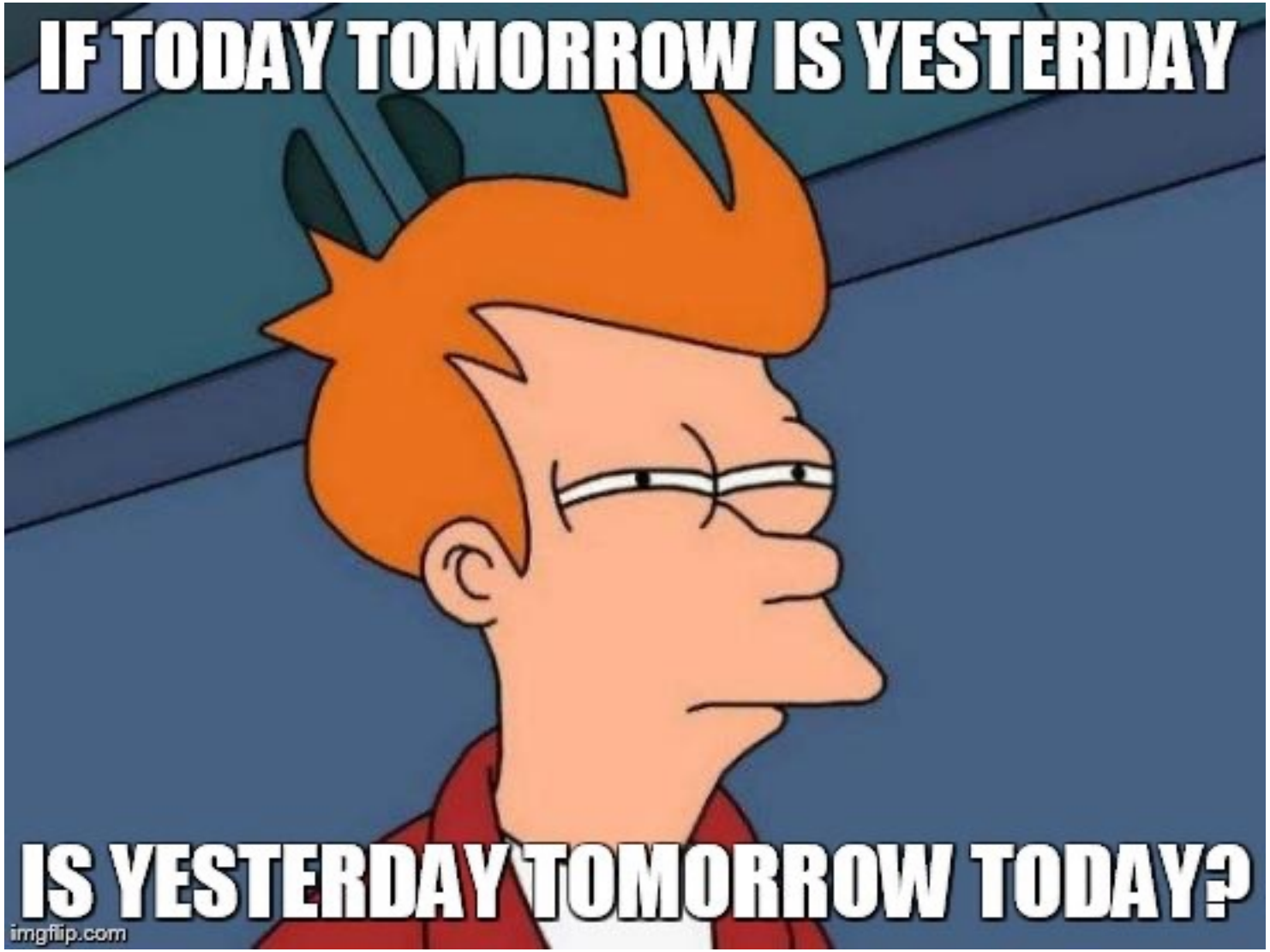
Methodologies for Structural Variant detection

Fritz Sedlazeck & Luis Paulin

Dec,12, 2023

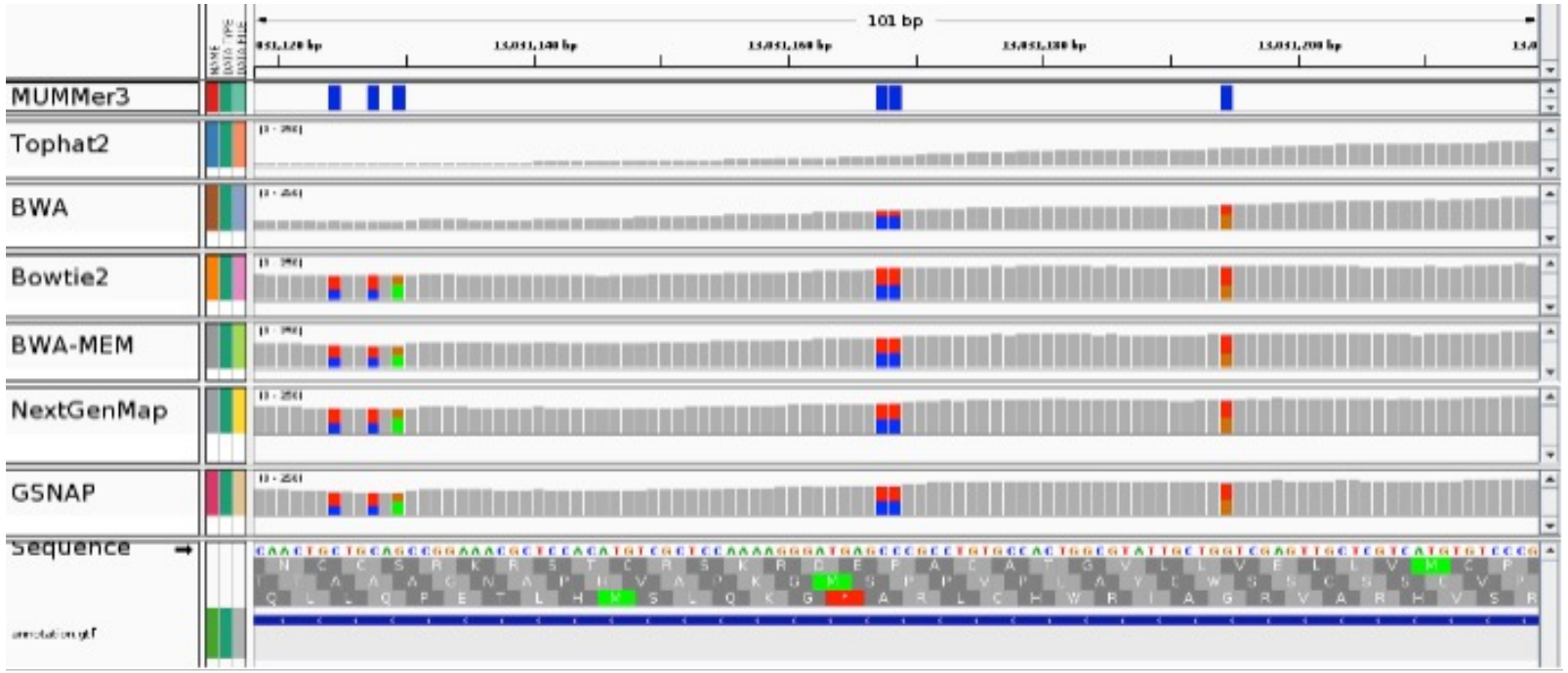
RICE

Recap from yesterday



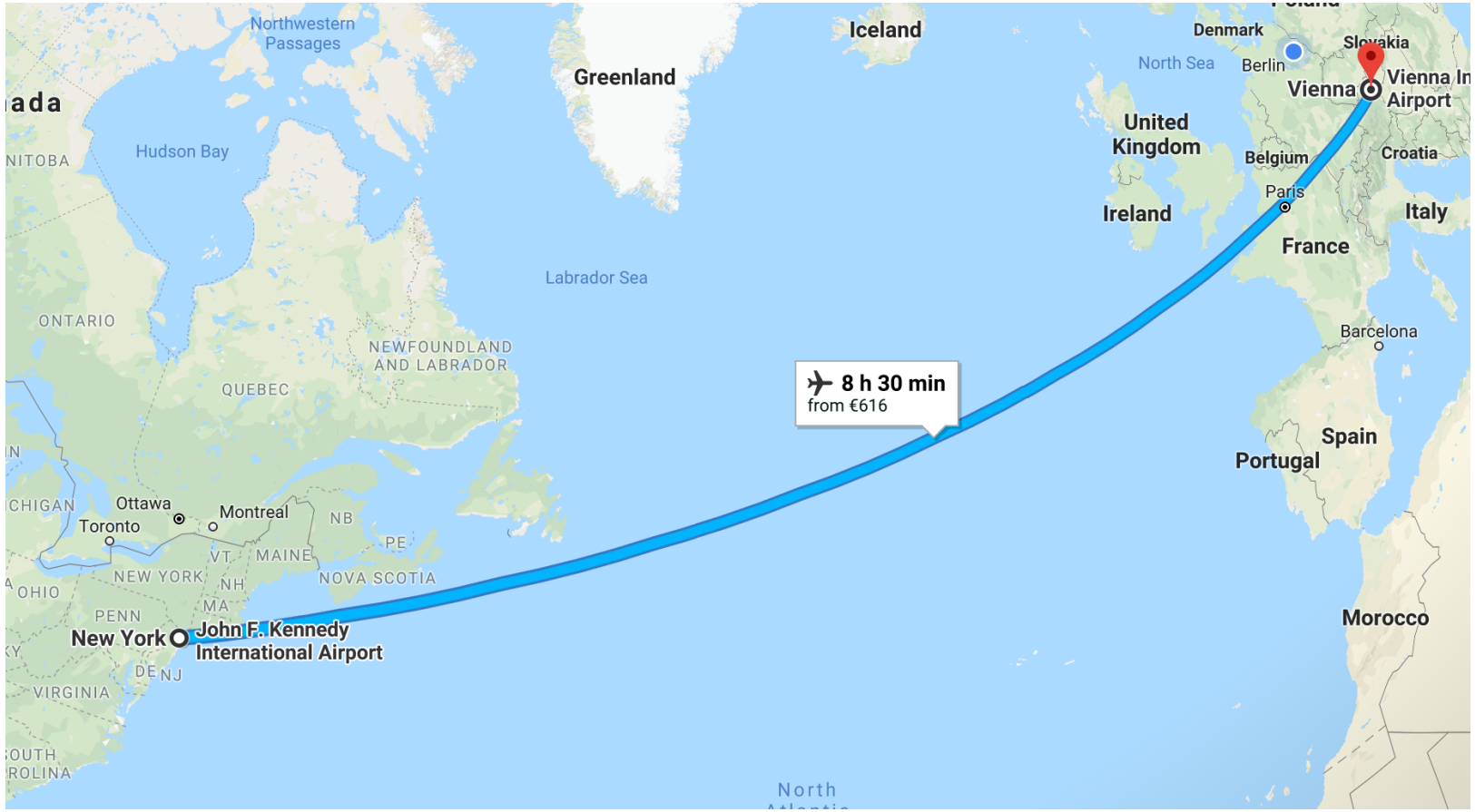
Are alignments solved?

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



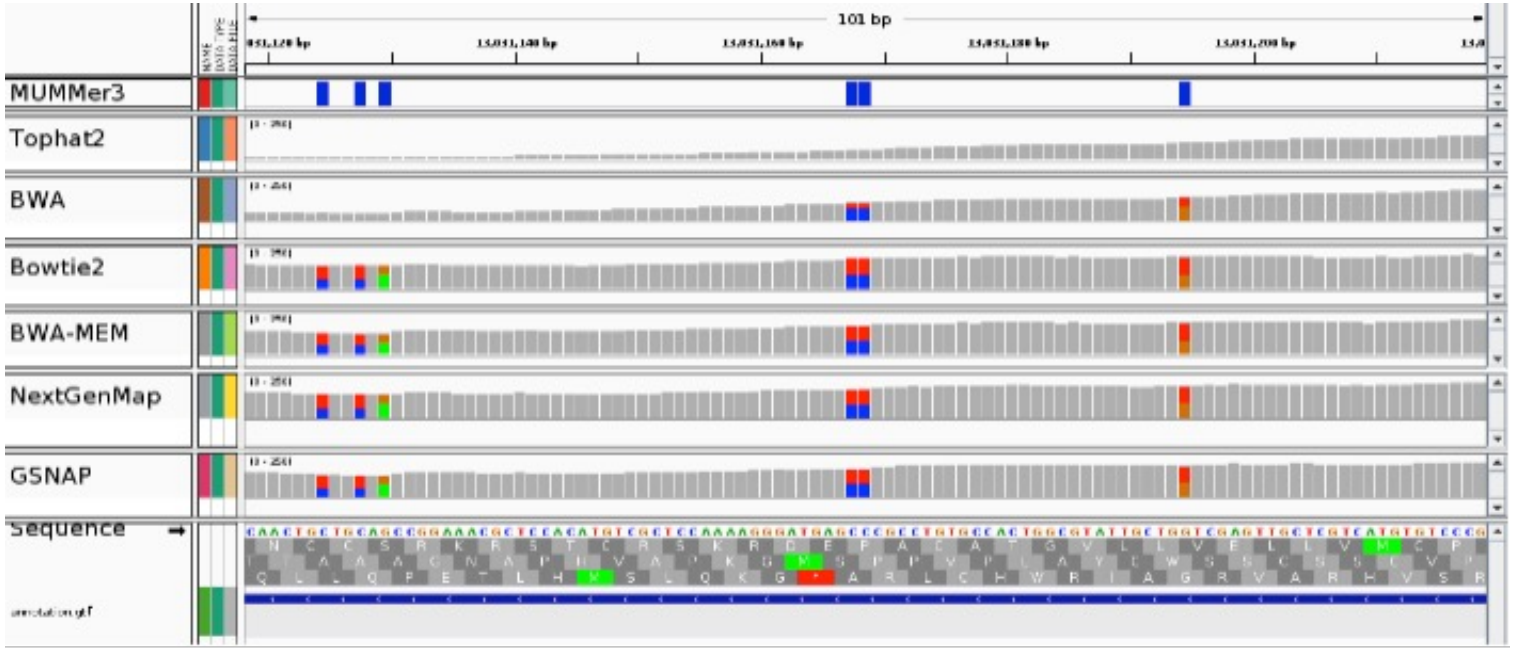
Mapping of reads

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



Mapping algorithms

- Short reads:
  - BWA
  - Bowtie2
  - Stampy
  - NGM
- Long reads
  - BlasR
  - Minimap2
  - NGMLR
  - Graphmap



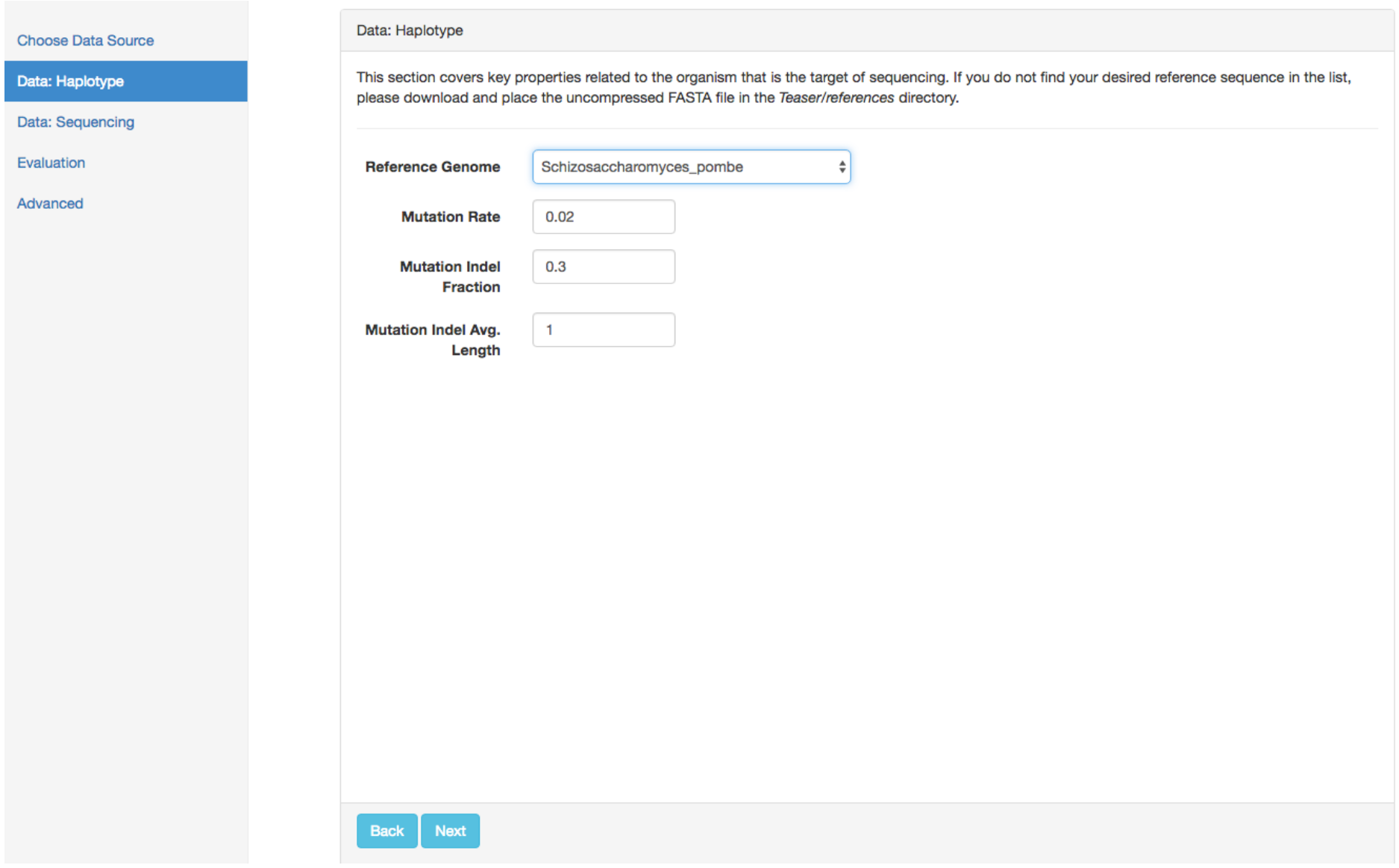
Teaser

Moritz Smolka

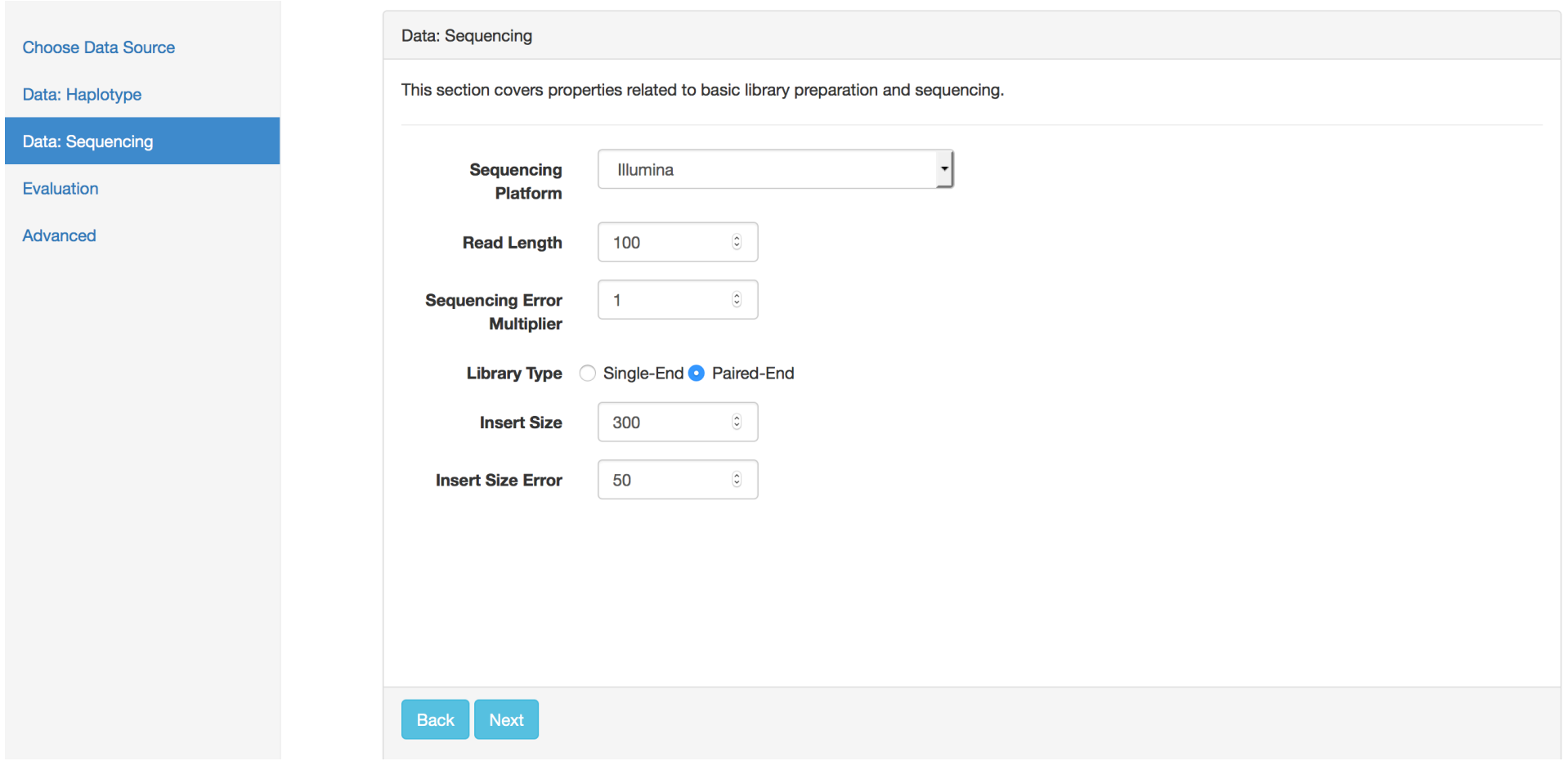
- Personalized benchmarking
  - Including benchmarking of parameter
- Short turnaround time
- Easy to extend/use

Smolka et al. (2015)

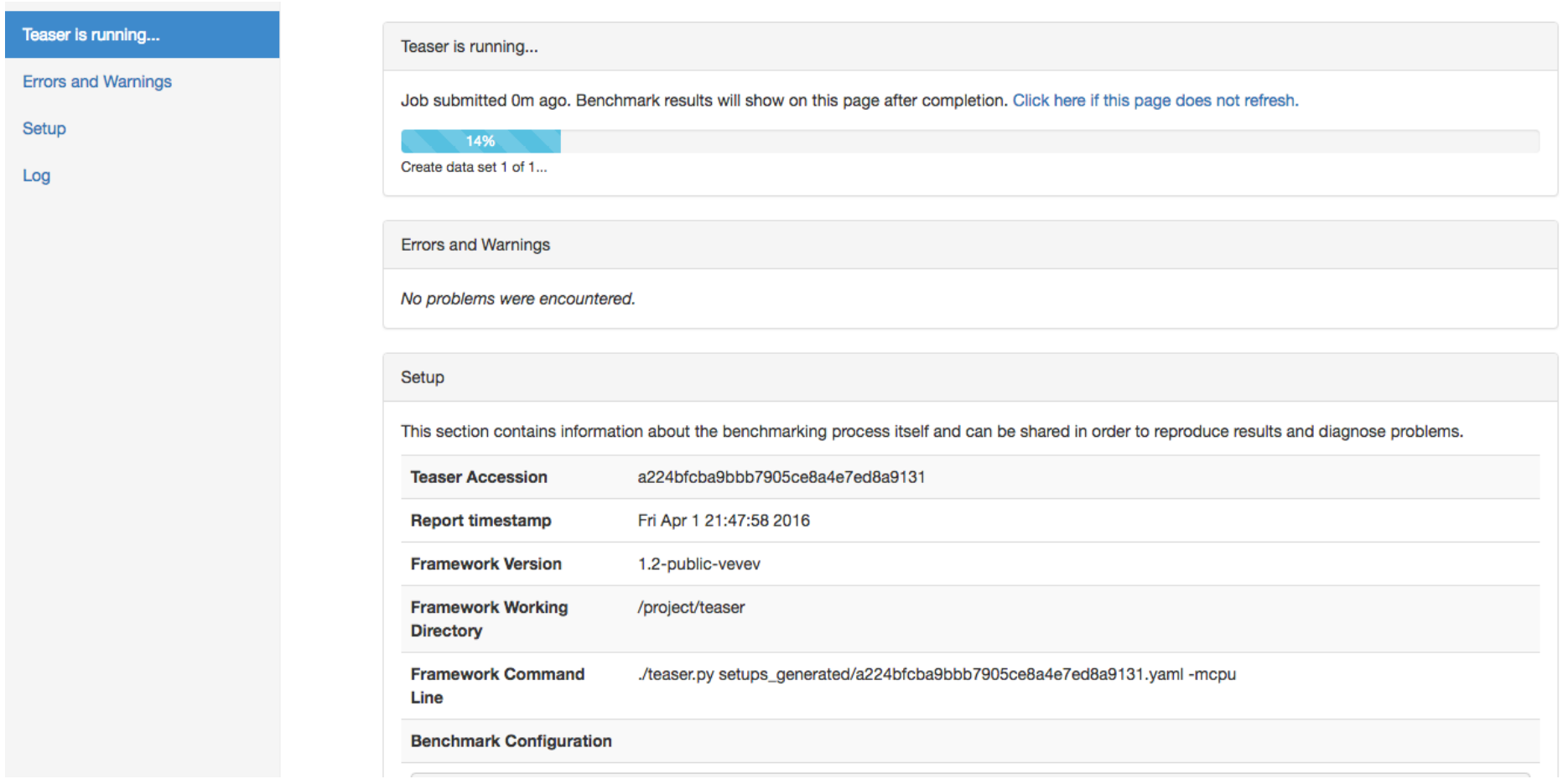
Teaser


Choose Data Source
Data: Haplotype
Data: Sequencing
Evaluation
Advanced
Data: Haplotype
This section covers key properties related to the organism that is the target of sequencing. If you do not find your desired reference sequence in the list, please download and place the uncompressed FASTA file in the Teaser/references directory.
Reference Genome
Schizosaccharomyces_pombe
Mutation Rate
0.02
Mutation Indel Fraction
0.3
Mutation Indel Avg. Length
1
Back
Next

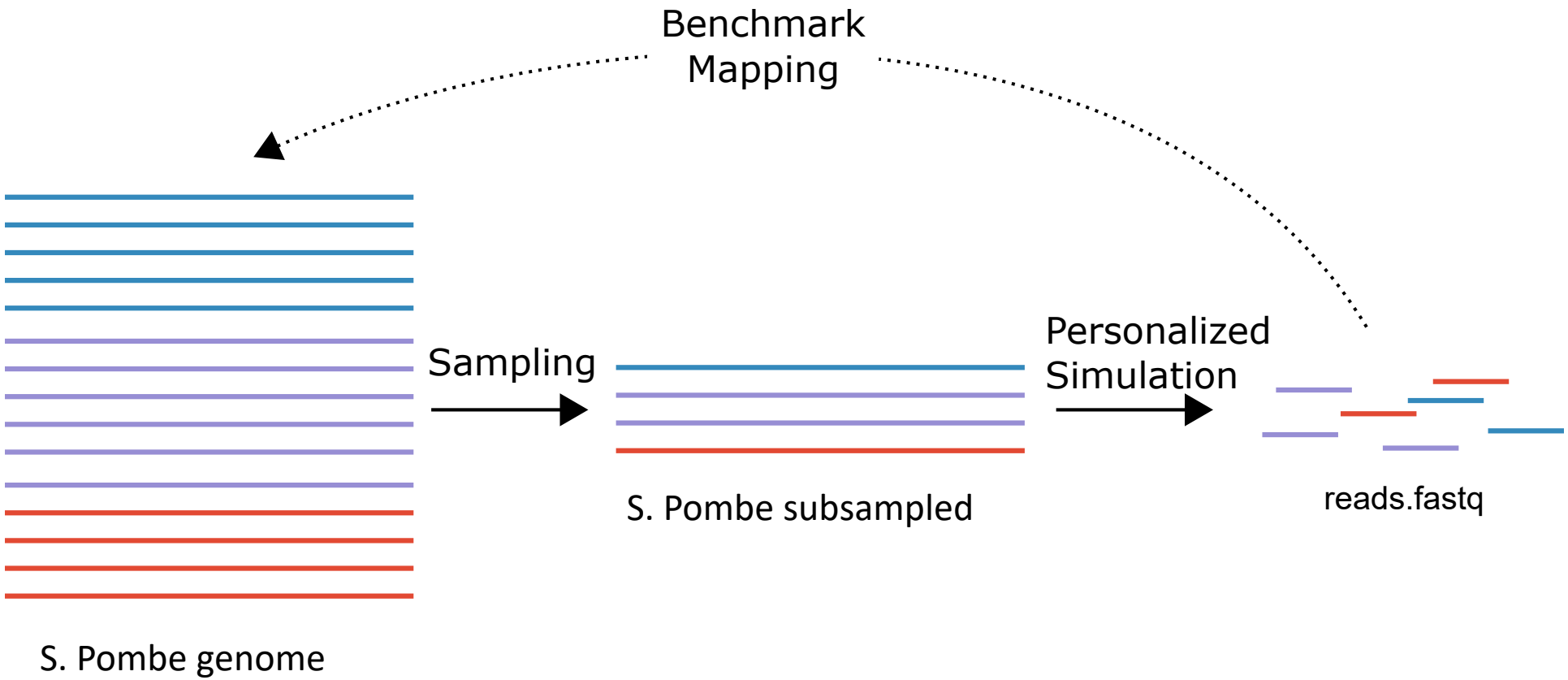
Teaser


Choose Data Source
Data: Haplotype
Data: Sequencing
Evaluation
Advanced
Data: Sequencing
This section covers properties related to basic library preparation and sequencing.
Sequencing Platform
Illumina
Read Length
100
Sequencing Error Multiplier
1
Library Type
Single-End
Paired-End
Insert Size
300
Insert Size Error
50
Back
Next

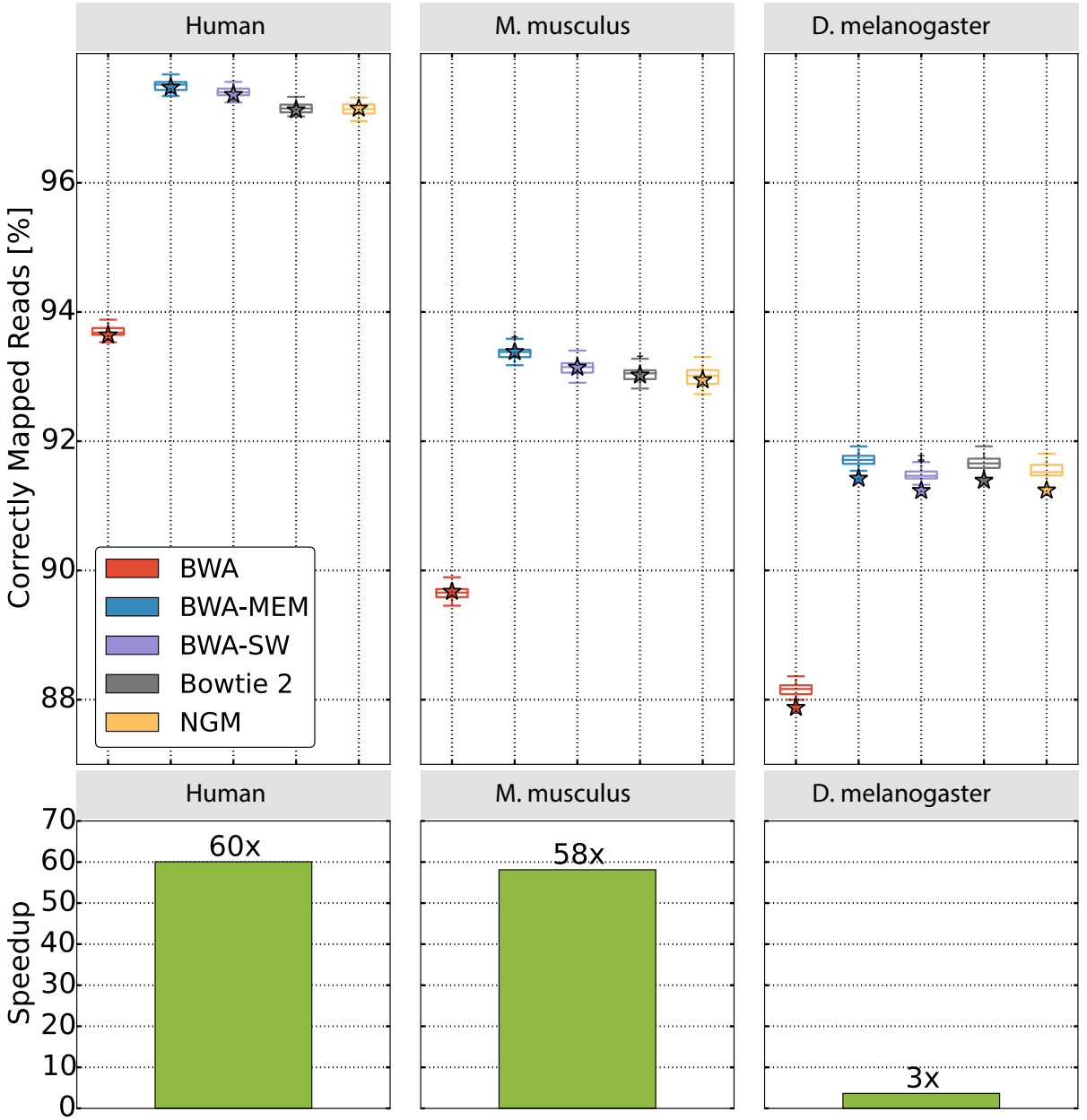
Teaser


Teaser is running...
Errors and Warnings
Setup
Log
Teaser is running...
Job submitted 0m ago. Benchmark results will show on this page after completion. Click here if this page does not refresh.
14%
Create data set 1 of 1...
Errors and Warnings
No problems were encountered.
Setup
This section contains information about the benchmarking process itself and can be shared in order to reproduce results and diagnose problems.
Teaser Accession
a224bfcba9bbb7905ce8a4e7ed8a9131
Report timestamp
Fri Apr 1 21:47:58 2016
Framework Version
1.2-public-vevev
Framework Working Directory
/project/teaser
Framework Command Line
./teaser.py setups_generated/a224bfcba9bbb7905ce8a4e7ed8a9131.yaml -mcpu
Benchmark Configuration

Teaser- cartoon

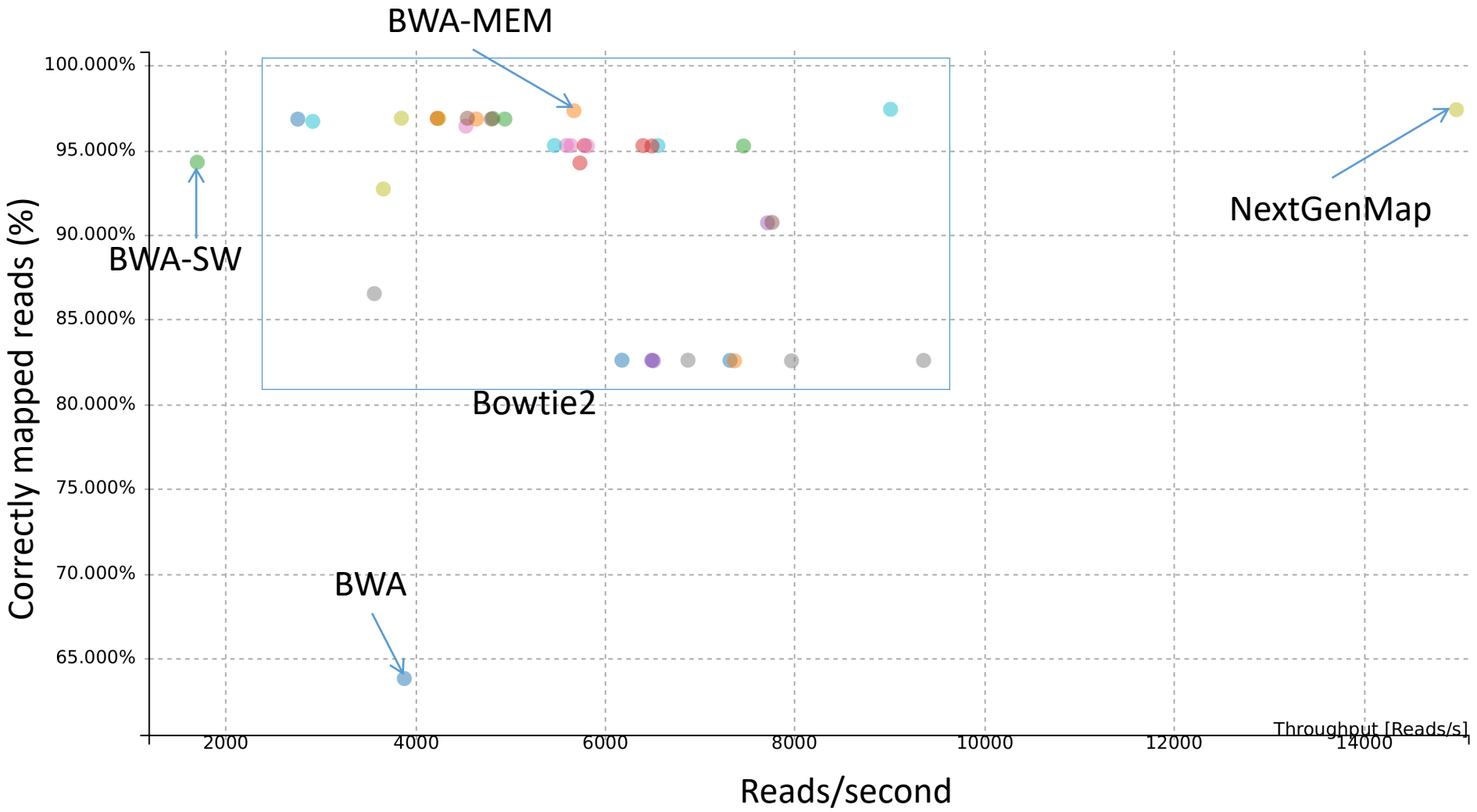


Teaser: subsampling



Teaser for *S. pombe*

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

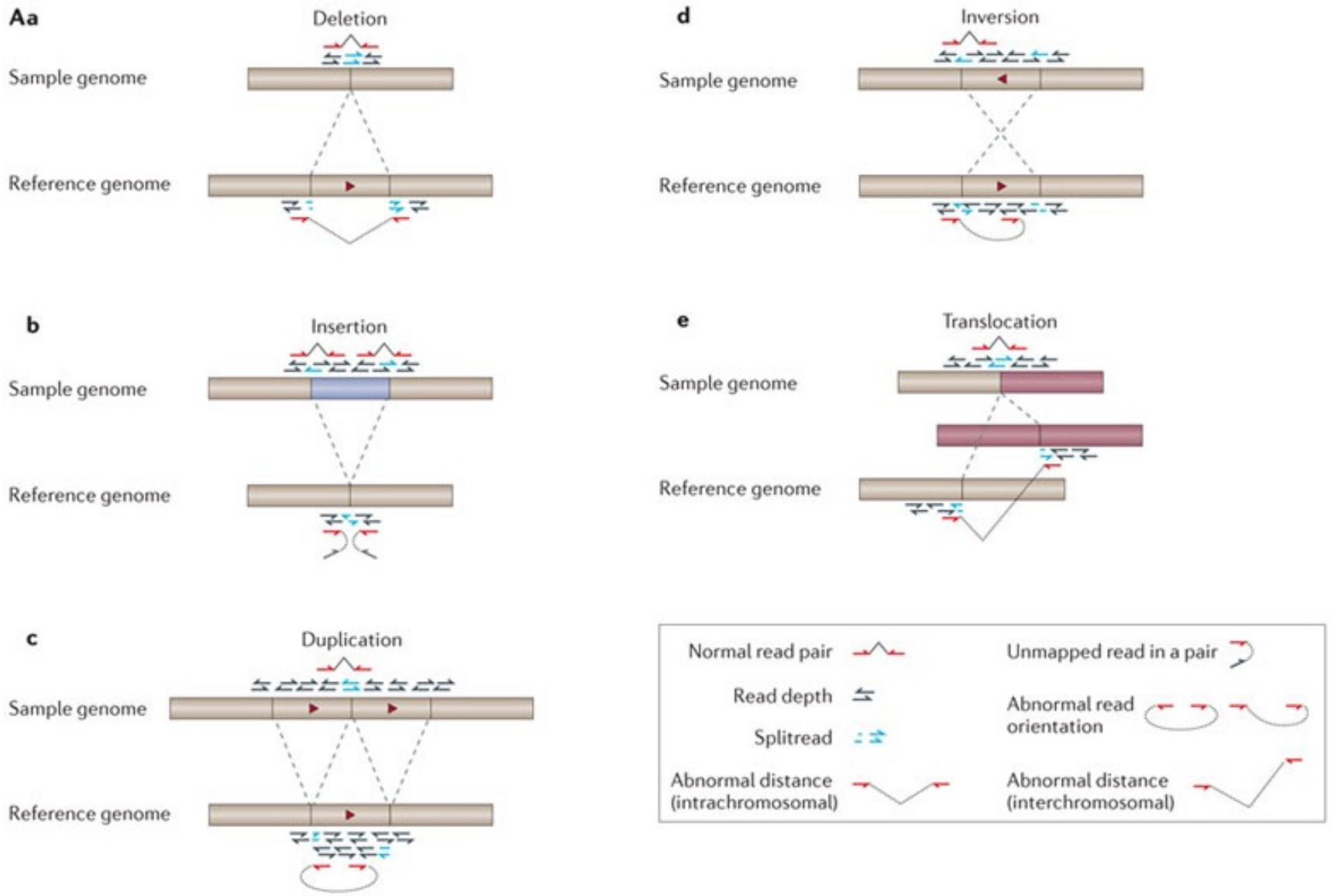


Variant calling

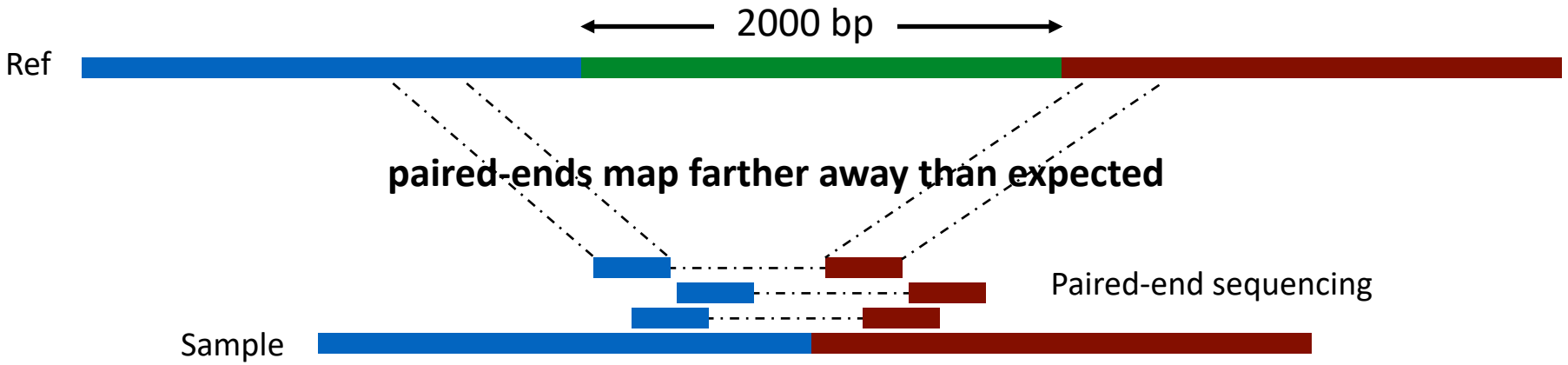
- SNV (no)
  - GATK[@] (meh), DRAGEN[@], DeepVariant[@,*], Clair[@,*]
- SV
  - DRAGEN[@], Manta[@], Sniffles[*], cuteSV[*], etc
- Tandem repeats
  - Expension Hunter[@], TRGT[*], Straglr[*], medaka[*] (new models)
- CNV
  - DRAGEN[@], Spectre[*],

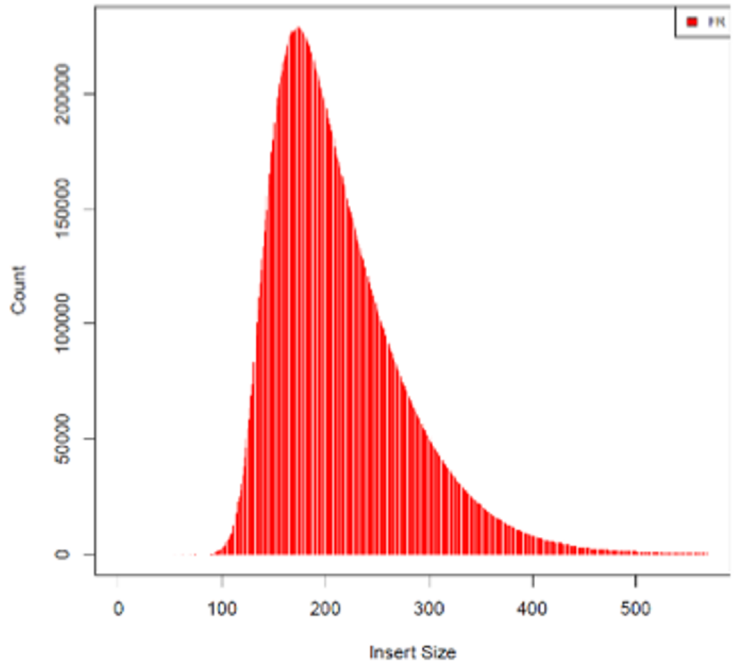
@: Short reads
* : long reads

How to detect Structural Variations



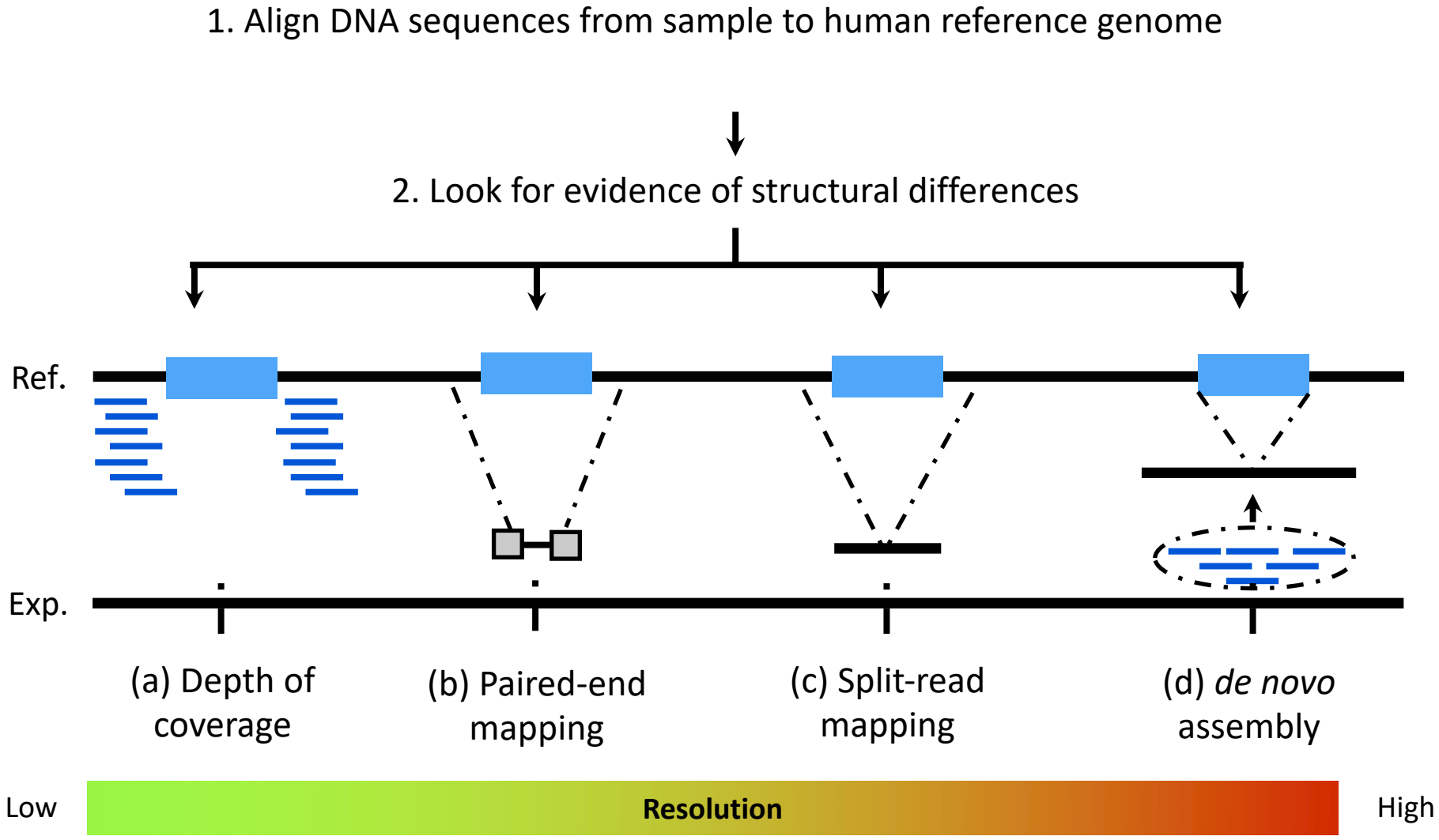
Looking for "discordant" paired-end fragments





Slide in collaboration with Ira Hall

Sequence alignment “signals” for structural variation



A probabilistic framework for SV discovery

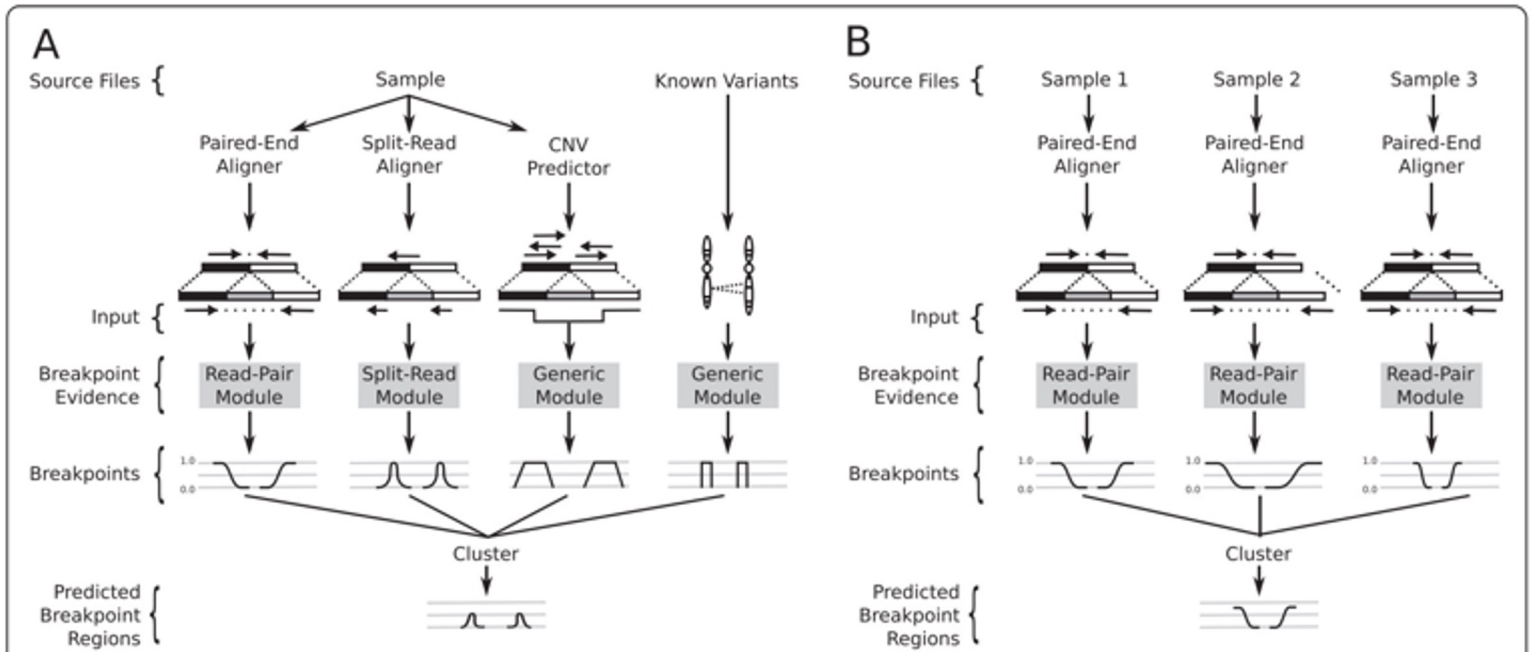


Figure 1 The LUMPY framework for integrating multiple structural variation signals. (A) A scenario in which LUMPY integrates three different sequence alignment signals (read-pair, split-read and read-depth) from a genome single sample. Additionally, sites of known variants are provided to LUMPY as prior 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.

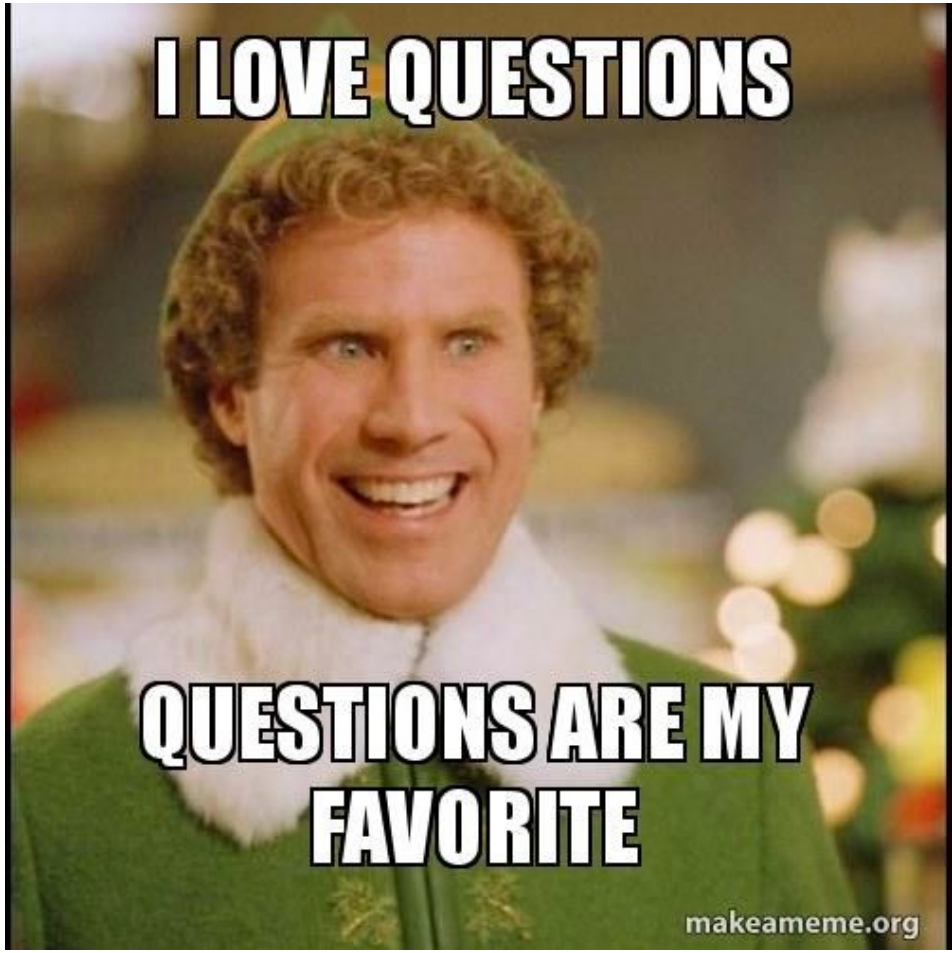
How to filter / choose the SV caller?

- 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

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

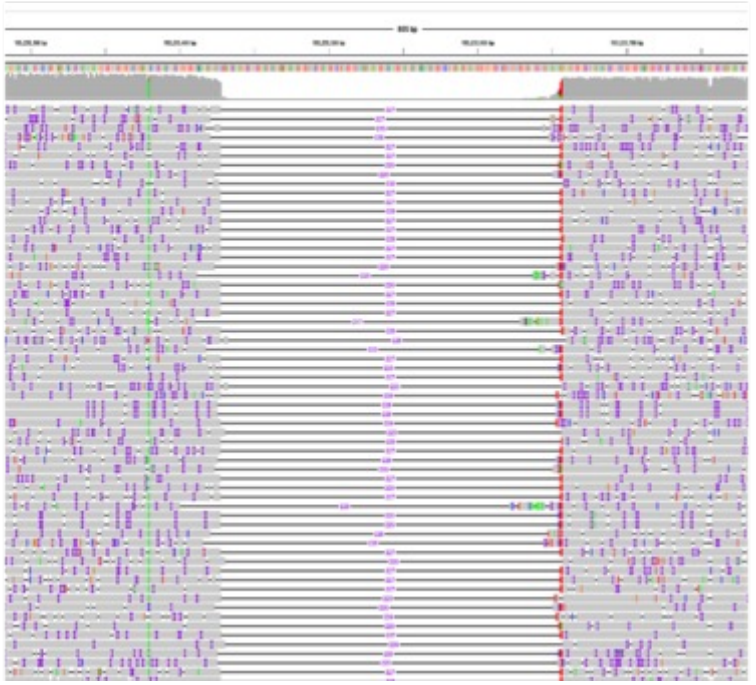
- (+) SVs in repetitive regions
- (+) Span SVs
- (+) Uniform coverage
- (+) Can identify more complex SVs

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

Mapping challenges

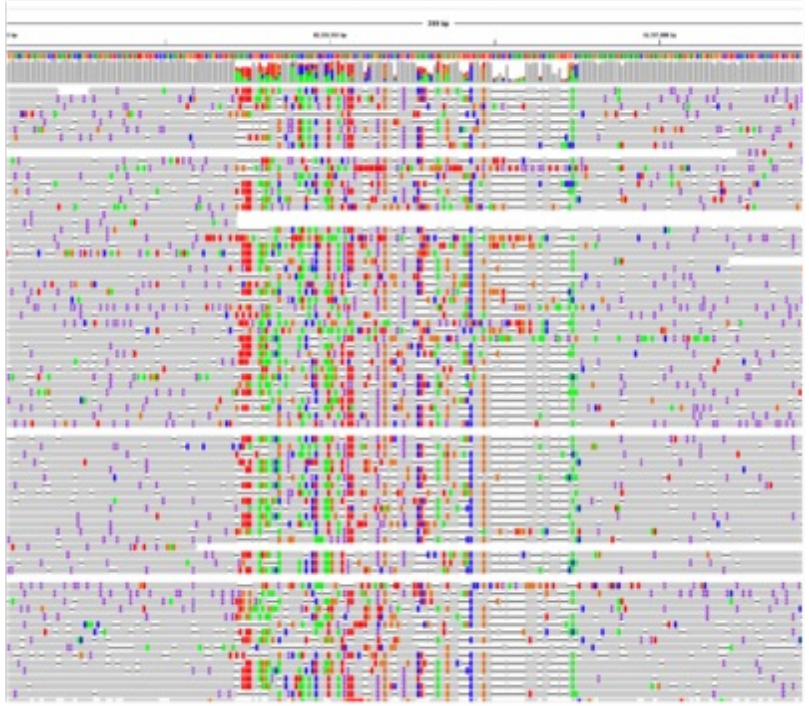
BWA-MEM:

NGMLR:



Mapping challenges

BWA-MEM:



NGMLR:

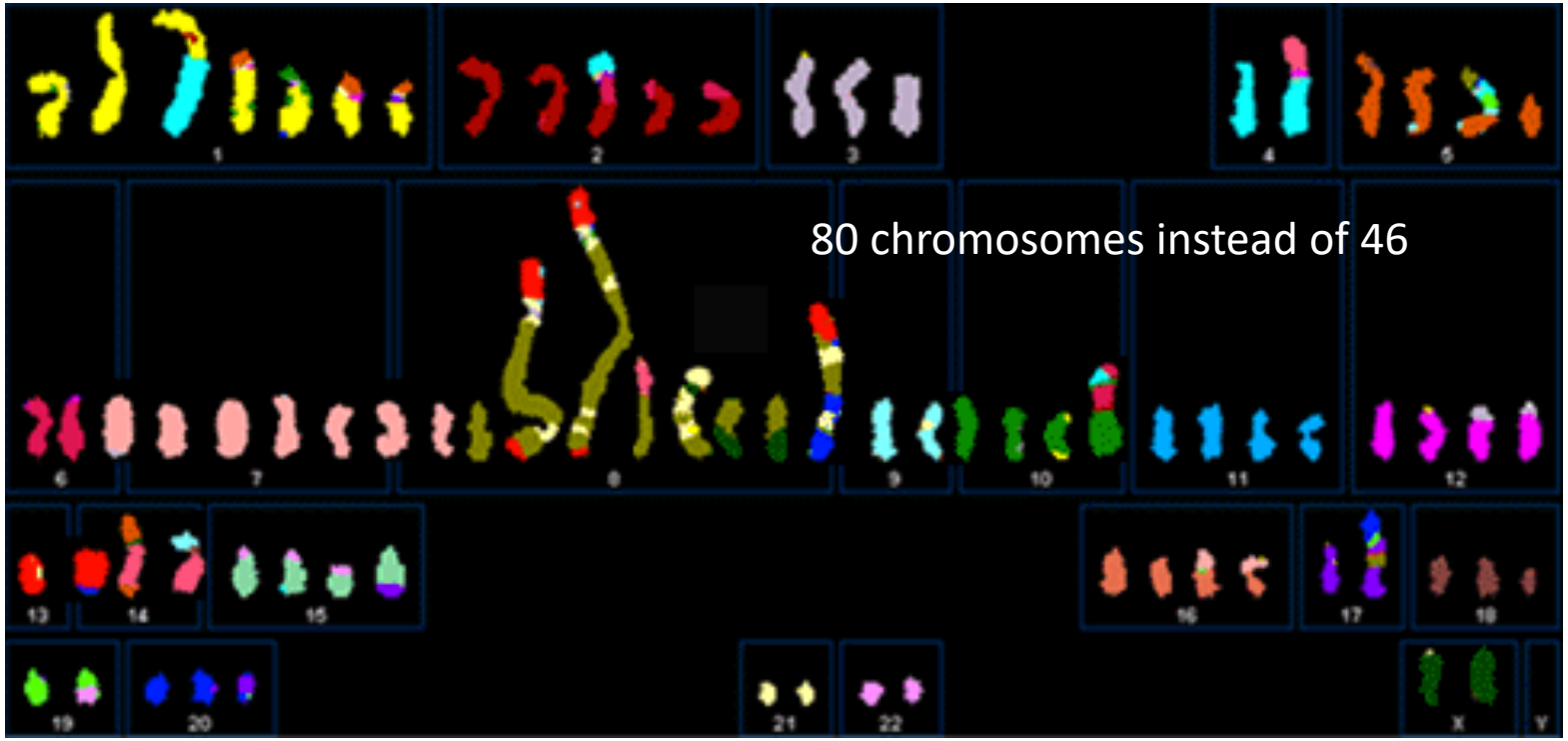
NGMLR + Sniffles

- 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

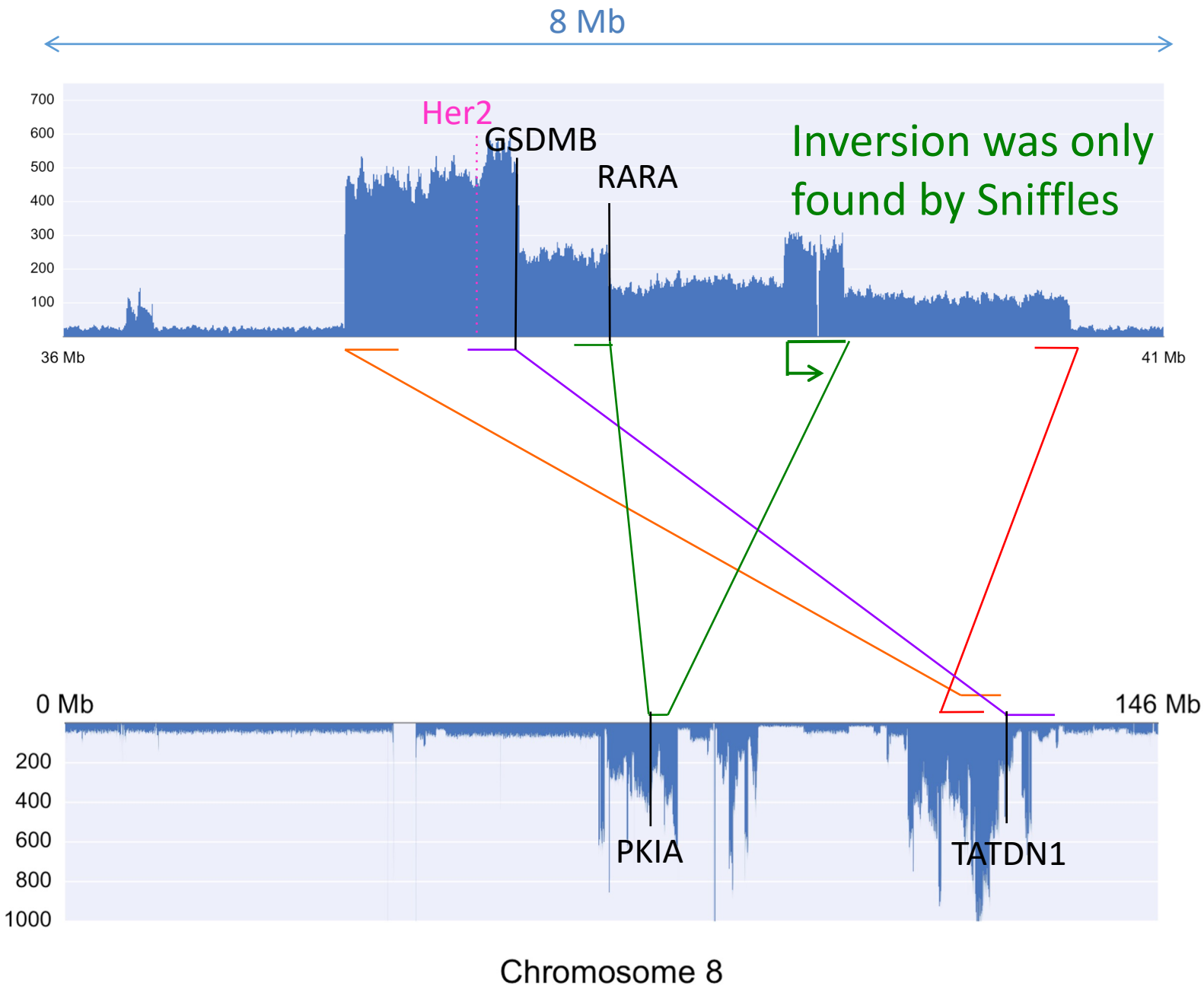
SKBR-3 using Pacbio

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

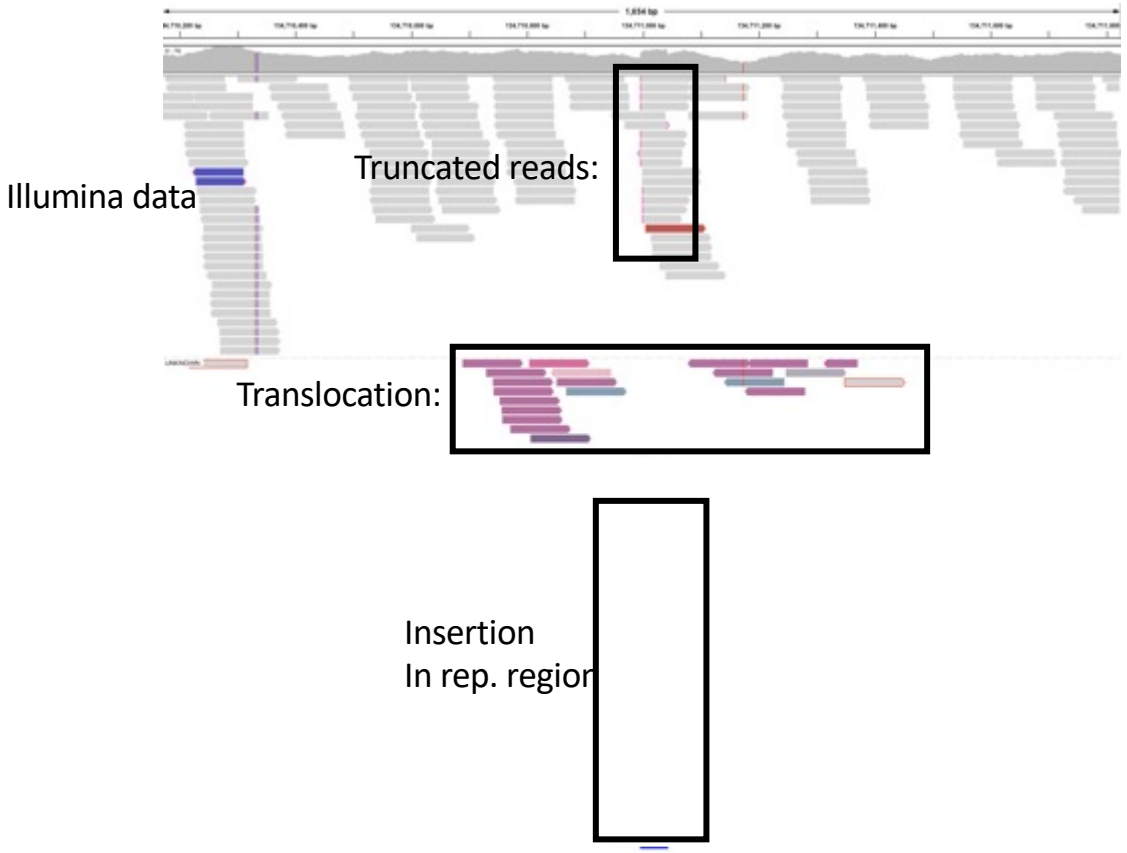
- Healthy female
- Gold standard in genomics
- Sequenced with many technologies independently:
  - Illumina, PacBio, Oxford Nanopore

3.2 NA12878: Deletion calling

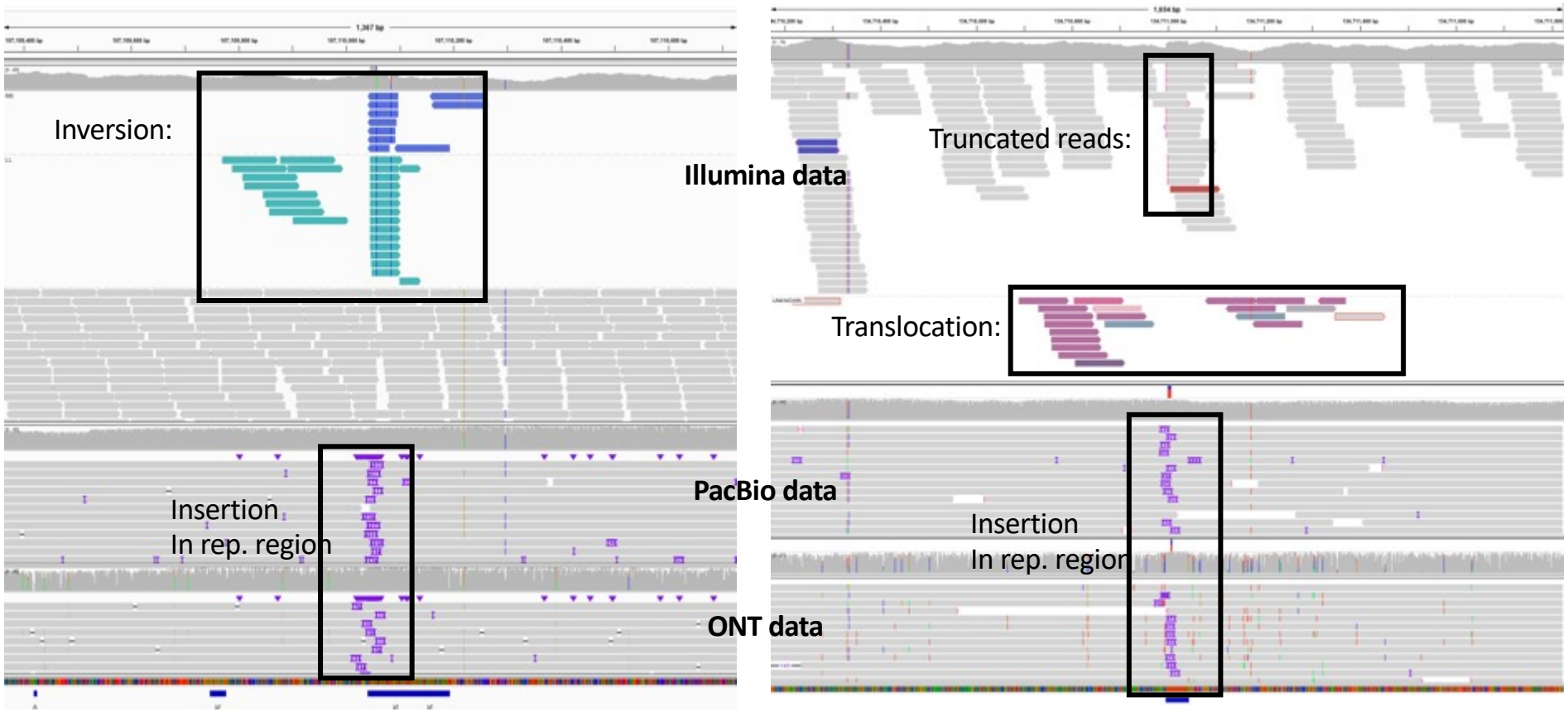
Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	**119**
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	**43**
Oxford Nanopore @Baylor	34x	4,982	12,596	7,102	169	113	5,166	**46**
Illumina	50x	2 x 101	7,275	3,744	731	553	0	**2,247**

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**

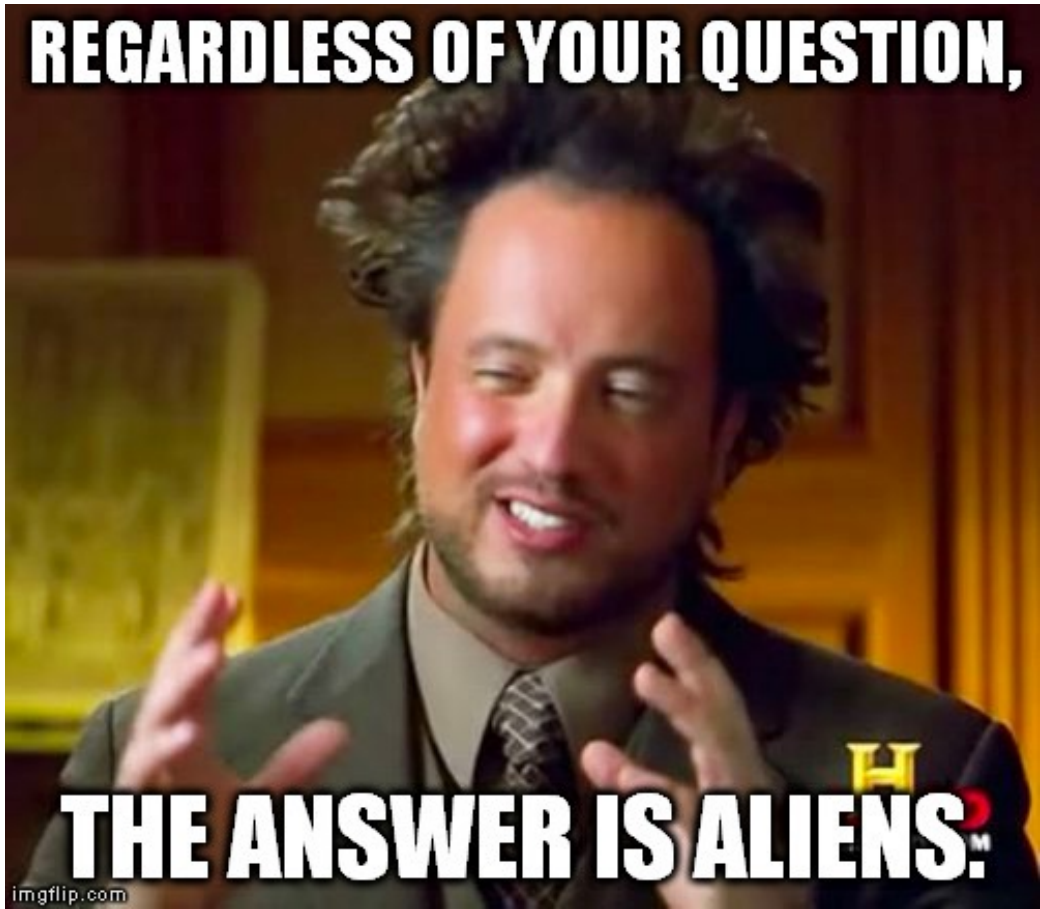


NA12878: check 2,247 TRA



Question: 3

What are the problems of long reads?



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

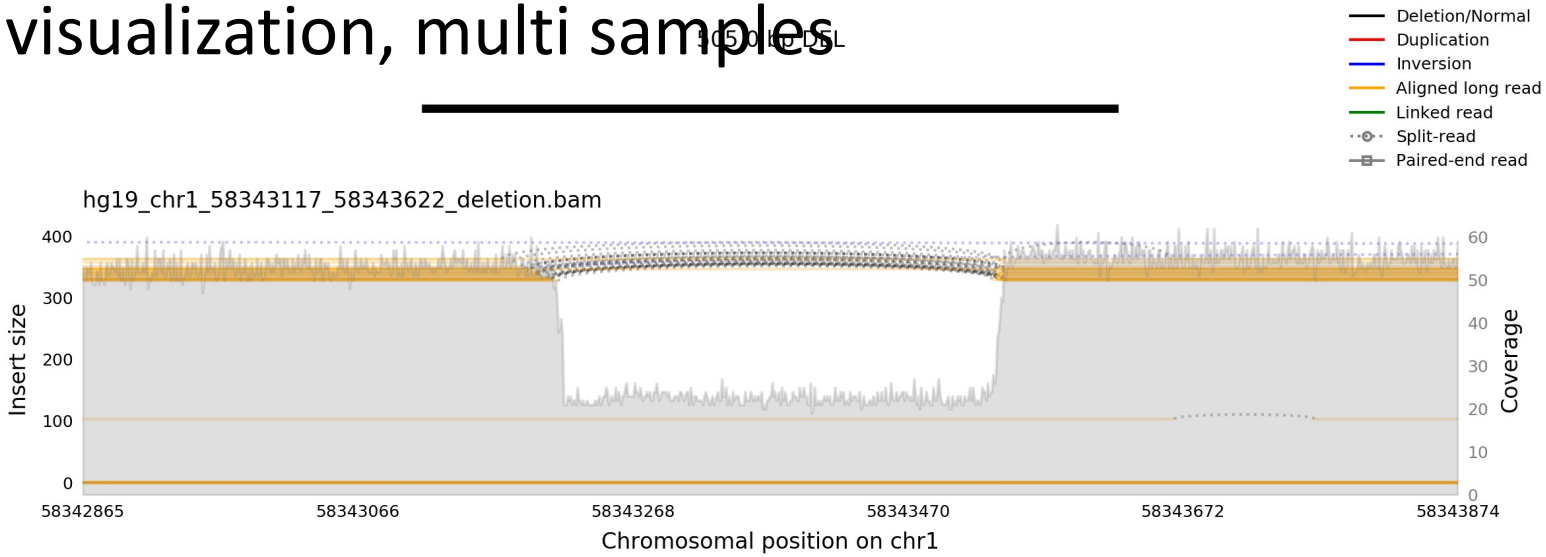
Filter

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 ~2010.
- 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 ->