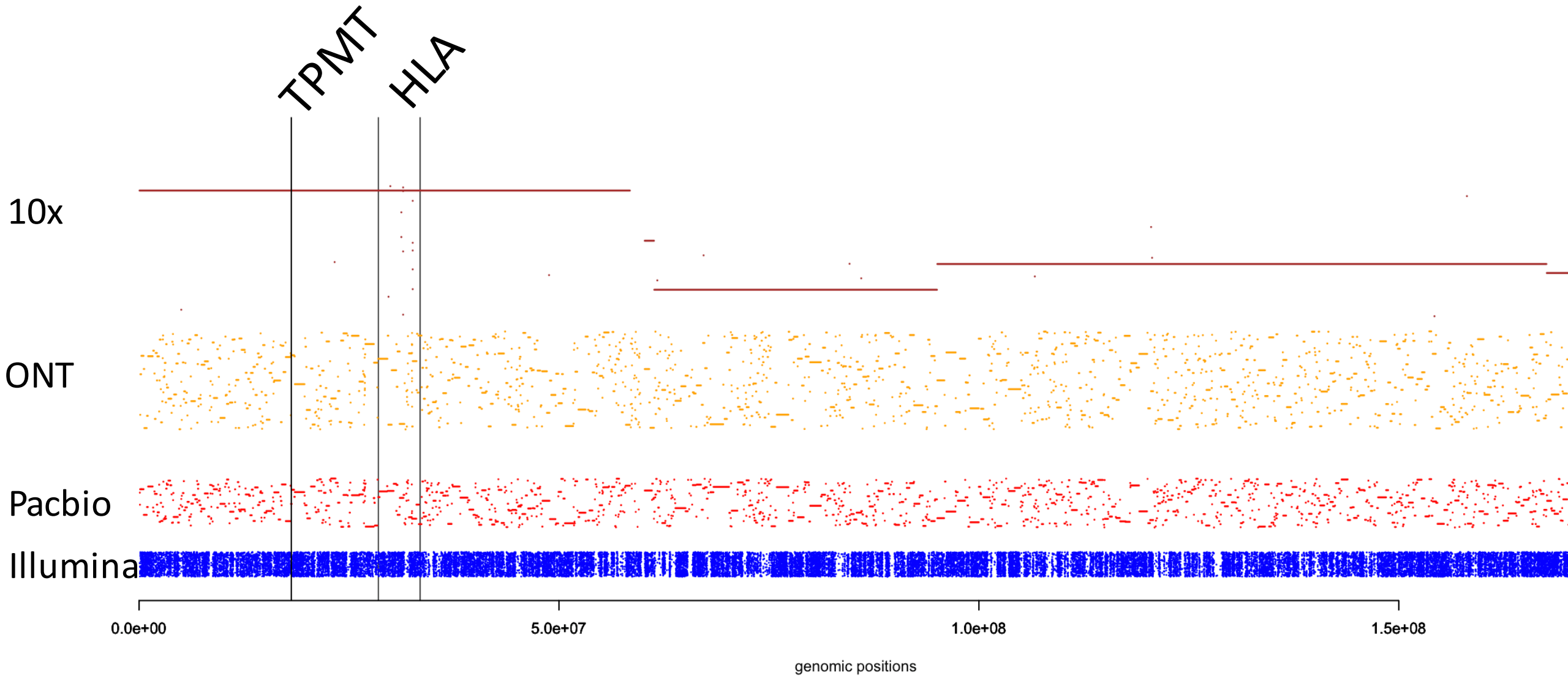
TPMT:HS1011


rs1800460
rs1142345
31 kb
18,130 kb
18,140 kb
18,150 kb
10x
ONT
PacBio
Illumina

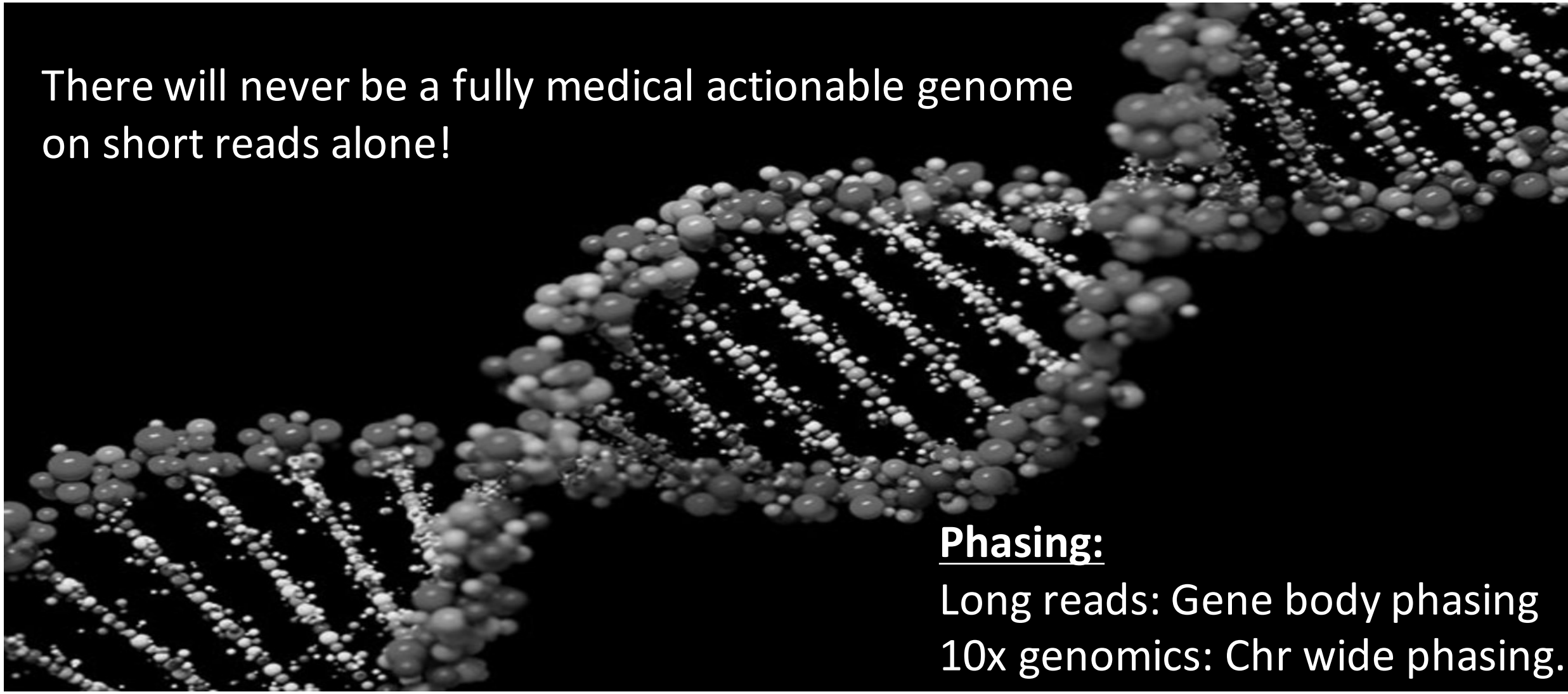
Phasing: TPMT



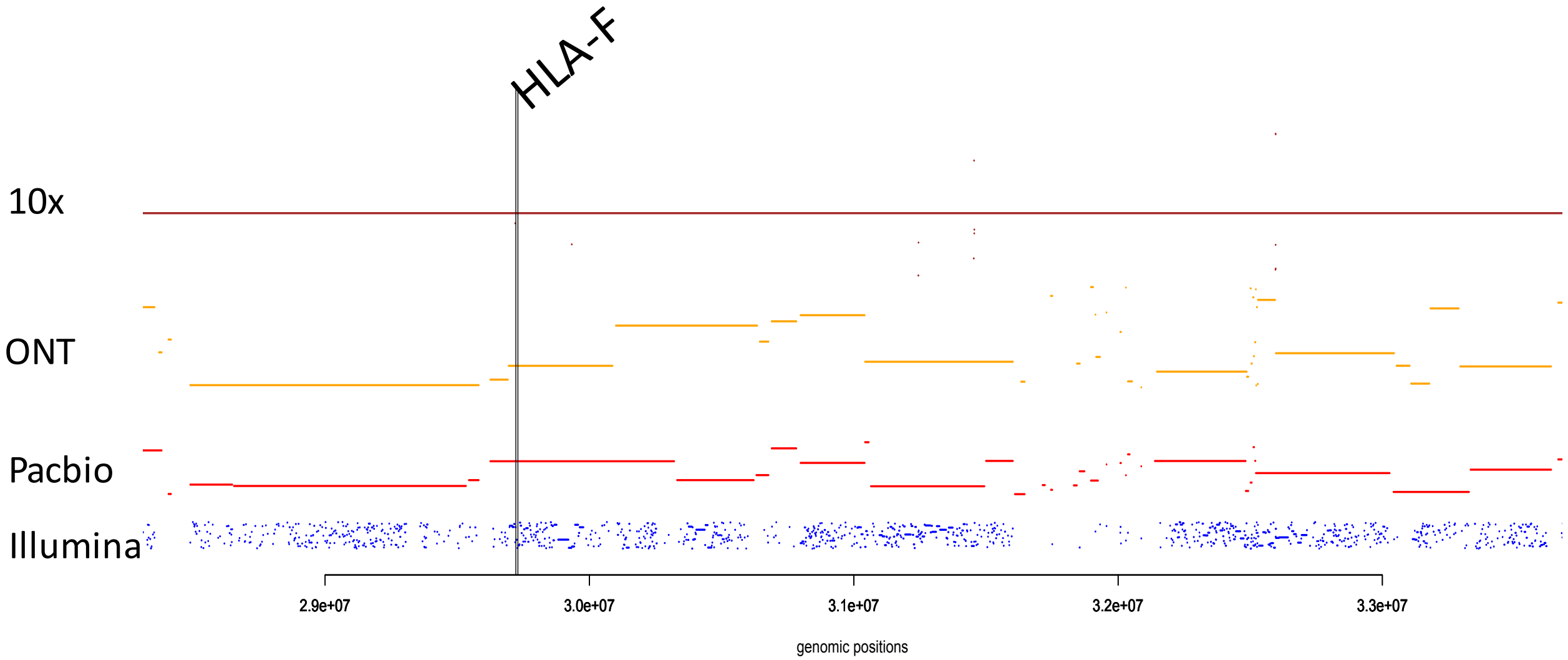
Phasing: Chr6



The dark genome



Phasing: HLA


HLA-F
10x
ONT
Pacbio
Illumina
2.9e+07
3.0e+07
3.1e+07
3.2e+07
3.3e+07
genomic positions

HLA-F

- Non-classical HLA class I heavy chain paralogues
- HLA-F regulates the immune system in pregnancy, infection, and autoimmunity by signaling through NK-cell receptors
- Important for cancer and tumors [1]
- Several diseases: hepatitis B, Systemic Lupus Erythematosus, and Type 1 diabetes (T1D). [1]

1: PMID 25862890, PMID 26332651, PMID 22544725, PMID 21561677, PMID 25435979, PMID 24350297, PMID 22015712..

HLA-F deletion

PacBio:
HS-1011

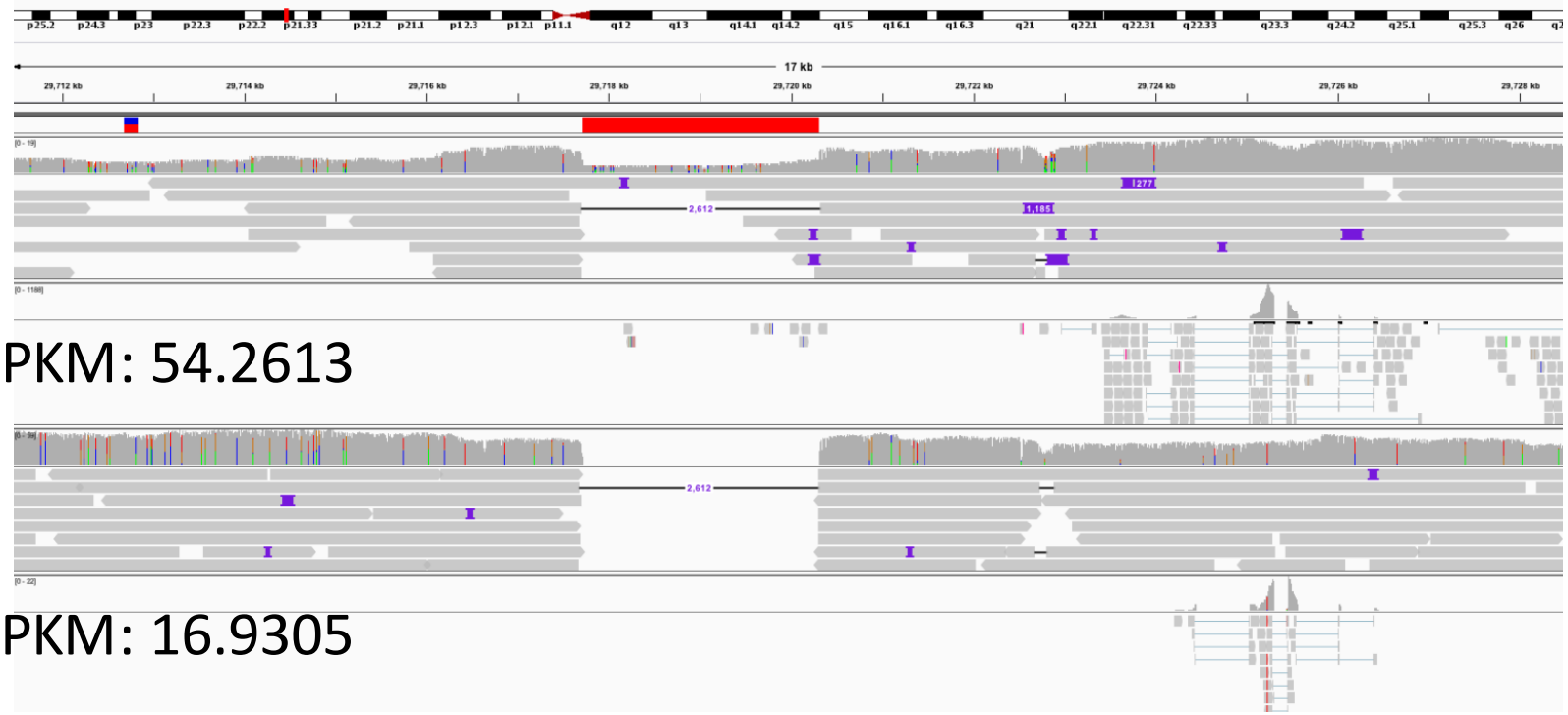
RNA-Seq
HS-1011

FPKM: 54.2613

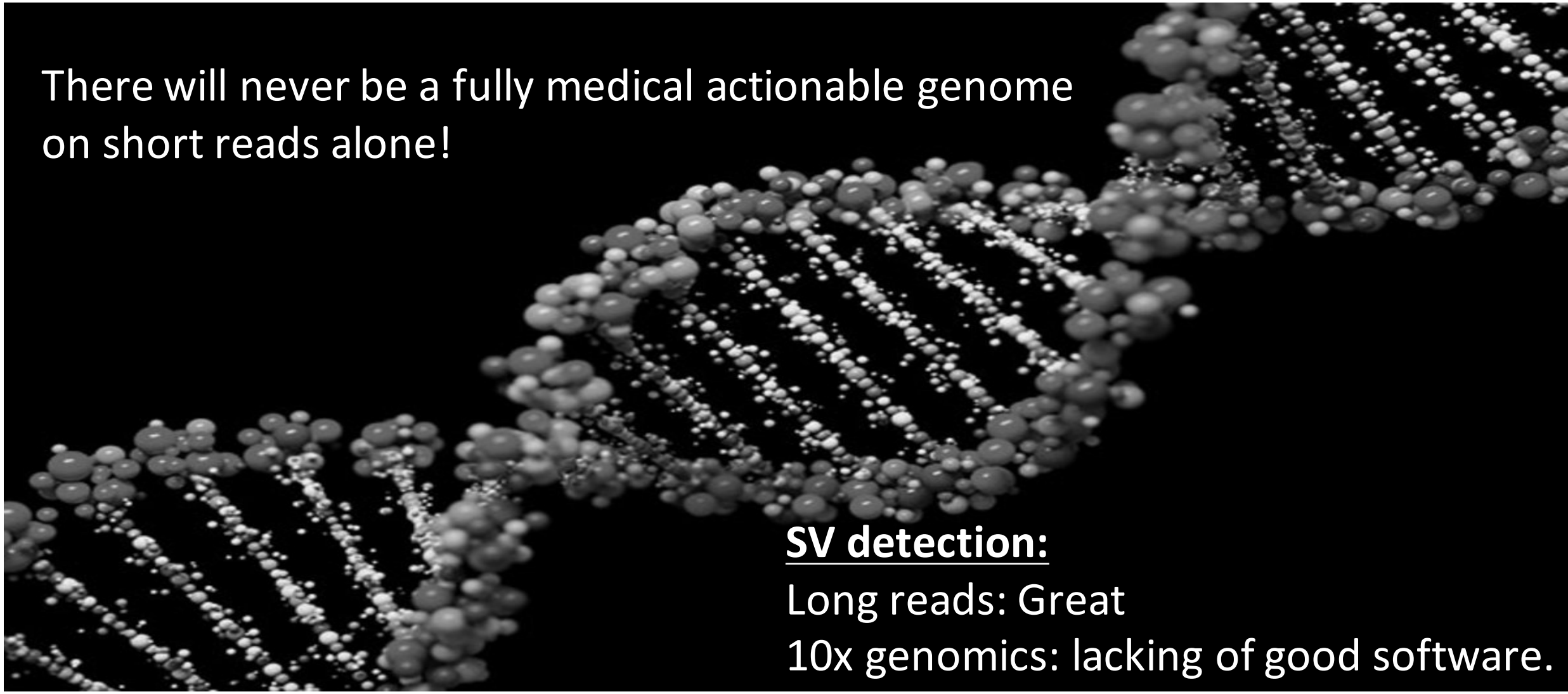
PacBio:
NA12878

RNA-Seq
NA12878

FPKM: 16.9305



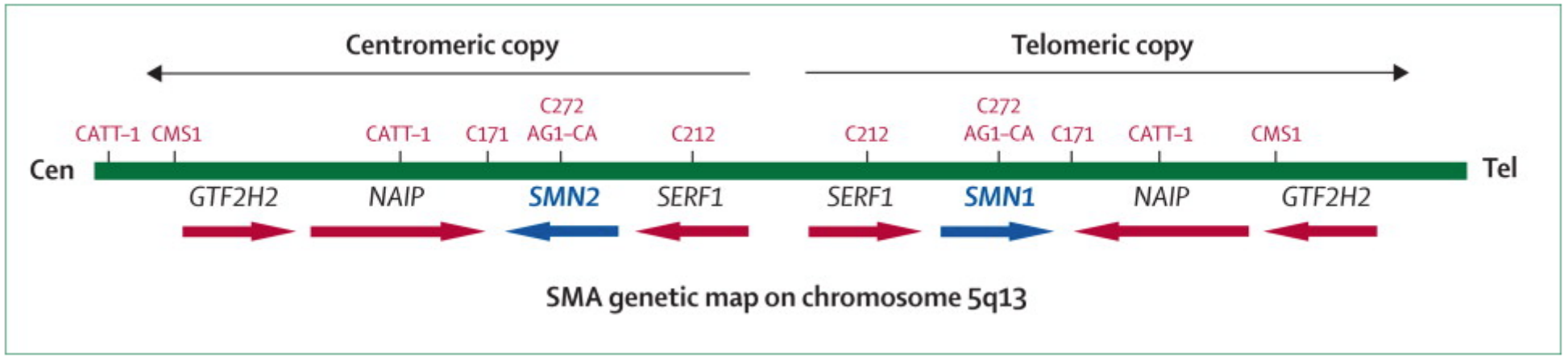
The dark genome



SMA: spinal muscular atrophy

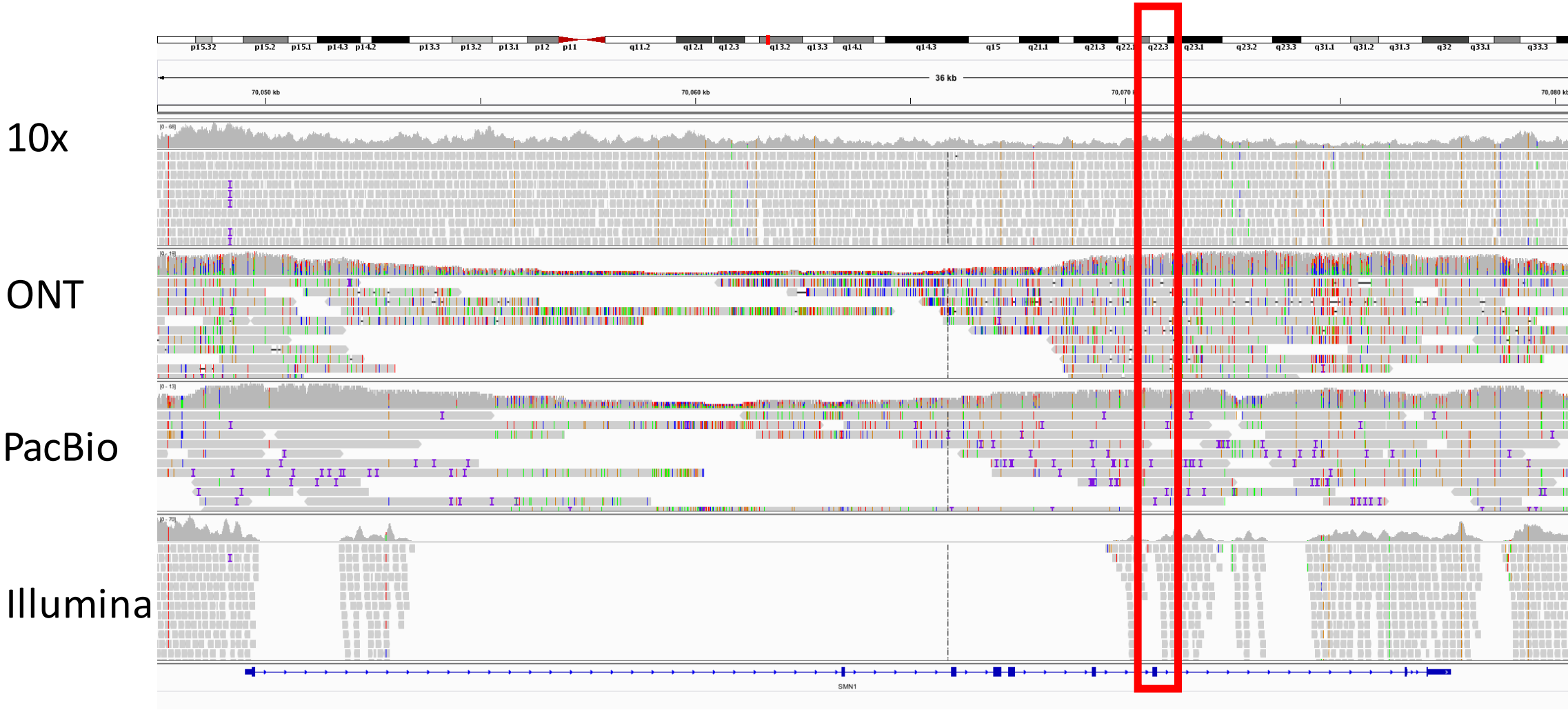
- SMN1 and SMN2 are highly similar and are ~850kbp apart.
  - 5 bases distinguish them

- 95% of individuals with spinal muscular atrophy have homozygous deletions of exon 7 in SMN1



https://www.togetherinsma-hcp.com/en_us/home/disease-education/spinal-muscular-atrophy.html

SMN1 + SMN2



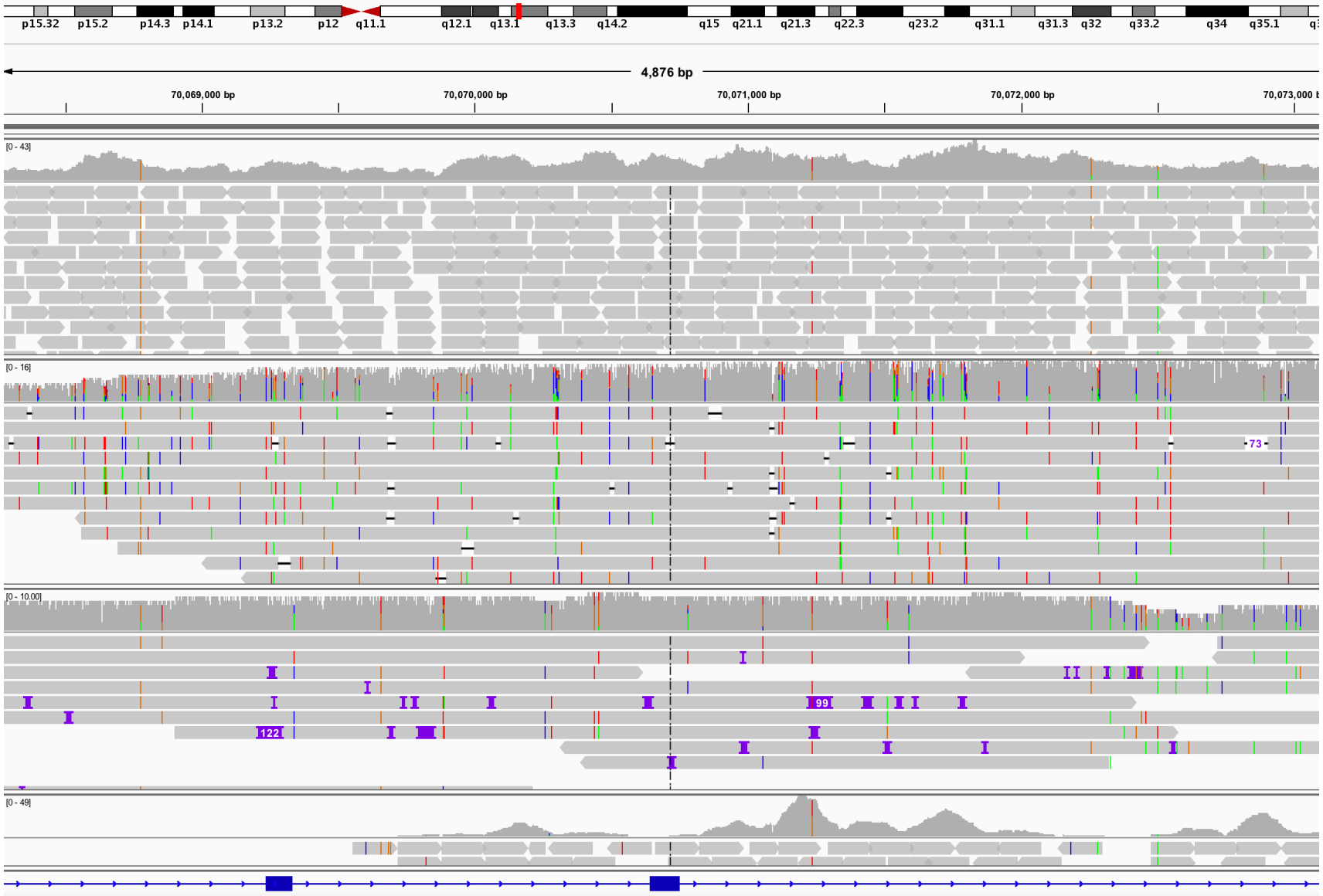
SMN1 exon7

10x

ONT

PacBio

Illumina



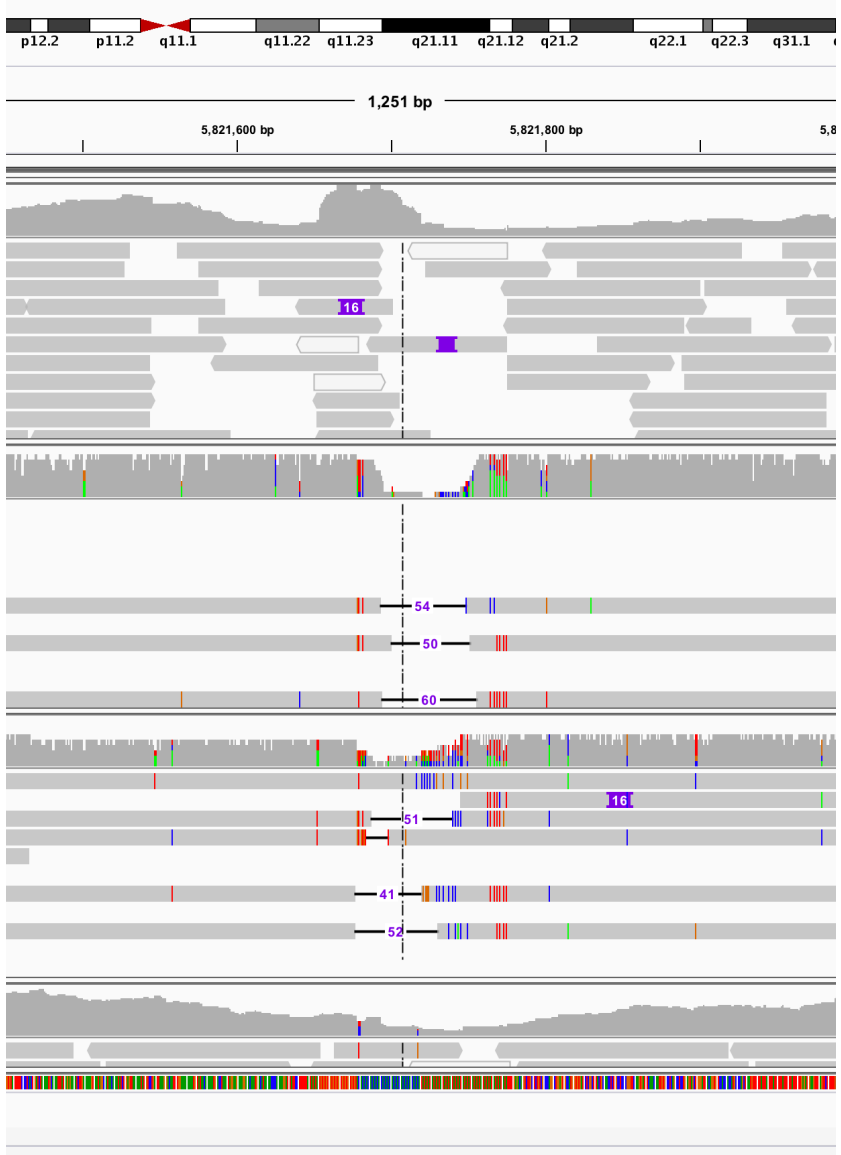
Repeat expansion/contraction

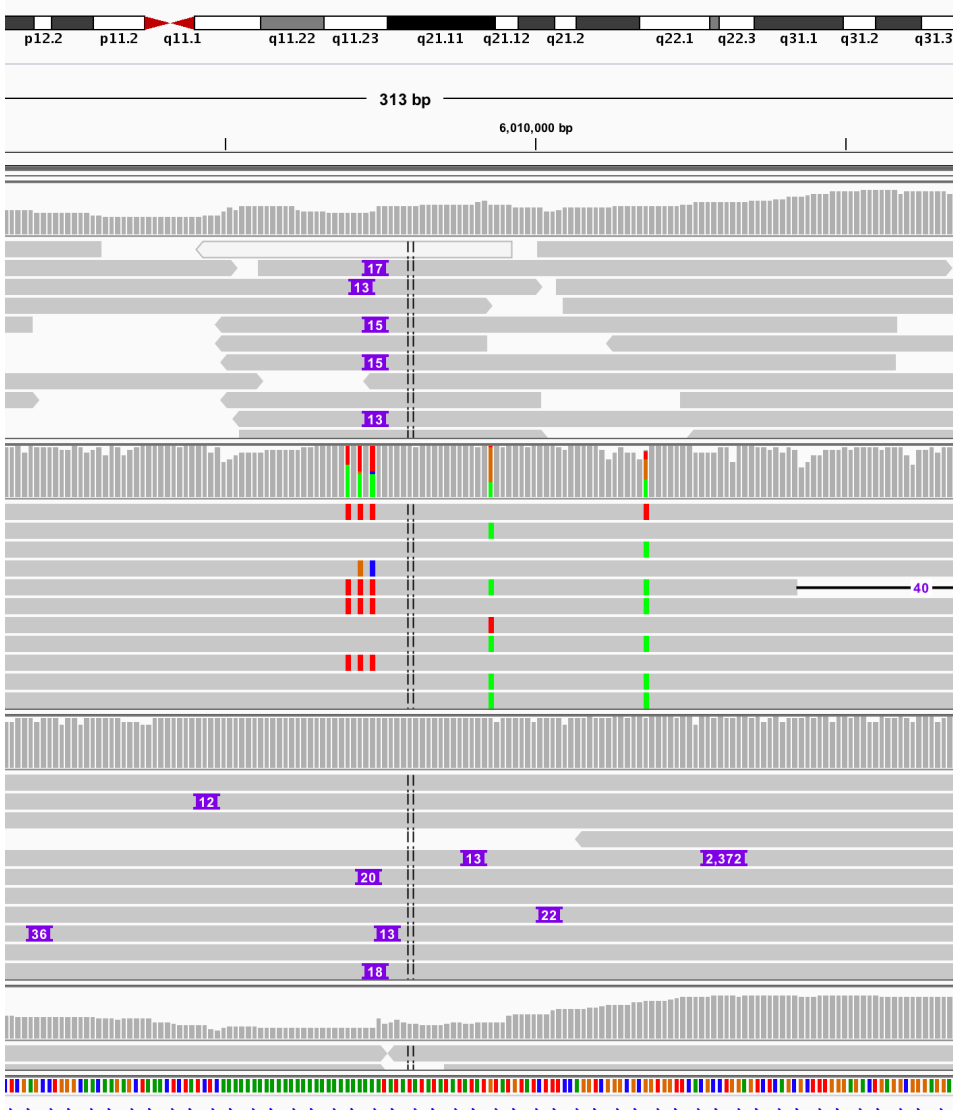
10x

ONT

PacBio

Illumina

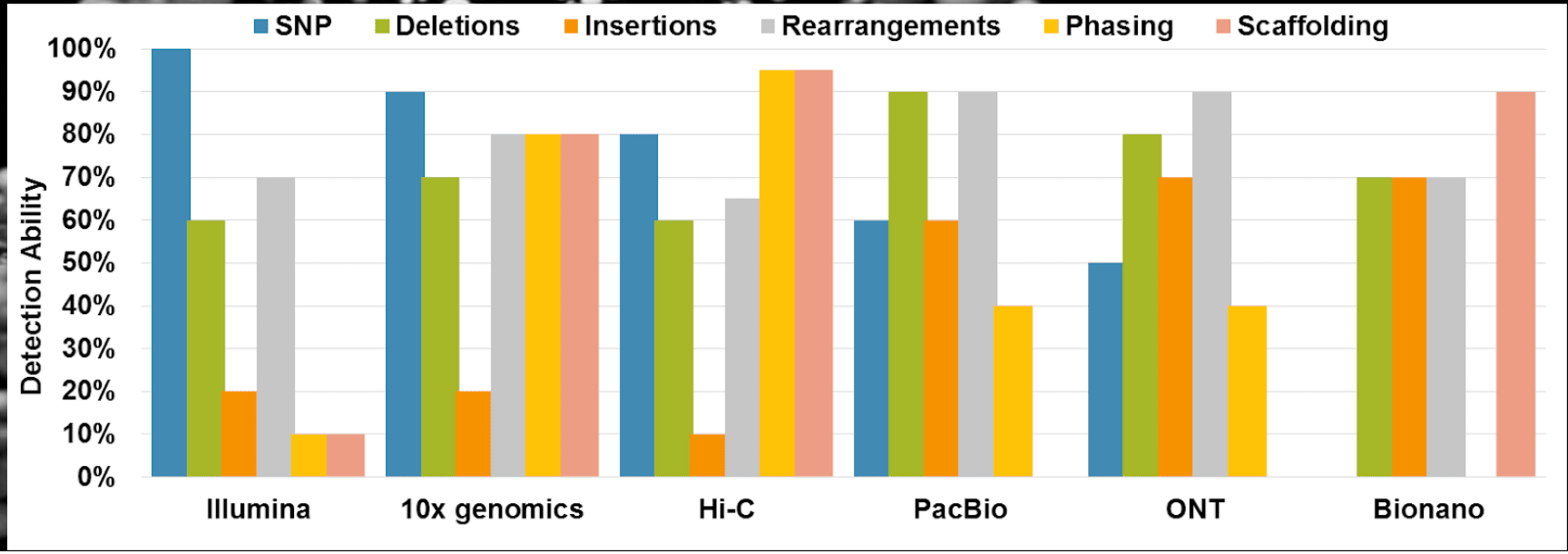




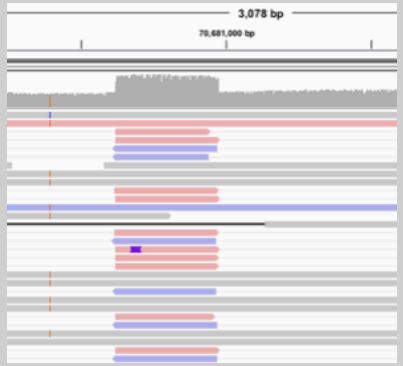
The dark genome: conclusion

There will never be a fully medical actionable genome on short reads alone!

What to choose?



Software



NGMLR
Sedlazeck et. al. (2018)

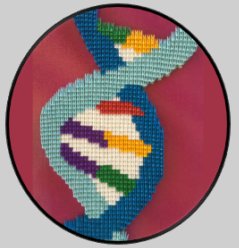
- Convex gap model
- Improved split reads
- Oxford Nanopore + PacBio
- github.com/philres/ngmlr

Sniffles
Sedlazeck et. al. (2018)

- Detection of all SV types
- Also nested/adjacent events
- Oxford Nanopore+ PacBio
- github.com/fritzsedlazeck/Sniffles

Clairvoyante
Lou et. al. (in review)

- Neural Network based SNP calling
- Illumina+ Oxford Nanopore + PacBio
- github.com/aquaskyline/Clairvoyante



Crosstich
Kirsche et.al.(in prep)

- Phases SV + SNPs
- Local assembly
- Uses HapCut2 (SNPs) + Sniffles(SVs)
- github.com/schatzlab/crossstitch

Acknowledgments

Medhat Helmi
Eric Venner
Adam Mansfield
David Murdock
William Salerno
Donna Muzny
Ginger Metcalf
Richard Gibbs

Michael Schatz
Schatz lab