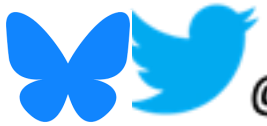
Methodologies for Structural Variant detection

Fritz Sedlazeck & Luis Paulin

Dec,3, 2025



@sedlazeck

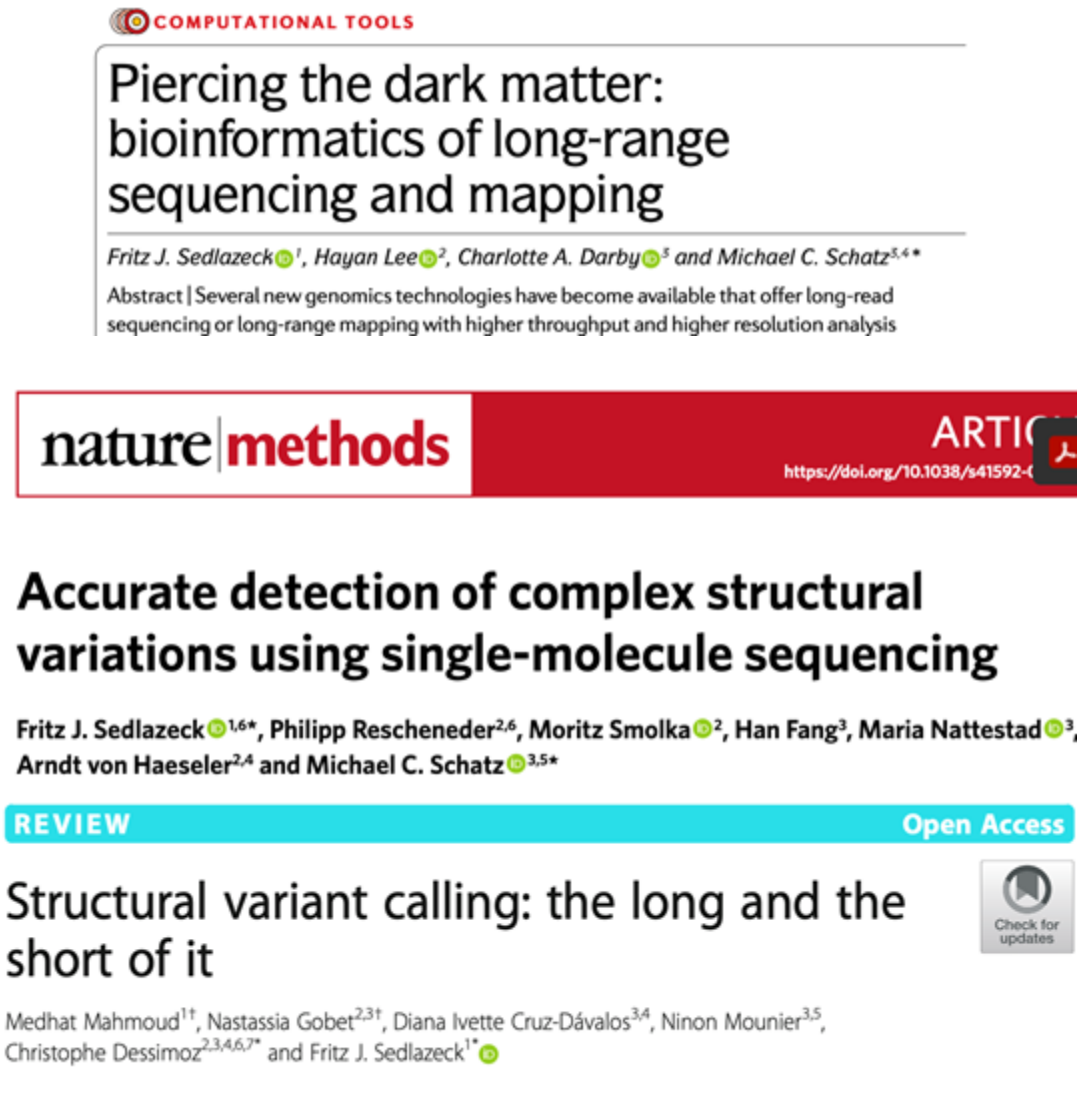
RICE

Recap from yesterday

- SV calling short & long of it

Long read sequencing Structural Variants

- More comprehensive
  - LR: 20-23k SV / human
  - SR: ~10-12k SV / human
- Access in repetitive regions
  - 193 medical genes (Mandelker 2019)
  - 386 medically relevant genes (Wagner 2022)
  - Centromere, telomeres (e.g. T2T)
- Assembly/ Phasing
  - N50, no gaps, phased, etc.


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]*

Abstract | Several new genomics technologies have become available that offer long-read sequencing or long-range mapping with higher throughput and higher resolution analysis

nature methods

ARTIC

https://doi.org/10.1038/s41592-

Accurate detection of complex structural variations using single-molecule sequencing

Fritz J. Sedlazeck[1,6]*, Philipp Rescheneder[2,6], Moritz Smolka[2], Han Fang[3], Maria Nattestad[3], Arndt von Haeseler[2,4] and Michael C. Schatz[3,5]*

REVIEW

Open Access

Structural variant calling: the long and the short of it

Check for updates

Medhat Mahmoud[1†], Nastassia Gobet[2,3†], Diana Ivette Cruz-Dávalos[3,4], Ninon Mounier[3,5], Christophe Dessimoz[2,3,4,6,7*] and Fritz J. Sedlazeck[1*]

New Applications in SV detection

1. Germline SV
2. **Population scale**

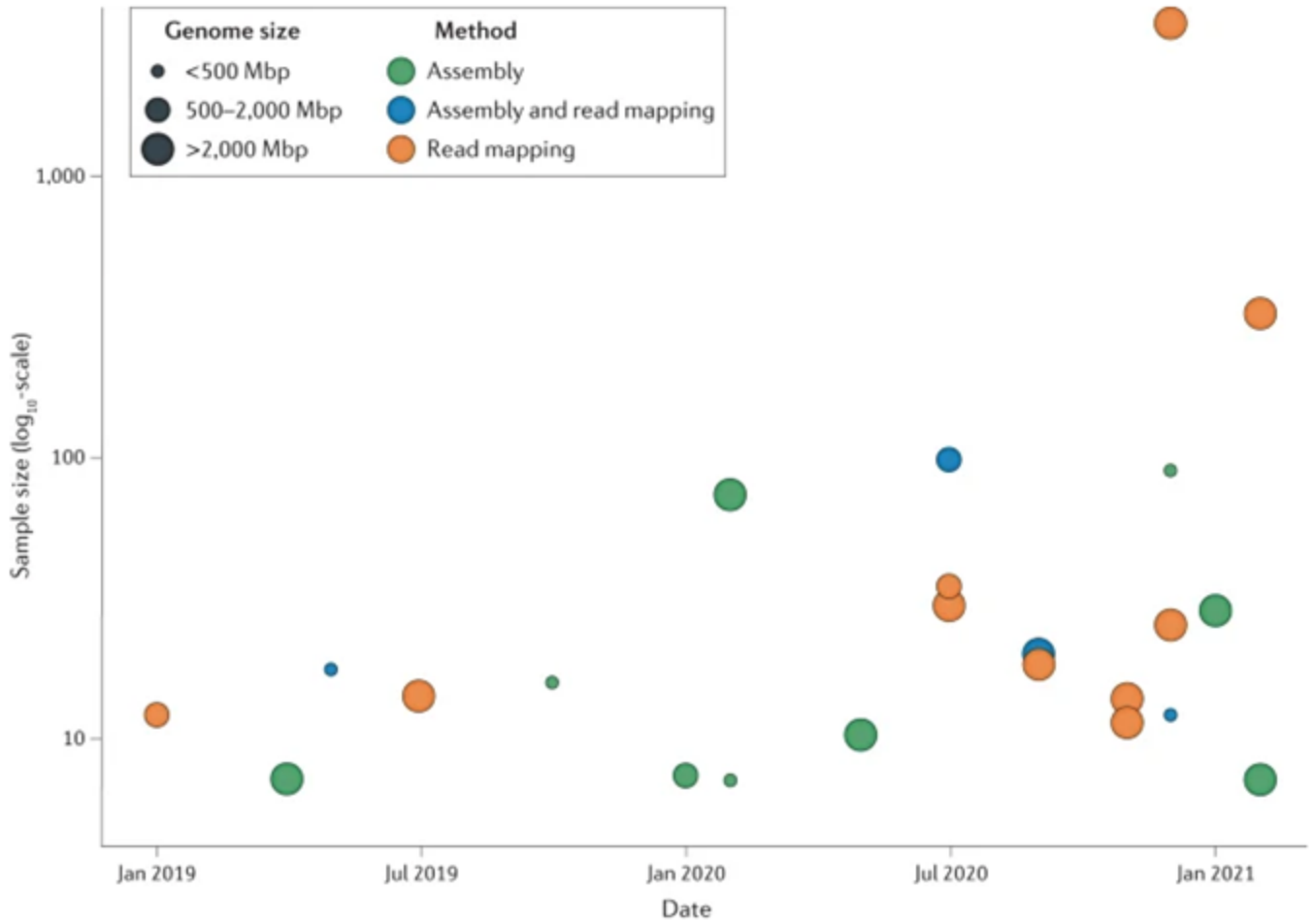
nature > nature reviews genetics > review articles > article

Review Article | Published: 28 May 2021

Towards population-scale long-read sequencing

Wouter De Coster, Matthias H. Weissensteiner & Fritz J. Sedlazeck

Fig. 1: Overview of population-scale studies using long-read sequencing.



Tumor vs normal: colo829

- Improves SV prioritization
- Mutations of PTEN are a step in the development of many cancers

Full SV Genotyping: From family to population scale

cuteSV: SV calling → merging → re-genotyping → merging → population VCF

~36 CPU hours

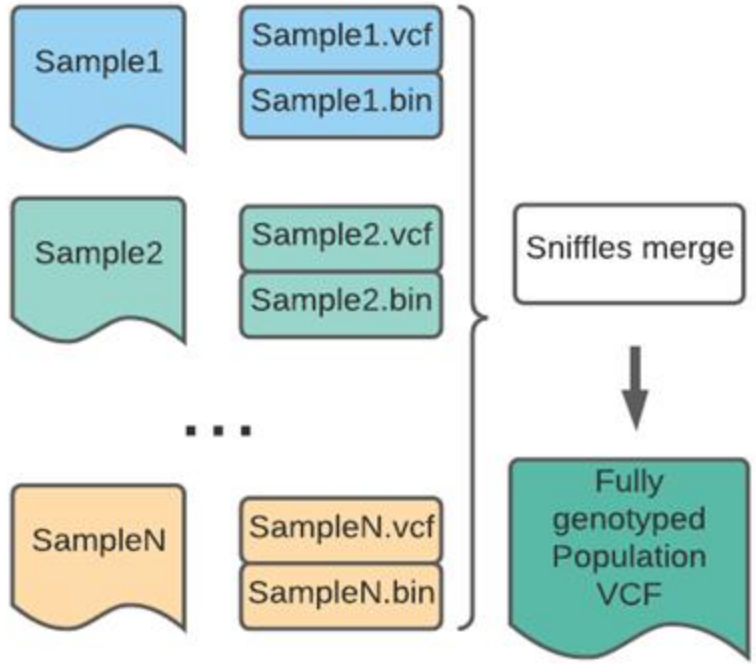
Sniffles2: SV calling → merging → population VCF

65 seconds **(>2000x faster merging)**

Solves n+1 problem

Scaling up to population level

Improves tumor vs. normal



Benchmark data: HG002/3/4 family trio, ONT.

Sniffles2 vs cuteSV: Family Genotyping Accuracy

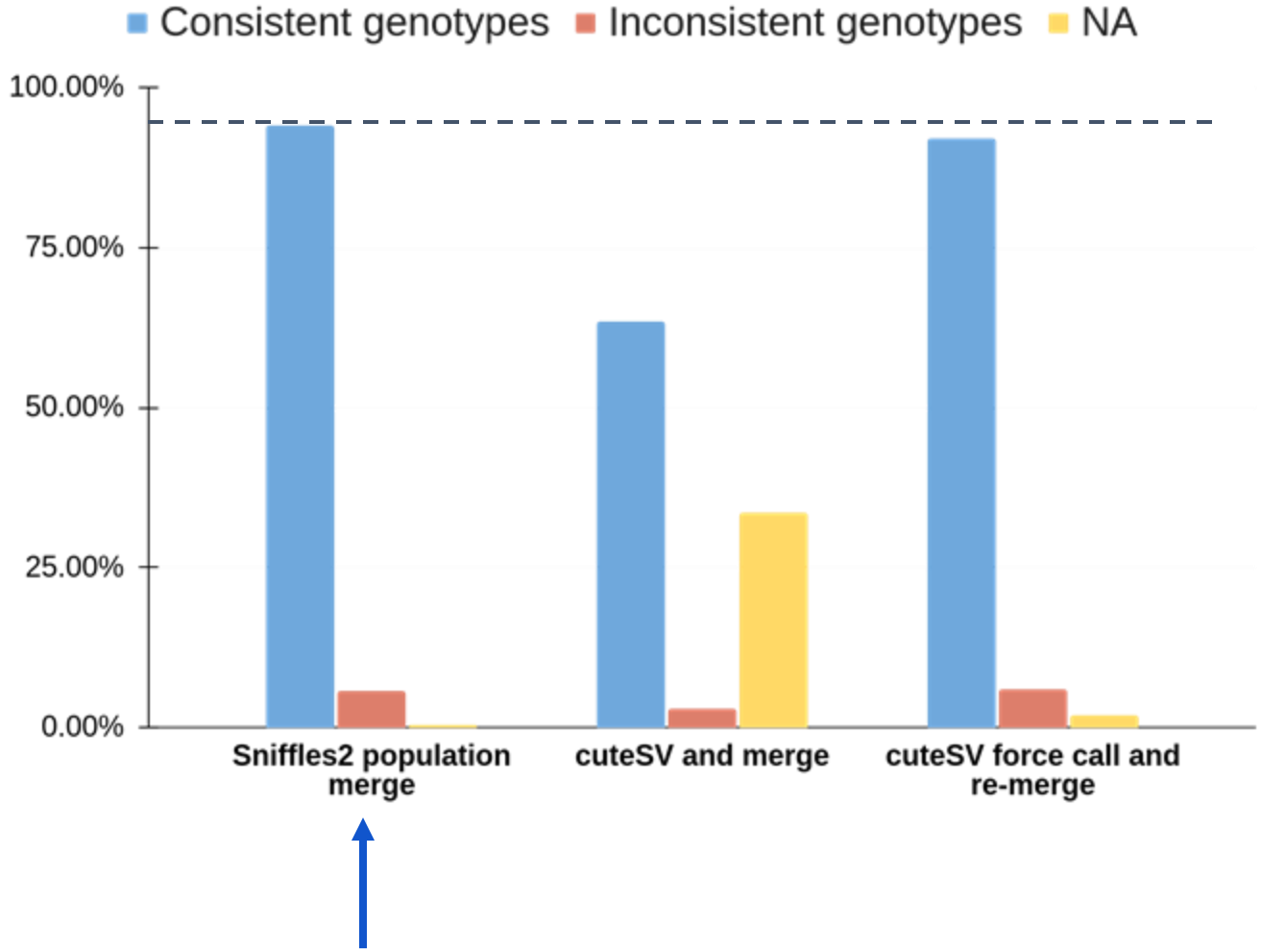
For a family trio*, **Sniffles2**:

- highest fraction of mendelian concordant genotypes (**blue**)
- fewer incomplete genotypes (**yellow**)
- Comparable, yet lower number of non-concordant genotypes (**red**)

Stress test: merging 768 genomes:

15.03 CPU hours

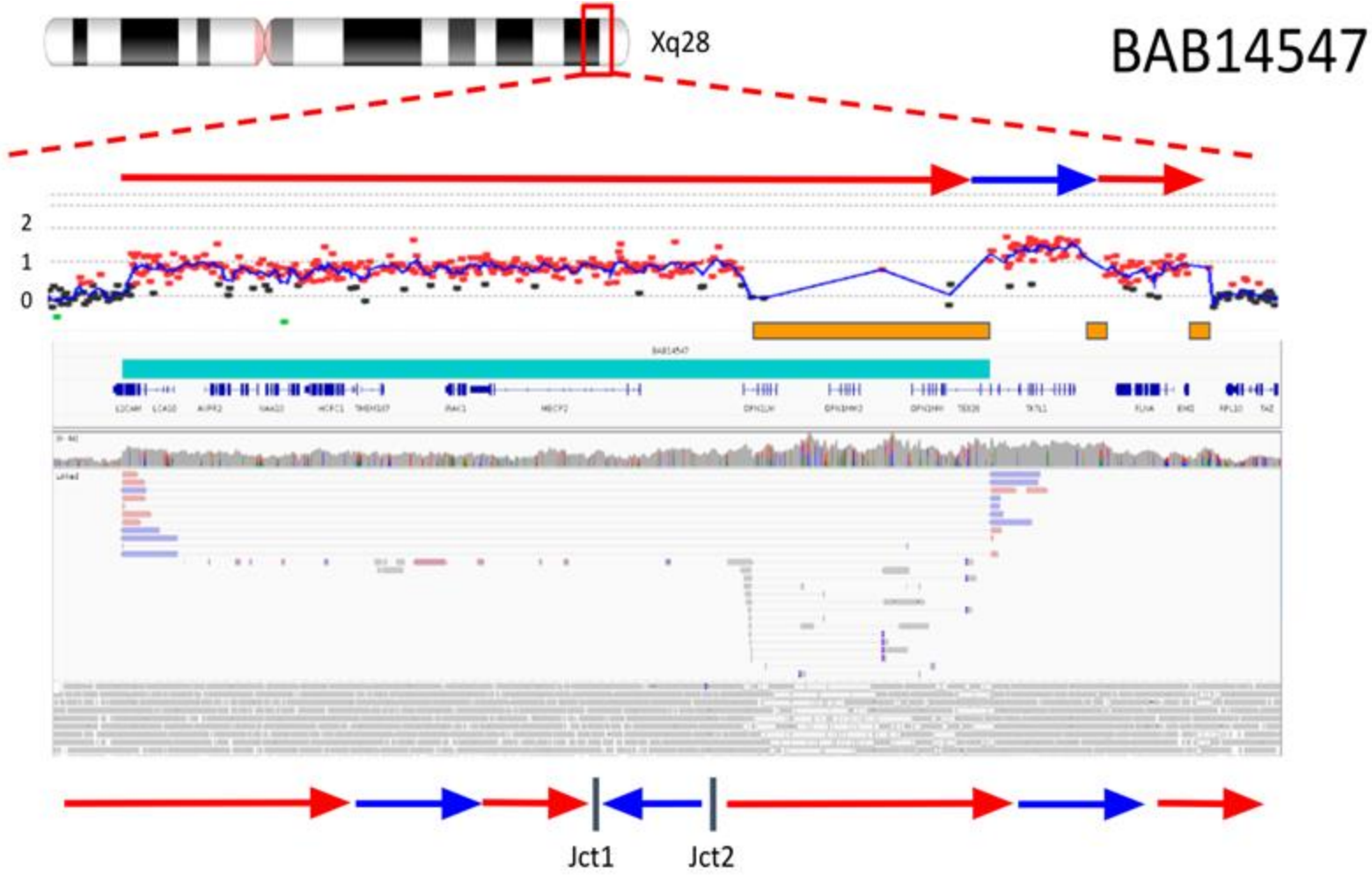
~2x faster than trio with cuteSV



*Benchmark data: HG002/3/4 family trio, ONT.

Sniffles2: Resolving SVs in *MECP2* Duplication Syndrome (MDS)

- *MECP2:* profound neurologic and developmental delay in affected males.
- SV resolution can improve M participant outcomes.



in collaboration with Claudia Carvalho (PNRI)

Annotation of SV Problem..

Closing the gap: Solving complex medically relevant genes at scale

Medhat Mahmoud, John Harting, Holly Corbitt, Xiao Chen, Shalini N. Jhangiani, Harsha Doddapaneni, Qingchang Meng, Tina Han, Christine Lambert, Siyuan Zhang, Primo Baybayan, Geoff Henno, Hua Shen, Jianhong Hu, Yi Han, Casey Riegler, Ginger Metcalf, Geoff Henno, Ivan K. Chinn, Michael A. Eberle, Sarah Kingan, Tim Farinholt, Claudia M.B. Carvalho, Richard A. Gibbs, Zev Kronenberg, Donna Muzny, Fritz J. Sedlazeck

doi: https://doi.org/10.1101/2024.03.14.24304179

- We need better ways to annotate SV with population frequency!
  - HG002 SNV: 99.16% annotatable with Gnomad
  - HG002 SV: 22.80% annotatable in GnomadSV
    - SV in CMRG: 10/217 SV annotatable in GnomadSV
- This hinders variant prioritization!
  - It doesn’t matter how good your calls are.. They are not useable. !?
  - DB are depending on version of caller , filtering , merging of variants..
- Extending STIX for long reads
  - Indexing reads directly instead of VCF files
  - No reference allele bias

Brief Communication | Open access | Published: 08 April 2022

Searching thousands of genomes to classify somatic and novel structural variants using STIX

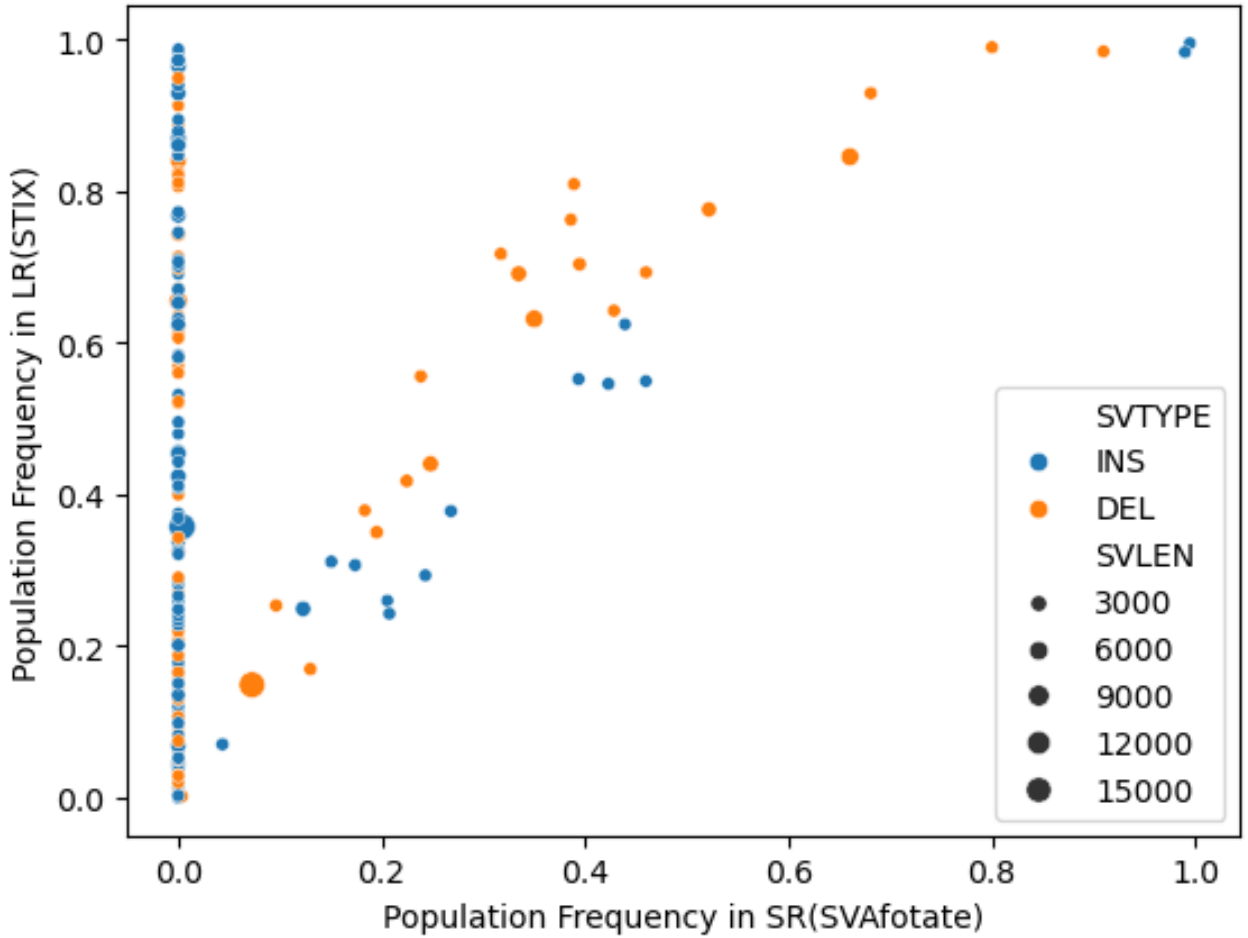
Murad Chowdhury, Brent S. Pedersen, Fritz J. Sedlazeck, Aaron R. Quinlan & Ryan M. Layer

Nature Methods **19**, 445–448 (2022) | Cite this article

1000 genome based annotation

Indexed 1108 ONT data sets that are publicly available.

- Significant & high concordance wit outliers.
- Can detect better SV than 220,000 WGS SR (GnomadSV, etc)
- Even in hard to assess medically relevant genes!



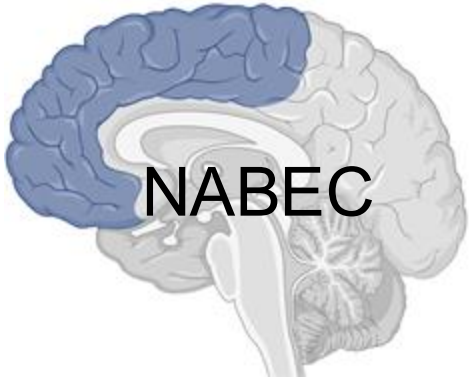
Applications: research groups

- Gregor
  - Solving unsolved mendelian diseases
  - Complex variants in hard to assess regions
- All of Us
  - 1 million Illumina clinical WGS & 2 million arra
  - Report findings back to participants
  - ONT will be applied on a subset for research
- Emirates (G42)
  - 85,000 ONT WGS genomes sequenced
  - Annotation resource
- CARD (NIH)
  - 4,000 brains across neuro dementia

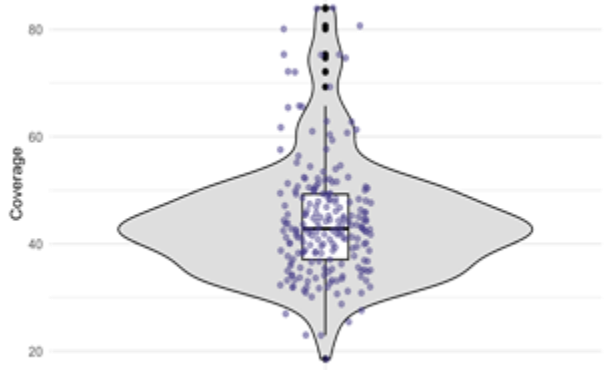
Sequenced hundreds of control human brains across two cohorts

Cohort 1

Sequenced **222** frontal cortex samples **R.9**

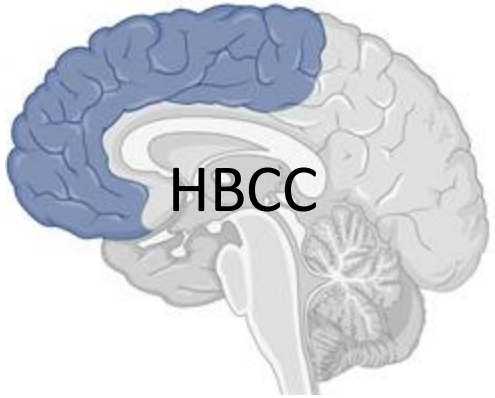


European ancestry
Average Coverage = 44X

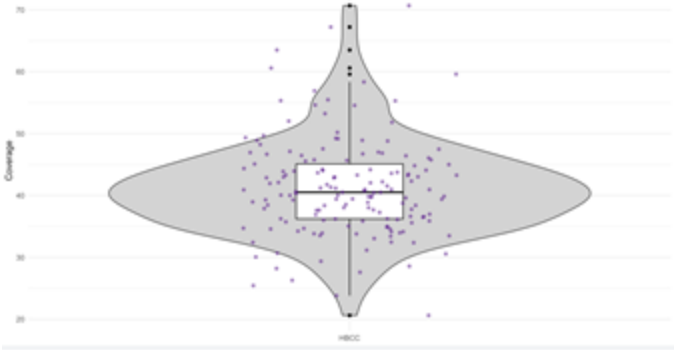


Cohort 2

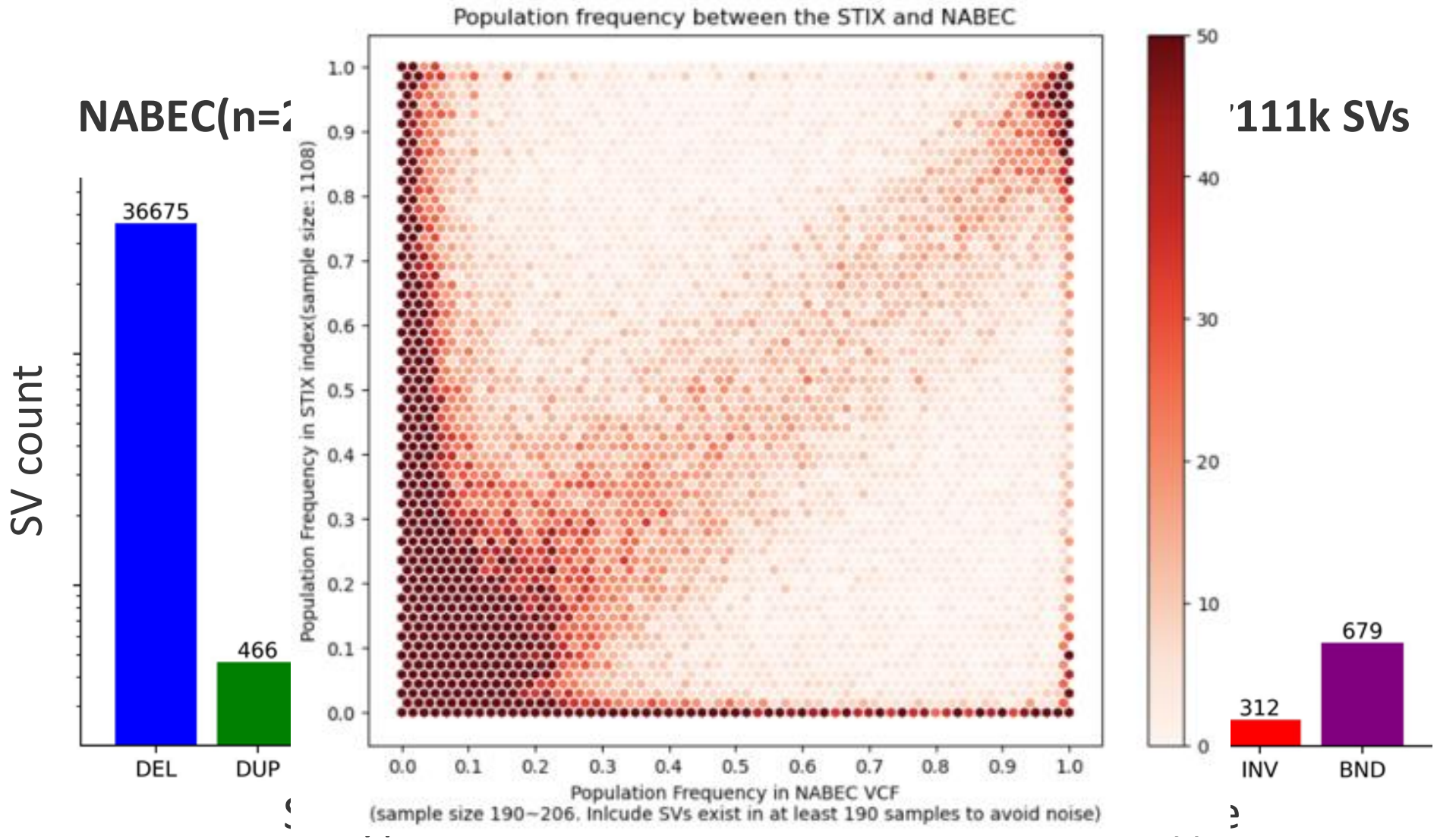
Sequenced **159** frontal cortex samples **R.10**



African ancestry
Average Coverage = 41X



Characterizing structural variation in the human brain

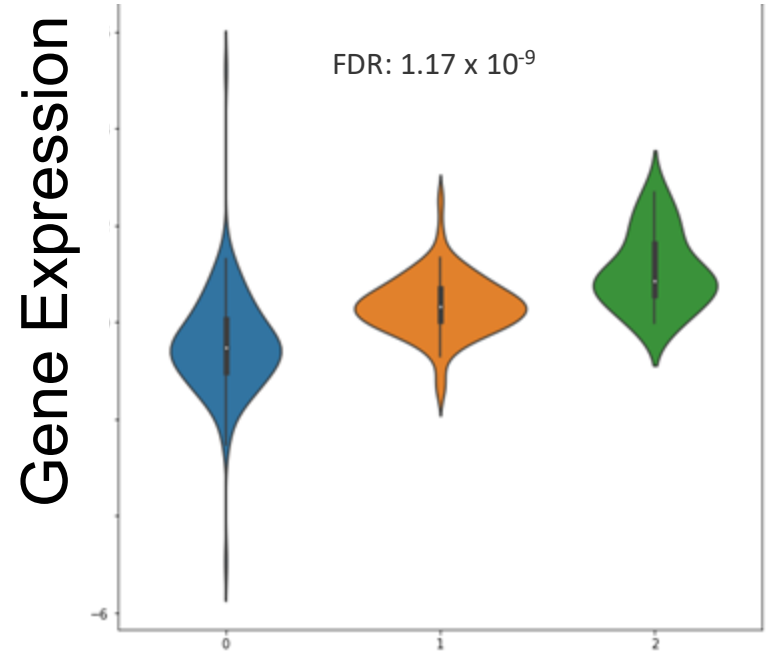


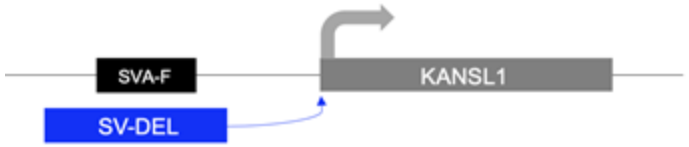
HBCC more SV's = more diverse ancestry + R.10 rather than R.9

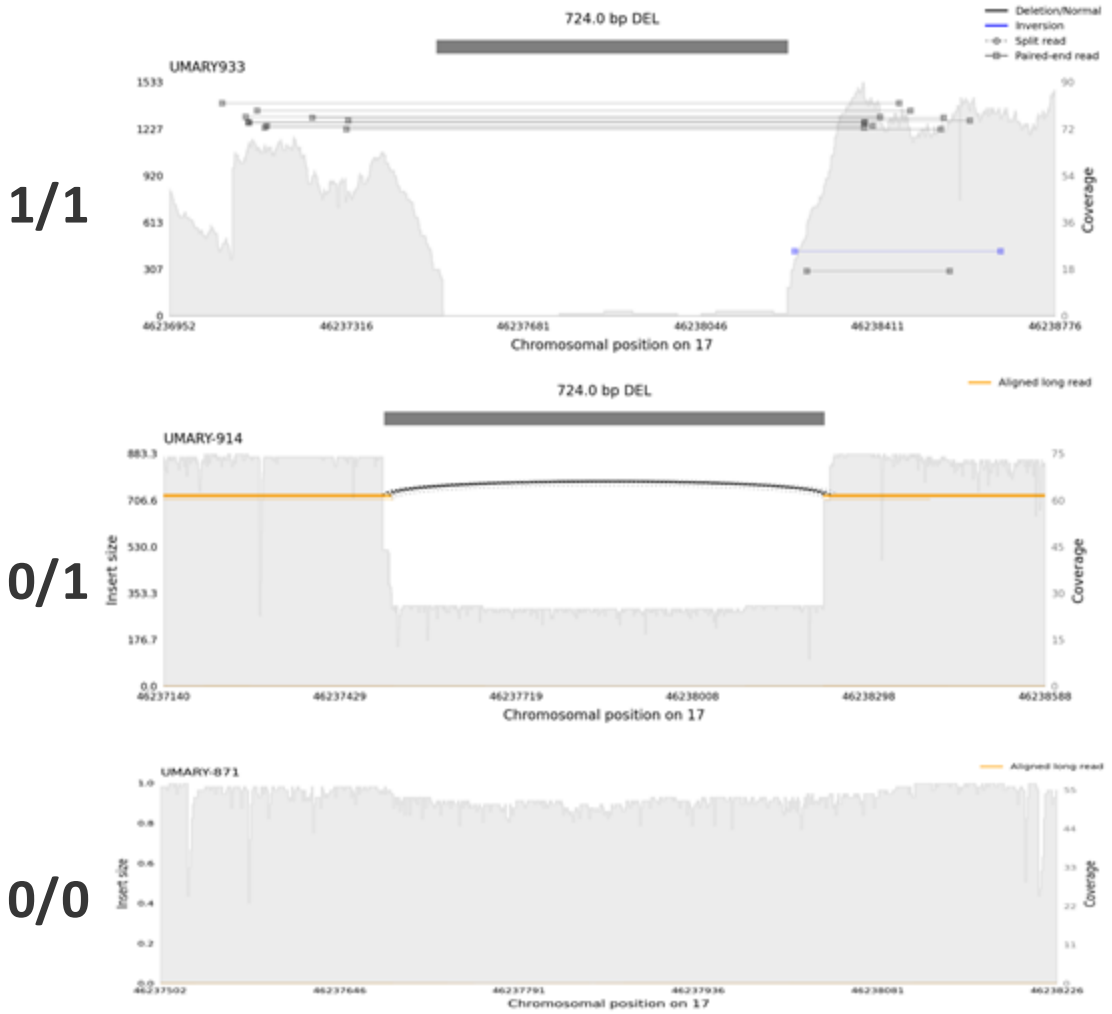
SV-eQTL in the human brain

Kensuke Daida
(CARD-NIH)

A ~700bp deletion is a eQTL for the gene *KANSL1*

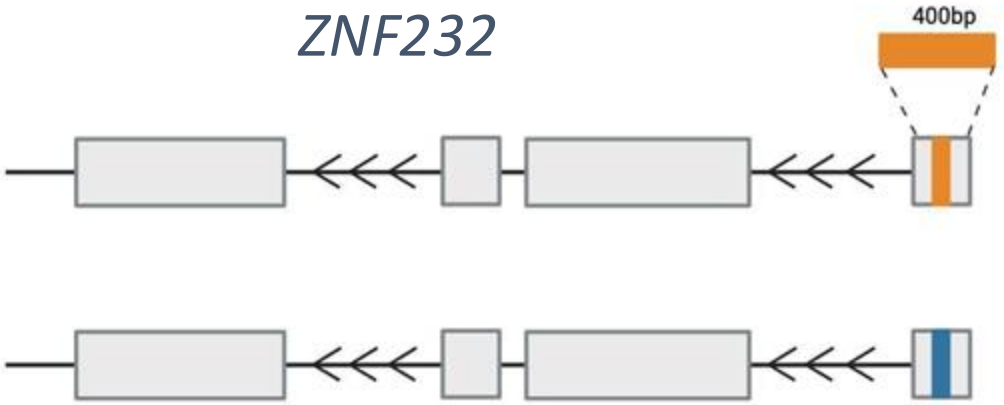


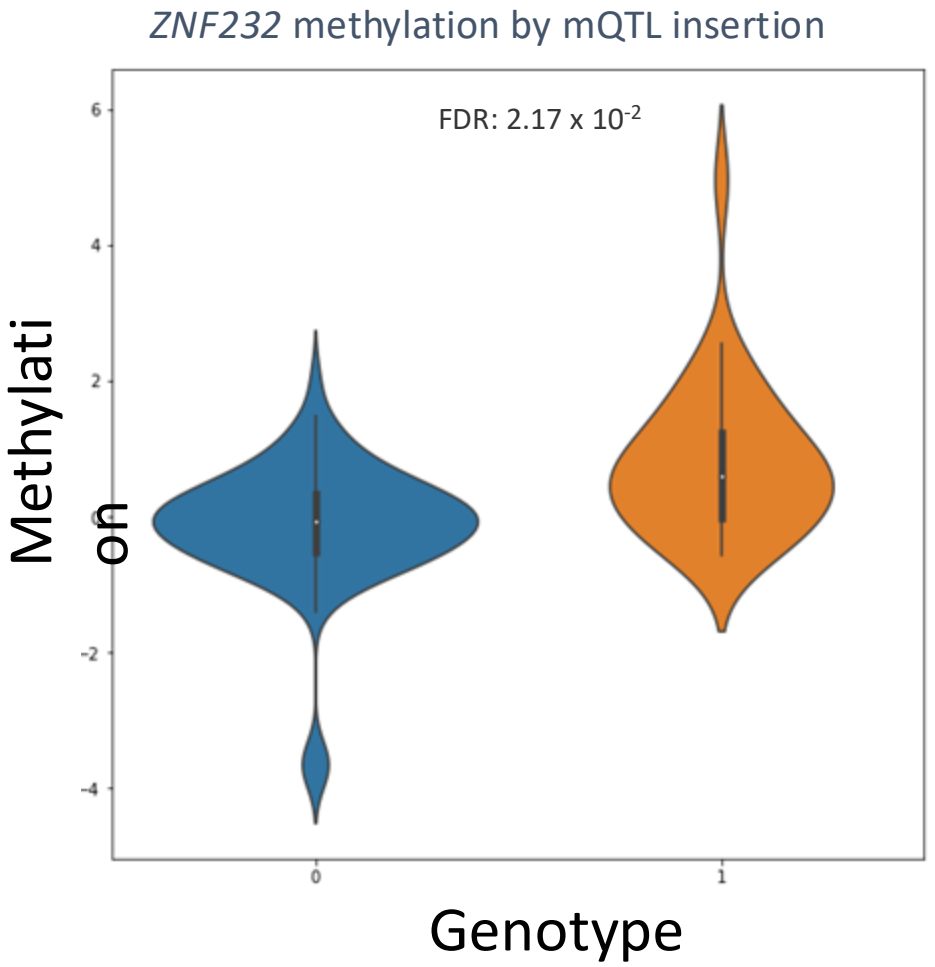




SV- methQTL in promoter of AD-related gene in NABEC brain samples

Kensuke Daida
(CARD-NIH)





Low variant fraction SV?

Somatic SVs and human disease:

→ Neurodegenerative disorders -

accounting for non-heritable disease risk?

→ Cancer drivers (subclonal level)

ELSEVIER

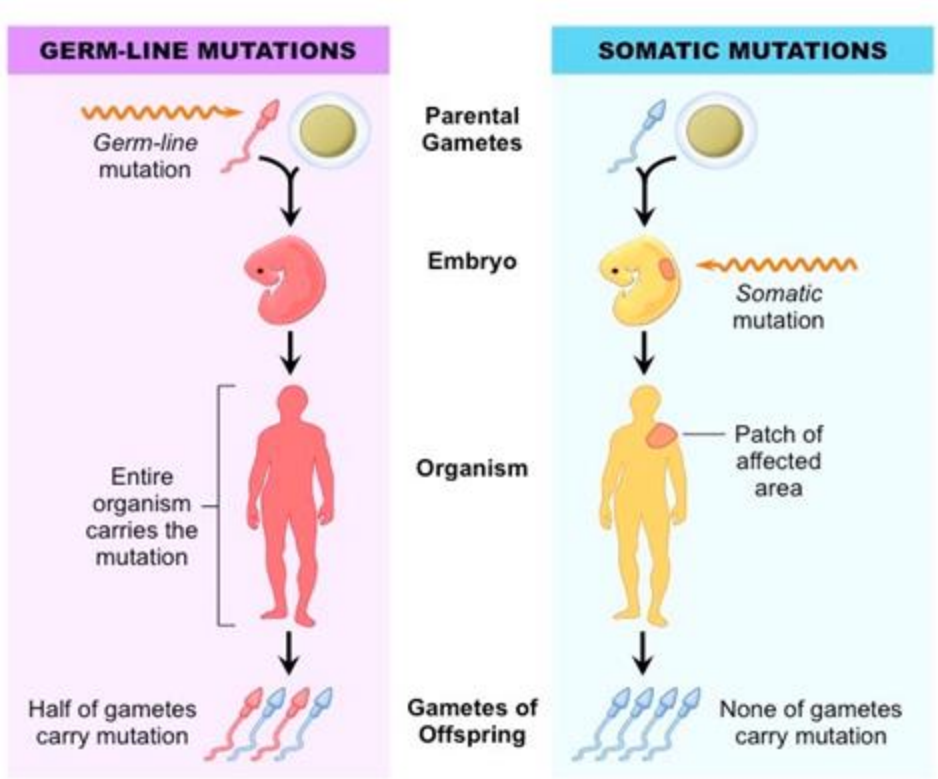
Neurobiology of Disease

Volume 144, October 2020, 105021

Review

Somatic mutations in neurodegeneration: An update

Christos Proukakis ✉



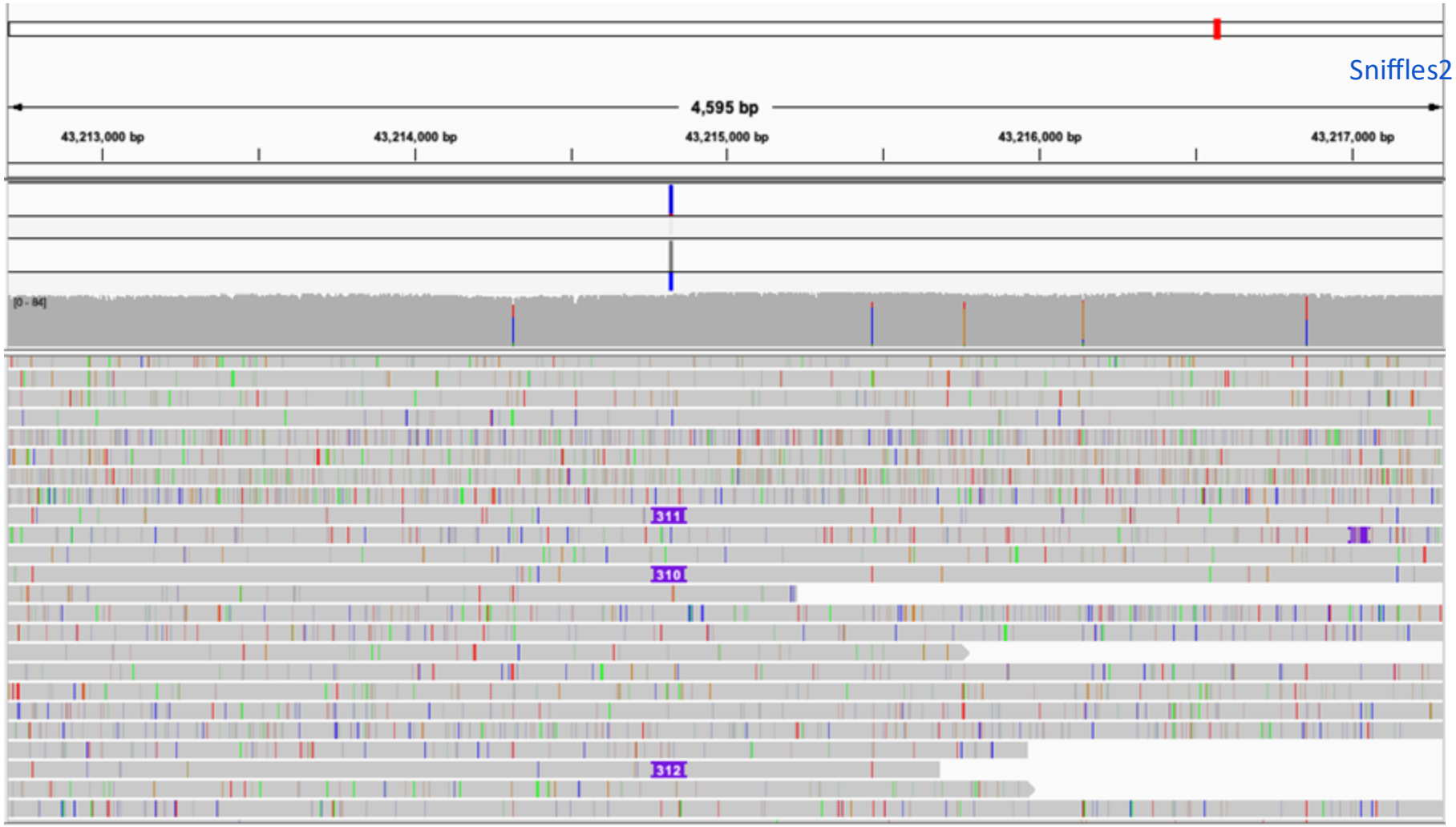
Detecting rare SVs with Sniffles2: Mosaic

Genomic Coordinates

Gold standard (GIAB)

Read Alignments

showing mosaic insertion (300bp)



Data: Real data spike-in mosaicism of HG002 into HG004

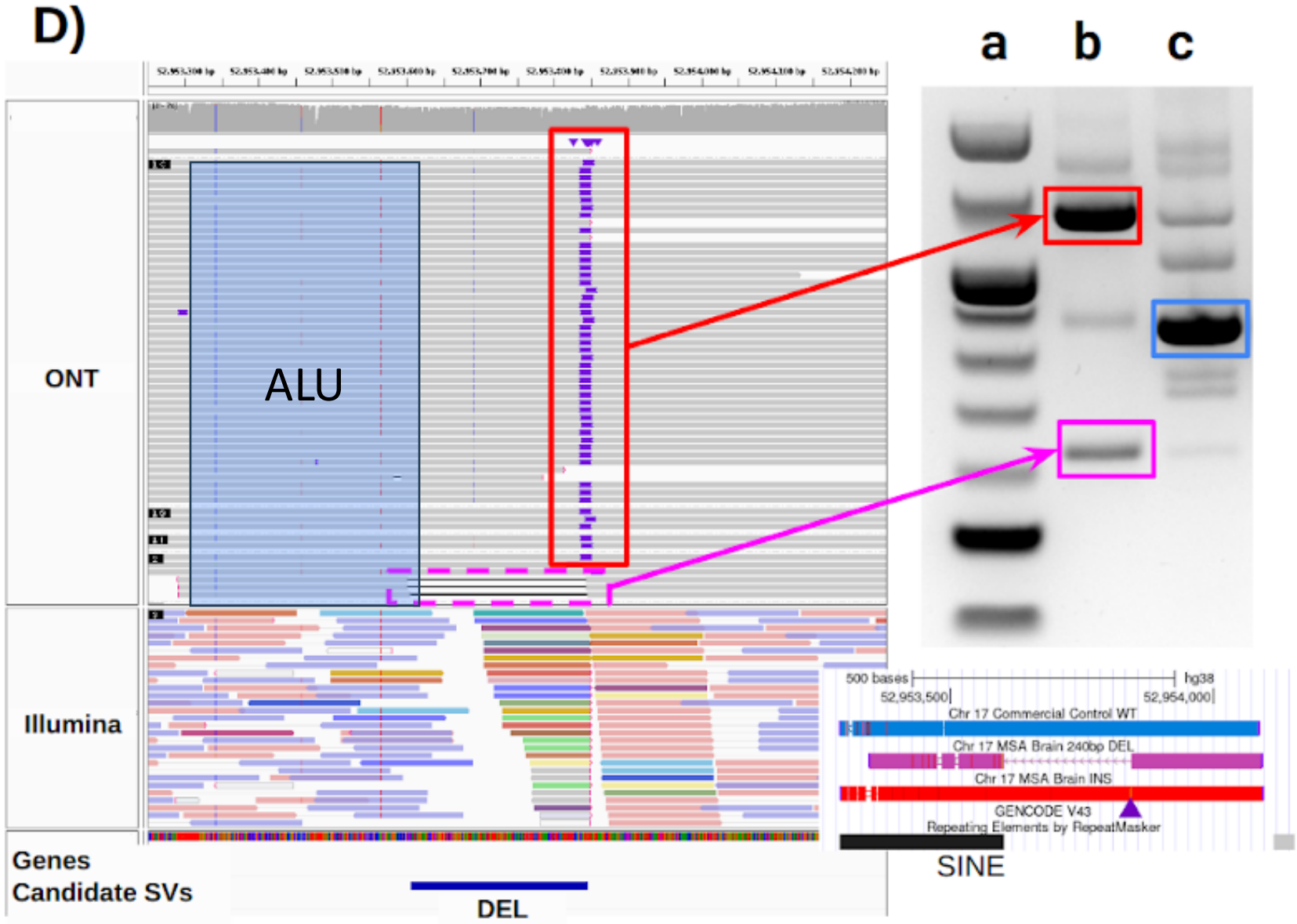
Sniffles 2 mosaic

55x MSA sample:

- Rare neurodegenerative disorder
- Progressive autonomic dysfunction
- Parkinson-like symptoms

•26 Alu -Alu -> mosaic del

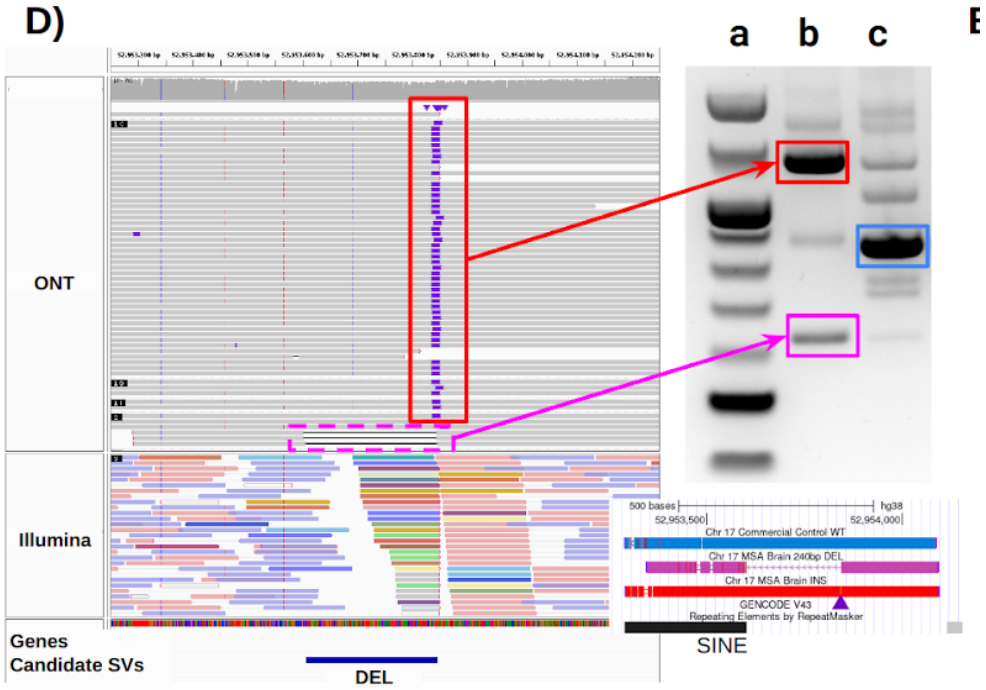
•125 Ins -> mosaic del

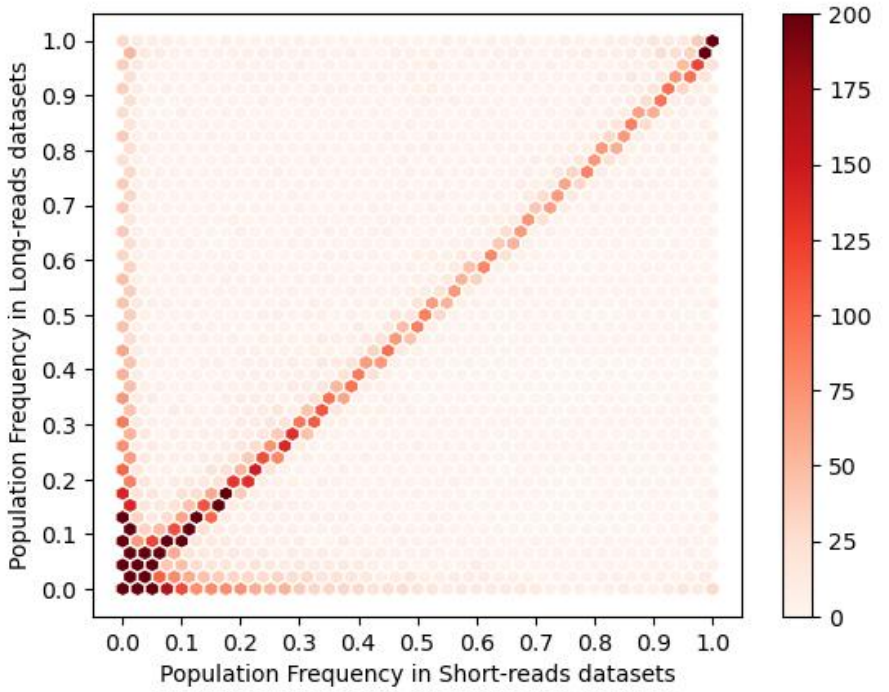


in collaboration with Christos Proukakis (UCL)

Population frequency?

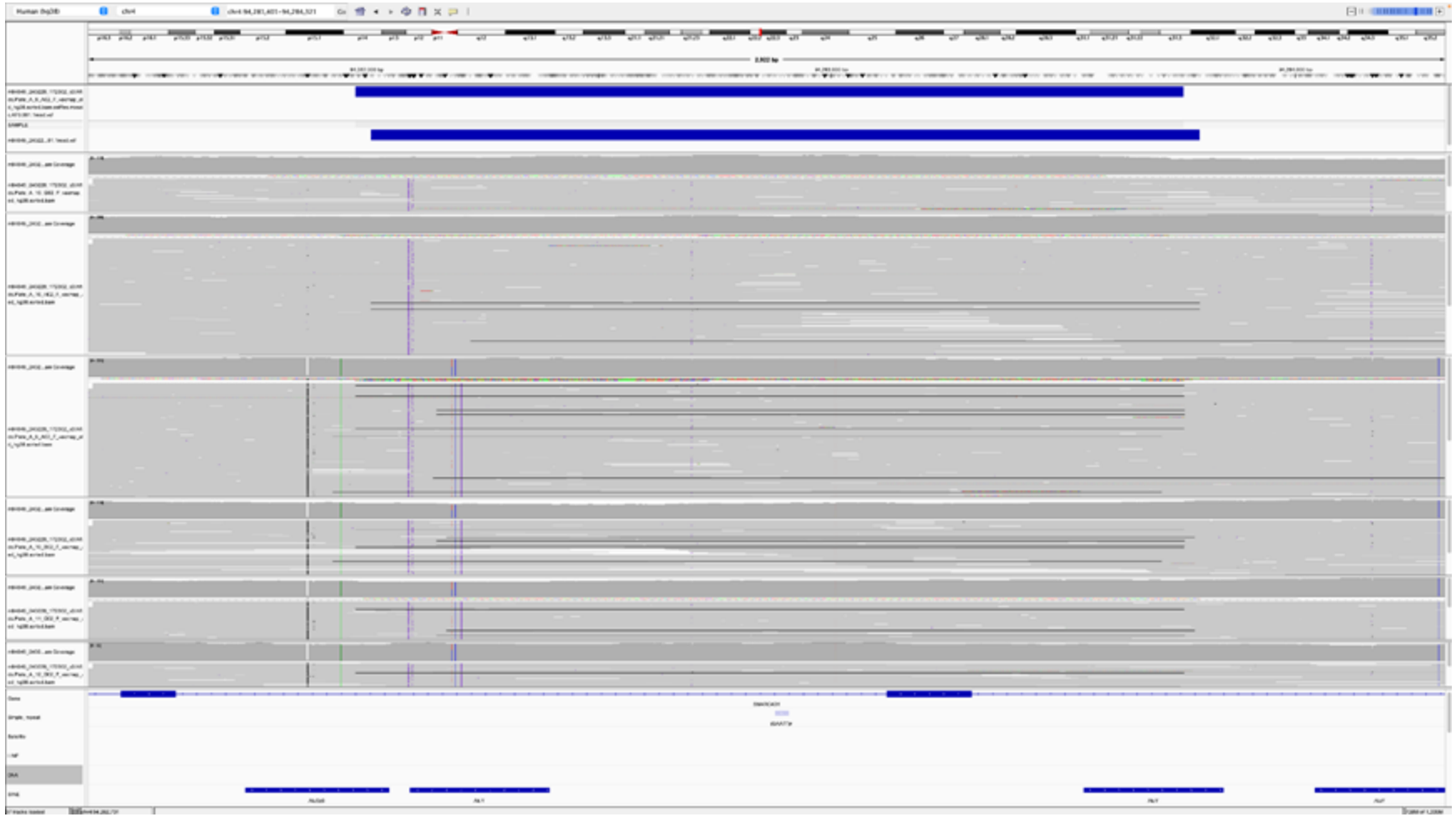
- Alu-Y Insertion: 53.24% in 1KGP
  - Common instability ?
- Mosaic deletion: 2.08% in 1KGP





Another cool example repeat recombination (UCL)

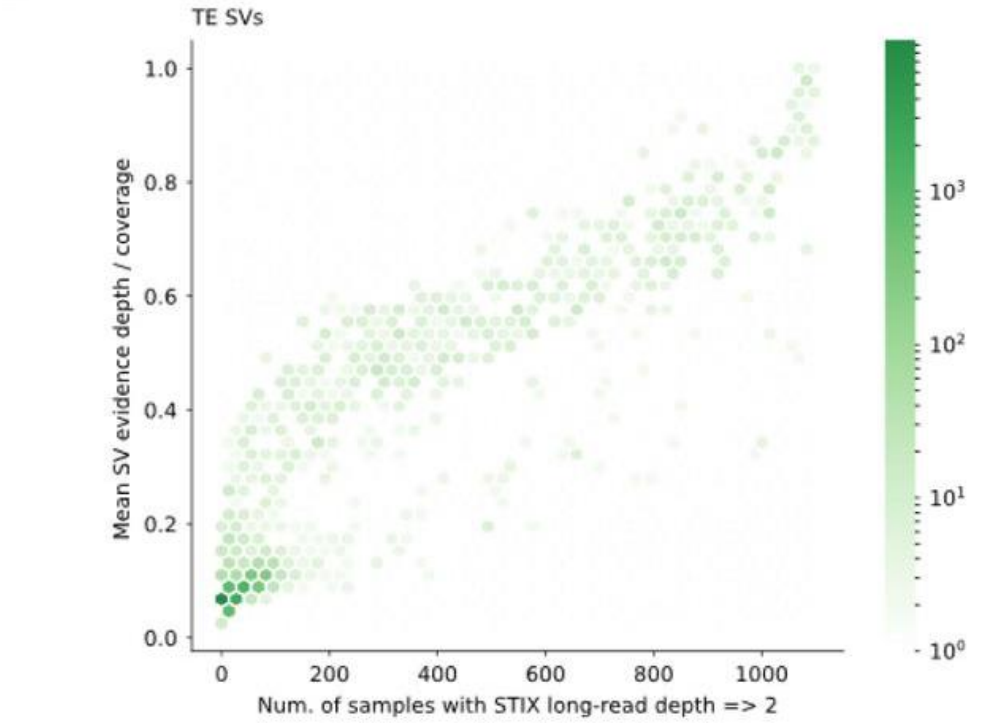
- Pacbio Twist capture
- Multiple regions of 2 MSA brains
- DEL takes out exon of DNA repair gene, which is highly expressed i brain and DEL in both MSA brains
- Christos Proukakis applied to become a SMAHT member but didn't hear back?



Collaboration with Christos Proukakis (UCL)

Are repeat recombinants getting fixed in population?

- Annotation of mosaic TR recombinants from cell paper
- ~10% could be found in 1000g data from STIX
  - Most in low frequency
  - Some rising to fixation/germline.



Applications/Collaborations

- Center for Alzheimer's and Related Dementias (CARD)
  - 4,000 ONT genomes
  - 3 different neurological diseases + 1 control group
  - Variant and epigenetic data resource

- Canada's Michael Smith Genome Sciences Centre (Marathon of Hope)
  - Hundreds of ONT cancer + normal samples
  - Illumina RNA seq

- Genomics England
  - Developing cancer pipeline

- SMAHT

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

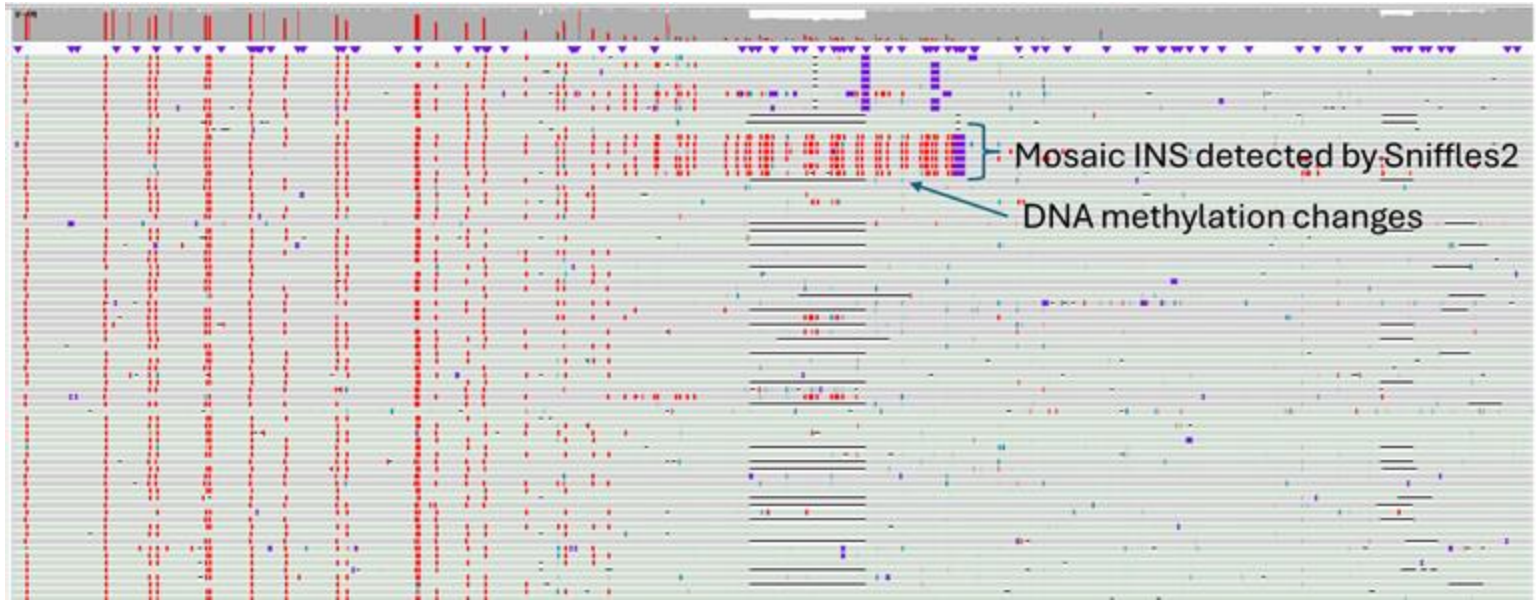
MaHT Somatic Mosaicism across Human Tissues

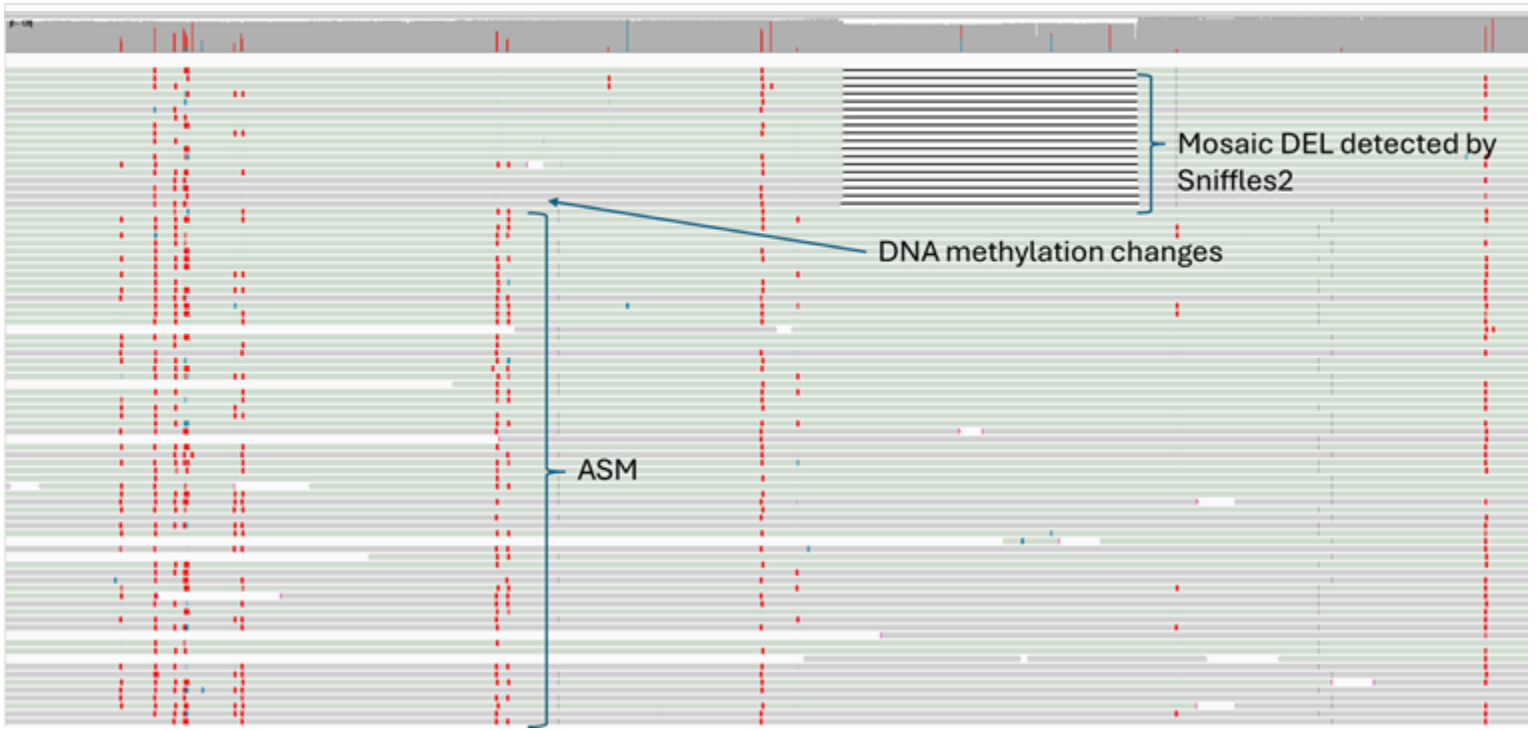
Genomics england

MARATHON OF HOPE CANCER CENTRES NETWORK

Sniff+Meth

- Scanning samples as they become available
- Also looking in Brain data that we have access to.
- Are these SV representing cell types?





Today hands on

- https://github.com/fritzsedlazeck/teaching_material/blob/main/2023_SV_workshop/Day3.md

- We will go over the individual sections.