Methodologies for Structural Variant detection

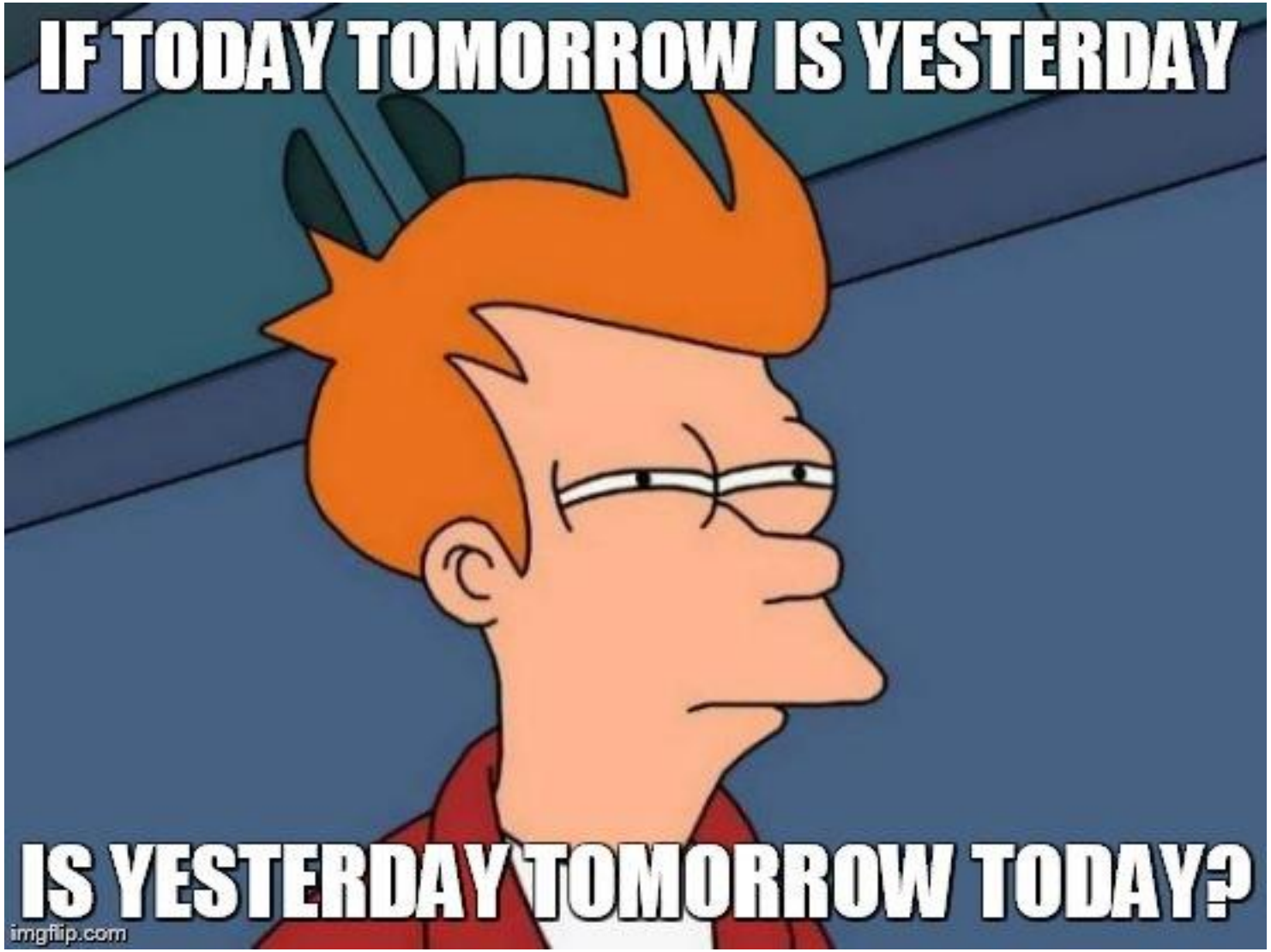
Fritz Sedlazeck & Luis Paulin

Dec,2, 2025

@sedlazeck

RICE

Recap from yesterday



What do we need for alignment based SV calling?

1. A reference that we can trust!
    1. Inspect the reference if you are not sure.
2. Alignment of reads
3. QC of alignment of reads
4. SV detection itself
5. Comparison of SV..

Human Genome Reference

- Human Genome project “completed” in 2003
- GRCh37 (hg19) and GRCh38 (hg38) mostly used
- T2T released CHM13
  - ~200Mbps more regions, resolves error in hg38
  - Lot of works to be done for annotation etc.
  - Liftover could be used (e.g. LevioSAM2)
- GIAB v3 GRCh38 reference
  - Masked false duplication
  - Decoy sequence for falsely collapsed
  - Masked contaminations
  - Refinement of our previous work

Article | Published: 30 November 2023

Improved sequence mapping using a complete reference genome and lift-over

Nae-Chyun Chen, Luis F. Paulin, Fritz J. Sedlazeck, Sergey Koren, Adam M. Phillippy & Ben Langmead

Nature Methods **21**, 41–49 (2024) | Cite this article

Method | Open access | Published: 21 February 2023

FixItFelix: improving genomic analysis by fixing reference errors

Sairam Behera, Jonathon LeFaive, Peter Orchard, Medhat Mahmoud, Luis F. Paulin, Jesse Farek, Daniela C. Soto, Stephen C. J. Parker, Albert V. Smith, Megan Y. Dennis, Justin M. Zook & Fritz J. Sedlazeck

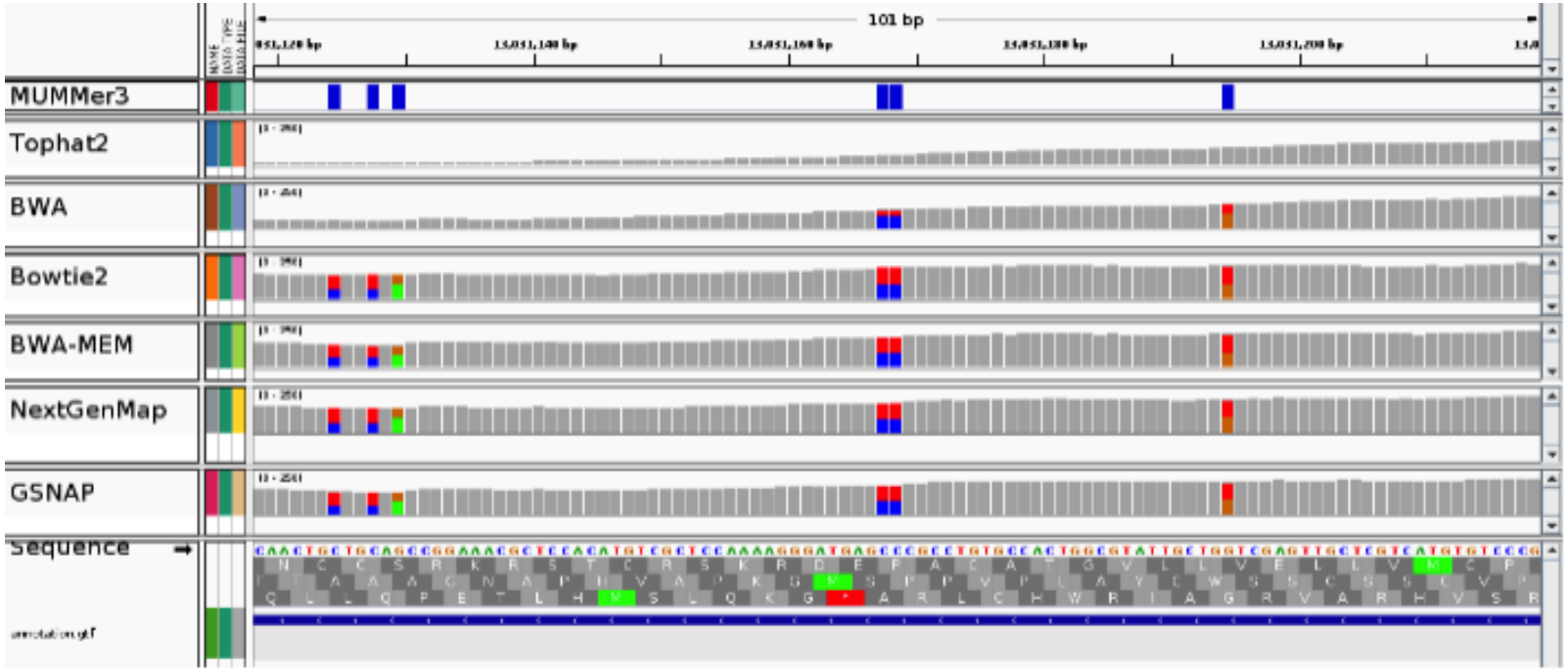
Genome Biology **24**, Article number: 31 (2023) | Cite this article

What do we need for alignment based SV calling?

1. A reference that we can trust!
    1. Inspect the reference if you are not sure.
2. Alignment of reads
3. QC of alignment of reads
4. SV detection itself
5. Comparison of SV….

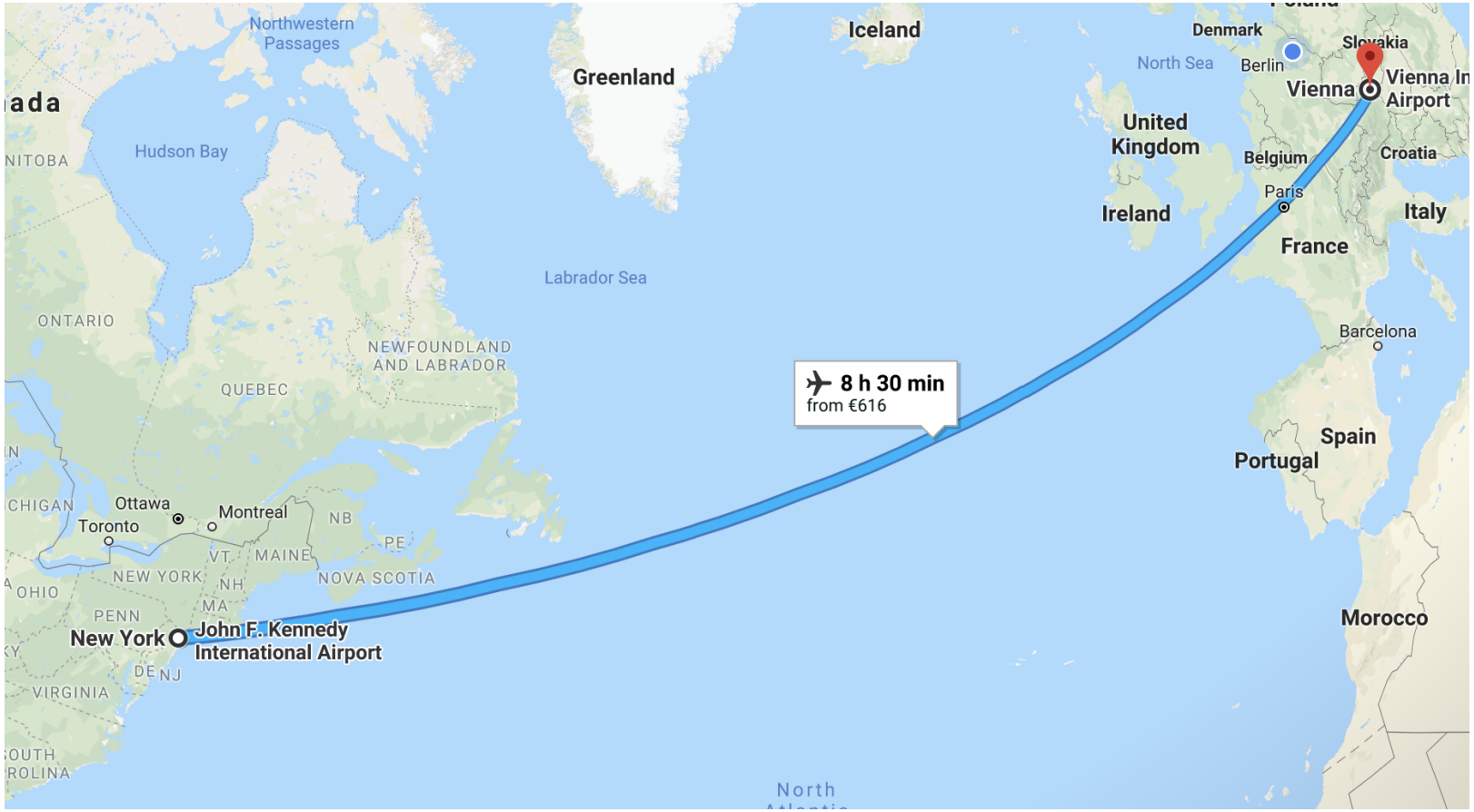
Are alignments solved?

- Limitations:
  - Read length
  - Repeats
    - SMN1+2
  - Polymorphism
    - KIRR



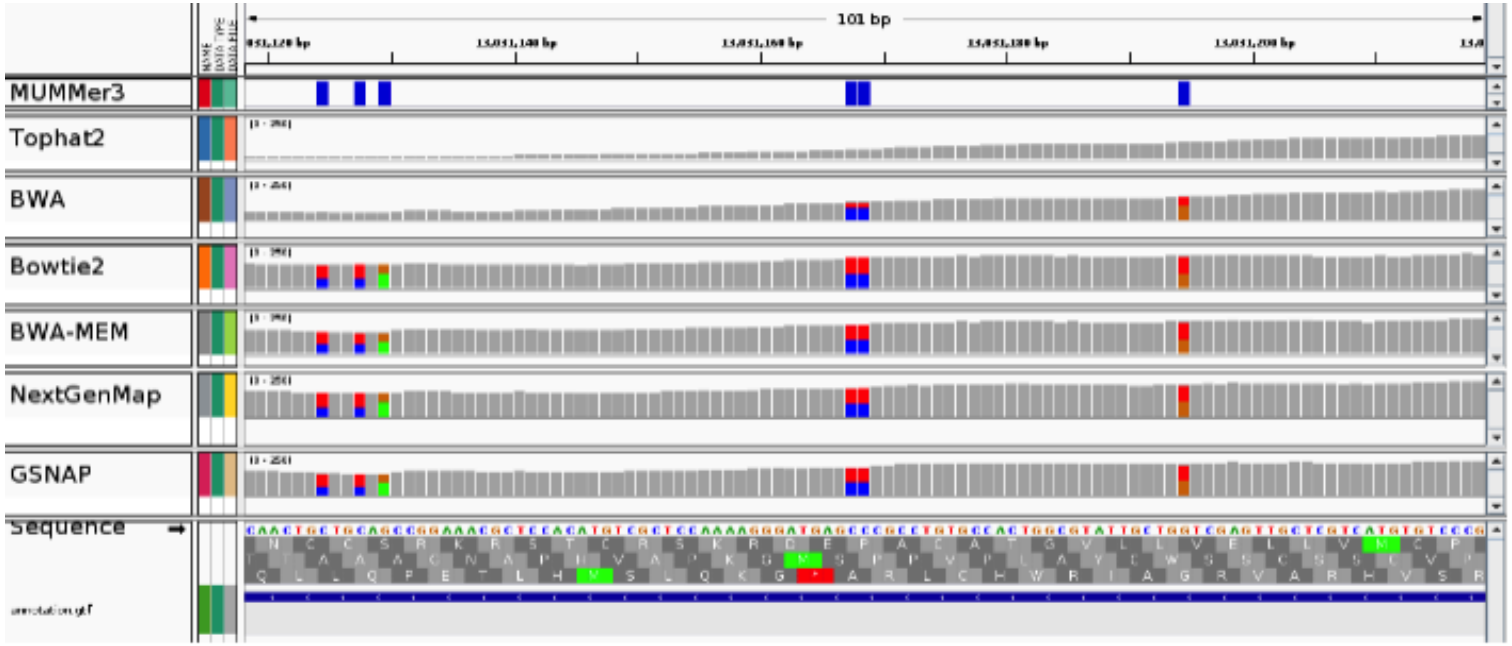
Mapping of reads

- The assignment of short reads to a region were they most likely origin.



Mapping algorithms

- Short reads:
  - BWA
  - Bowtie2
  - Stampy
  - NGM
- Long reads
  - Minimap2
  - NGMLR
  - VACMap



Teaser

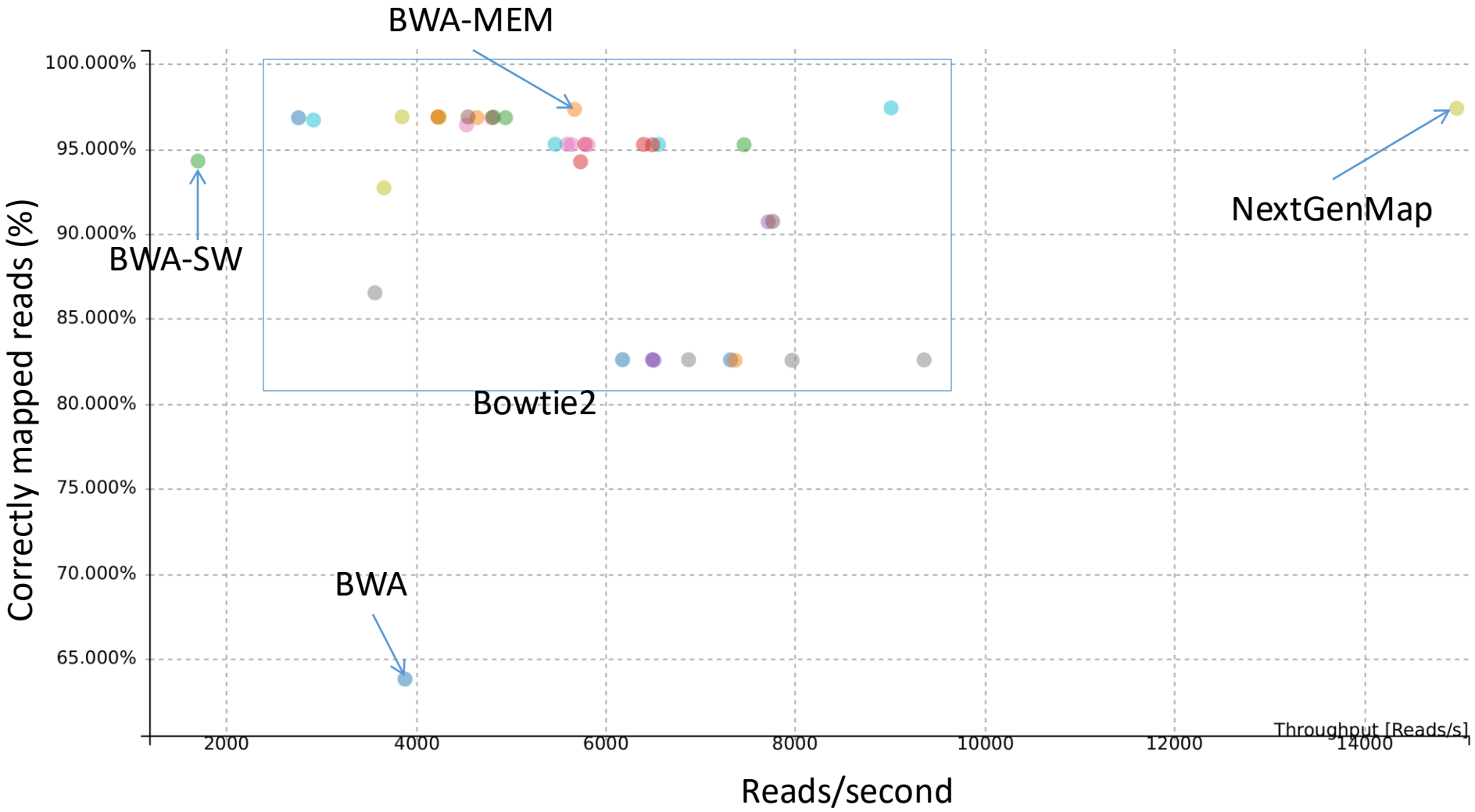
Moritz Smolka

- Personalized benchmarking
  - Including benchmarking of parameter
- Short turnaround time
- Easy to extend/use

Smolka et al. (2015)

Teaser for *S. pombe*

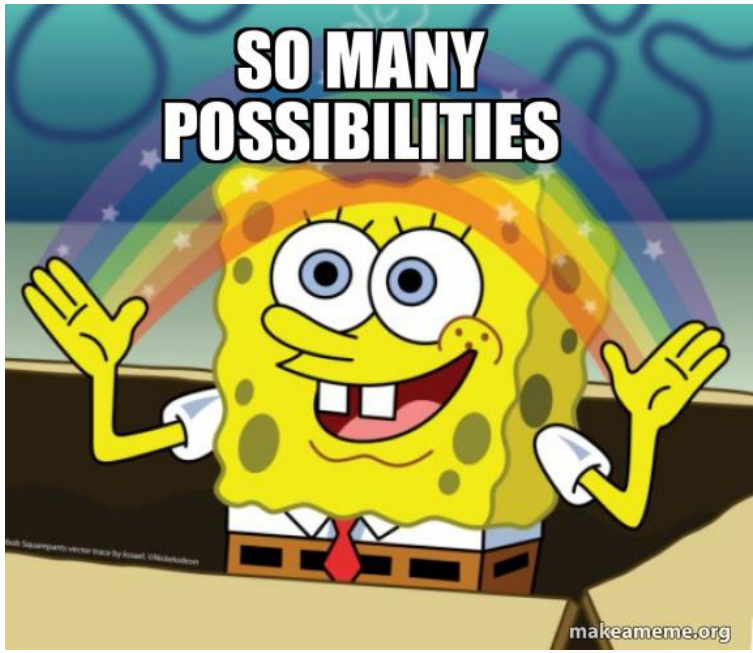
(teaser.cibiv.univie.ac.at: benchmarking 49 mappings in 18 min)



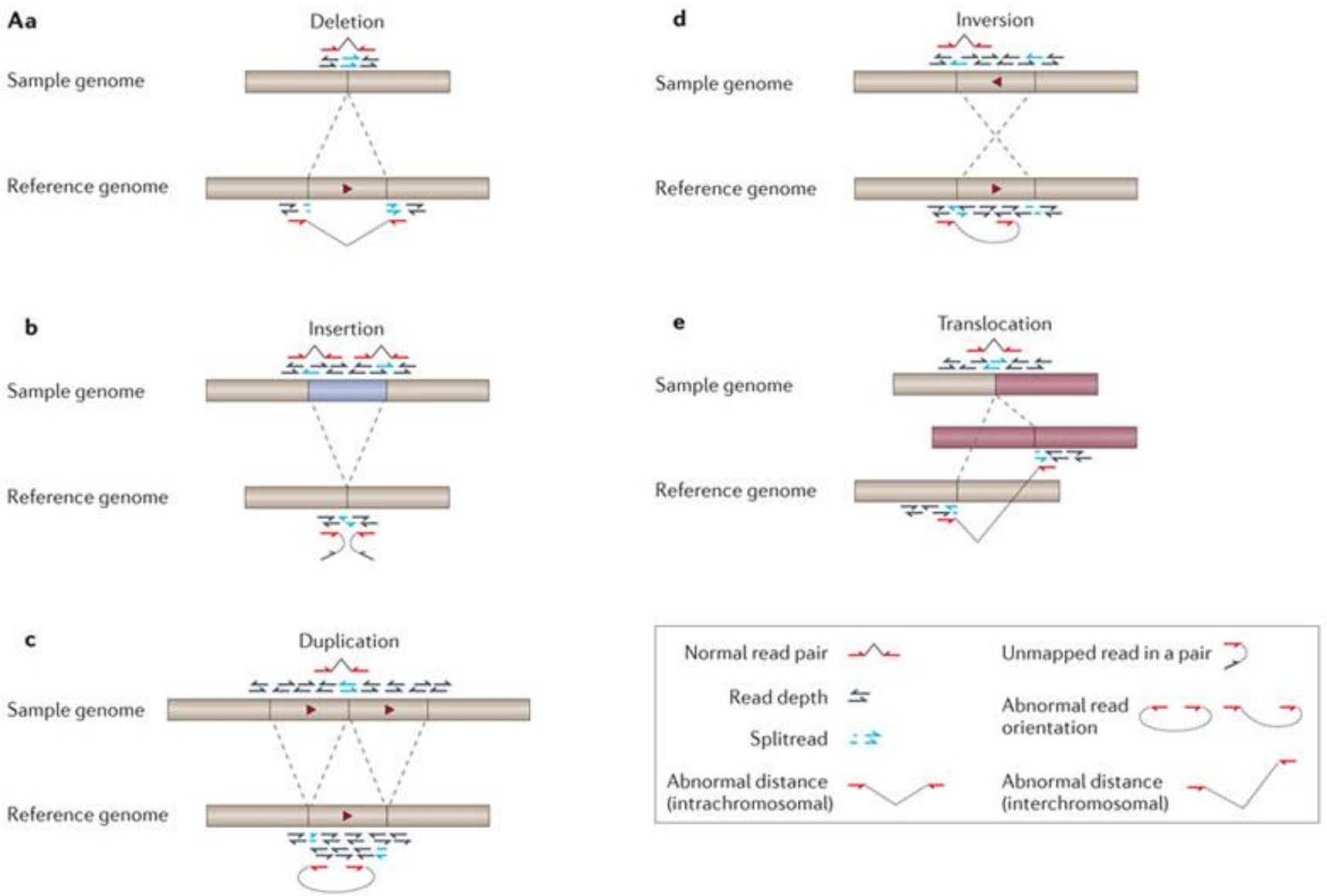
Variant calling

- SNV (no)
  - GATK[@] (meh), DRAGEN[@], DeepVariant[@,*], Clair[@,*]
- SV
  - DRAGEN[@], Manta[@], Sniffles[*], cuteSV[*], etc
- Tandem repeats
  - Expension Hunter[@], TRGT[*], Straglr[*], medaka[*] (new models)
- CNV
  - DRAGEN[@], Spectre[*],

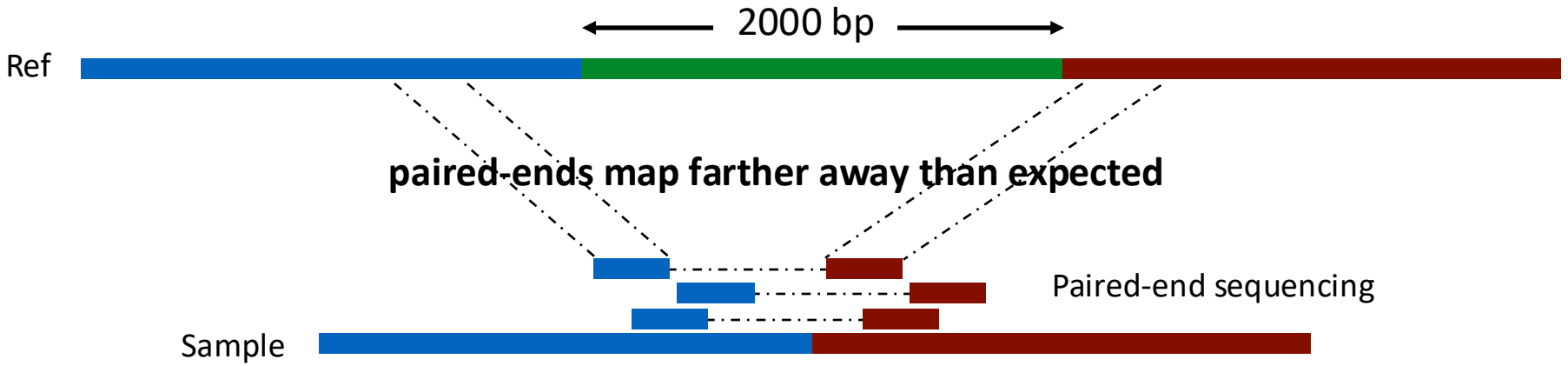
@: Short reads
* : long reads

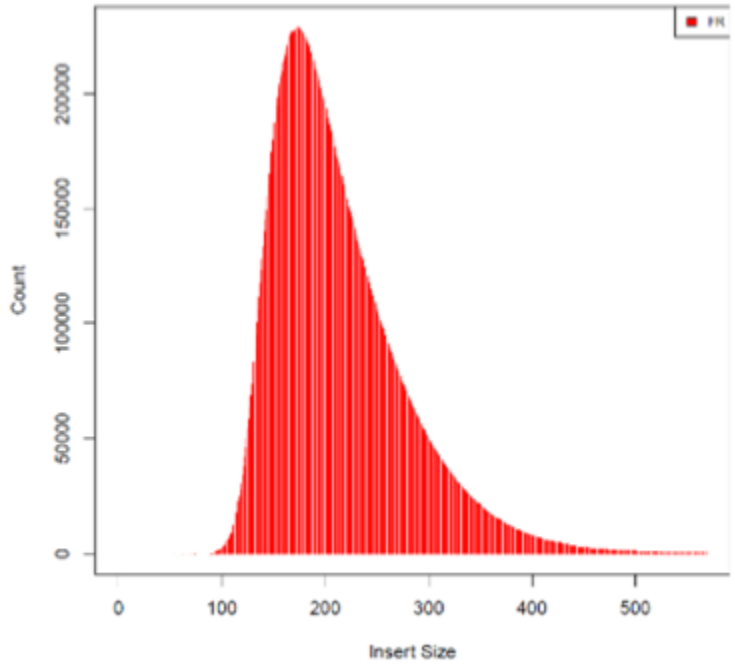


How to detect Structural Variations



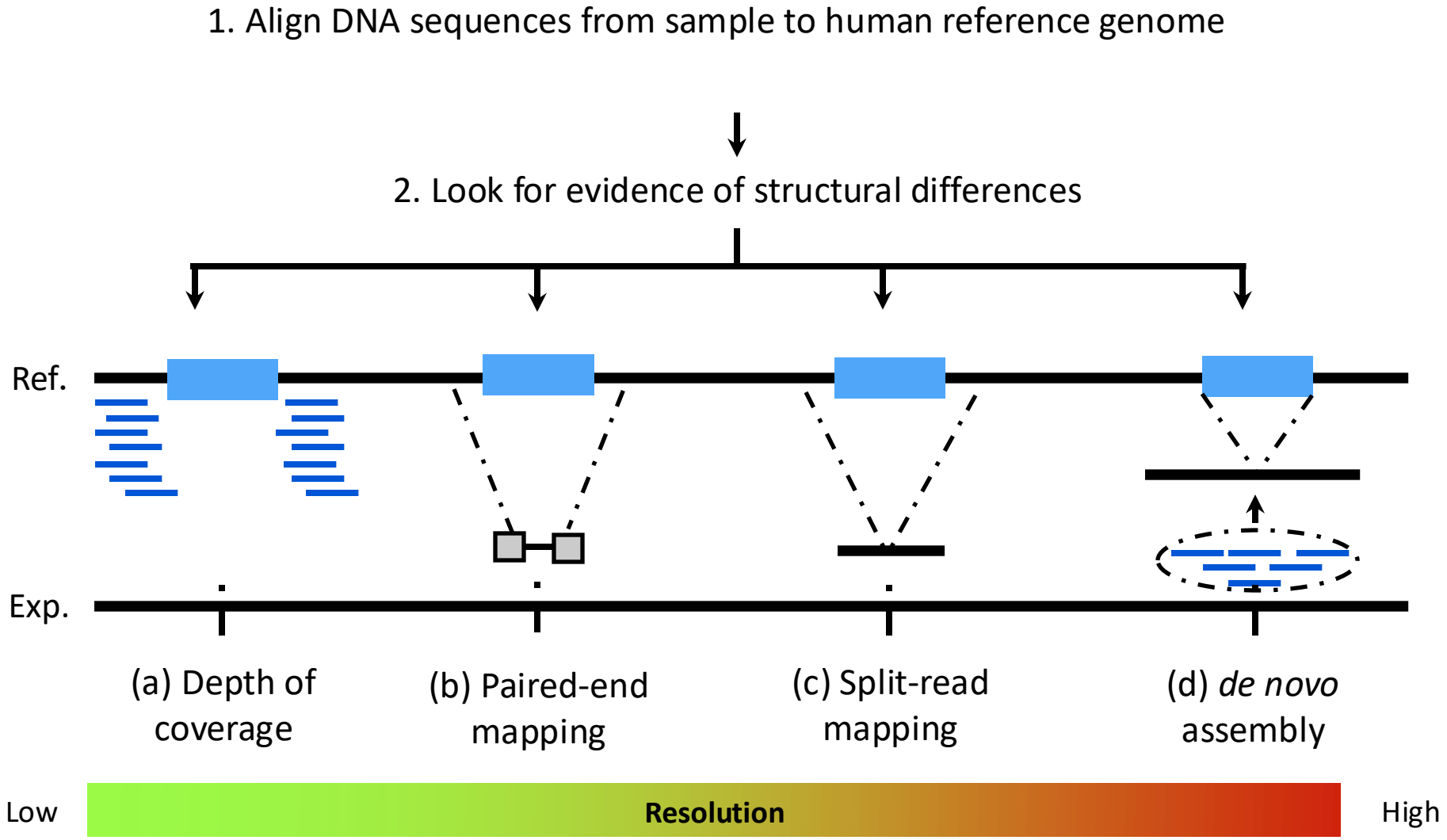
Looking for "discordant" paired-end fragments





Slide in collaboration with Ira Hall

Sequence alignment “signals” for structural variation



A probabilistic framework for SV discovery

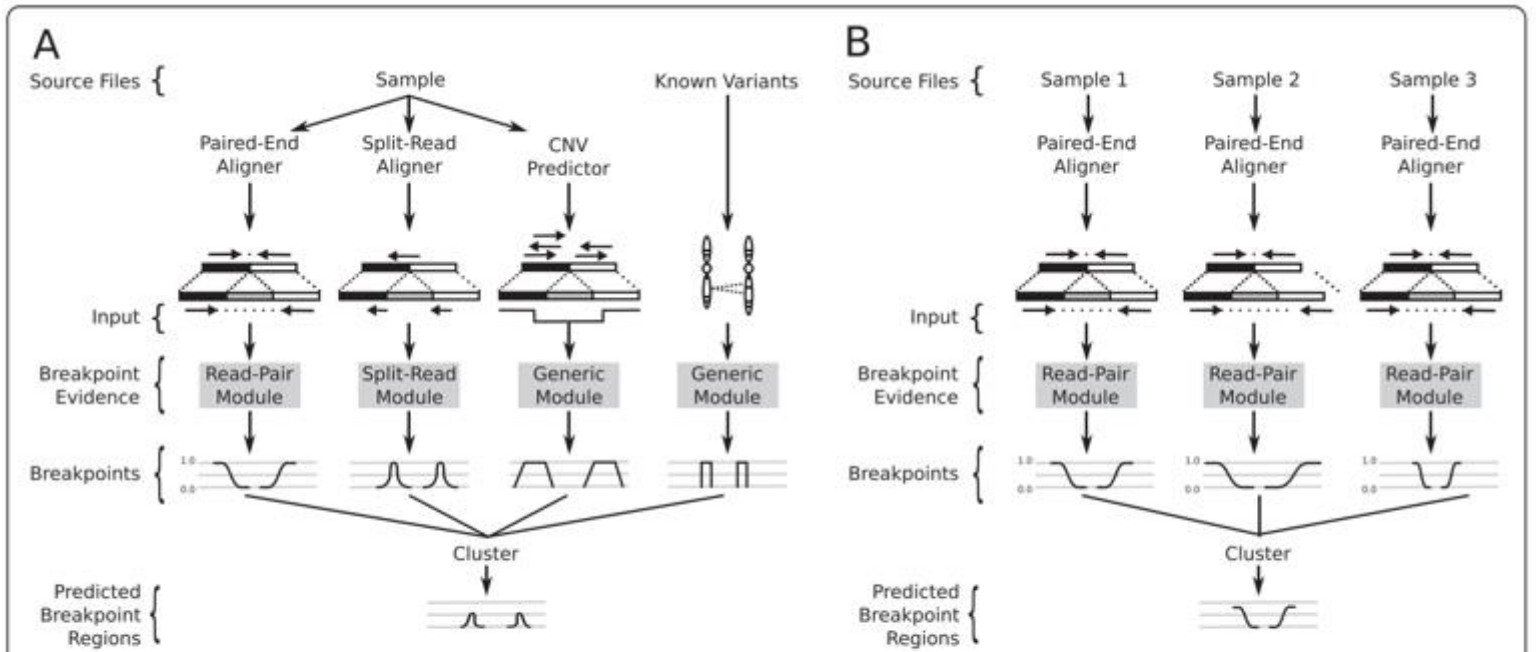


Figure 1 The LUMPY framework for integrating multiple structural variation signals. (A) A scenario in which LUMPY integrates three different sequence alignment signals (read-pair, split-read and read-depth) from a genome single sample. Additionally, sites of known variants are provided to LUMPY as prior knowledge in order to improve sensitivity. **(B)** A single signal type (in this case, read-pair) that is integrated from three different genome samples. We present these as example scenarios and emphasize that multi-signal and multi-sample workflows are not mutually exclusive. CNV, copy number variation.

Lumpy integrates paired-end mapping, split-read mapping, and depth of coverage for better SV discovery accuracy

Ryan Layer

Layer et al, 2014

Problem #1: Often many false positives

- Short reads + heuristic alignment + rep. genome **= systematic alignment artifacts (false calls)**
- Chimeras and duplicate molecules
- Ref. genome errors (e.g., gaps, mis-assemblies)
- **ALL** SV mapping studies use strict filters for above

Problem #2: The false negative rate is also typically high

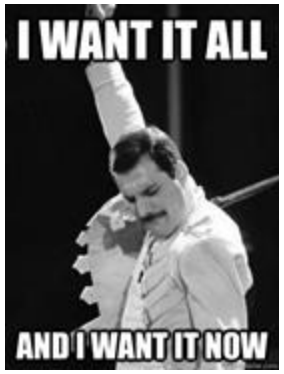
- Most current datasets have low to moderate ***physical*** coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- FILTERING!
- The false negative rate is usually hard to measure, but is thought to be extremely high for most paired-end mapping studies (>30%)
- When searching for spontaneous mutations in a family or a tumor/normal comparison, a false negative call in one sample can be a false positive somatic or de novo call in another.

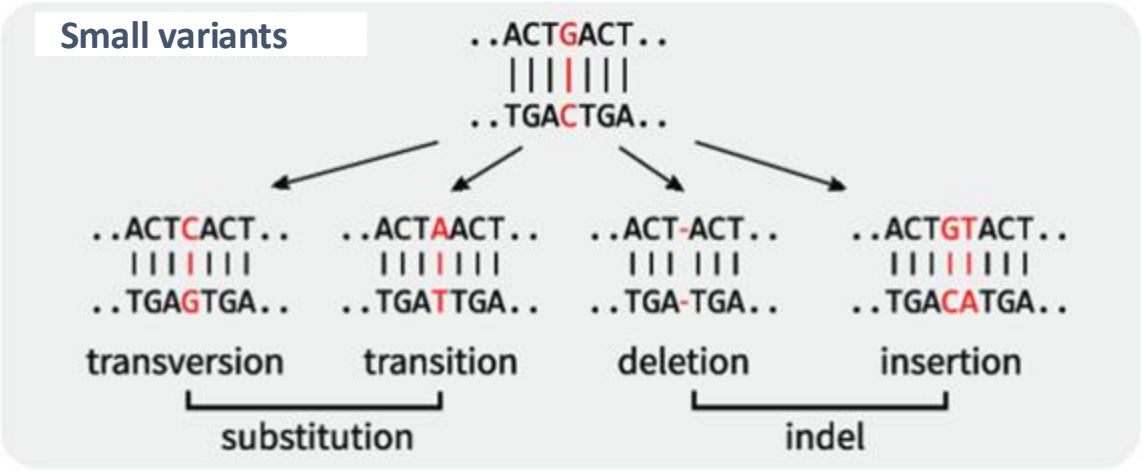
How to filter / choose the SV caller?

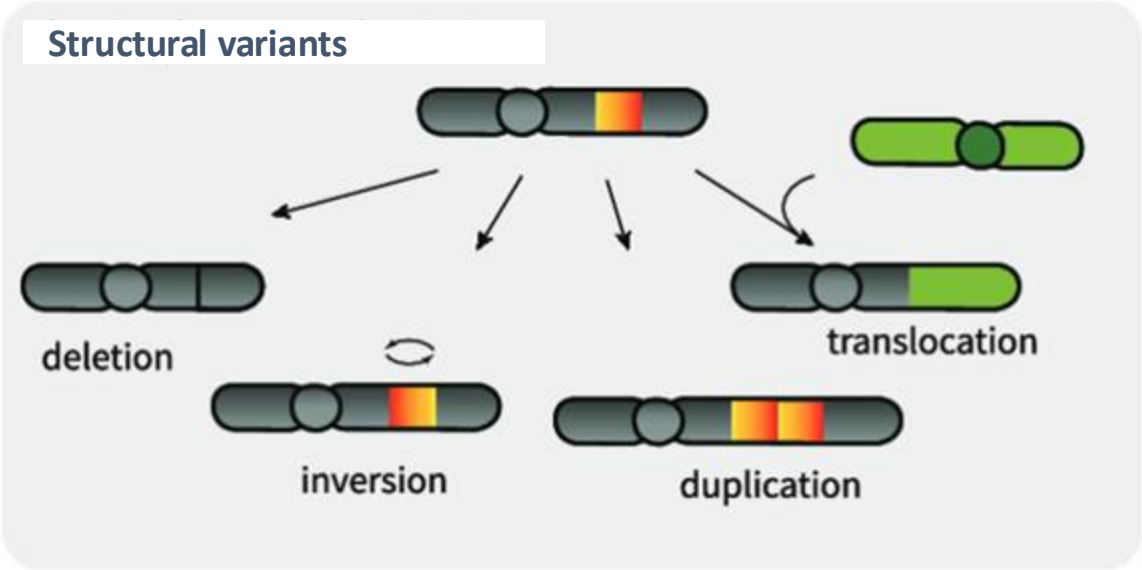
- Each method applies its own heuristics.

Method	# Sim. SV	avg FDR	avg Sensitivity
DELLY	33-198	0.13	**0.75**
LUMPY	33-198	0.06	0.62
Pindel	33-198	0.04	0.55
SURVIVOR	33-198	**0.01**	0.70

Genomic variations



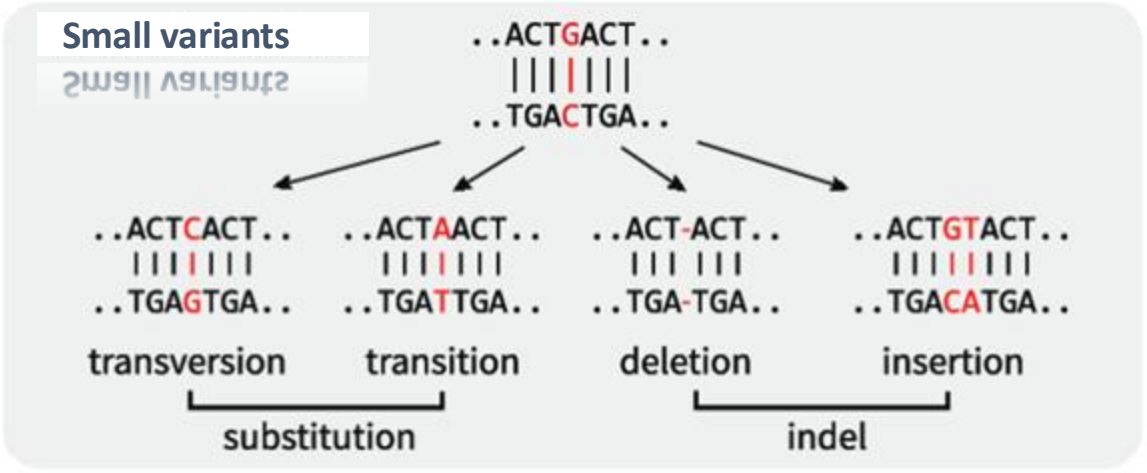


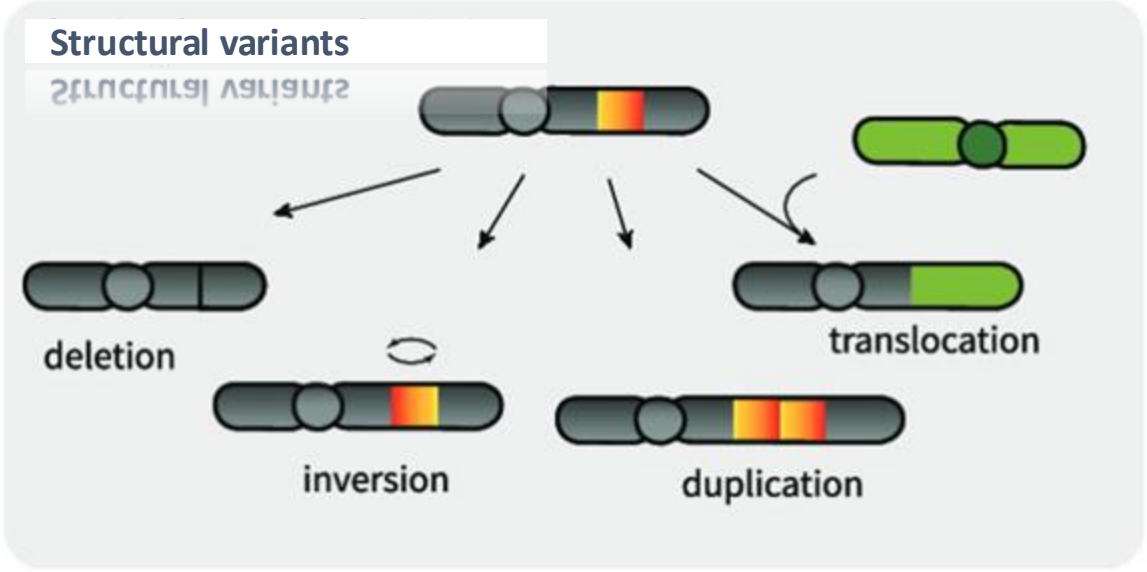


Credit: Kijong Yi et. al

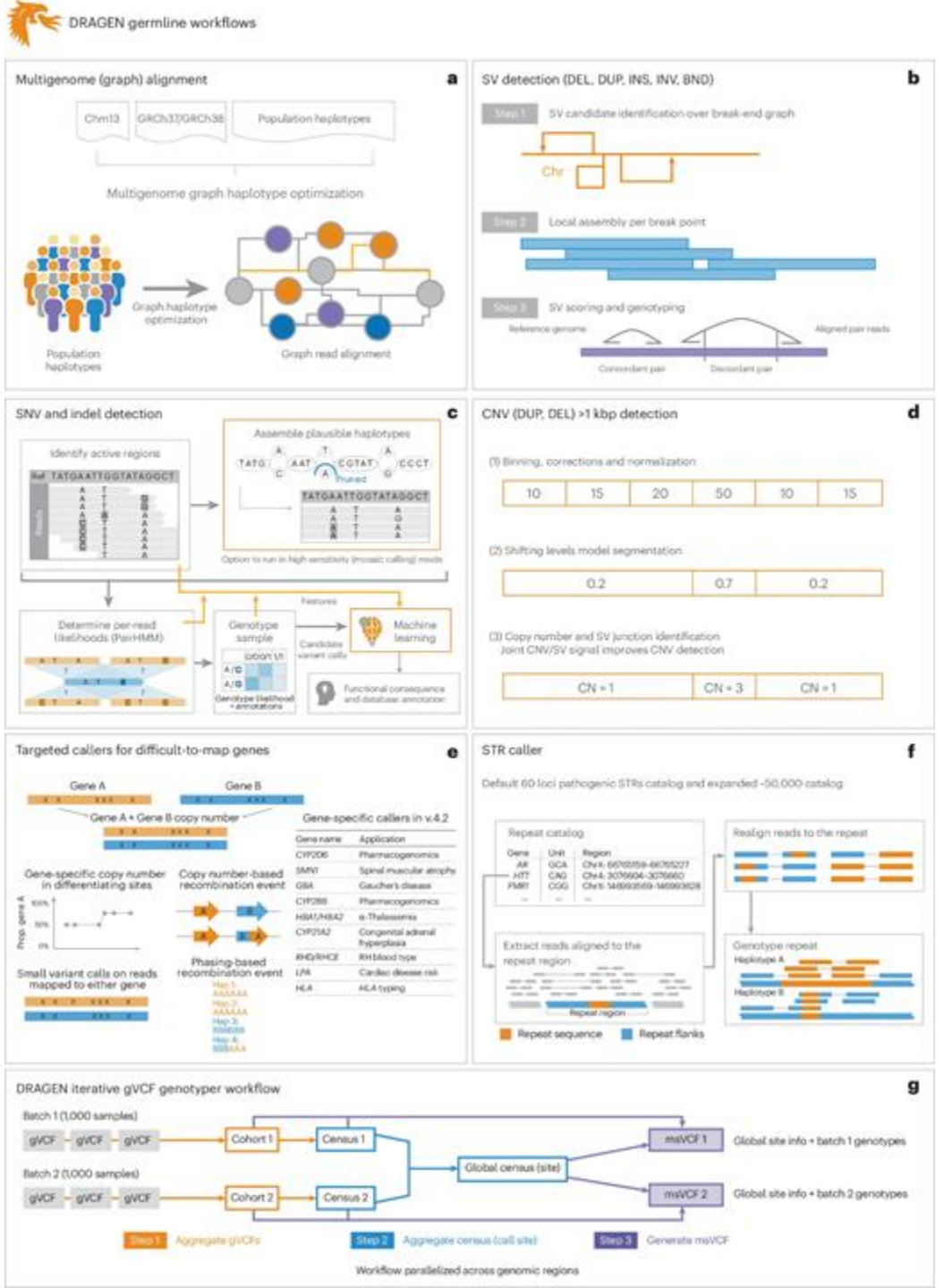
- Small variants (<50bp)
  - Associated with many diseases
- Structural Variants (>=50bp)
  - Neuro and cancer related
- Copy number variations (CNV)
  - Cardio, cancer etc.
- Short tandem repeats (STR)
  - Neuro related e.g. Huntington etc.

Genomic variations

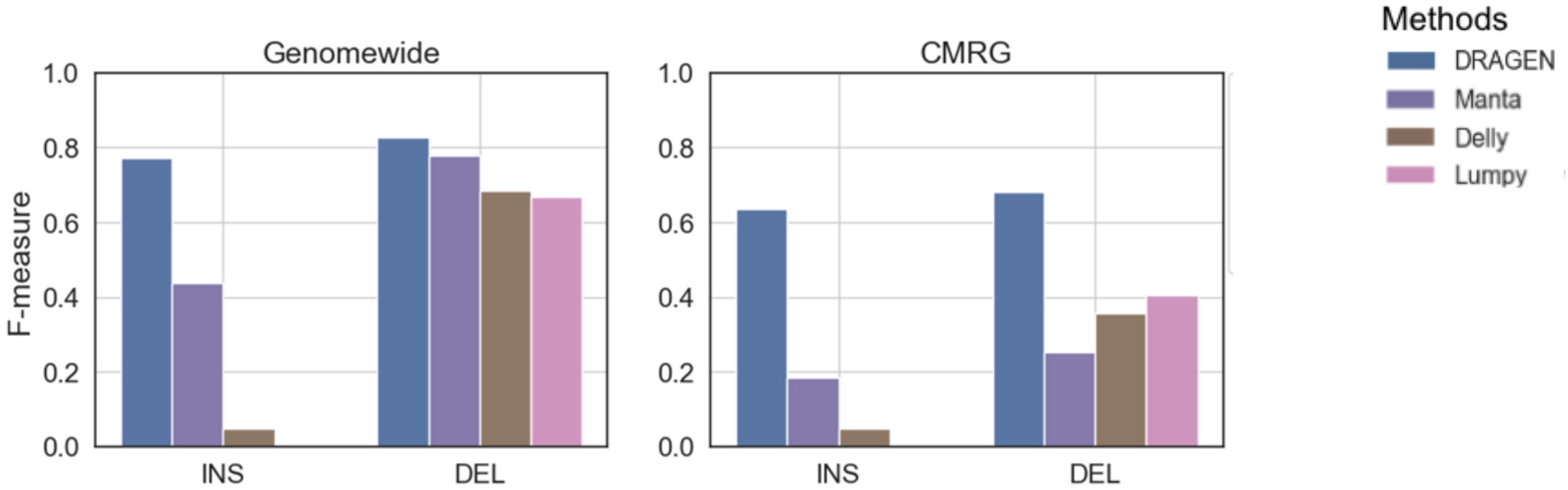




Credit: Kijong Yi et. al



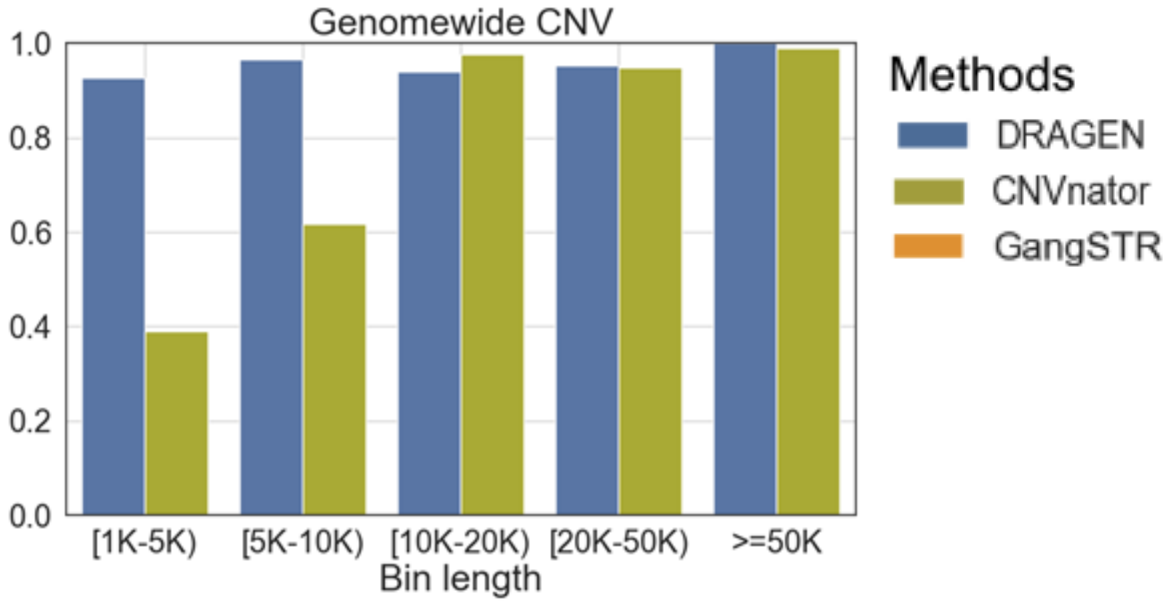
Structural variants



- Benchmark sets: GIAB v0.6 for genomwide and v1.00 for CMRG regions
- Reference used: hg19
- CMRG results showcase DRAGEN's ability to detect SVs in repetitive regions

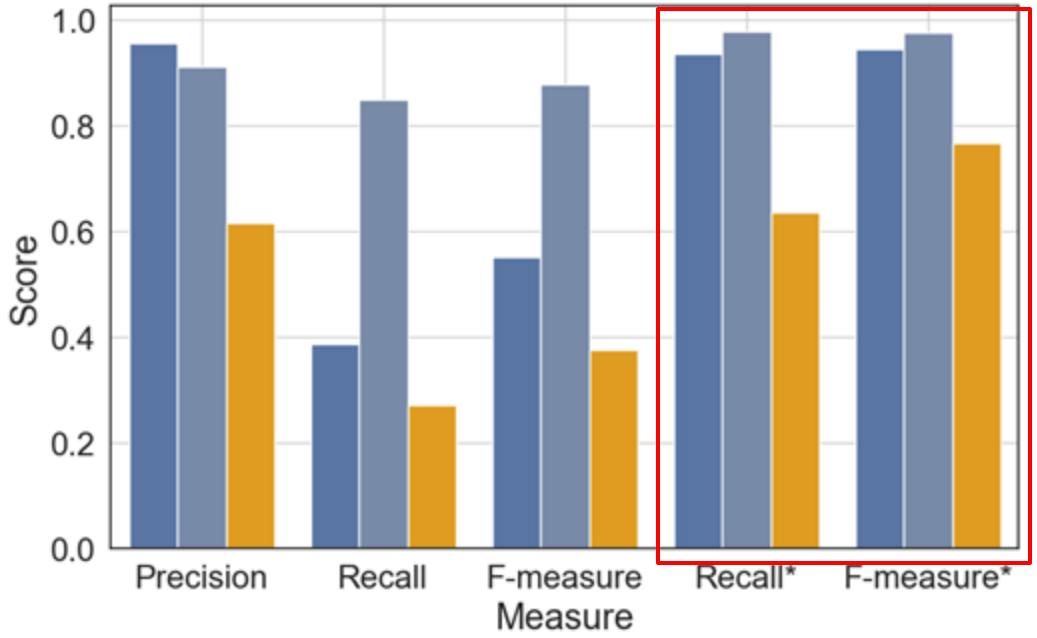
Copy number and short tandem repeats

- Uses discordant and split read signals from SV calling
- Utilize a modified shifting levels model (SLM)



GIAB SV deletion variants >= 10kbp

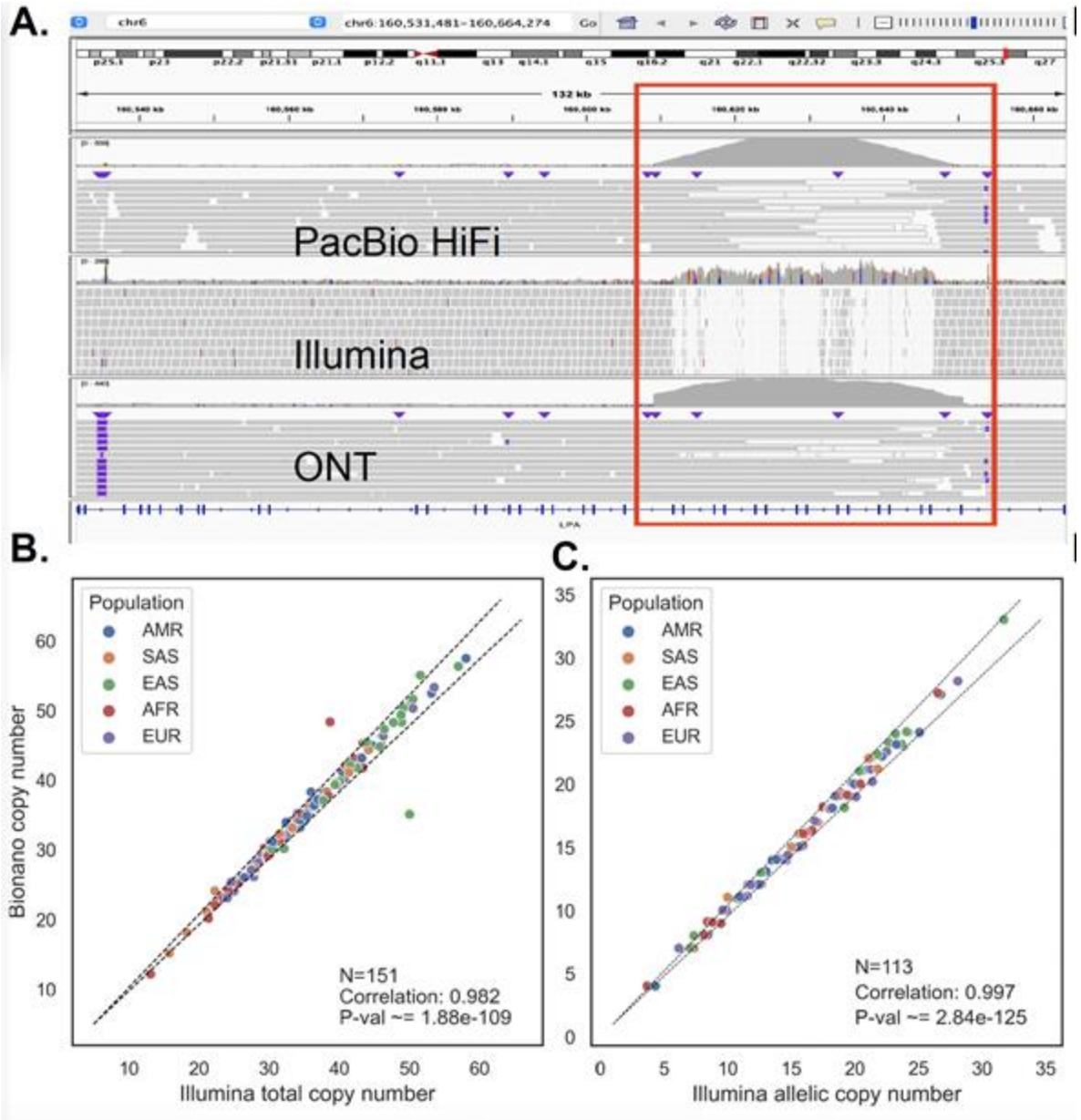
Short tandem repeat regions



- The light-blue is for DRAGEN indel calls
- Calls outside of individual catalog regions are not penalized

DRAGEN targeted caller: *LPA*

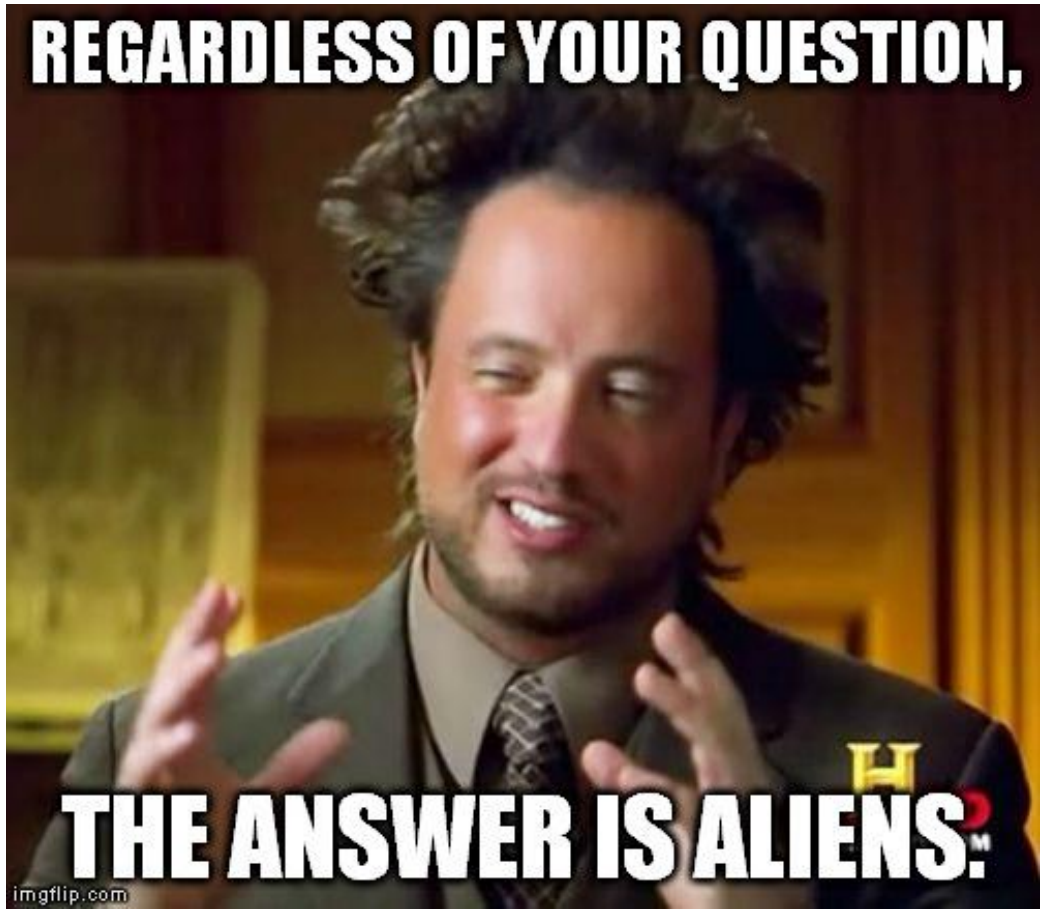
- Low Lp(a) associated with risk of **cardiovascular diseases**
- High number of KIV-2 repeats in LPA is associated with low serum Lp(a) concentrations
- Diagnostic SNV only work in Europeans
- DRAGEN call benchmarked against bionano



A) Mapping to LPA KIV-2 regions B) and C) benchmark against bionano

Question: 2

What is the difference between a CNV and SV duplication?



Exercise Part 2: Short read based

- Utilize short read mapping to call SV
  - We will use Manta

- Go to: Day 2, Part 1
  https://github.com/fritzsedlazeck/teaching_material

    - Remember files are also available locally

PacBio / ONT sequencer

Advantage:

- Long reads,

Disadvantage:

- Throughput/yield
- Costs
- High error rates

Long Read Technologies

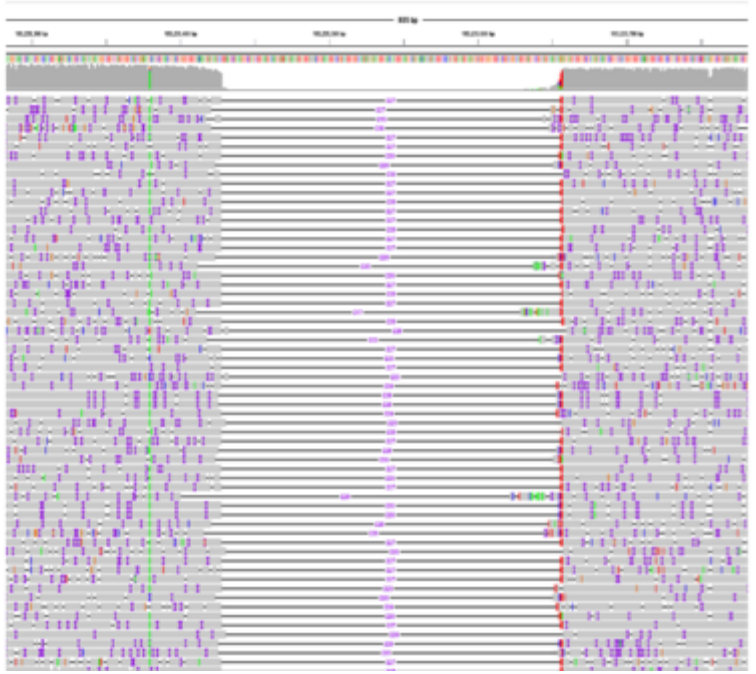
- (+) SVs in repetitive regions
- (+) Span SVs
- (+) Uniform coverage
- (+) Can identify more complex SVs

- (-) Higher seq. error rate
- (-) Hard to align

Mapping challenges

BWA-MEM:

NGMLR:



Mapping challenges

BWA-MEM:

NGMLR:

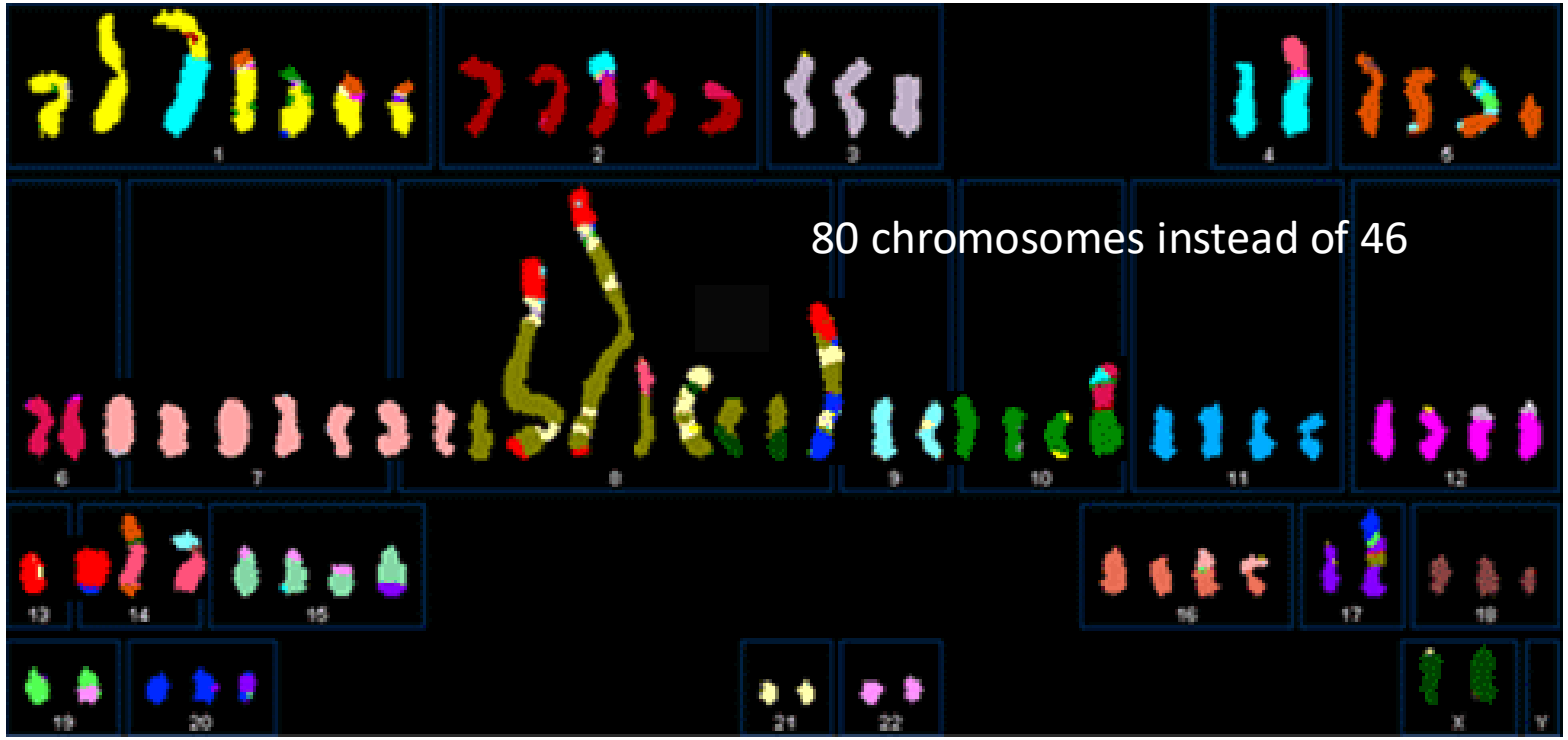
NGMLR + Sniffles

- NGMLR
  - Convex gap cost model to better distinguish seq. error vs. signal
  - Novel method for split read alignment.

- Sniffles
  - Includes multiple statistical models to distinguish noise vs. signal

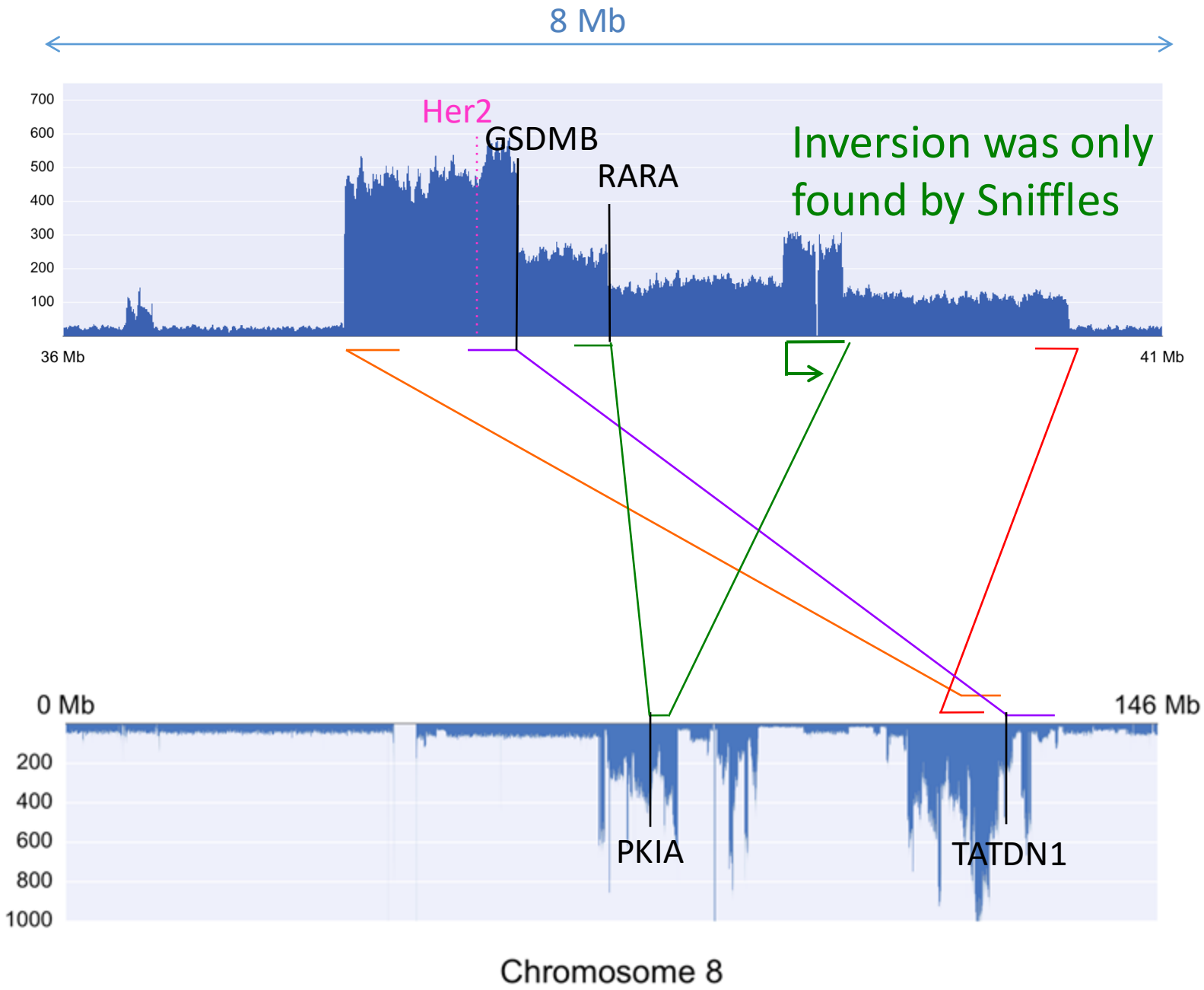
SKBR-3 using Pacbio

Most commonly used Her2-amplified breast cancer cell line



Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.

(Davidson et al, 2000)


8 Mb
700
600
500
400
300
200
100
Her2
GSDMB
RARA
Inversion was only found by Sniffles
36 Mb
41 Mb
0 Mb
146 Mb
200
400
600
800
1000
PKIA
TATDN1
Chromosome 8

3.2 NA12878

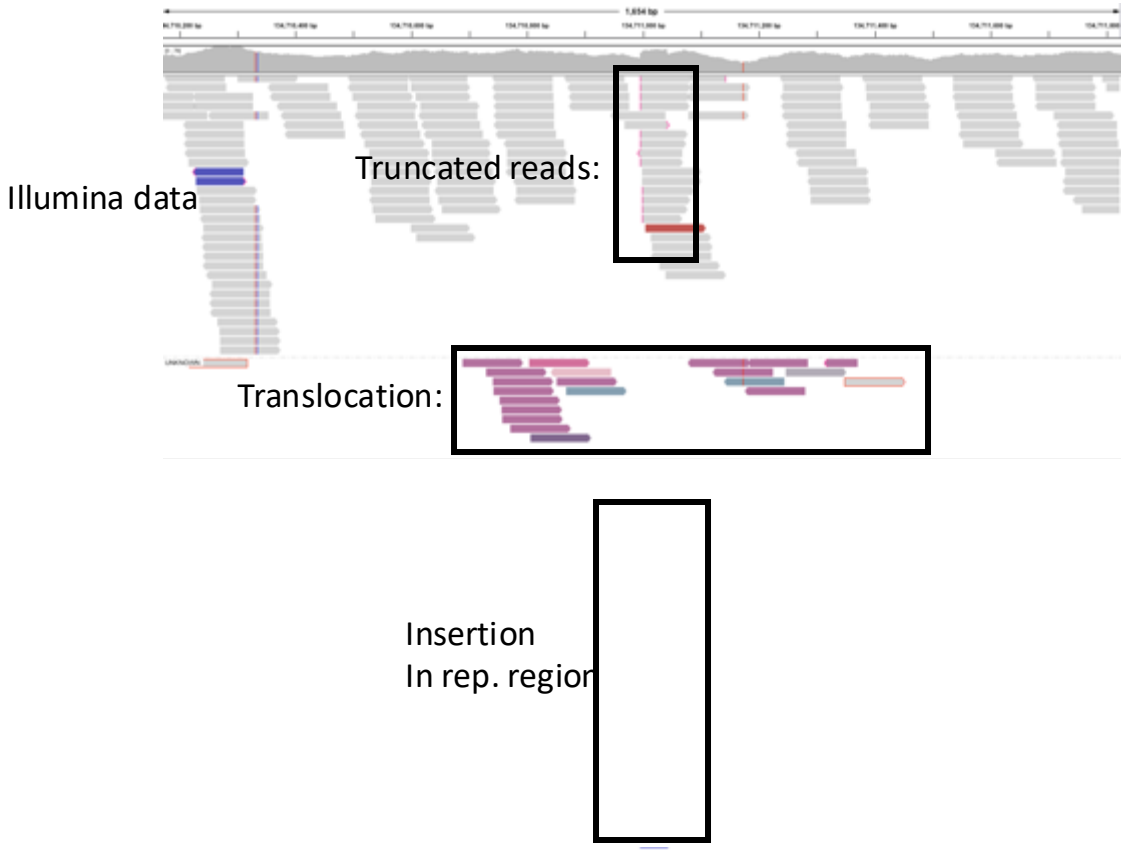
- Healthy female
- Gold standard in genomics
- Sequenced with many technologies independently:
  - Illumina, PacBio, Oxford Nanopore

3.2 NA12878: Deletion calling

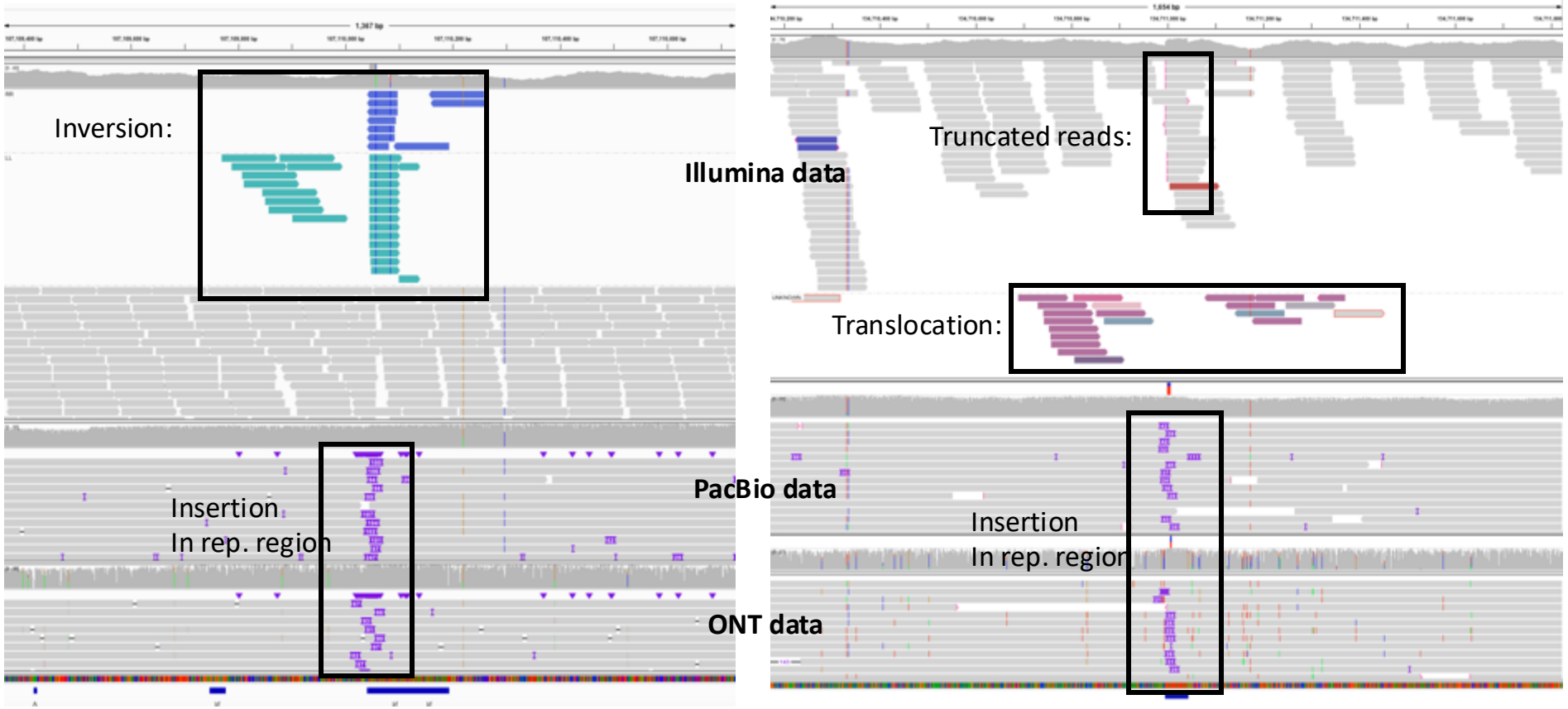
Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	**119**
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	**43**
Oxford Nanopore @Baylor	34x	4,982	12,596	7,102	169	113	5,166	**46**
Illumina	50x	2 x 101	7,275	3,744	731	553	0	**2,247**

3.2 NA12878: check 2,247 vs 119 TRA

Overlap	Illumina TRA(%)
Translocations	7.74
Insertions	53.05
Deletions	12.06
Duplications	0.57
Nested	0.31
High coverage	1.87
Low complexity	9.79
Explained	**85.40**



NA12878: check 2,247 TRA



Sniffles: HG002 v0.6 and CMRG

nature biotechnology

RESOURCE

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

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[4], 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]

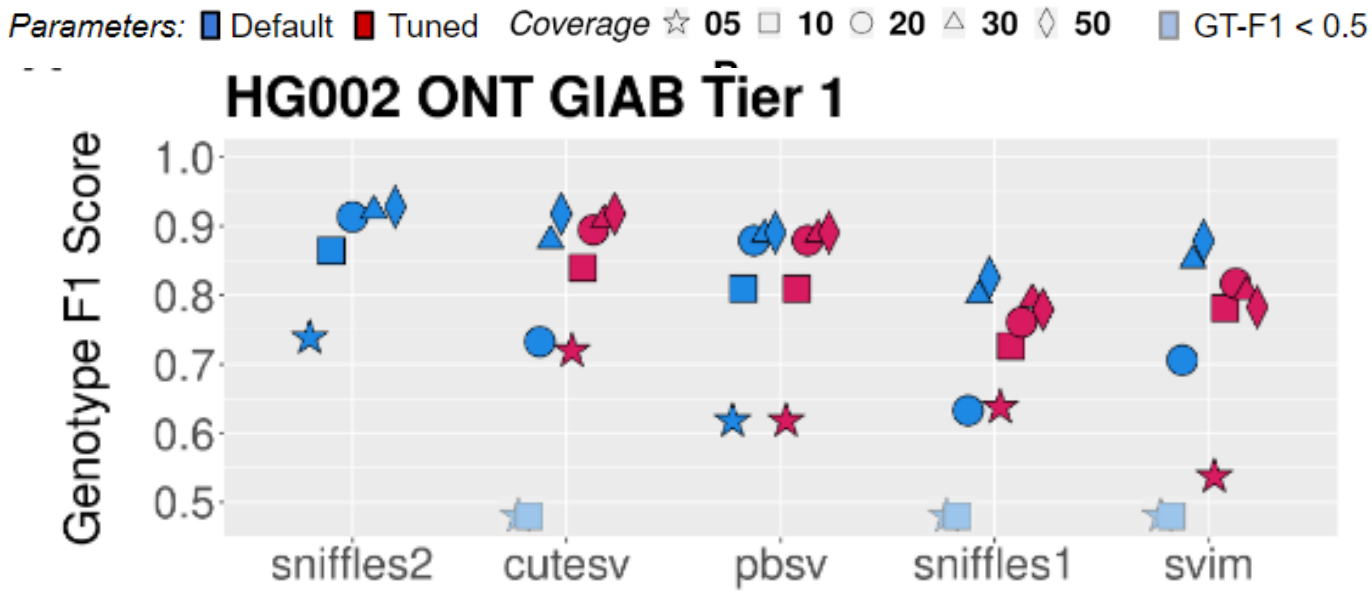
ARTICLES

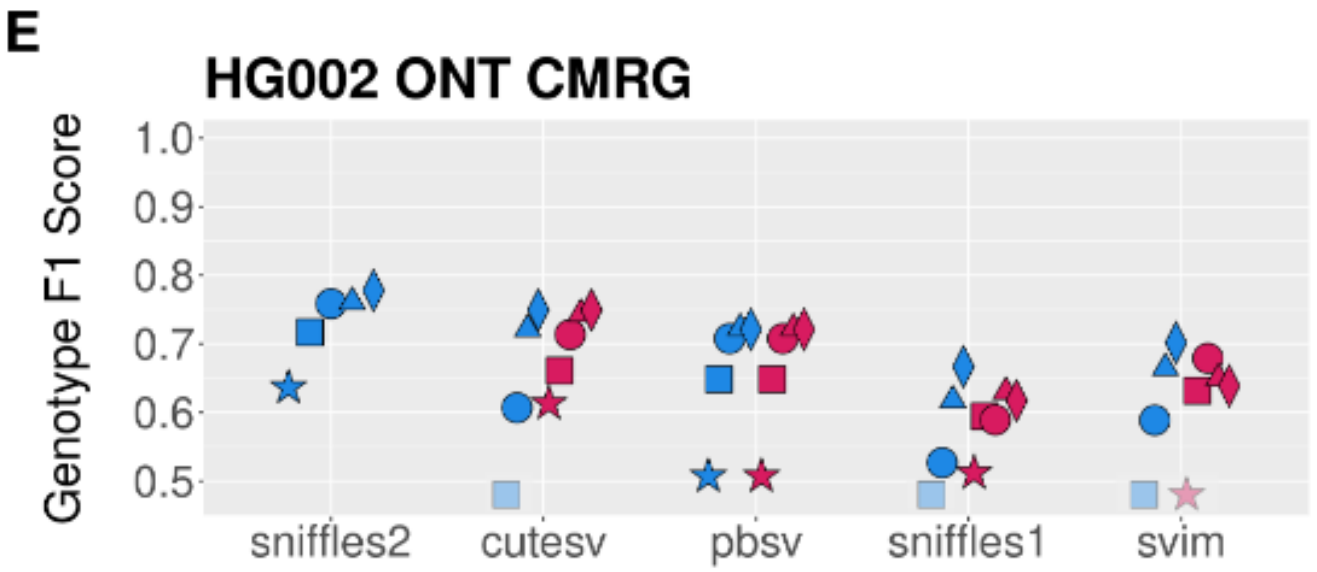
https://doi.org/10.1038/s41587-021-01158-1

nature biotechnology

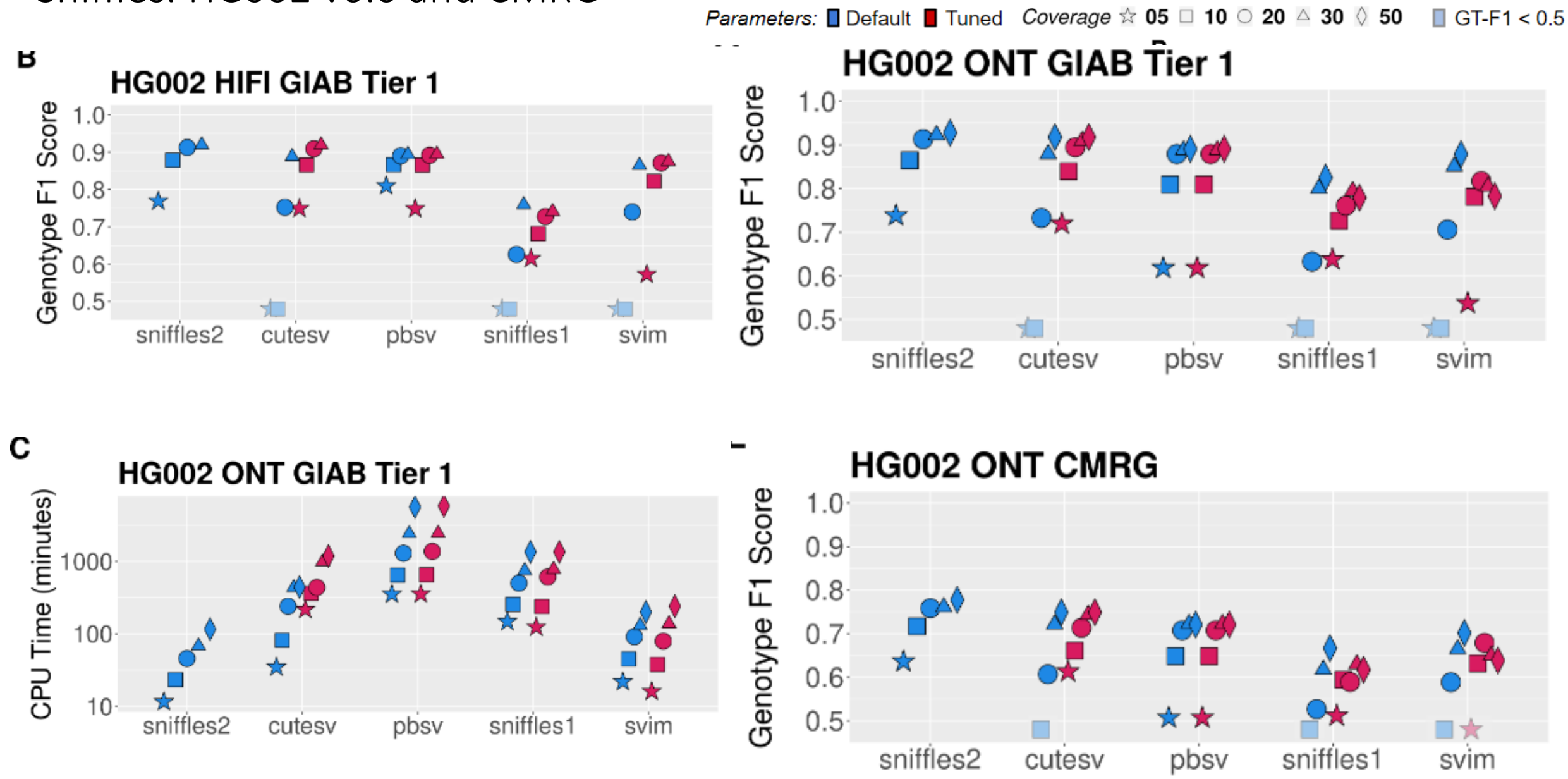
Curated variation benchmarks for challenging medically relevant autosomal genes

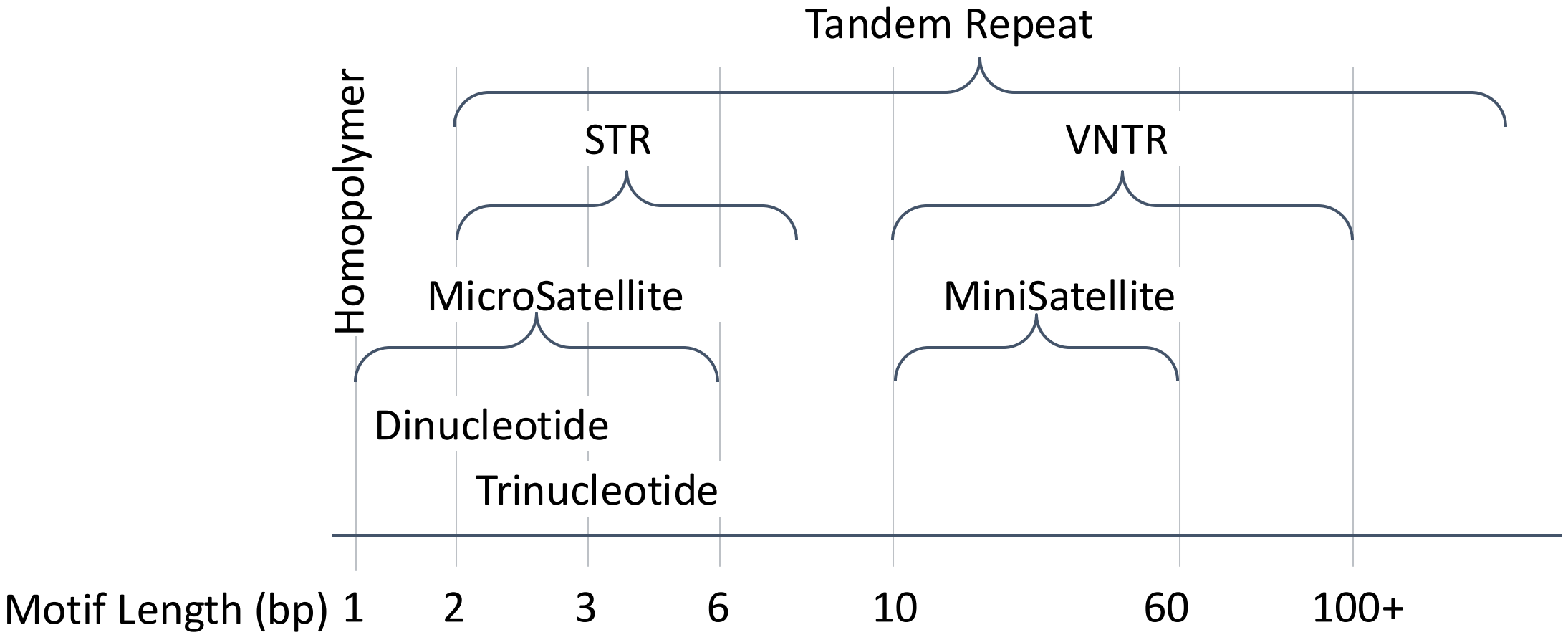
Justin Wagner[1], Nathan D. Olson[1], Lindsay Harris[1], Jennifer McDaniel[1], Haoyu Cheng[2], Arkarachai Fungtammasan[3], Yih-Chii Hwang[3], Richa Gupta[3], Aaron M. Wenger[4], William J. Rowell[4], Ziad M. Khan[5], Jesse Farek[5], Yiming Zhu[5], Aishwarya Pisupati[5], Medhat Mahmoud[5], Chunlin Xiao[6], 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✉]



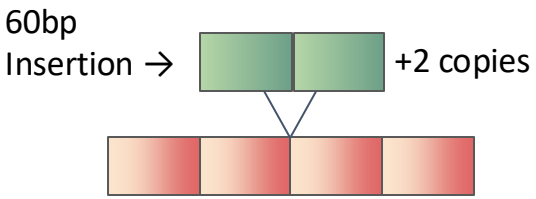


Sniffles: HG002 v0.6 and CMRG





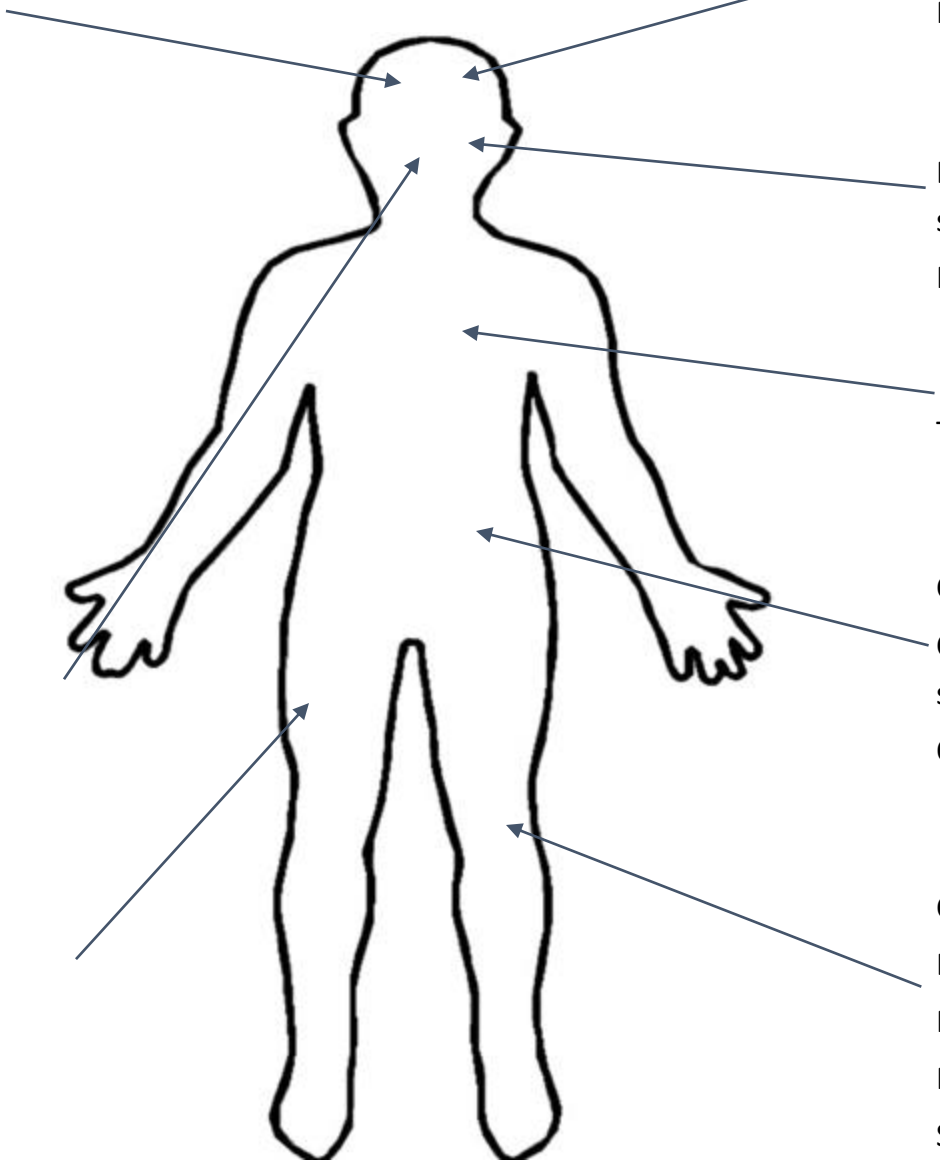

= 30bp repeat motif



Repeat Motif >= 2bp
Variant Sequence (INDEL) >= 5bp

An **insertion/deletion** represents repeat **expansion/contraction**

Known Pathogenic Tandem Repeats



Neurological Disorders

Amyotrophic lateral sclerosis and/or frontotemporal dementia

Dentatorubral-pallidoluysian atrophy

Episodic ataxia

Friedreich ataxia

Hereditary essential tremor type 6

Huntington's disease

Myotonic dystrophy 1

Neuronal intranuclear inclusion disease

Spinocerebellar ataxia

Spinal and bulbar muscular atrophy

Panhypopituitarism and Growth Hormone Deficiency

X-linked mental retardation with isolated growth hormone

X-linked panhypopituitarism

Muscular Dystrophies

Duchenne muscular dystrophy

Oculopharyngeal muscular dystrophy

*** Deduped/Grouped by ChatGPT**

Epilepsy and Seizure Disorders

Developmental and epileptic encephalopathy

Familial adult myoclonic epilepsy

Mental retardation, FRA12A type

Ophthalmological Disorders

Blepharophimosis, ptosis, and epicanthus inversus syndrome

Fuchs endothelial corneal dystrophy-3

Cardiovascular Disorders

Tetralogy of Fallot

Myopathy

CANVAS syndrome

Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome

Oculopharyngodistal myopathy 1

Connective Tissue Disorders

Cleidocranial dysplasia

Desbuquois dysplasia-2

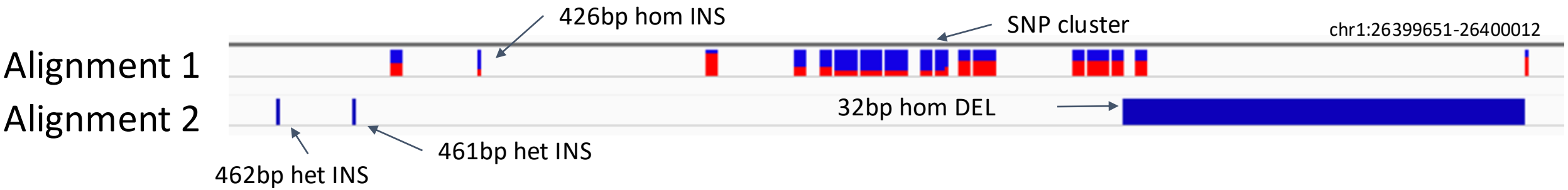
Multiple epiphyseal dysplasia

Pseudoachondroplasia

Synpolydactyly

Variant Representation

- Identical HPRC HG002 input assembly
- Different minimap2 alignment parameters



Benchmarking Pipeline

Three commands for comparing VCFs to the GIABTR benchmark

1) Truvari bench
2) Truvari refine
3) Laytr giabTR

nature biotechnology

Article

https://doi.org/10.1038/s41587-024-02225-z

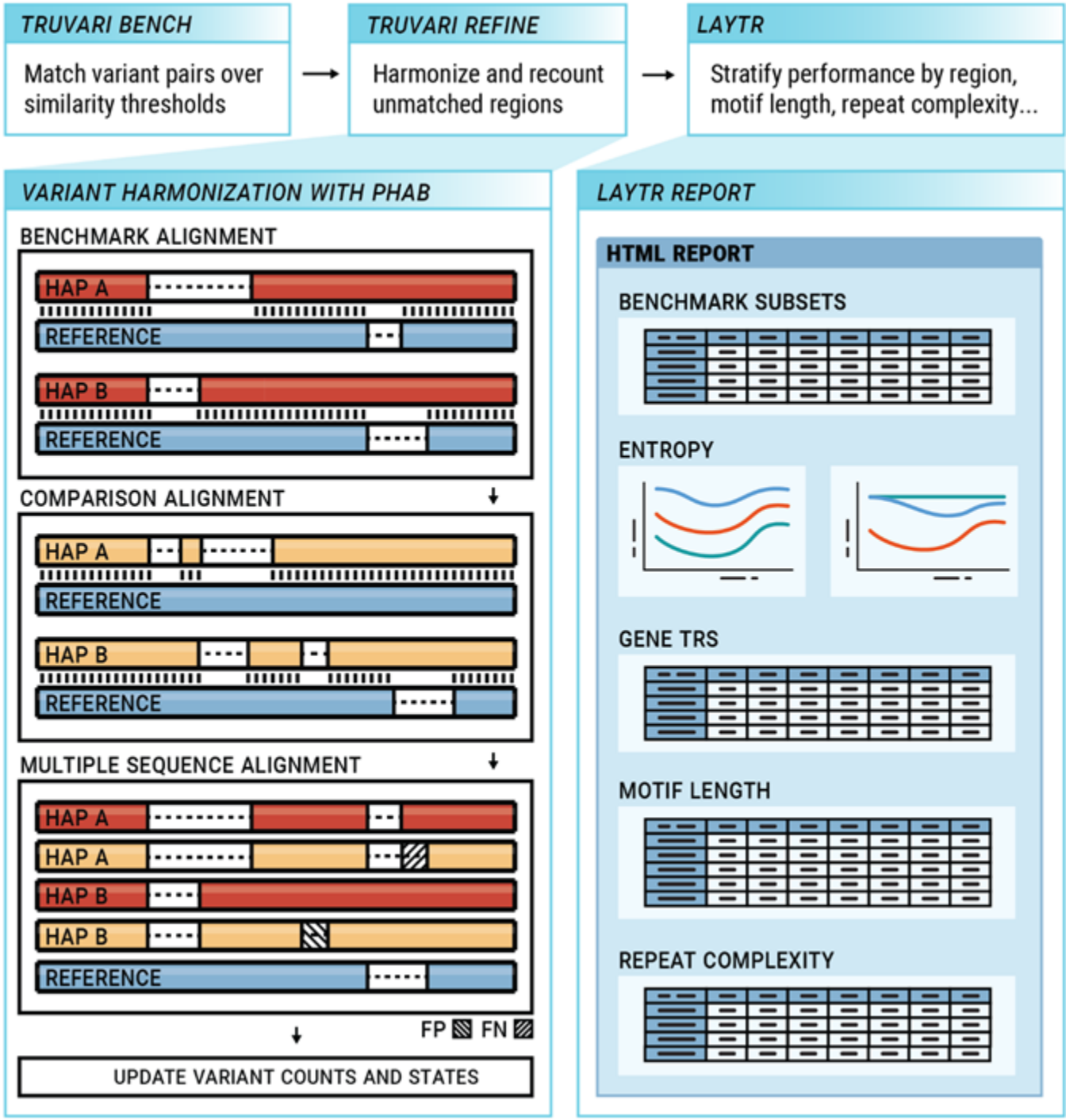
Analysis and benchmarking of small and large genomic variants across tandem repeats

Received: 30 October 2023

Accepted: 28 March 2024

Published online: 26 April 2024

Adam C. English[1], Egor Dolzhenko[2], Helyaneh Ziaei Jam[3], Sean K. McKenzie[4], Nathan D. Olson[5], Wouter De Coster[6,7], Jonghun Park[3], Bida Gu[8], Justin Wagner[5], Michael A. Eberle[2], Melissa Gymrek[3,9], Mark J. P. Chaisson[8], Justin M. Zook[5,12] & Fritz J. Sedlazeck[1,10,11,12]



TRGT: Tandem repeat variation det

Article | Published: 02 January 2024

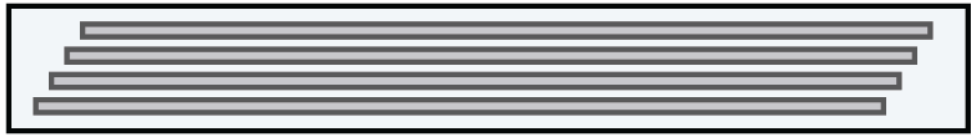
Characterization and visualization of tandem repeats at genome scale

Egor Dolzhenko, Adam English, Harriet Dashnow, Guilherme De Sena Brandine, Tom Mokveld, William J. Rowell, Caitlin Karniski, Zev Kronenberg, Matt C. Danzi, Warren A. Cheung, Chengpeng Bi, Emily Farrow, Aaron Wenger, Khi Pin Chua, Verónica Martínez-Cerdeño, Trevor D. Bartley, Peng Jin, David L. Nelson, Stephan Zuchner, Tomi Pastinen, Aaron R. Quinlan, Fritz J. Sedlazeck & Michael A. Eberle

Nature Biotechnology **42**, 1606–1614 (2024) | Cite this article

9816 Accesses | **9** Citations | **202** Altmetric | Metrics

ALIGNED HIFI READS



REPEAT DEFINITION

chrX 147912050 147912110 (CGG)n

Coordinates and structure of the repeat region

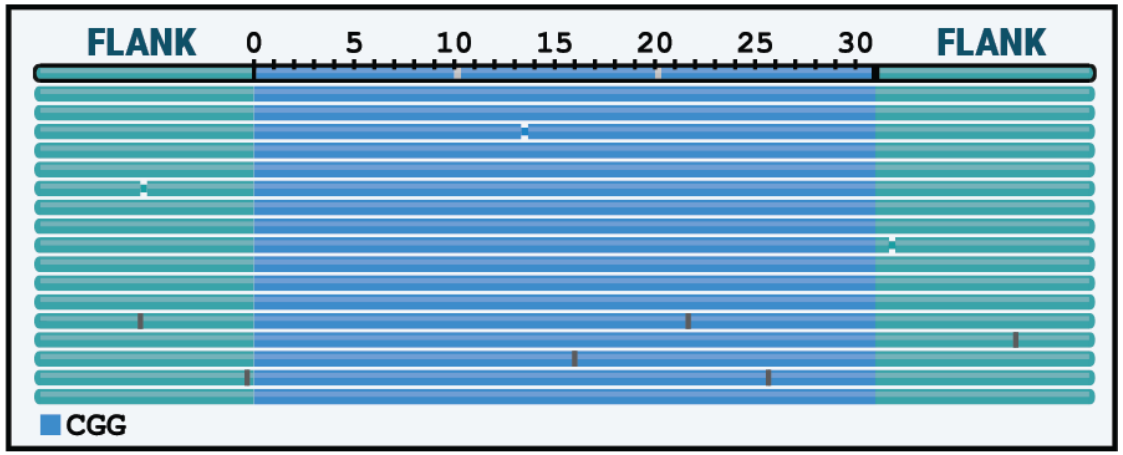
Tandem repeat genotyping tool (TRGT)

RESULTS

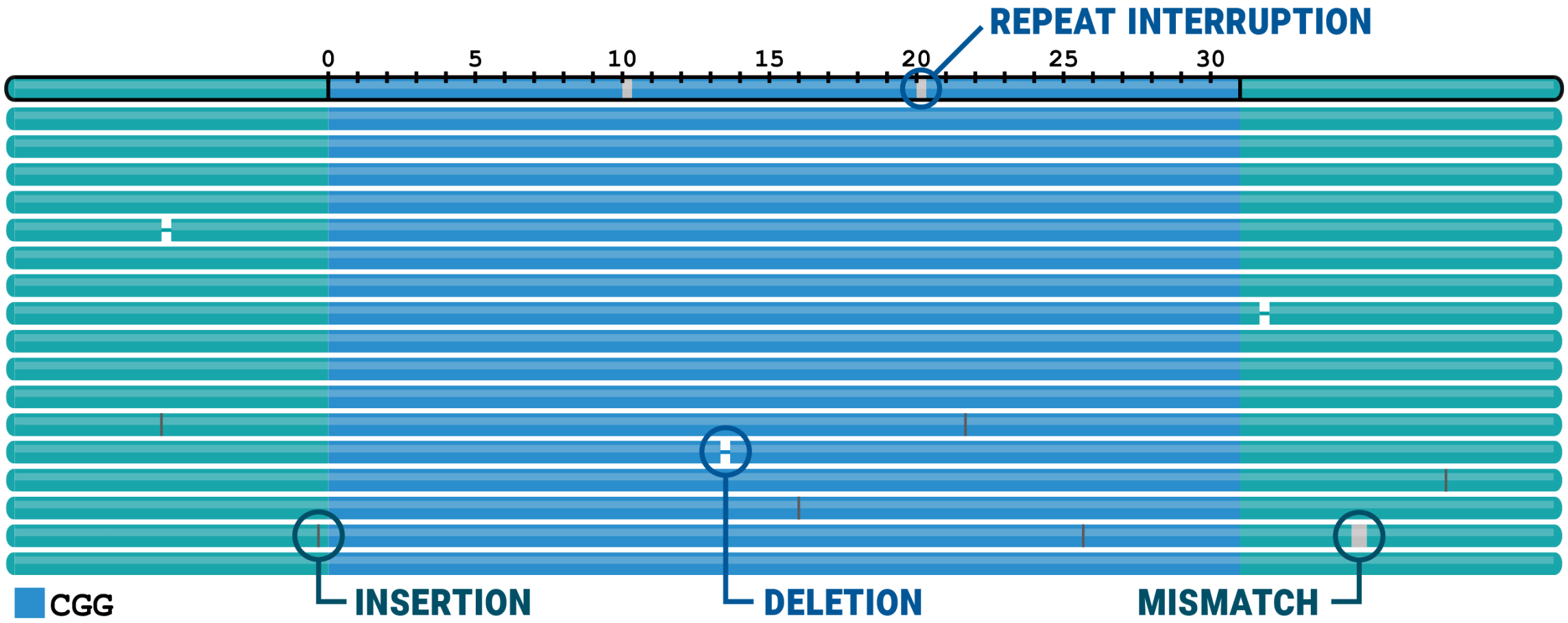
- Repeat sequences
- Repeat lengths
- Methylation levels
- and more!

Tandem repeat visualizer (TRVZ)

TRVZ PLOT



A TRVZ plot of an *FMR1* repeat spanning 31 copies of CGG motif in HG002

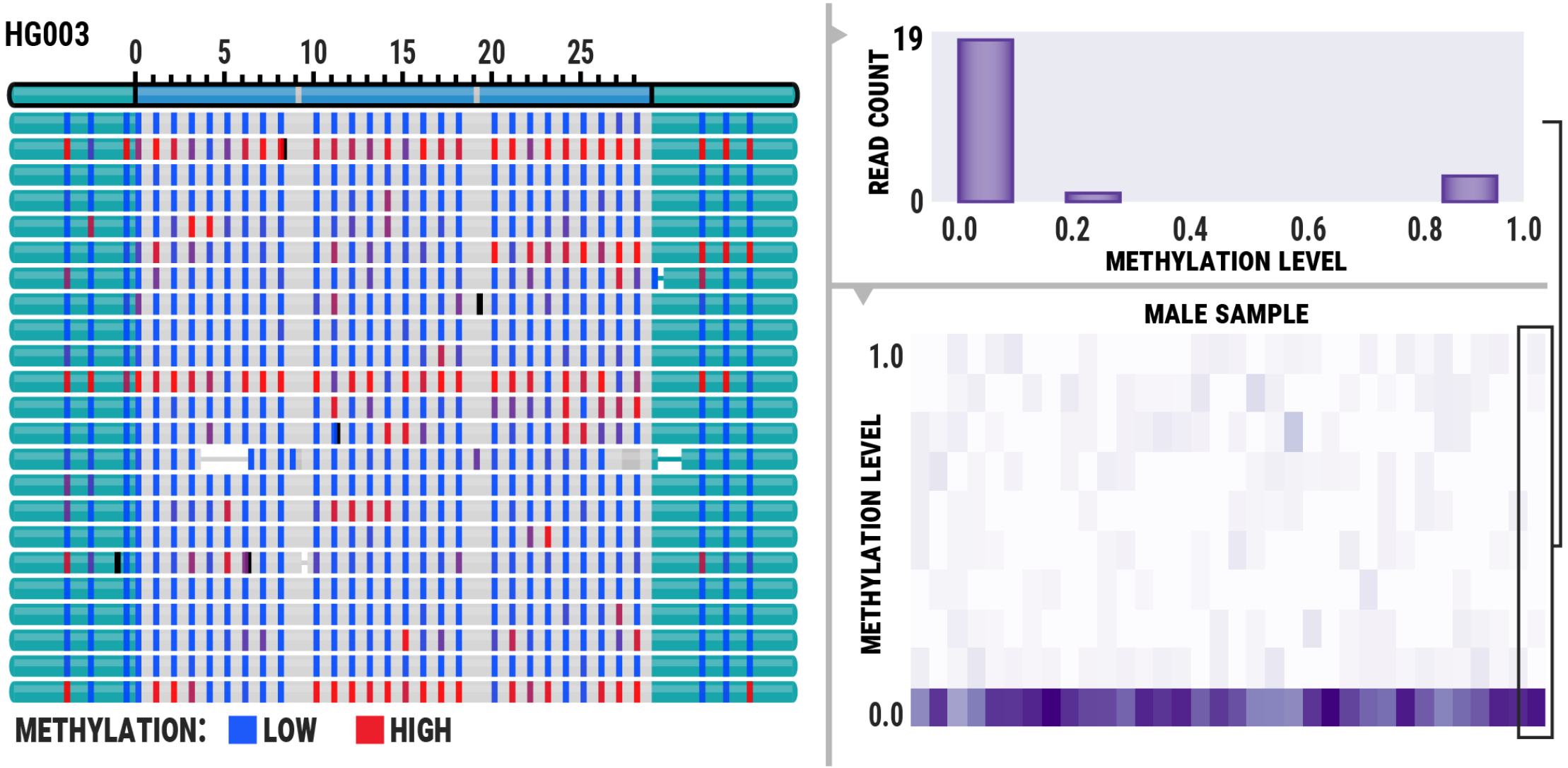


FMR1 repeat

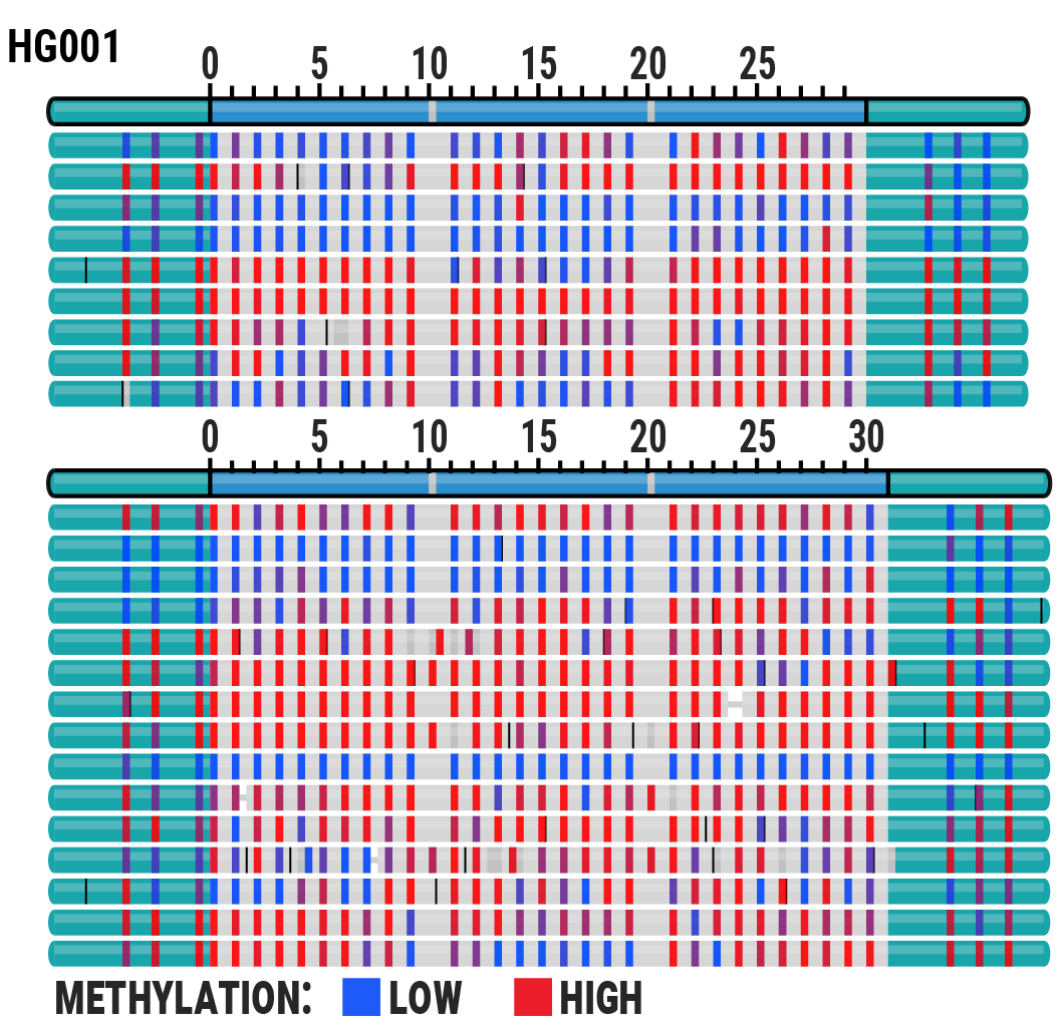
- Chromosome X located repeat
- Impact fragile X syndrome
  - Length exceeding 200 CGGs & methylation of the repeat and the promoter
  - Pre mutations: 55-200 repeats
  - Full mutations: 200+ repeats
- Highly polymorphic / mosaic if larger allele

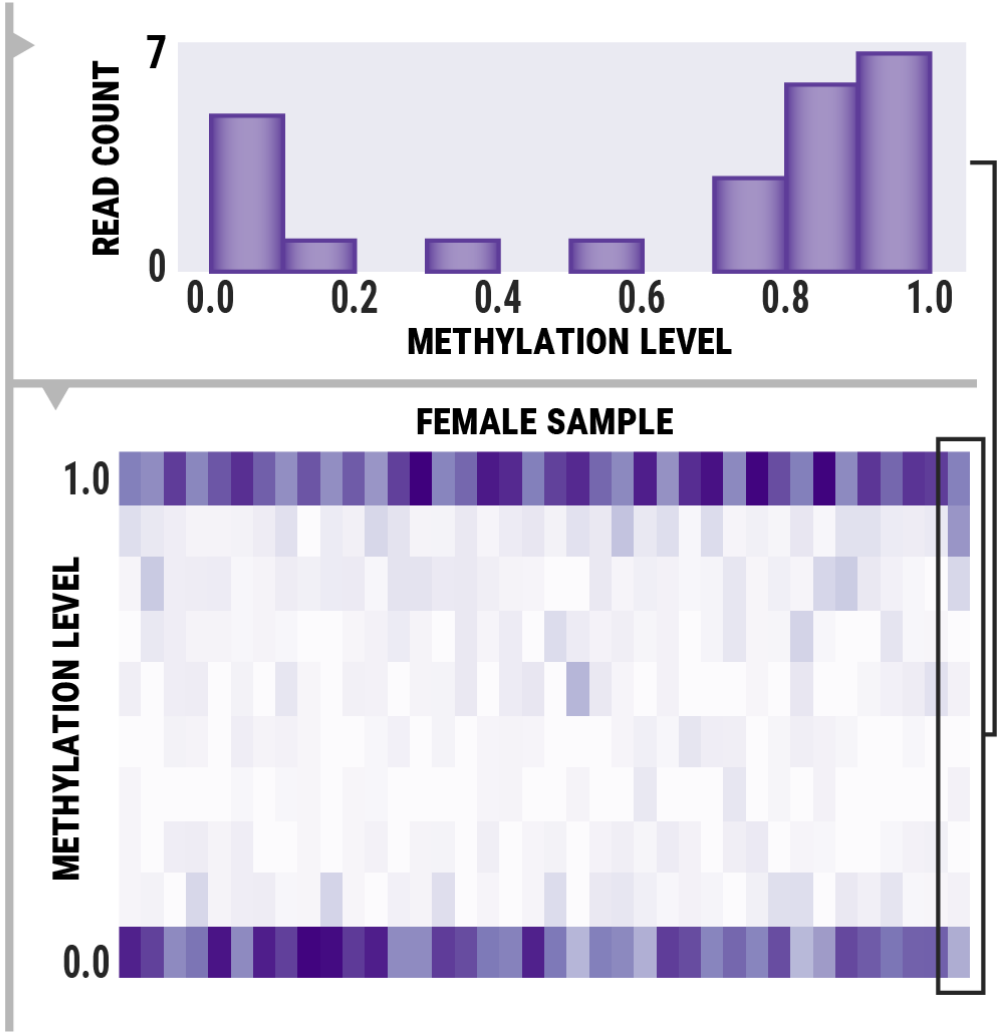
- Sequencing brains from Pre + Full mutation carriers with PacBio high coverage to understand motive, methylation & mosaic mutations better.
  - Peng Jin, David Nelson, Emily Allen & HGSC

FMR1: Male

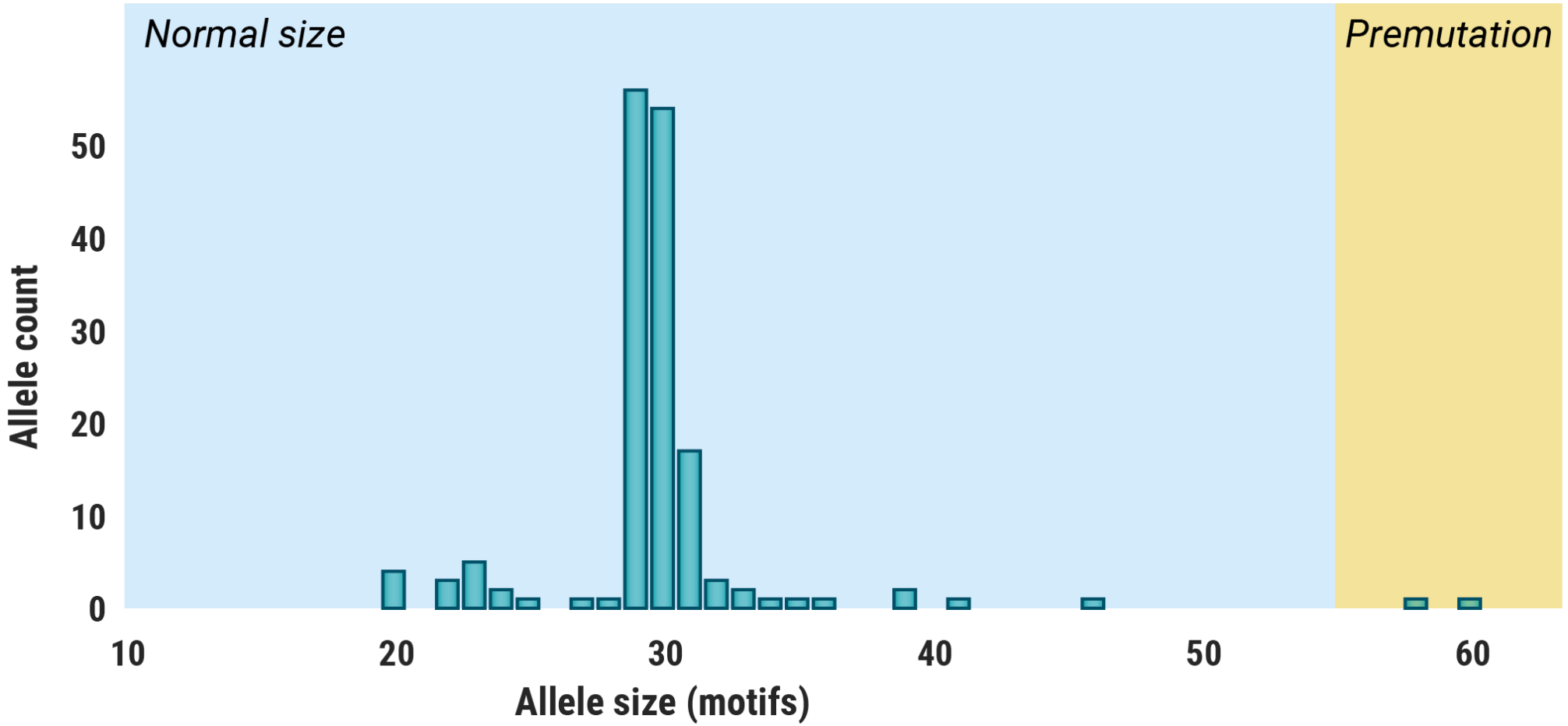

HG003
0
5
10
15
20
25
METHYLATION: LOW HIGH
19
READ COUNT
0
0.0
0.2
0.4
0.6
0.8
1.0
METHYLATION LEVEL
MALE SAMPLE
1.0
METHYLATION LEVEL
0.0

FMR1: Female

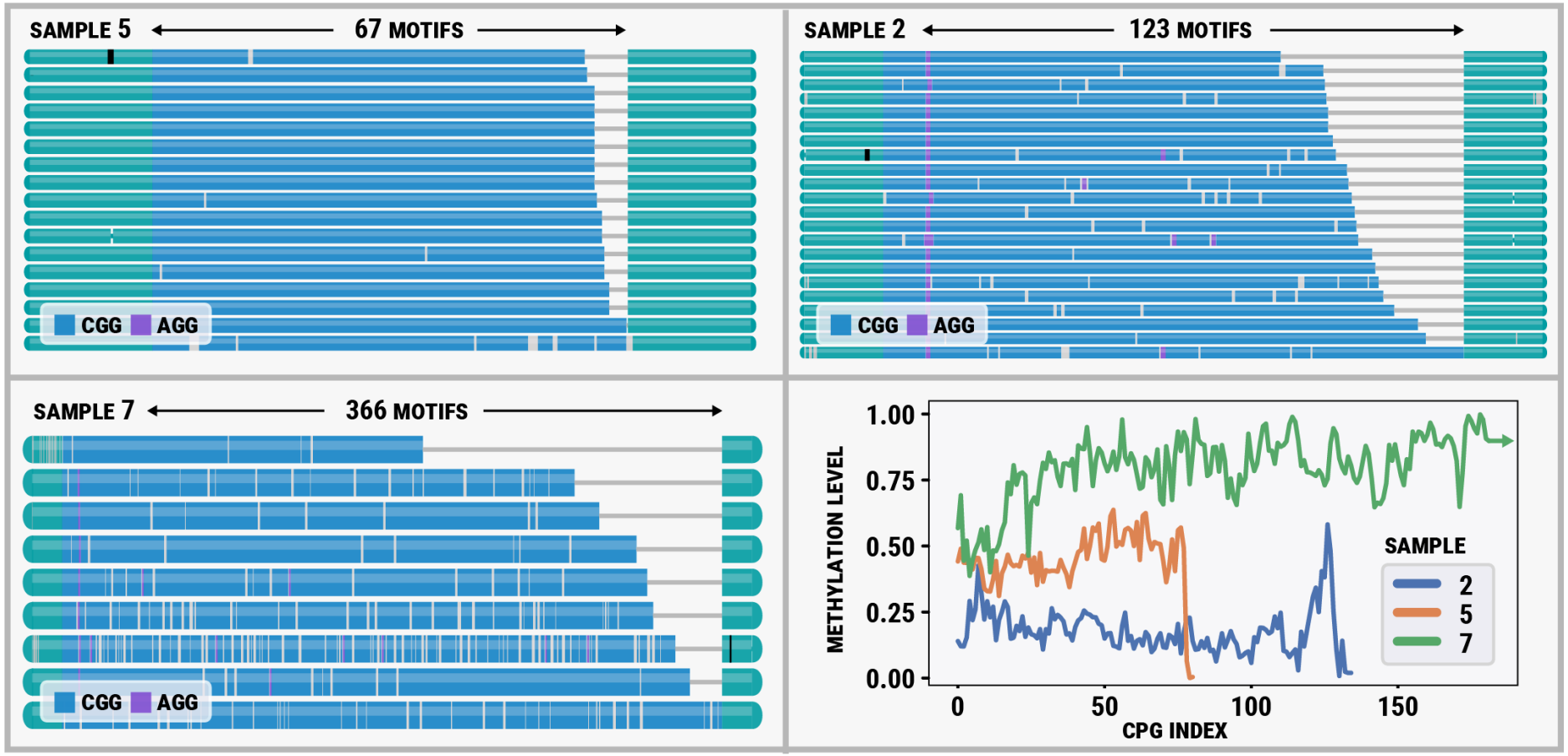

HG001
0
5
10
15
20
25
0
5
10
15
20
25
30
METHYLATION:
LOW
HIGH


READ COUNT
7
0
0.0
0.2
0.4
0.6
0.8
1.0
METHYLATION LEVEL
FEMALE SAMPLE
1.0
METHYLATION LEVEL
0.0

We identified two *FMR1* premutations in 100 HPRC samples

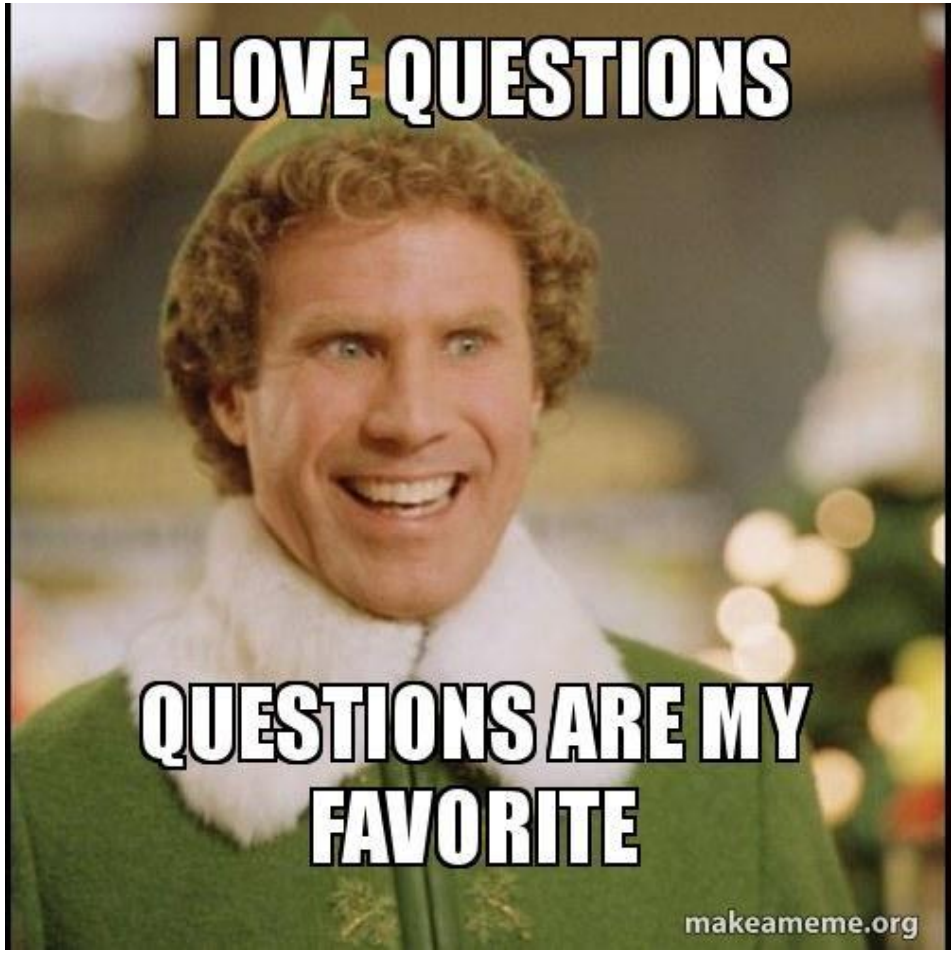


FMR1 pre-mutations and a full-mutation in samples from prefrontal cortex



Question: 3

What are the problems of long reads?

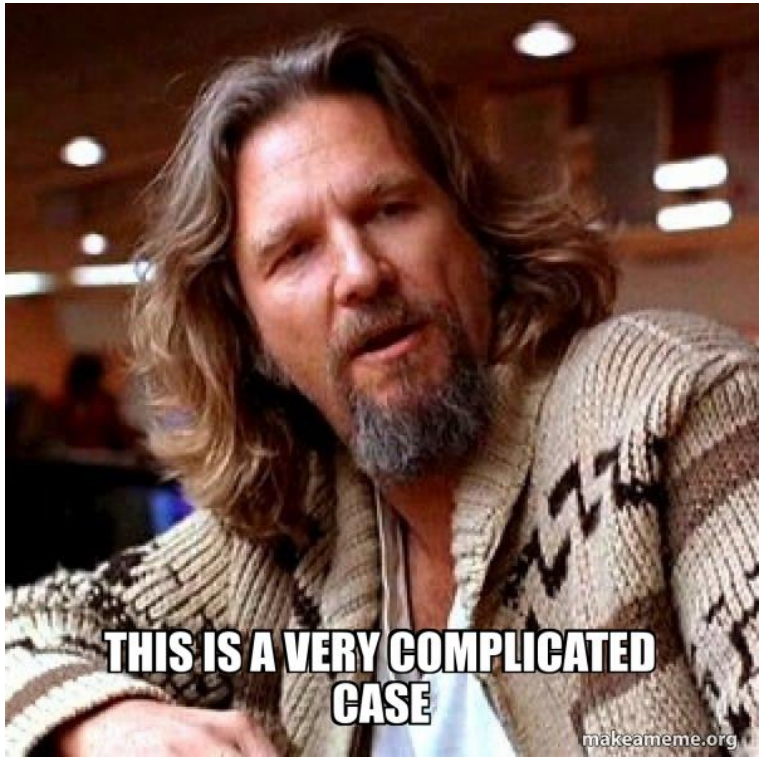


So… when are two SV the same?

- SNV/Indel: differences per bp.
  - Often used methods:
    - bcftools
- SV: majority occur in TR
  - We need to allow for some flexibility/differences
  - Different strategies across the years..
    - Reciprocal overlap (motivated by CNV)
    - Breakpoint agreement
    - Length agreement
    - Sequencing agreement
  - Often used methods:
    - SURVIVOR: https://github.com/fritzsedlazeck/SURVIVOR
    - Truvari: https://github.com/ACEnglish/truvari

So how to get a population catalog?

- SNV/ indels: gVCF files!
- SV: its complicated..
  - Calling -> merging -> re-genotyping -> merging (hours)
  - Sniffles-> merging (snf) (minutes)
- Re-genotyping
  - Short reads: Paragraph, Svtyper (easiest), STIX*
  - Long reads: Kanpig, SVJedi



Article | Open access | Published: 04 April 2025

K-mer analysis of long-read alignment pileups for structural variant genotyping

Adam C. English, Fabio Cunial, Ginger A. Metcalf, Richard A. Gibbs & Fritz J. Sedlazeck

Nature Communications **16**, Article number: 3218 (2025) | Cite this article

5413 Accesses | **6** Citations | **28** Altmetric | Metrics

JOURNAL ARTICLE

Evaluation of computational genotyping of structural variation for clinical diagnoses

Varuna Chander, Richard A Gibbs, Fritz J Sedlazeck

GigaScience, Volume 8, Issue 9, September 2019, giz110,
https://doi.org/10.1093/gigascience/giz110

Published: 08 September 2019 **Article history**

Alignment format

- SAM / BAM files are standard to report read alignments
- It includes information about aligned segments.
- The file is separated:
  - Header: starts with @
  - Body: each line 1 segment

Alignment format: header

- @HD:
  - first line includes information about the format and the version of the format.
- @SQ:
  - Reference sequence dictionary. Includes the information about the reference sequence that were used (SN:Name \t LN:length)
- @PG: Program information
  - Lists information about the programs, filters that were used including the parameter.
- @RG: Read group

Alignment format: entry

```
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
```

1. Read name
2. SAM Tag (next slide)
3. Reference (eg. Chromosome)
4. Position on the reference
5. MapQ
6. CIGAR string (coming)
7. Rnext (chr of mate/paired read)
8. Pnext (position of mate/paired read)
9. Tlen (distance between the two pairs)
10. SEQ sequence of the read
11. Quality values
12. Additional fields: eg. NM, MD string (optional)

Alignment format: SAM tags

- A combination of bitwise flags
  - https://broadinstitute.github.io/picard/explain-flags.html
- The combination of flags forms an integer number indicating the properties of the read:
  - 65: Paired + first pair + mapped
  - 73: 65+ mate is unmapped
  - 0: mapped, plus strand
  - 4: unmapped
  - 5: paired, unmapped read

Alignment format: CIGAR string

Examples:

- 25M1I19M6S
- 3M1D26M1D13M9S

- Encodes indels and match/mismatches
- M=match/mismatch
- I= insertion, D=deletion
- S=soft clipped, H=hard clipped

- Problem: we need further information to identify the substitutions -> MD string!

Alignment format: MD string

Examples:

- 25M1I19M6S MD:Z:44
- 3M1D26M1D13M9S MD:Z:3^A26^C13

- Indicates reference alleles for substitutions and deletions (insertions are encoded in the sequence tag)
- Numbers: matches
- ^A: a deletion with A as a reference allele
- Only A: Substitution

Alignment format: SAM/BAM

- Samtools package to:
  - Conversion/ compression
  - Manipulate
  - Sort
  - Query
  - Smaller operations
    - Depth , variants etc.

```
@HD     VN:1.6  SO:coordinate
@SQ     SN:1    LN:249250621
@SQ     SN:2    LN:243199373
@SQ     SN:3    LN:198022430
@SQ     SN:4    LN:191154276
@SQ     SN:5    LN:180915260
@SQ     SN:6    LN:171115067
@SQ     SN:7    LN:159138663
@SQ     SN:8    LN:146364022
@SQ     SN:9    LN:141213431
@SQ     SN:10   LN:135534747
@SQ     SN:11   LN:135006516
@SQ     SN:12   LN:133851895
@SQ     SN:13   LN:115169878
@SQ     SN:14   LN:107349540
@SQ     SN:15   LN:102531392
@SQ     SN:16   LN:90354753
@SQ     SN:17   LN:81195210
@SQ     SN:18   LN:78077248
@SQ     SN:19   LN:59128983
@SQ     SN:20   LN:63025520
@SQ     SN:21   LN:48129895
@SQ     SN:22   LN:51304566
@SQ     SN:X    LN:155270560
@SQ     SN:Y    LN:59373566
@SQ     SN:MT   LN:16569
@PG     ID:minimap2     PN:minimap2     CL:minimap2 -ax map-ont --MD -t 10 /users/mmahmoud
7f29a893-53cd-4875-954d-2dd33556da1a    256     22      16050001        0       370S56M1I1
cff9a442-8ecf-4ea5-a12a-912b49e956c5    256     22      16050001        0       3263S46M1D
947943f2-9f83-483c-a496-f3b9584b1fa6    256     22      16050001        0       1731S47M1I
90439624-8cba-4368-b8bc-7ee11842dbcc    256     22      16050001        0       17846S25M1
1e40e3dc-62e3-4eed-9658-b172484a5bc9    256     22      16050001        0       3682S46M1D
7c1dd982-4085-4270-9020-dd79d70e4e87    256     22      16050001        0       471S10M1D3
a42252ba-2369-42cb-99ae-20cf43d98b23    256     22      16050001        0       637S9M1I3M
91b9eab1-344e-4e74-9a0e-f93c1f0839fc    256     22      16050001        0       3763S25M1D
7950735d-0740-4a7e-8bad-93ef382be663    256     22      16050001        0       57561S10M1
58627a9c-ac6c-4e26-a37a-9ff9827e3ffb    256     22      16050001        0       30438S32M1
0c85d27d-eebe-489b-ad12-53ef571140a6    256     22      16050001        0       1032S108M1
```

File Formats:

- VCF file: (main format)
  - Tab separated text file
  - Header holds information on what means what.
  - Body: 1 entry per variant/position

- Bedpe file:
  - Tab separated text file, 12 defined columns.
  - 1 entry per variant/position

Hands on: VCF-Header

```
##fileformat=VCFv4.2
##source=LUMPY
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=STRANDS,Number=.,Type=String,Description="Strand orientation of the adjacency in BEDPE format (DEL:+-, DUP:-+, INV:++/--)">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS95,Number=2,Type=Integer,Description="Confidence interval (95%) around POS for imprecise variants">
##INFO=<ID=CIEND95,Number=2,Type=Integer,Description="Confidence interval (95%) around END for imprecise variants">
##INFO=<ID=MATEID,Number=.,Type=String,Description="ID of mate breakends">
##INFO=<ID=EVENT,Number=1,Type=String,Description="ID of event associated to breakend">
##INFO=<ID=SECONDARY,Number=0,Type=Flag,Description="Secondary breakend in a multi-line variants">
##INFO=<ID=SU,Number=.,Type=Integer,Description="Number of pieces of evidence supporting the variant across all samples">
##INFO=<ID=PE,Number=.,Type=Integer,Description="Number of paired-end reads supporting the variant across all samples">
##INFO=<ID=SR,Number=.,Type=Integer,Description="Number of split reads supporting the variant across all samples">
##INFO=<ID=BD,Number=.,Type=Integer,Description="Amount of BED evidence supporting the variant across all samples">
##INFO=<ID=EV,Number=.,Type=String,Description="Type of LUMPY evidence contributing to the variant call">
##INFO=<ID=PRPOS,Number=.,Type=String,Description="LUMPY probability curve of the POS breakend">
##INFO=<ID=PREND,Number=.,Type=String,Description="LUMPY probability curve of the END breakend">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=DUP:TANDEM,Description="Tandem duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=SU,Number=1,Type=Integer,Description="Number of pieces of evidence supporting the variant">
##FORMAT=<ID=PE,Number=1,Type=Integer,Description="Number of paired-end reads supporting the variant">
##FORMAT=<ID=SR,Number=1,Type=Integer,Description="Number of split reads supporting the variant">
##FORMAT=<ID=BD,Number=1,Type=Integer,Description="Amount of BED evidence supporting the variant">
```

Holds important information about the data listed below, file format

Hands on: VCF entries

```
1    10196130    11153_2  N    [2:11128811[N    .    .    SVTYPE=BND;STRANDS=--:4;SECONDARY;EVENT=1115
1    10196158    11154_2  N    N]2:11129218]    .    .    SVTYPE=BND;STRANDS=++:7;SECONDARY;EVENT=1115
1    10199540    7653_1   N    [1:16717319[N    .    .    SVTYPE=BND;STRANDS=--:6;EVENT=7653;MATEID=76
1    10199552    7654_1   N    N]1:16717620]    .    .    SVTYPE=BND;STRANDS=++:6;EVENT=7654;MATEID=76
1    10271879    7020     N    <DEL>    .    .    SVTYPE=DEL;STRANDS=+-:11;SVLEN=-256;END=10272135;CIP
1    10272132    7021     N    <DUP>    .    .    SVTYPE=DUP;STRANDS=-+:6;SVLEN=9059;END=10281191;CIPO
1    10274057    7022     N    <DUP>    .    .    SVTYPE=DUP;STRANDS=-+:9;SVLEN=9644;END=10283701;CIPO
1    10274072    7023     N    <DEL>    .    .    SVTYPE=DEL;STRANDS=+-:6;SVLEN=-9299;END=10283371;CIP
```

Start chromosome

Start position

Variant ID

Reference allele

Alternative allele

Quality

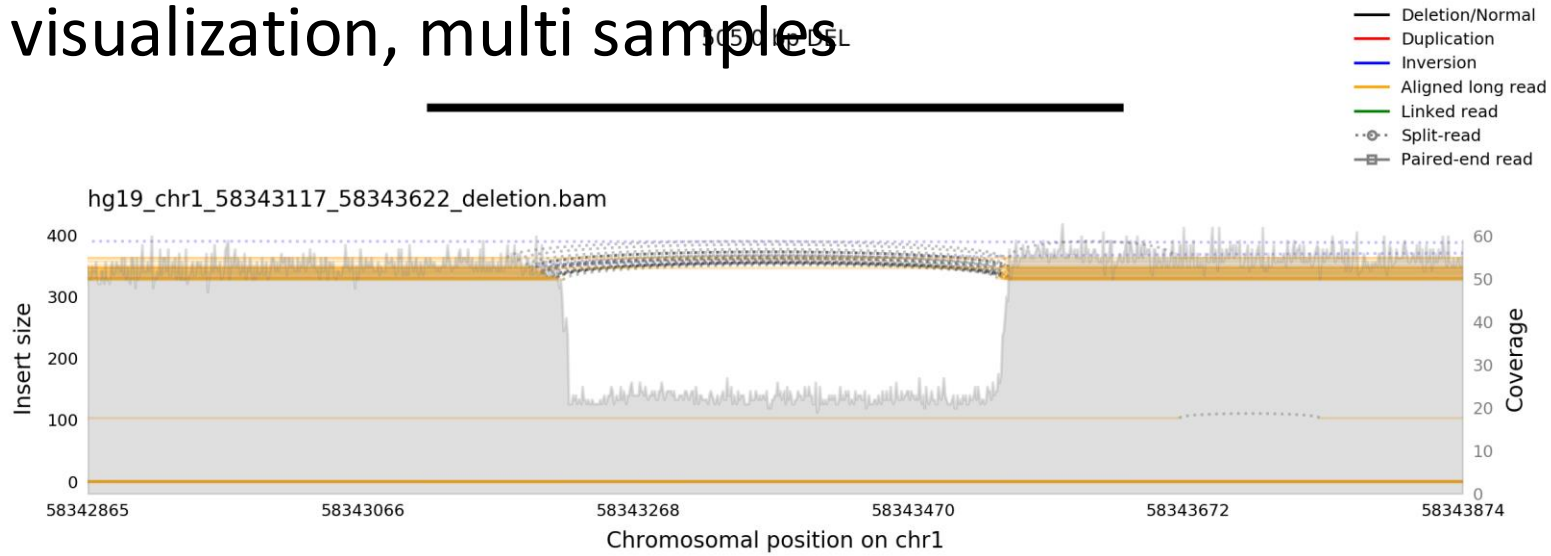
Filter

Additional information defined in the header

Visualization?

- IGV: comprehensive visualization
- - Complex and sometimes not very intuitive.
- + It shows everything! ☺

- Samplot: https://github.com/ryanlayer/samplot
- + program in terminal , nice visualization, multi samples
- - no INS, not interactive



Exercise Part 3: Long read based

- Utilize Oxford Nanopore Technology to identify SV
  - We will use Sniffles v2

- Go to: Day2 https://github.com/fritzsedlazeck/teaching_material
  - Files are also available locally. If you don’t find a file I have included download links.

Thank you

- The choice of the mapper matters!
- SV calling is SNP calling of ~2011.
- SV: Reads are typically shorter than the allele.
- Lot of noise in the data

Hands on section: Variant calling

- Keep in mind to check out the methods and not just copy paste!
- Ask questions, but try to think first if the path is ok etc.
- Take breaks! Otherwise ->