Methodologies for Structural Variant detection

Fritz Sedlazeck & Luis Paulin

Dec,13, 2023

@sedlazeck

RICE

Recap from yesterday

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*]

Sniffles: HG002 v0.6 and CMRG

nature biotechnology

RESOURCE

https://doi.org/10.1038/s41587-020-0538-8

Check for updates

A robust benchmark for detection of germline large deletions and insertions

Justin M. Zook[1]✉, Nancy F. Hansen[2], Nathan D. Olson[1], Lesley Chapman[1], James C. Mullikin[2], Chunlin Xiao[3], Stephen Sherry[3], Sergey Koren[2], Adam M. Phillippy[2], Paul C. Boutros[4], Sayed Mohammad E. Sahraeian[5], Vincent Huang[6], Alexandre Rouette[7], Noah Alexander[8], Christopher E. Mason[9,10,11,12], Iman Hajirasouliha[9], Camir Ricketts[9], Joyce Lee[13], Rick Tearle[14], Ian T. Fiddes[15], Alvaro Martinez Barrio[15], Jeremiah Wala[16], Andrew Carroll[17], Noushin Ghaffari[18], Oscar L. Rodriguez[19], Ali Bashir[19], Shaun Jackman[20], John J. Farrell[21], Aaron M. Wenger[22], Can Alkan[23], Arda Soylev[24], Michael C. Schatz[25], Shilpa Garg[26], George Church[26], Tobias Marschall[27], Ken Chen[28], Xian Fan[29], Adam C. English[30], Jeffrey A. Rosenfeld[31,32], Weichen Zhou[33], Ryan E. Mills[33], Jay M. Sage[34], Jennifer R. Davis[34], Michael D. Kaiser[34], John S. Oliver[34], Anthony P. Catalano[34], Mark J. P. Chaisson[35], Noah Spies[36], Fritz J. Sedlazeck[37] and Marc Salit[36]

ARTICLES

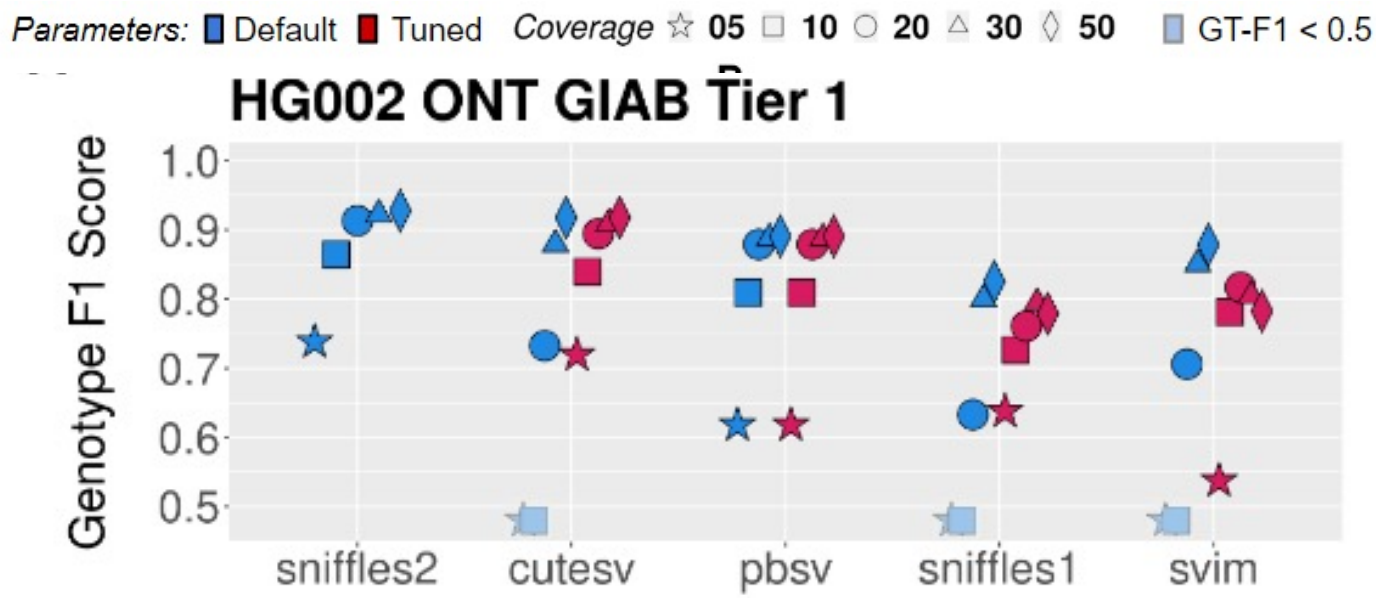
https://doi.org/10.1038/s41587-021-01158-1

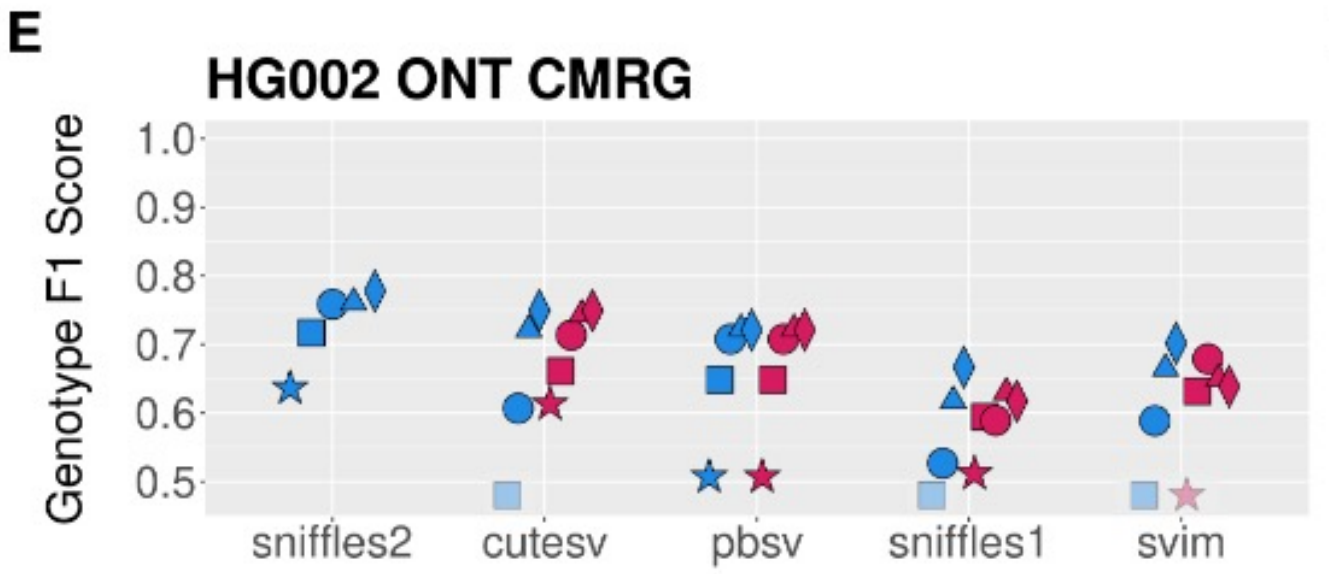
nature biotechnology

Check for updates

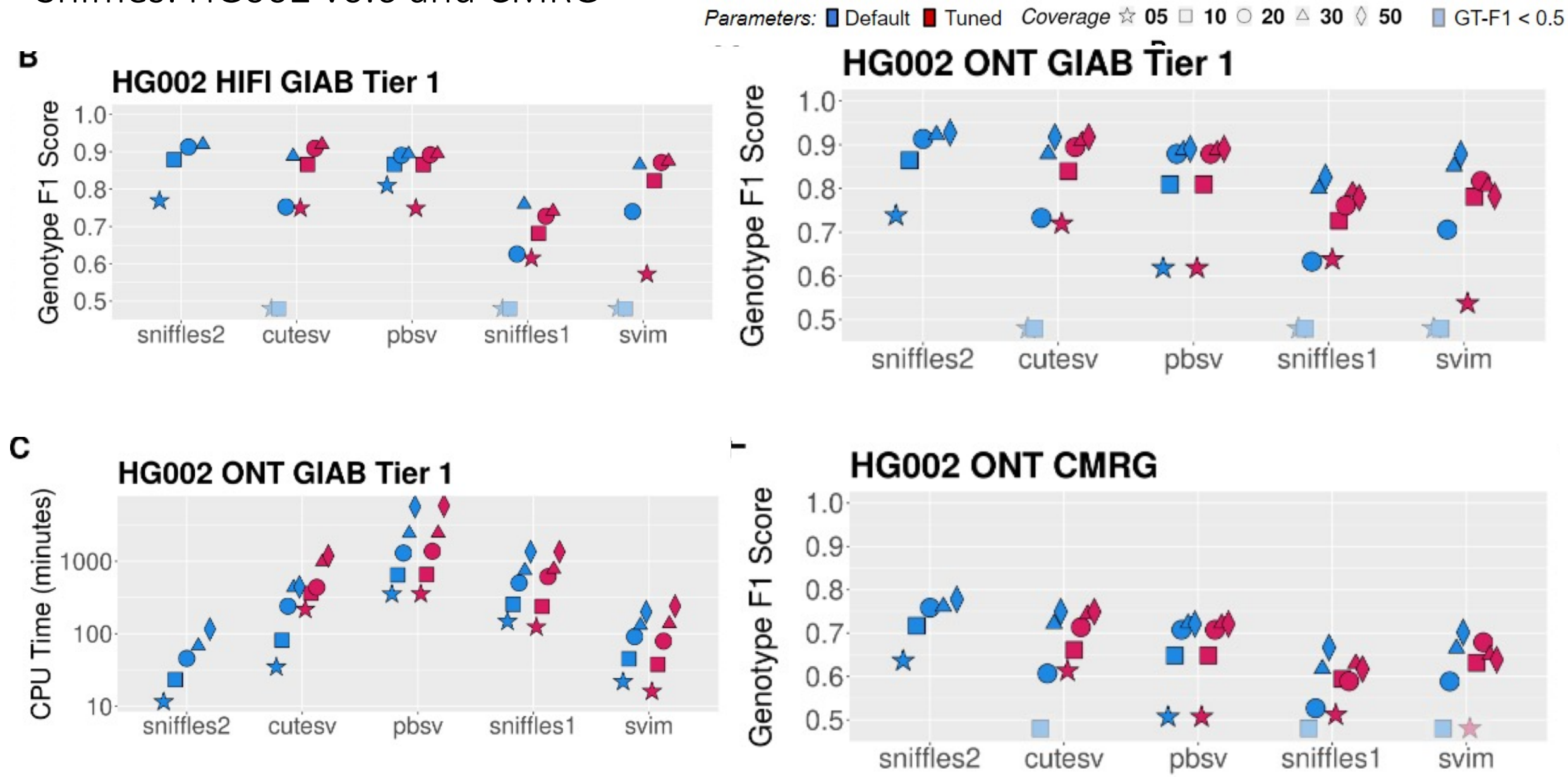
Curated variation benchmarks for challenging medically relevant autosomal genes

Justin Wagner[1], Nathan D. Olson[1], Lindsay Harris[1], Jennifer McDaniel[1], Haoyu Cheng[2], Arkarachai Fungtammasan[3], Yih-Chii Hwang[3], Richa Gupta[3], Aaron M. Wenger[4], William J. Rowell[4], Ziad M. Khan[5], Jesse Farek[5], Yiming Zhu[5], Aishwarya Pisupati[5], Medhat Mahmoud[5], Chunlin Xiao[6], Byunggil Yoo[7], Sayed Mohammad Ebrahim Sahraeian[8], Danny E. Miller[9,10], David Jáspez[11], José M. Lorenzo-Salazar[11], Adrián Muñoz-Barrera[11], Luis A. Rubio-Rodríguez[11], Carlos Flores[11,12,13], Giuseppe Narzisi[14], Uday Shanker Evani[14], Wayne E. Clarke[14], Joyce Lee[15], Christopher E. Mason[16], Stephen E. Lincoln[17], Karen H. Miga[18], Mark T. W. Ebbert[19,20,21], Alaina Shumate[22,23], Heng Li[2], Chen-Shan Chin[3,24]✉, Justin M. Zook[1,24]✉ and Fritz J. Sedlazeck[5,24]✉





Sniffles: HG002 v0.6 and CMRG



New Applications in SV detection

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

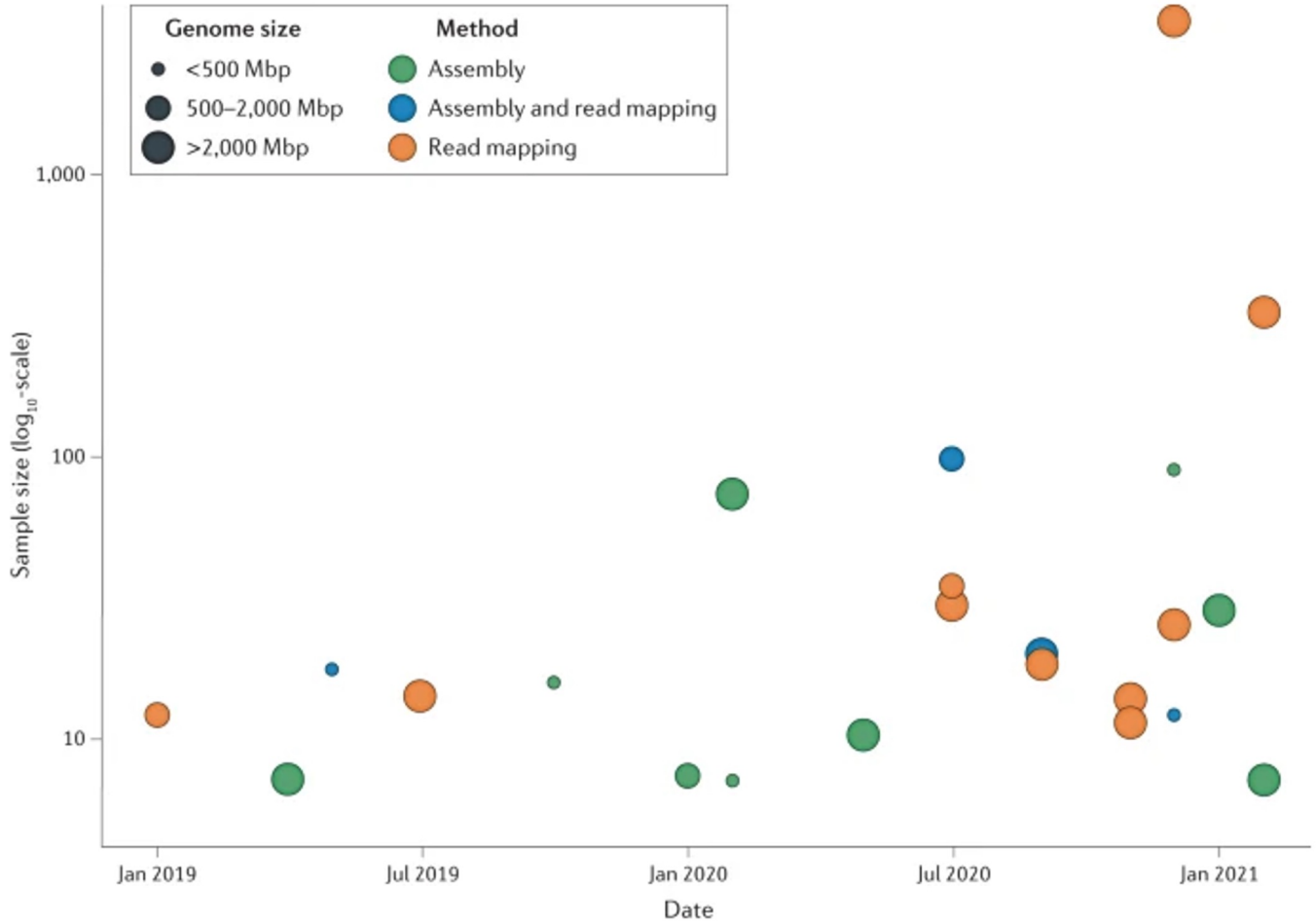
nature › nature reviews genetics › review articles › article

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.



Full SV Genotyping: From family to population scale

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

~36 CPU hours

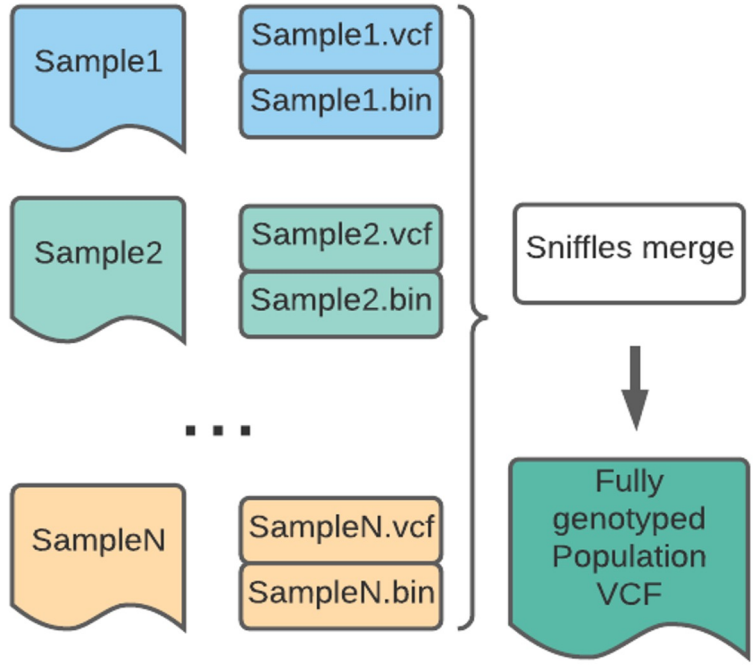
Sniffles2: SV calling → merging → population VCF

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

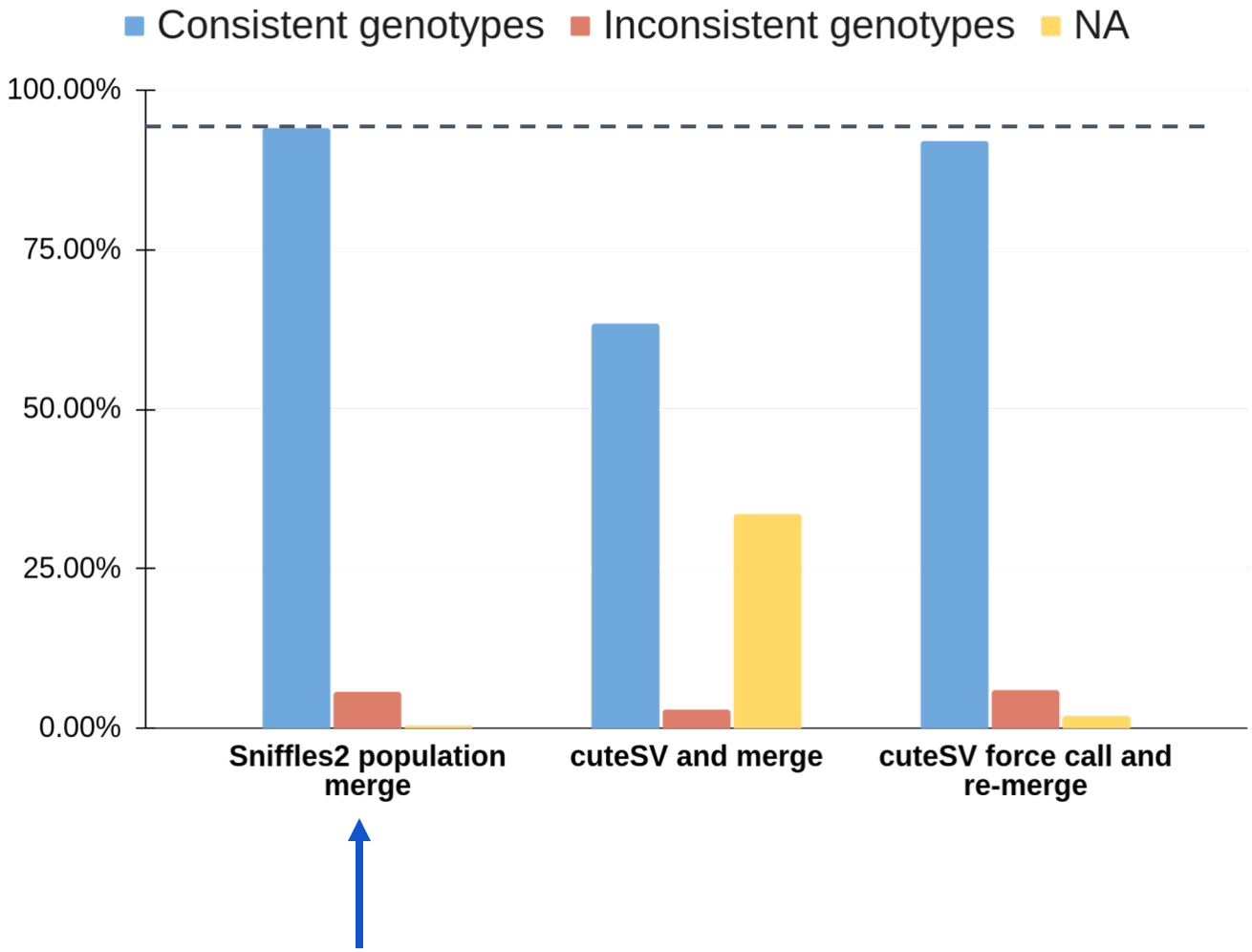
For a family trio*, **Sniffles2**:

- 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

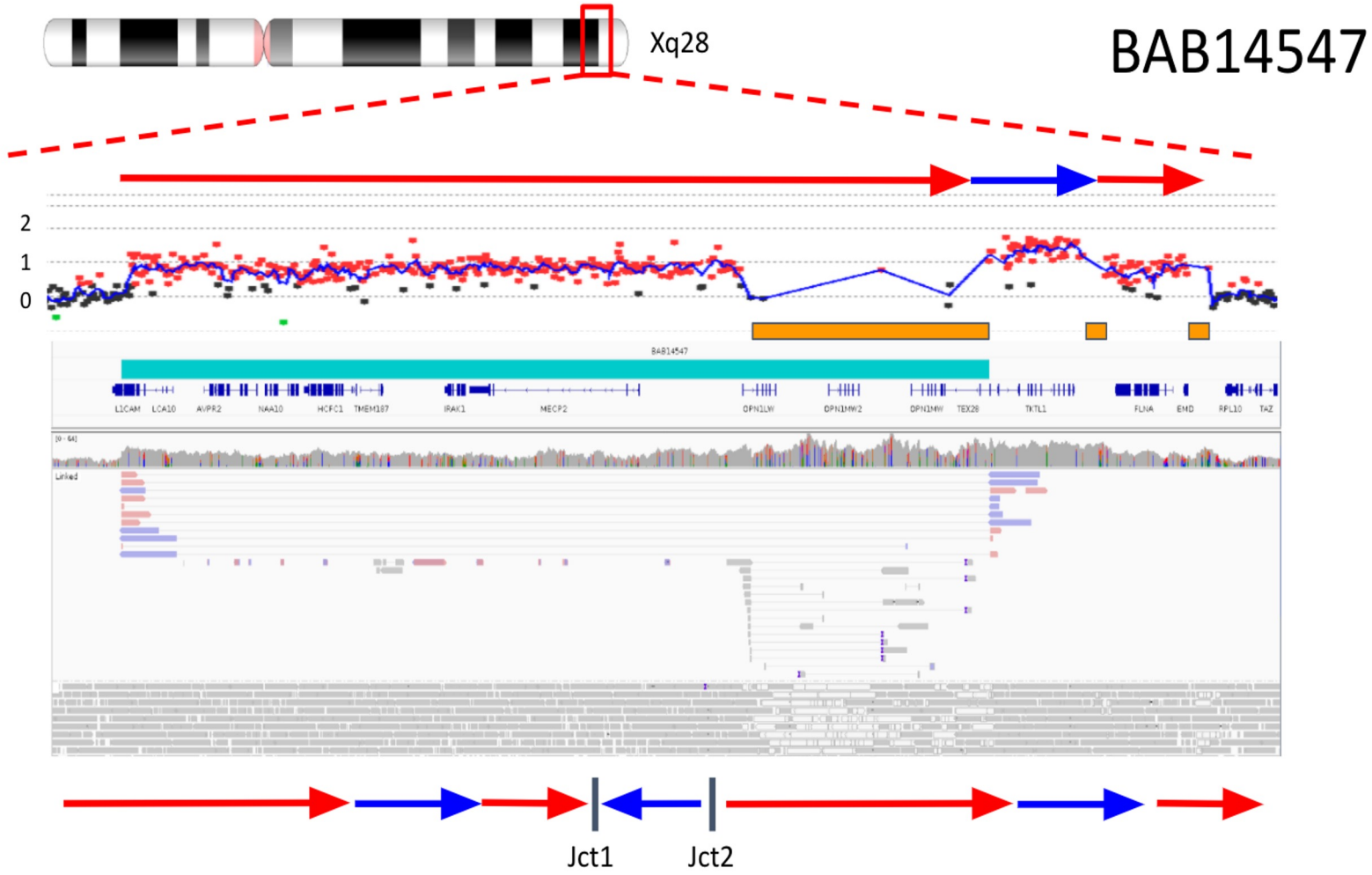
~2x faster than trio with cuteSV



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

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

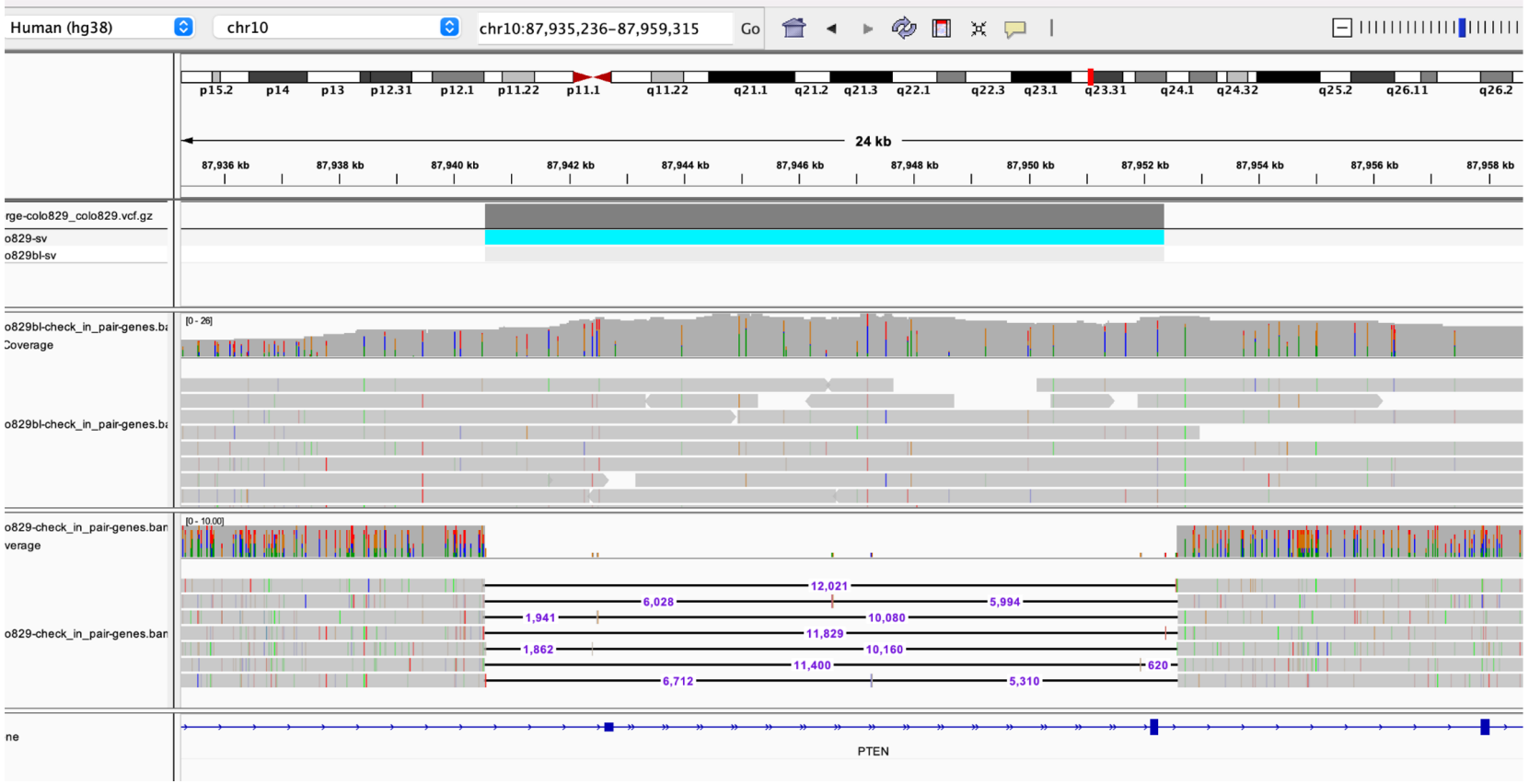
- *MECP2:* profound neurologic and developmental delay in affected males.
- SV resolution can improve M participant outcomes.



in collaboration with Claudia Carvalho (PNRI)

Tumor vs normal: colo829

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



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

New Applications in SV detectio

1. Germline SV
2. Population scale
3. **Somatic SVs and human disease:**

   → Neurodegenerative disorders -

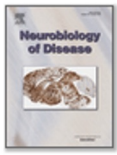
   <u>accounting for non-heritable disease risk?</u>

   → Cancer drivers (subclonal level)

ELSEVIER

Neurobiology of Disease

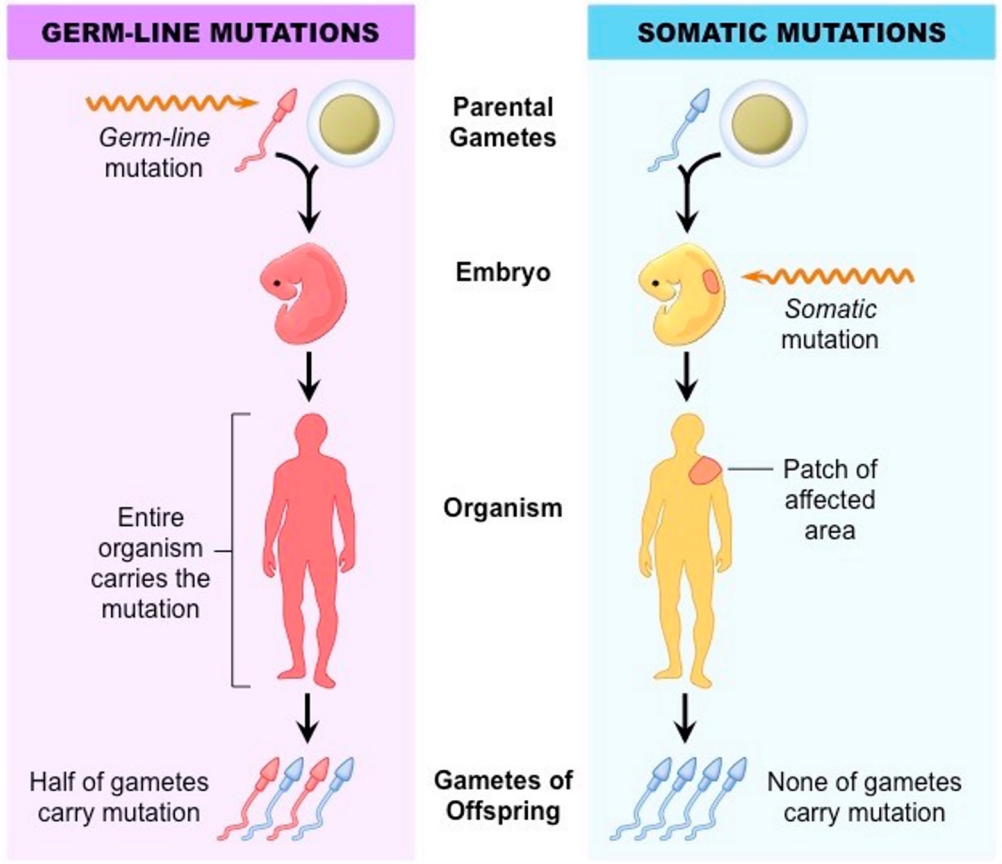
Volume 144, October 2020, 105021



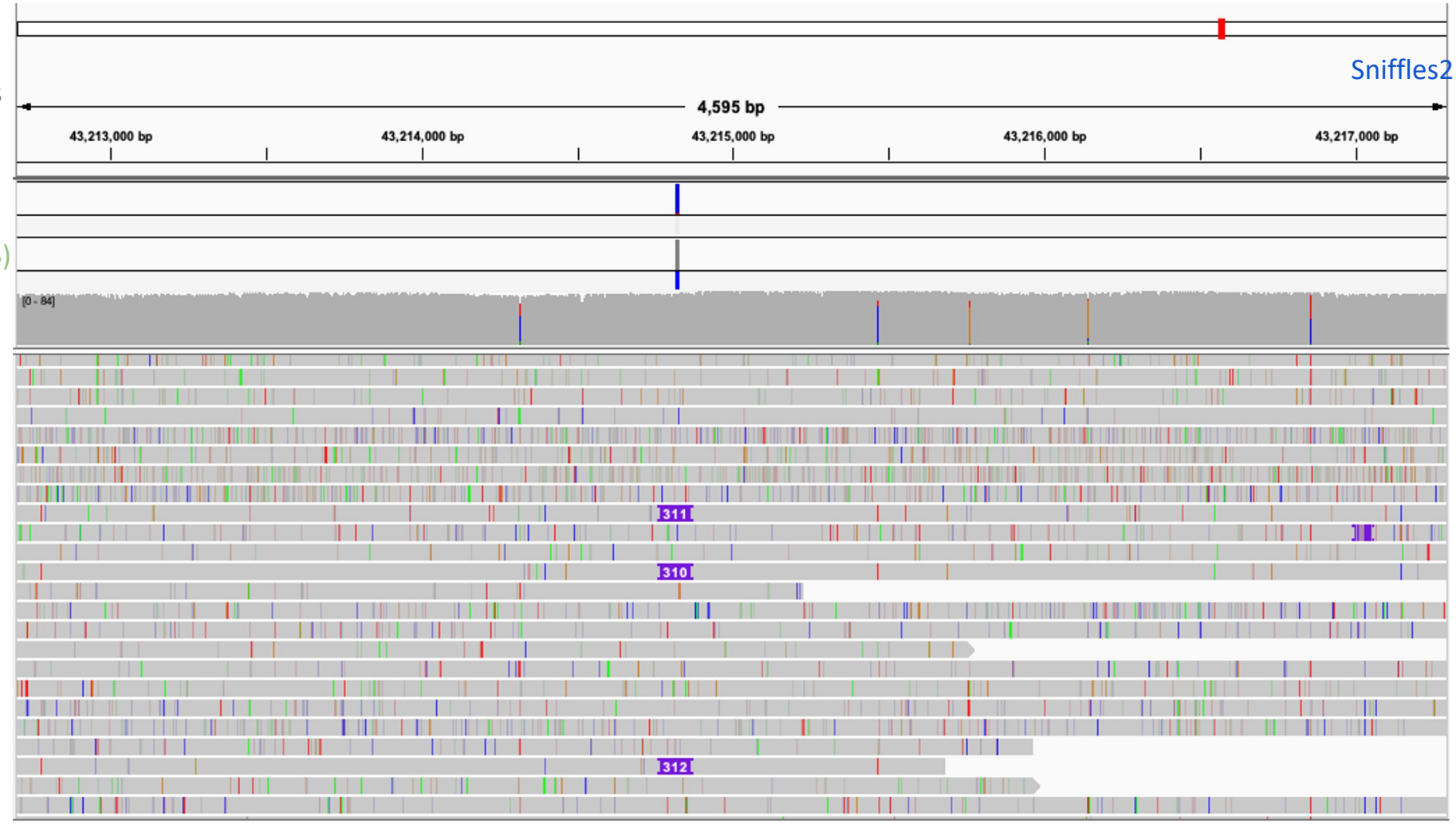
Review

Somatic mutations in neurodegeneration: An update

Christos Proukakis ✉



Detecting rare SVs with Sniffles2: Mosaic

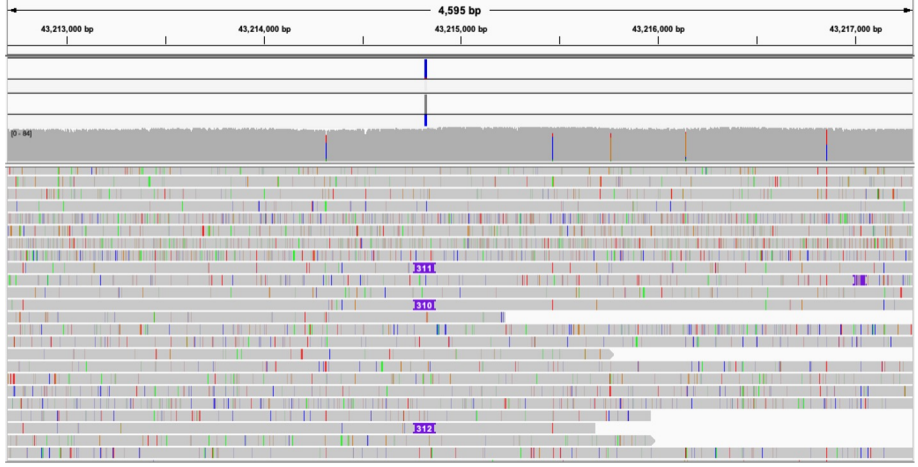


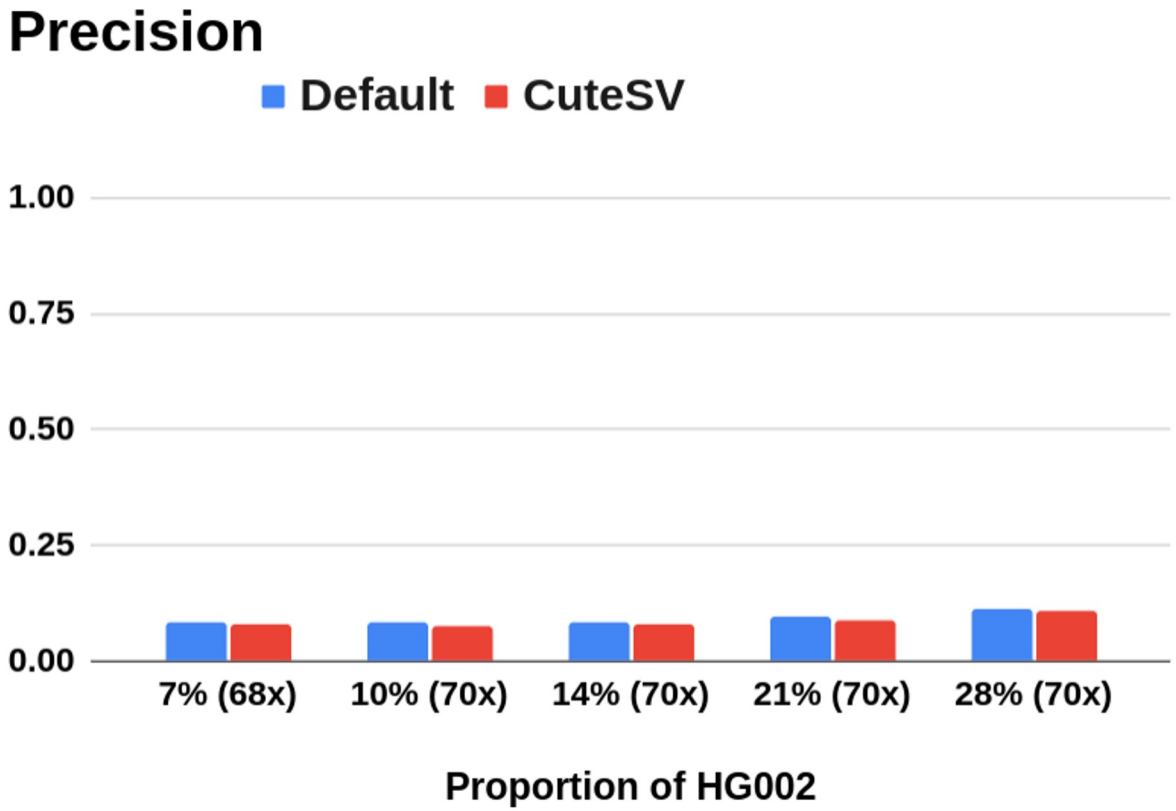
showing mosaic insertion (300bp)

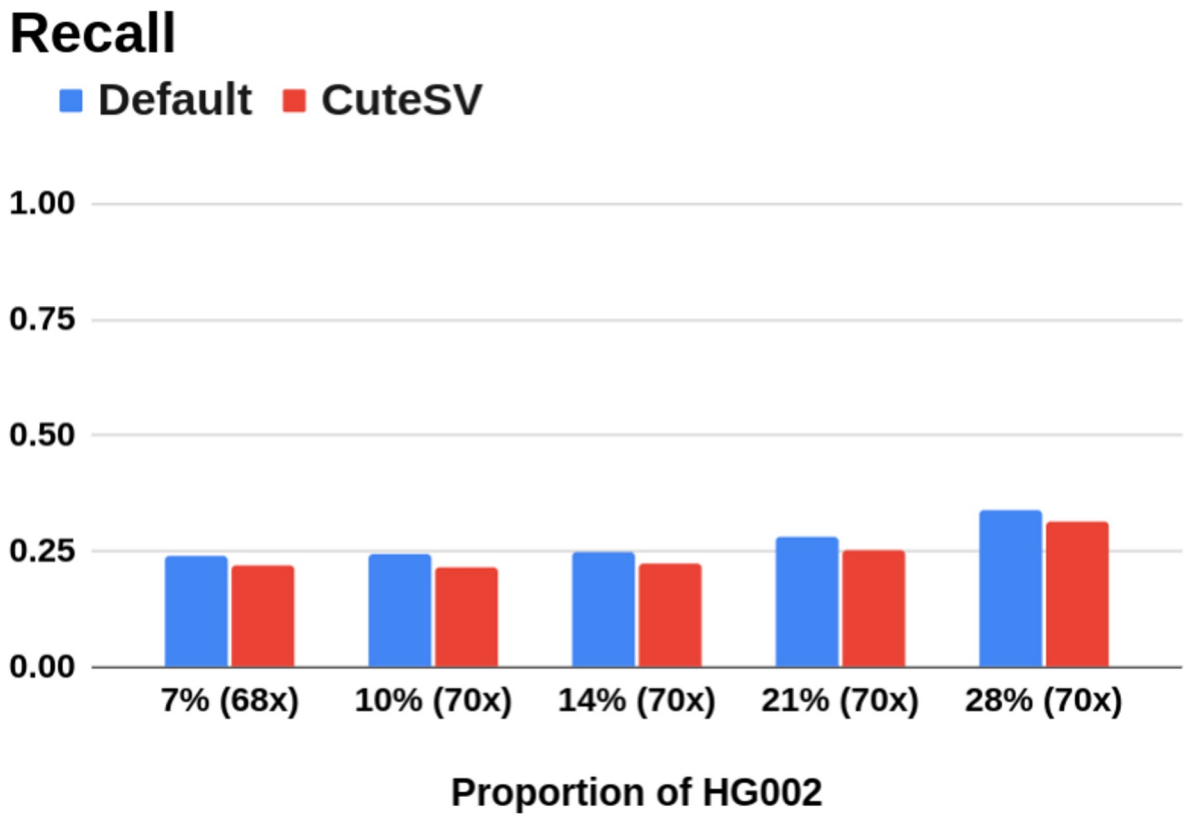
Data: Real data spike-in mosaicism of HG002 into HG004

Sniffles 2 mosaic

HG002 (low coverage) HG00733 (high coverage)

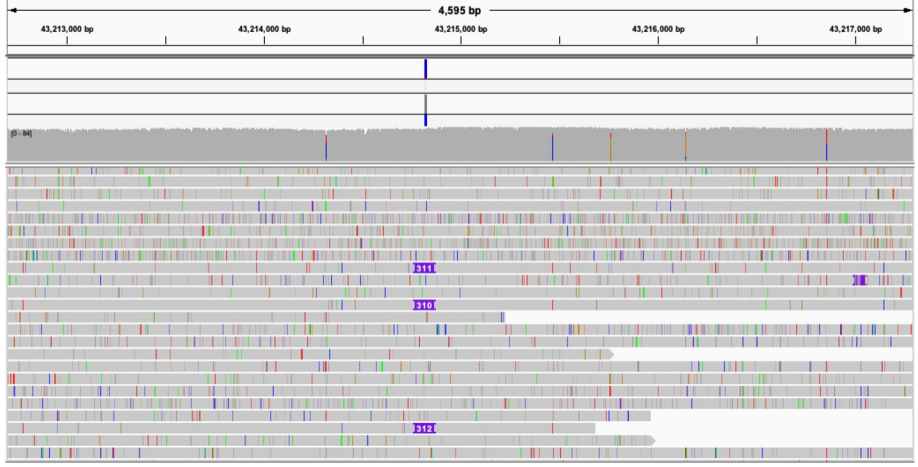


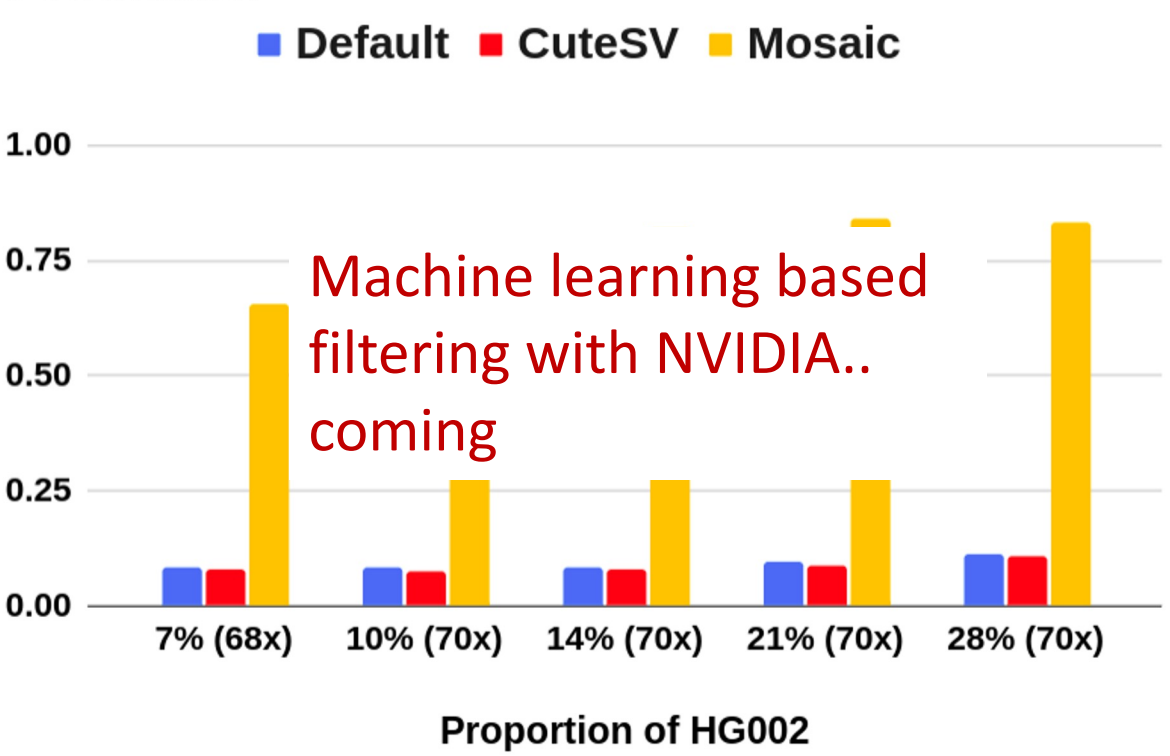


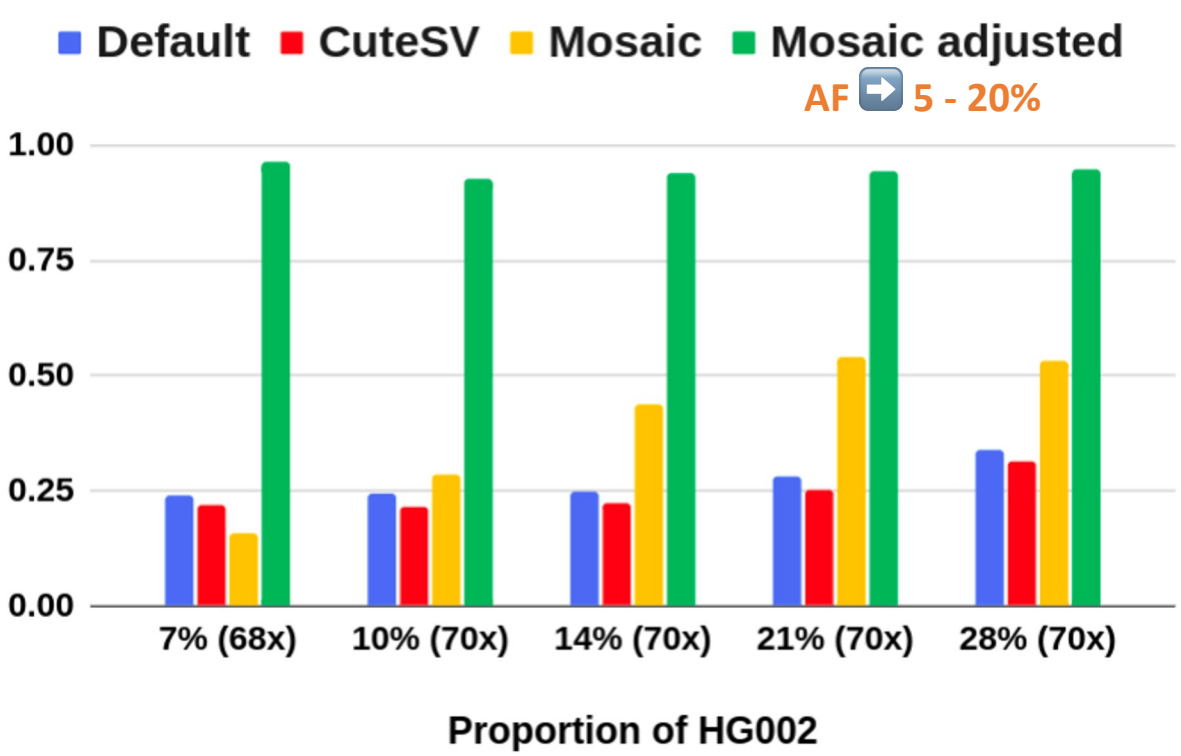


Sniffles 2 mosaic

HG002 (low coverage) HG00733 (high coverage)
Mosaic mode identifies rare SVs (down to 2-3 reads only)







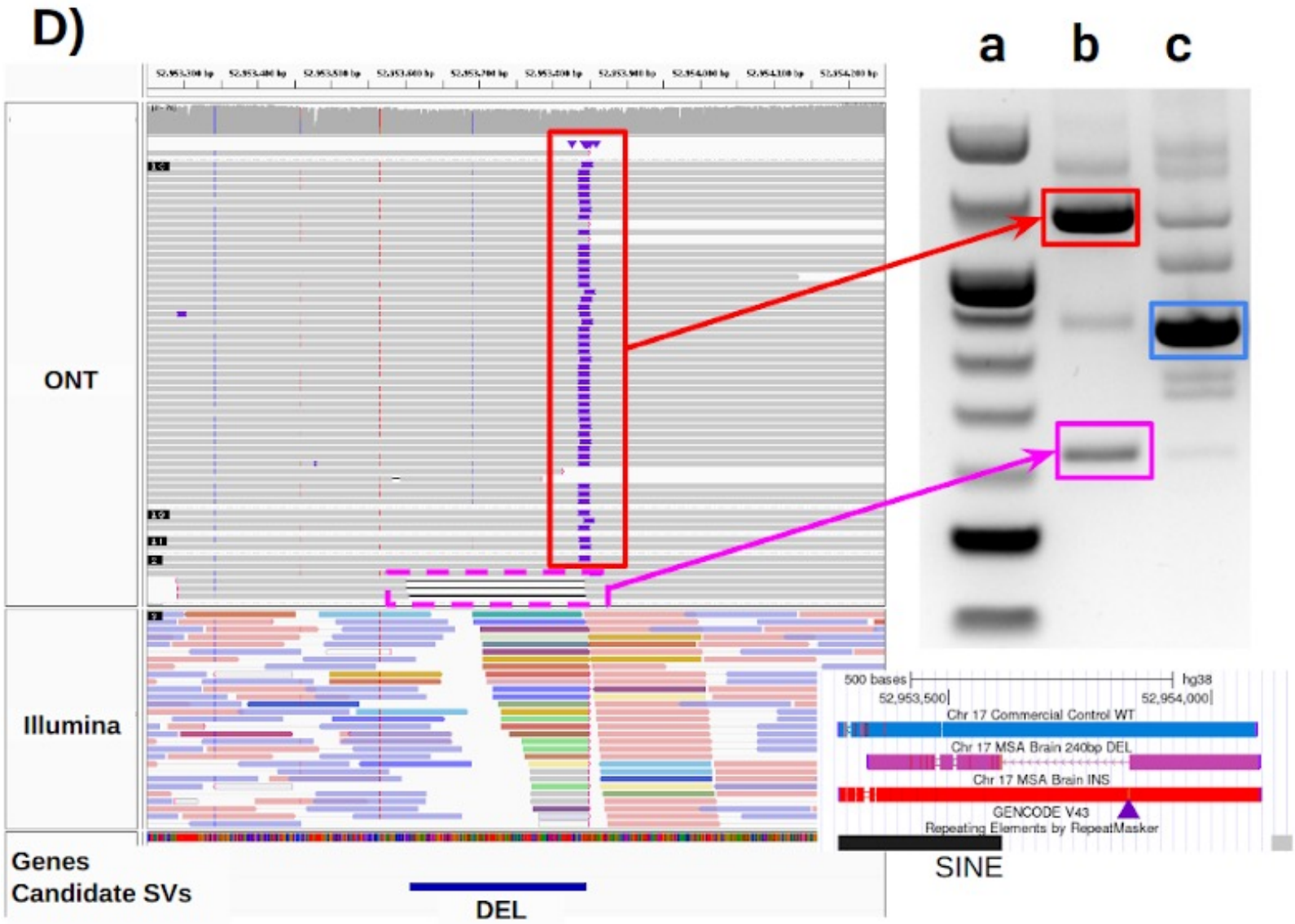
Sniffles 2 mosaic

55x MSA sample:

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

•26 Alu -Alu -> mosaic del

•125 Ins -> mosaic del



Applications: research groups including somatic variants

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

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

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

- Genomics England
  - Developing cancer pipeline

Genomics england

MARATHON OF HOPE CANCER CENTRES NETWORK

Acknowledgments

Today hands on

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

- We will go over the individual sections.