Detecting variants in DNA sequencing data

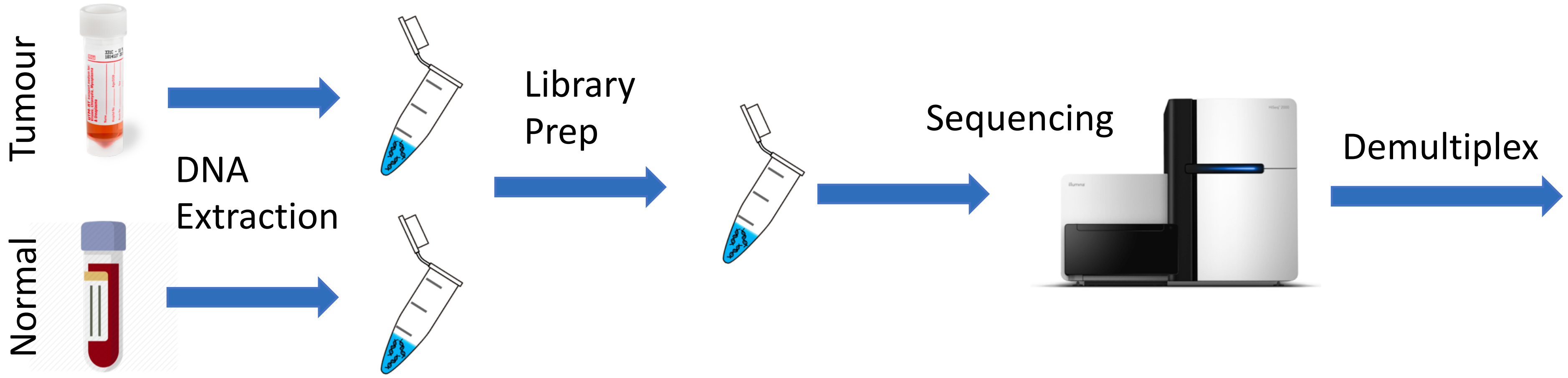
MBP Bioinformatics Tech Talk

Jeff Bruce, PhD

jbruce@uhnresearch.ca

UHN Princess Margaret Cancer Centre

Brief Overview: Sequencing and Alignment



Tumour

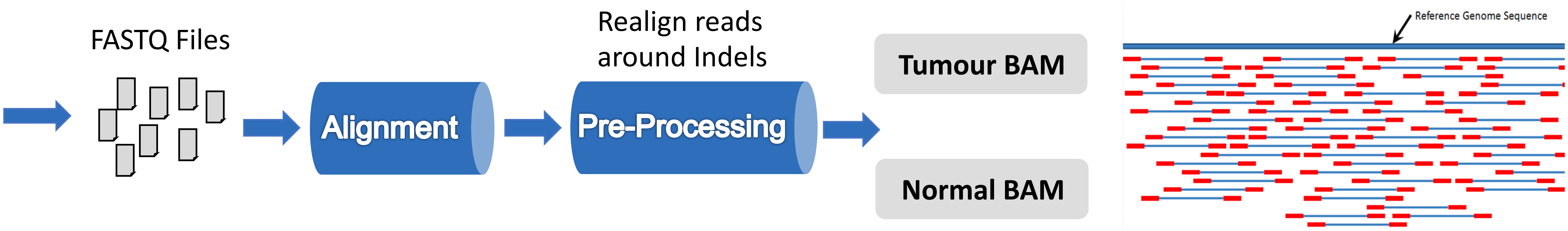
Normal

DNA Extraction

Library Prep

Sequencing

Demultiplex



FASTQ Files

Alignment

Realign reads around Indels

Pre-Processing

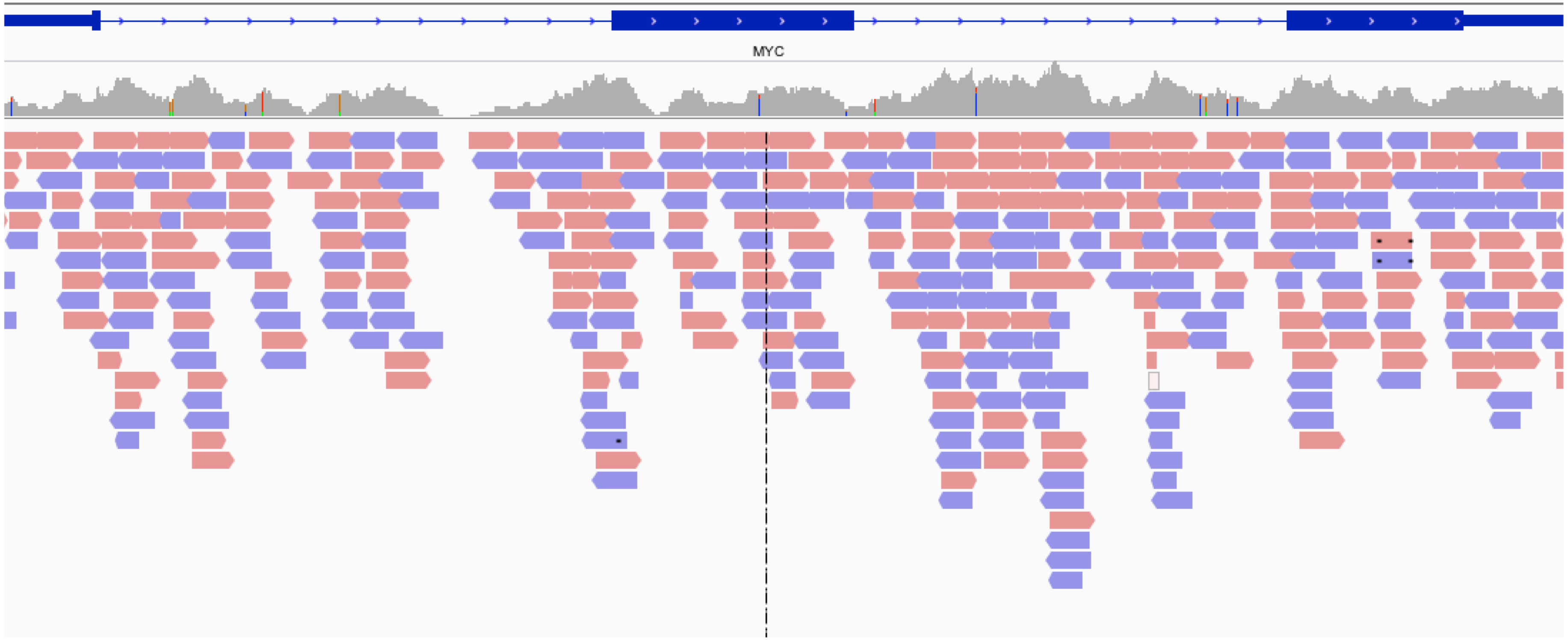
Tumour BAM

Normal BAM

Brief Overview: Sequencing and Alignment

Sequence Alignment/Map (SAM)

Binary representation of SAM = BAM



Gene annotation

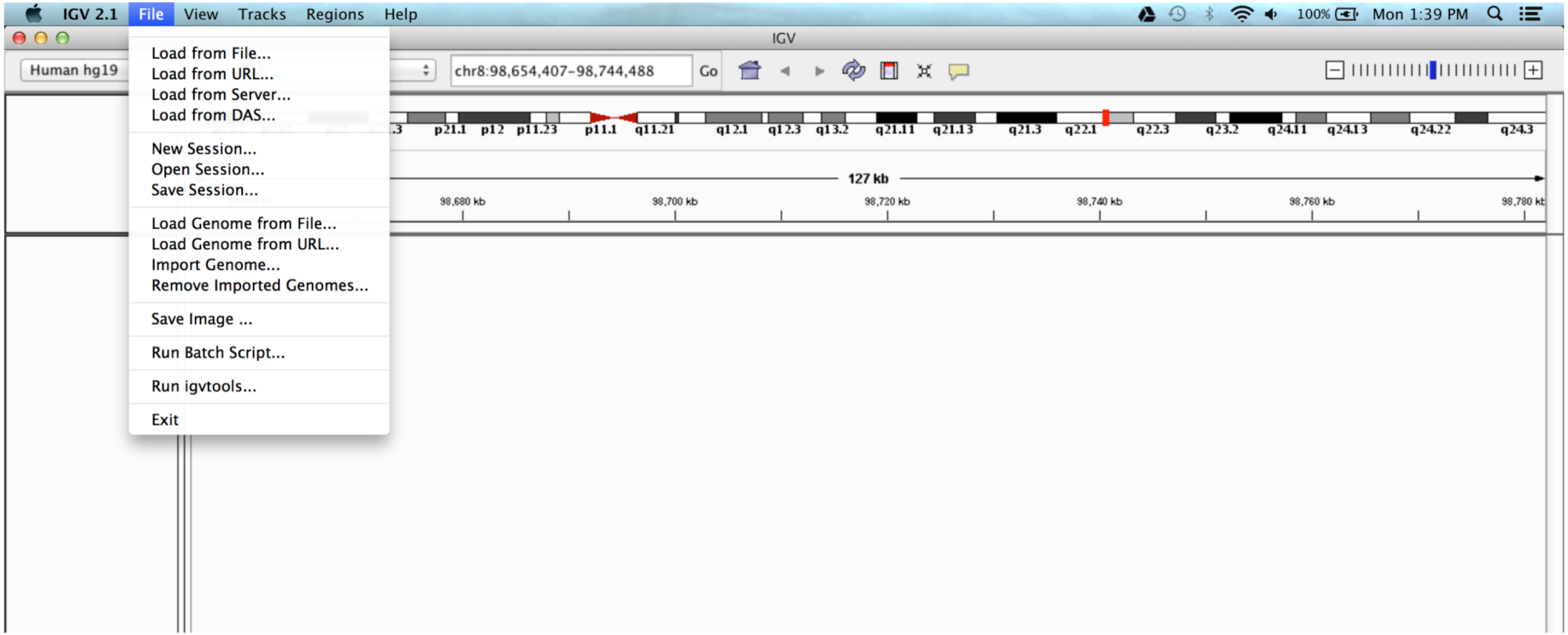
Coverage

Read-Pairs

Pink: left read

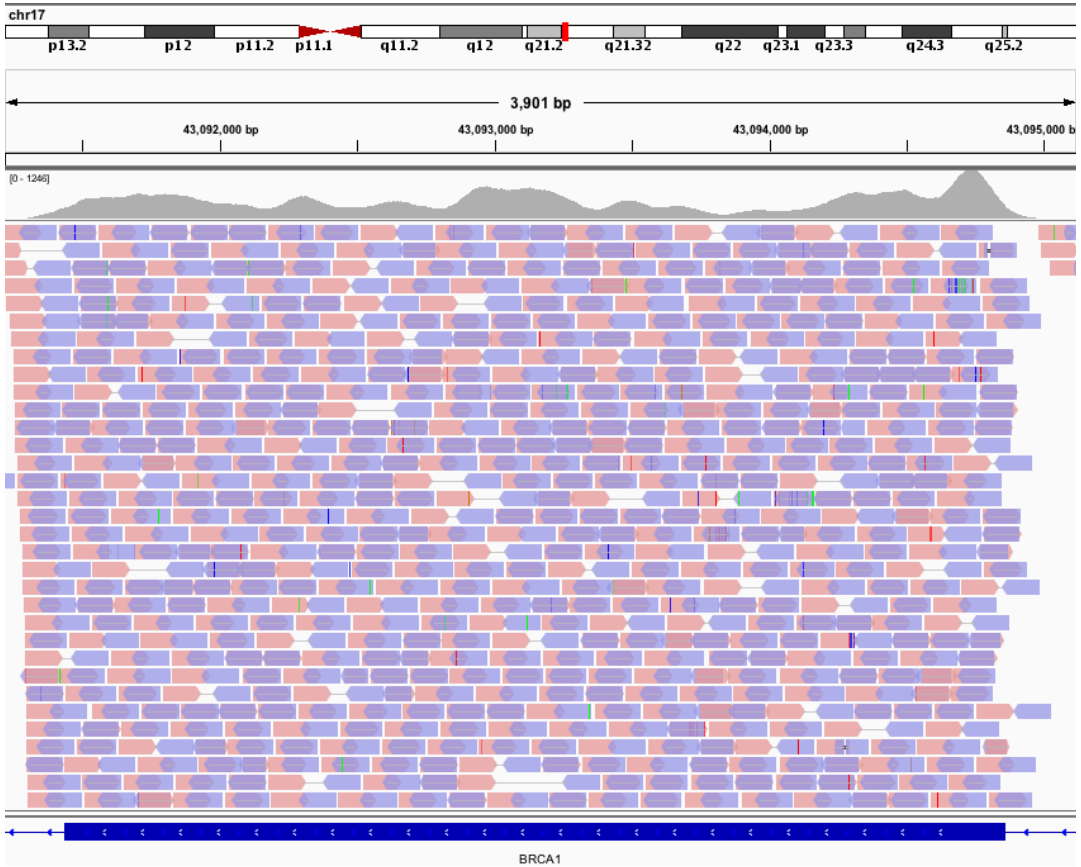
Blue: right read

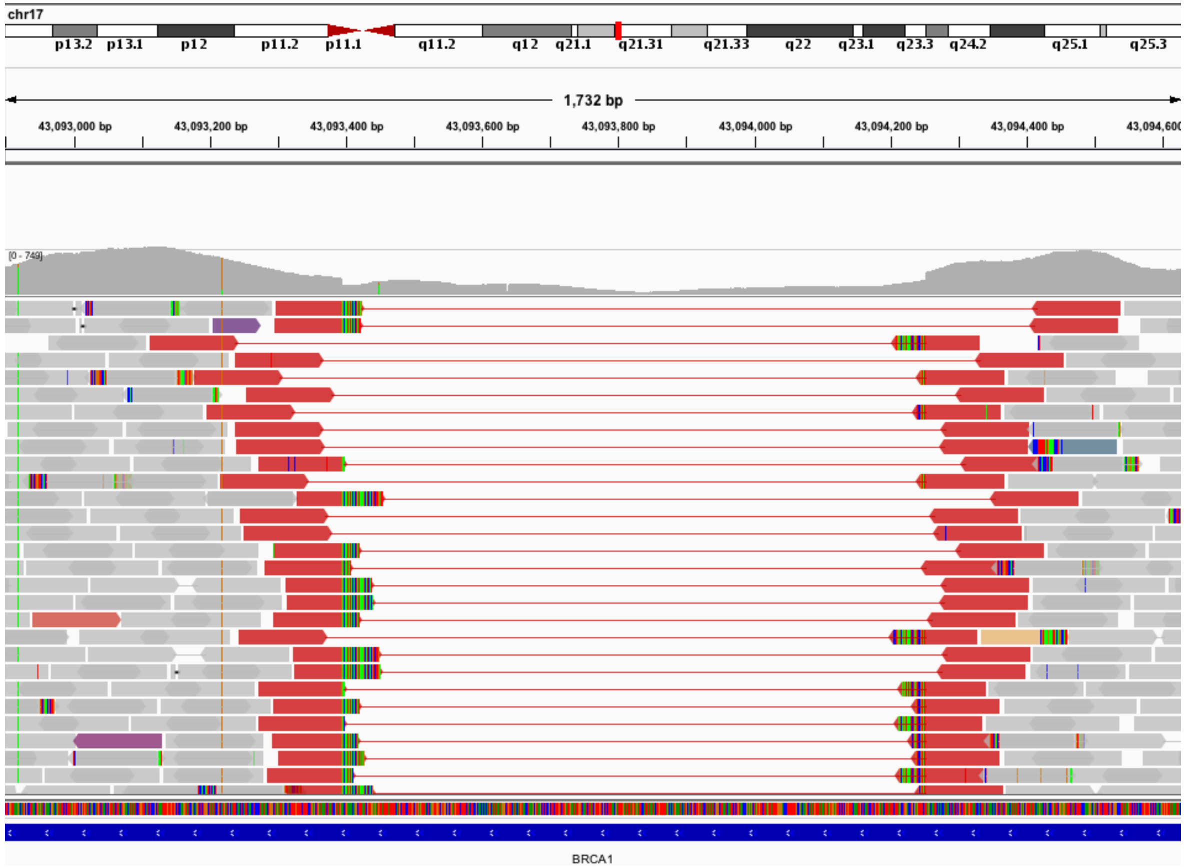
Brief Overview: Integrative Genomics Viewer



http://www.broadinstitute.org/igv/

Brief Overview: Integrative Genomics Viewer (IGV)





Brief Overview: HPC/Linux/command line

- **Some HPC options in Toronto/Ontario**
  - Mordor, HPC4Health, SciNet, SickKids HPF, SharkNet
- **Cloud HPC options**
  - AWS (amazon)/Google cloud/Microsoft Azure
  - Cancer Genomics Cloud
  - FireCloud (broad institute)
- **Why use command-line tools for computational biology?**

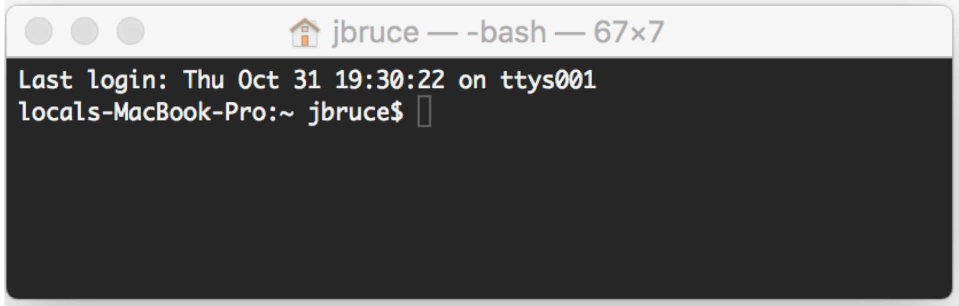
The field is always changing so we need tools that are:

- *Relatively* simple/quick/inexpensive to create
- Open source/easily modifiable
- Distributable across HPC cluster/cloud nodes
- Lightweight and portable

Brief Overview: HPC/Linux/command line

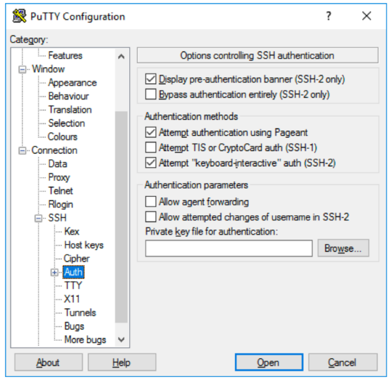
Tools to access remote HPC clusters

- **MacOS and Linux:** Terminal



```
Last login: Thu Oct 31 19:30:22 on ttys001
locals-MacBook-Pro:~ jbruce$ 
```

- **Windows:** PuTTY



Options controlling SSH authentication

☑ Display pre-authentication banner (SSH-2 only)
☐ Bypass authentication entirely (SSH-2 only)

Authentication methods

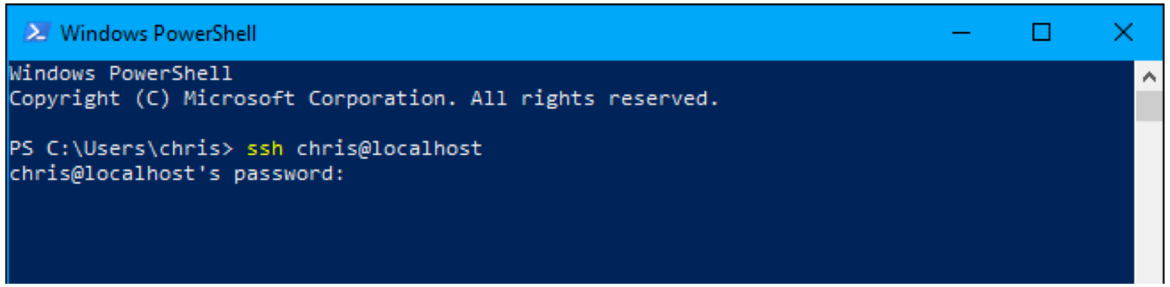
☑ Attempt authentication using Pageant
☐ Attempt TIS or CryptoCard auth (SSH-1)
☑ Attempt "keyboard-interactive" auth (SSH-2)

Authentication parameters

☐ Allow agent forwarding
☐ Allow attempted changes of username in SSH-2

Private key file for authentication:

Windows PowerShell



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

PS C:\Users\chris> ssh chris@localhost
chris@localhost's password:
```

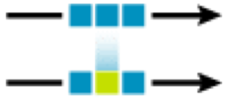
Terminology:

- **Variant:** Any genomic sequence that differs from a given reference
- **Germline Variant:** Inherited; present in all* of the cells of the individual
- **Somatic Variant:** Variants that are not inherited or passed on to offspring through the germline. In cancer, these are tumour specific
- **Mutation:**
  - Germline - Based on population frequency (<1% of a given population)
  - **The physical event resulting in a change to the genome**
  - Sometimes used Interchangeably with “variant”
- **Polymorphism:**
  - Germline variants present in >1% of the population

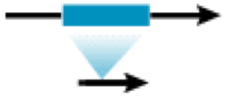
* Or a large proportion in the case of mosaicism

Types of Variants

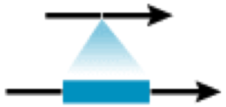
Single Nucleotide Variant



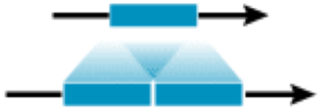
Deletion



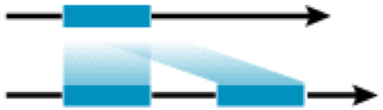
Insertion



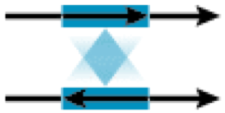
Tandem Duplication



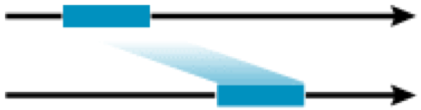
Interspersed Duplication



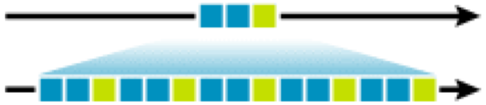
Inversion



Translocation

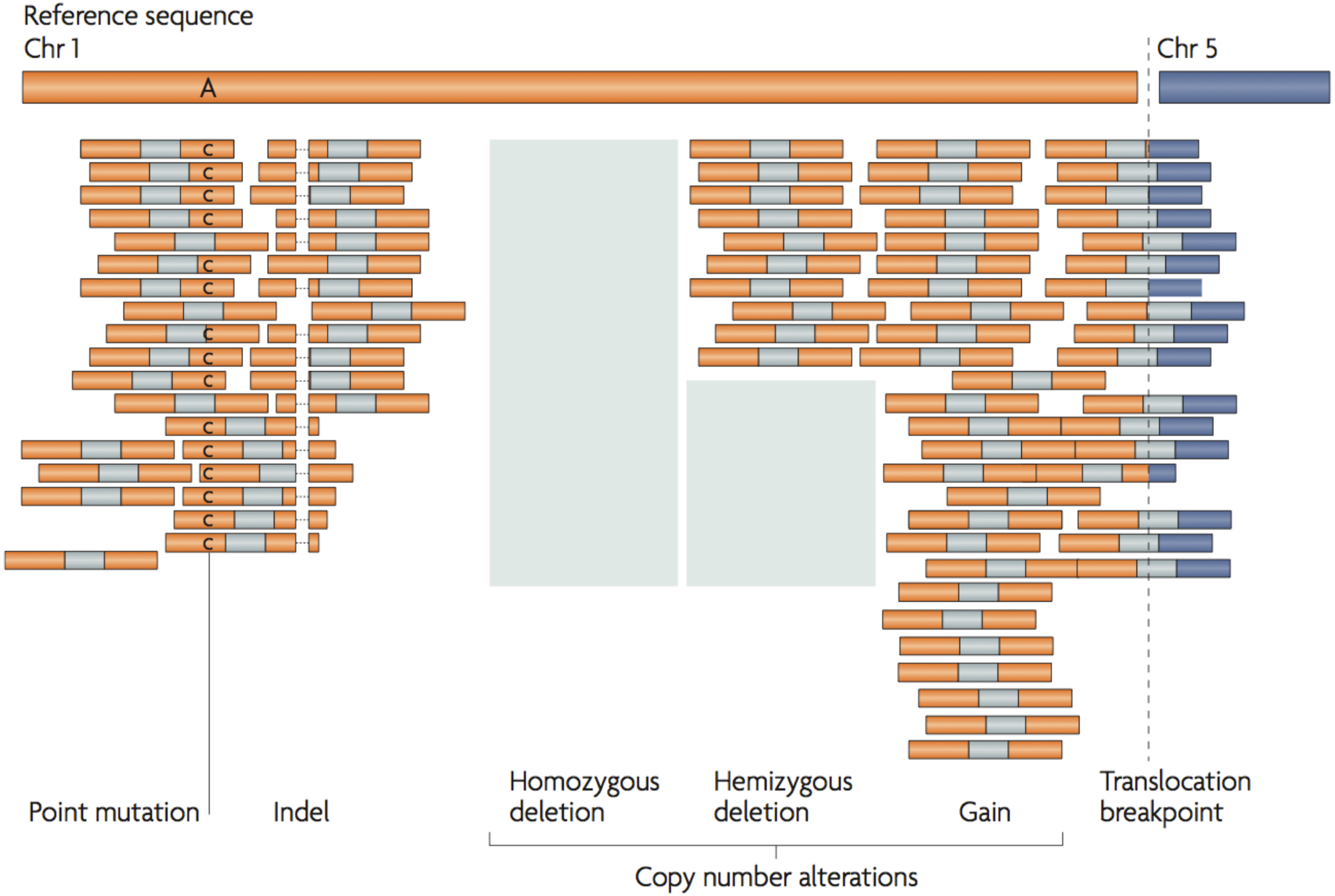


Copy Number Variant



https://www.pacb.com/

Types of Variants



Meyerson *et al. Nat Rev Genet.* 2010 Oct;11(10):685-96.

Sequencing Techniques

Genomic footprint
< 1kb
3Gb +

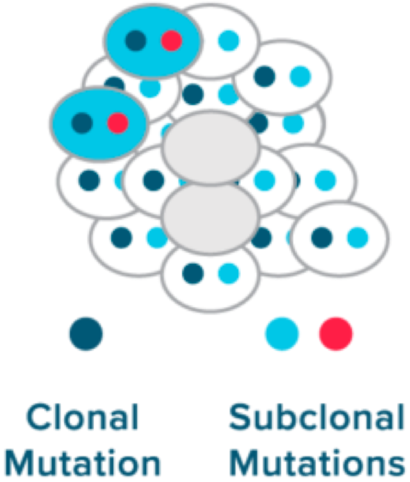
Sequencing Methodology	Detectable Variant Types				
	Coding SNVs/Indels	Non-Coding SNVs/Indels	Exon/Gene CNVs	Mid-Large/ Complex SVs	Large CNVs
Sanger					
Targeted Panel		*	*	*	*
Whole Exome		*			
Whole Genome					

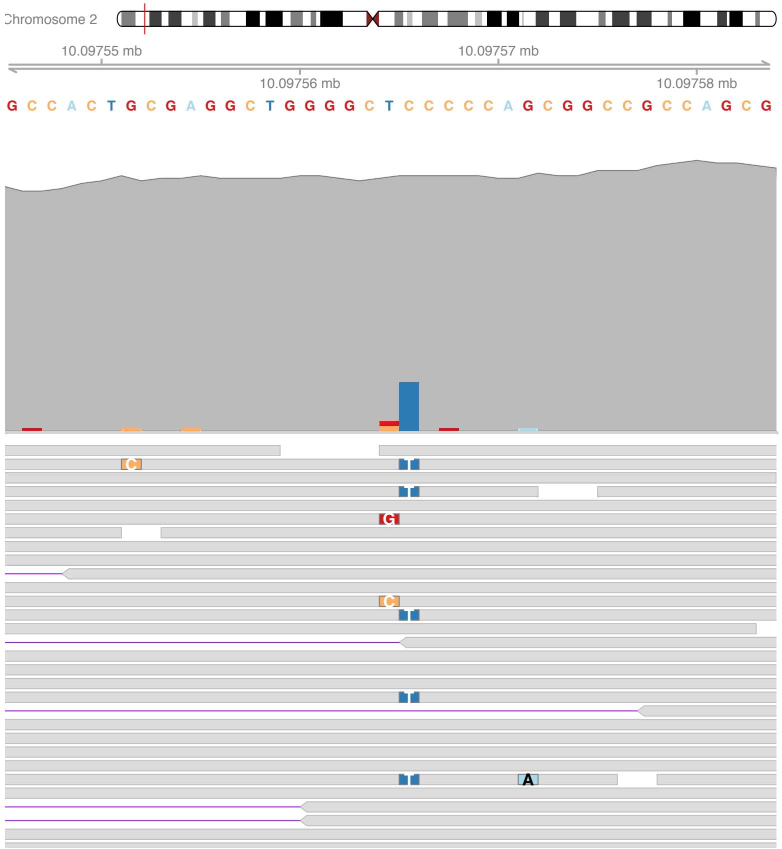
Yes

* With Caveats

Challenges for variant detection:

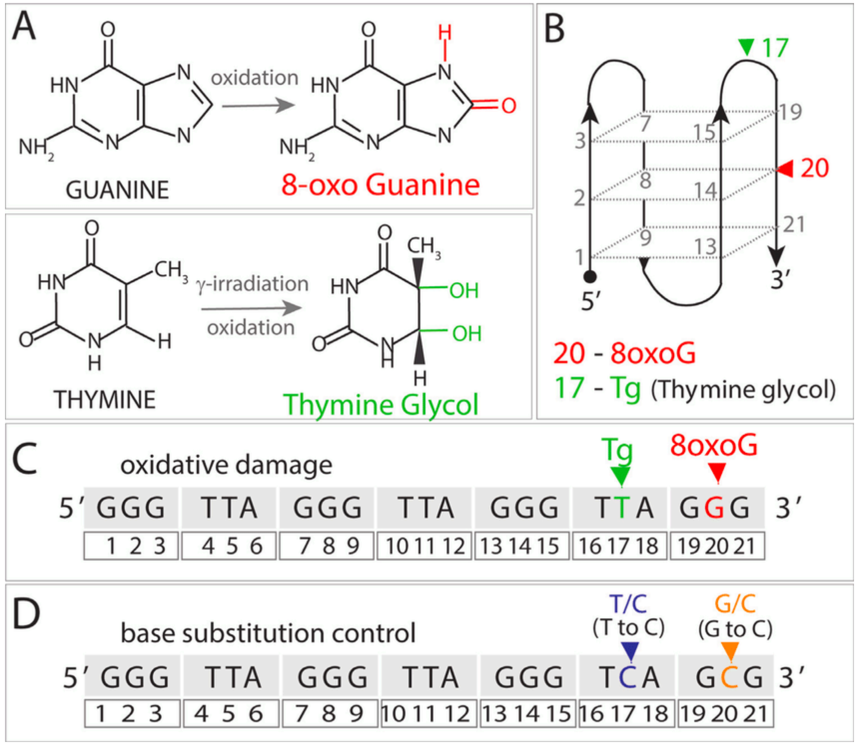
- **Low variant read proportion:**
  - Depth of sequencing
  - Normal contamination
  - Subclonality
  - Overlapping copy-number variants
  - Difficult read mapping
- **Sources of false positives:**
  - DNA damage due to processing
  - PCR errors
  - Sequencing errors
  - Mapping artifacts
- **Sources of false negatives:**
  - Poor mapping regions
  - Depth of sequencing
  - Local error rate
  - Variant complexity and size




Chromosome 2
10.09755 mb
10.09757 mb
10.09756 mb
10.09758 mb
G C C A C T G C G A G G C T G G G G C T C C C C C A G C G G C C G C C A G C G
C
G
C
A

Challenges for variant detection:

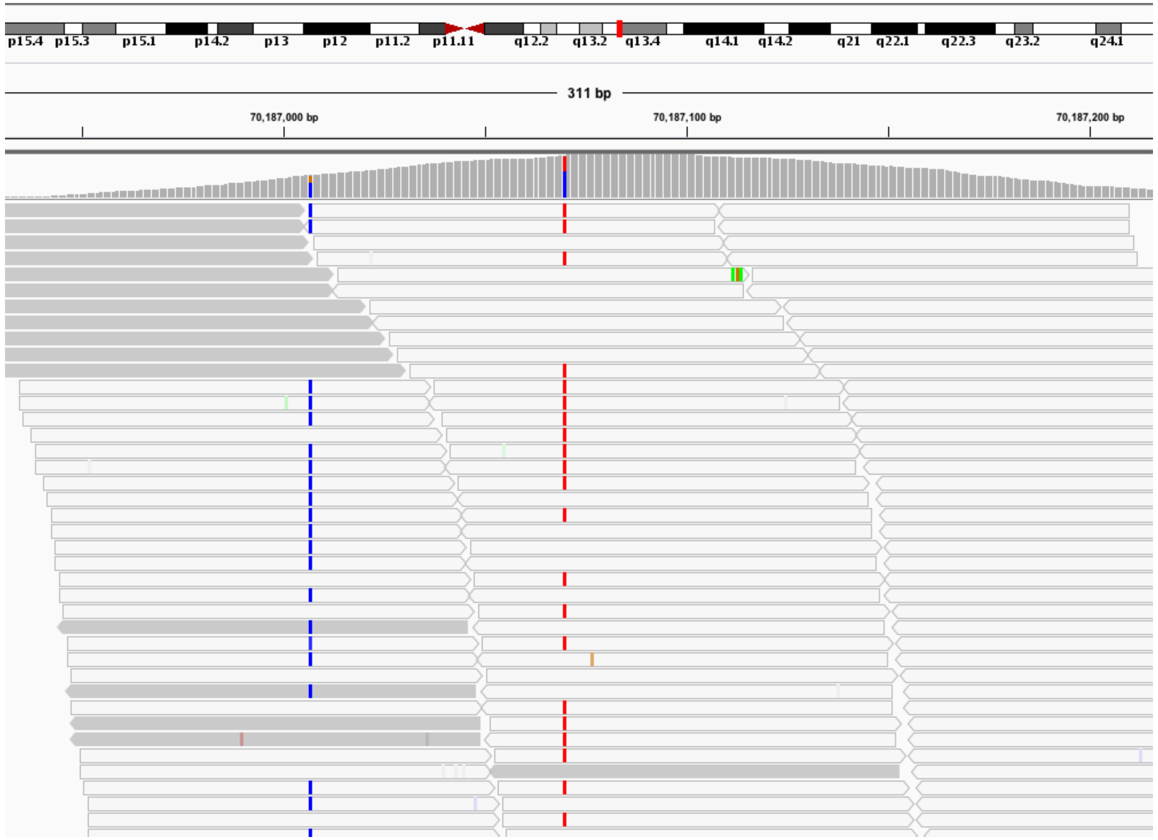
- **Low variant read proportion:**
  - Depth of sequencing
  - Normal contamination
  - Subclonality
  - Overlapping copy-number variants
  - Difficult read mapping
- **Sources of false positives:**
  - DNA damage due to storage and processing
  - PCR errors
  - Sequencer errors
  - Mapping artifacts
- **Sources of false negatives:**
  - Poor mapping regions
  - Depth of sequencing
  - Local error rate
  - Variant complexity and size



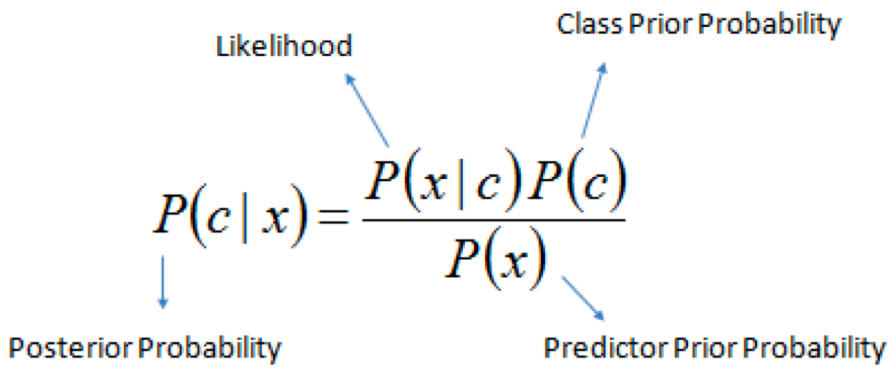
Nucleic Acids Research 45(20) · September 2017

Challenges for variant detection:

- **Low variant read proportion:**
  - Depth of sequencing
  - Normal contamination
  - Subclonality
  - Overlapping copy-number variants
  - Difficult read mapping
- **Sources of false positives:**
  - DNA damage due to processing
  - PCR errors
  - Sequencing errors
  - Mapping artifacts
- **Sources of false negatives:**
  - Poor mapping regions
  - Depth of sequencing
  - Local error rate
  - Variant complexity and size



These challenges, or “priors” lend themselves to the application of Bayesian statistics



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Likelihood

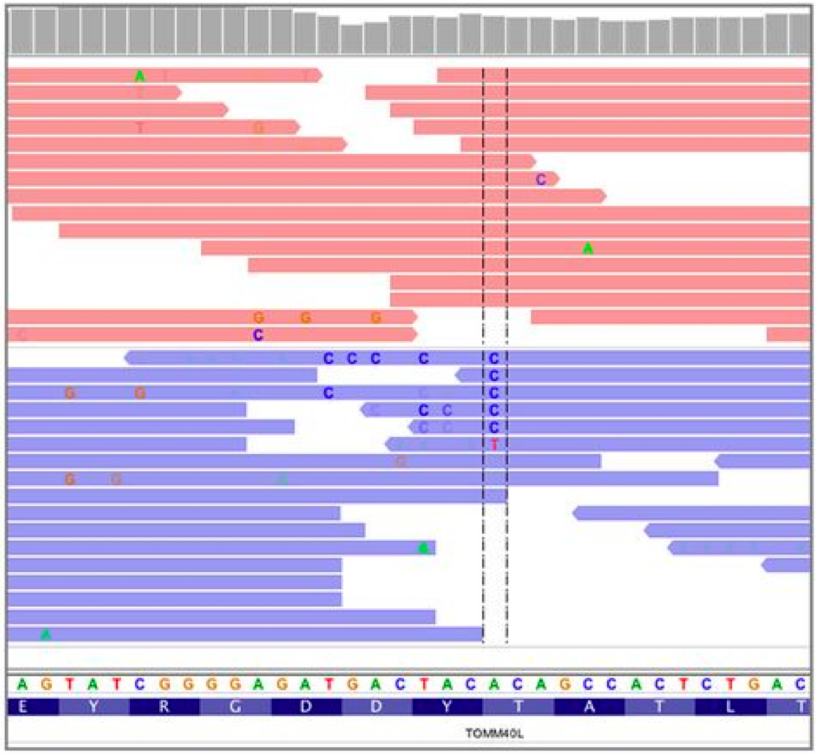
Class Prior Probability

Posterior Probability

Predictor Prior Probability

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

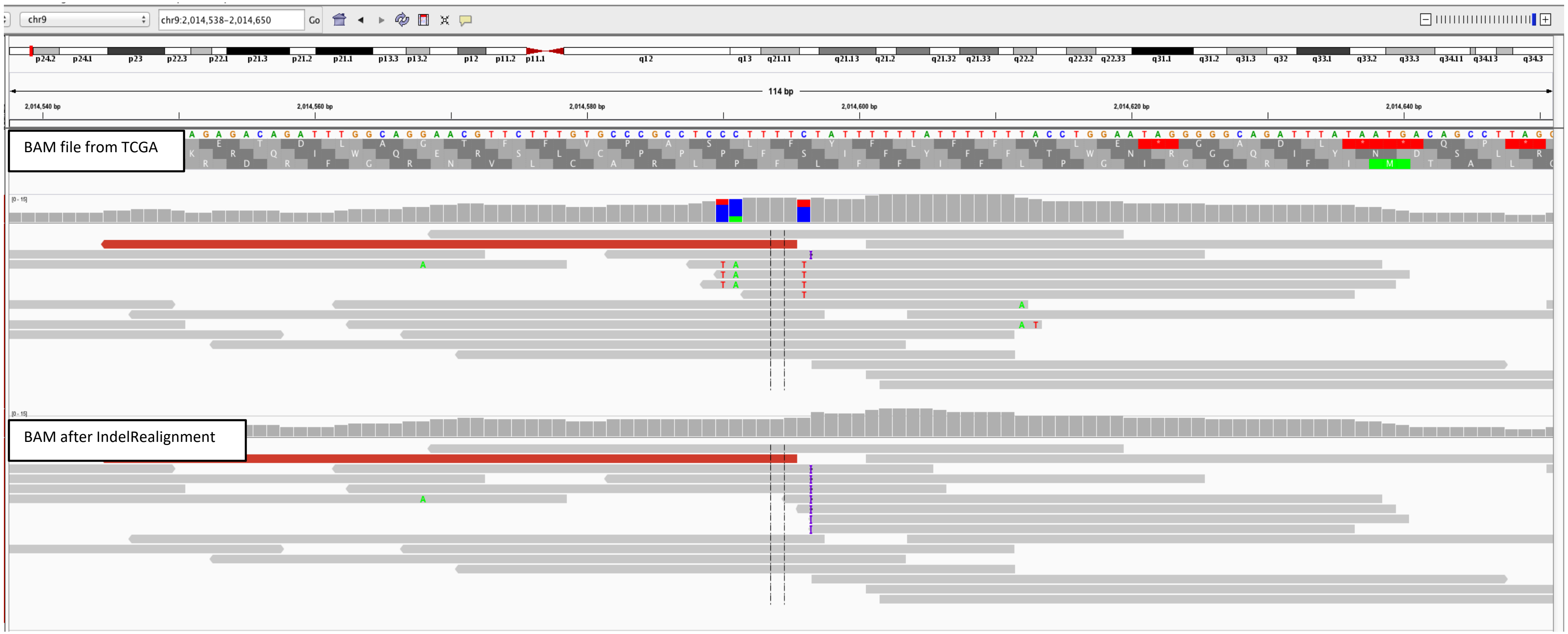
wikipedia.org


A G T A T C G G G G A G A T G A C T A C A C A G C C A C T C T G A C
E Y R G D D Y T A T L T
TOMM40L

Robinson, *et al.* Cancer Research 2017

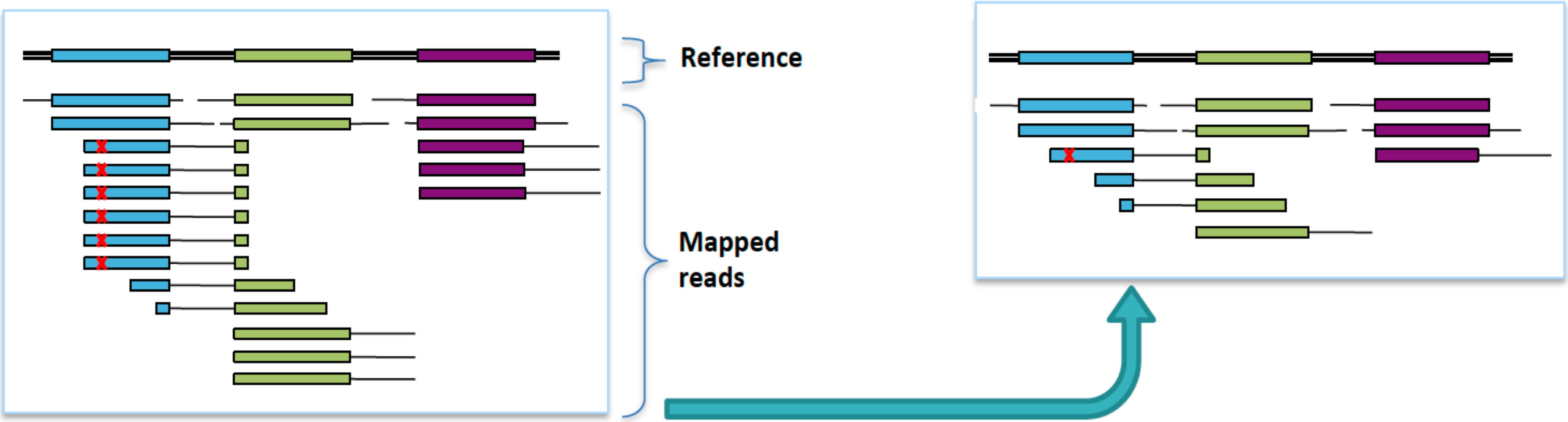
Some strategies to remove/limit errors

- Indel realignment (Local/genome-wide)



Some strategies to remove/limit errors

- Indel realignment (Local/genome-wide)
- Duplicate-read marking/filtering

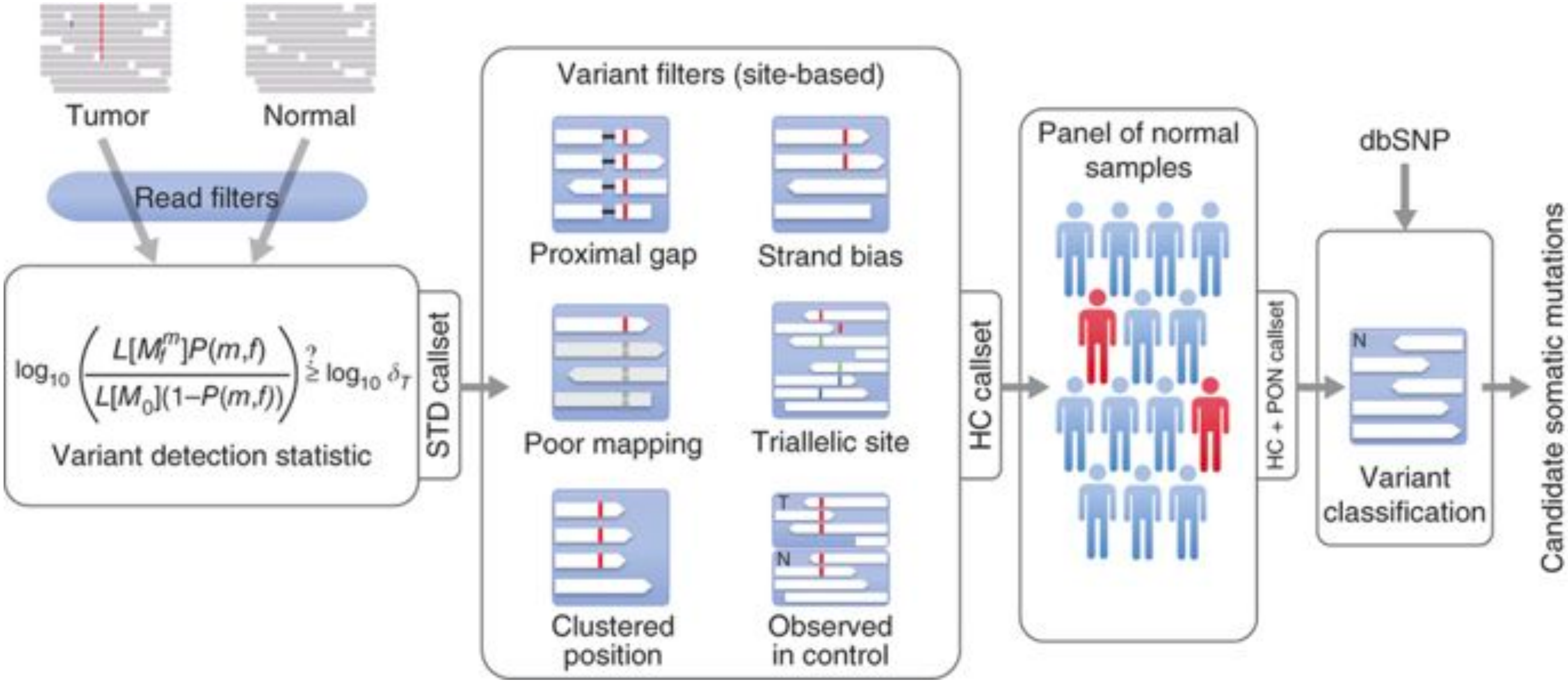


✖ = sequencing error propagated in duplicates

Some strategies to remove/limit errors

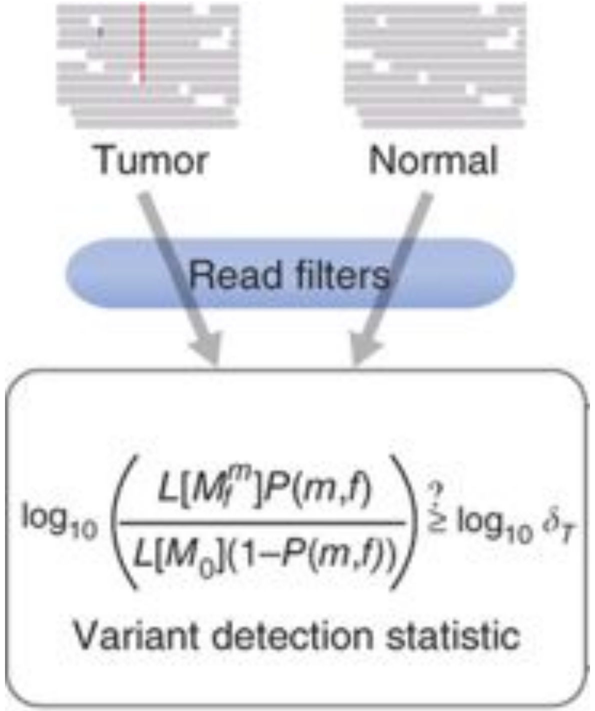
- Indel realignment (Local/genome-wide)
- Duplicate-read marking/filtering
- Base quality recalibration
- Joint calling (germline)
- Panel of Normals (somatic)

SNV detection algorithm example: MuTect



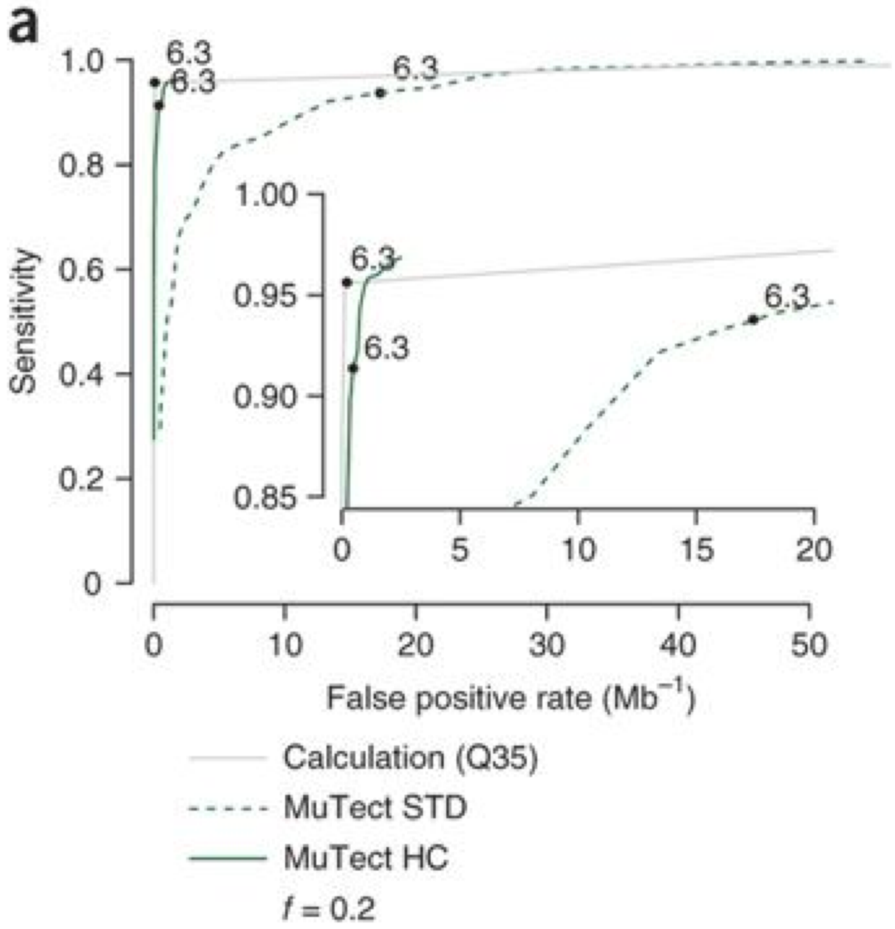
Cibulskis *et al.* Nature Biotechnology 2013

SNV detection algorithm example: MuTect



$$\log_{10}\left(\frac{L[M_f^m]P(m,f)}{L[M_0](1-P(m,f))}\right) \overset{?}{\geq} \log_{10}\delta_T$$

Variant detection statistic



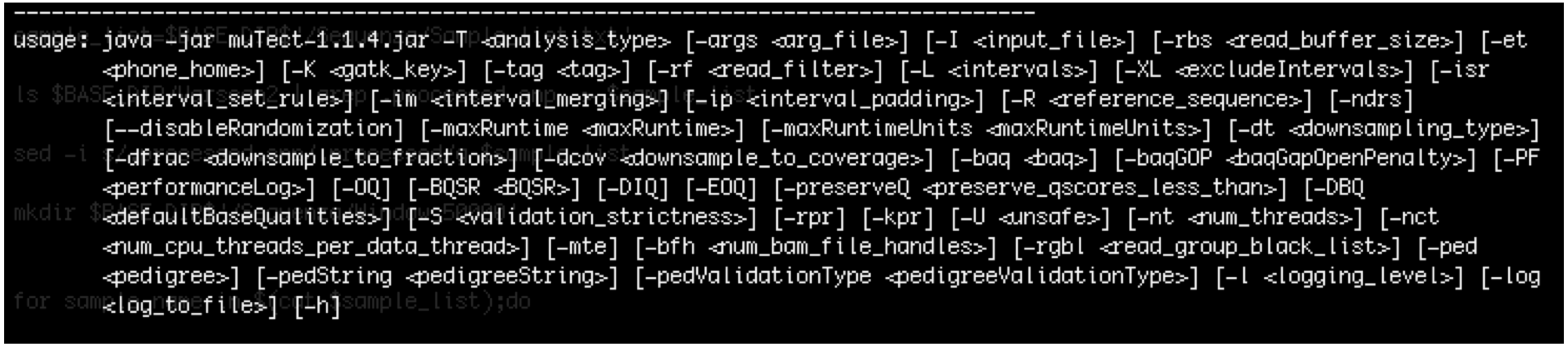
M_0 : reference model, alt allele is due to error

M_1 : variant model, alt allele is a true variant

Cibulskis *et al.* Nature Biotechnology 2013

Practical application example: MuTect v.1.1.4

Available Parameters:



```
--------------------------------------------------------------------------------
usage: java -jar muTect-1.1.4.jar -T <analysis_type> [-args <arg_file>] [-I <input_file>] [-rbs <read_buffer_size>] [-et
       <phone_home>] [-K <gatk_key>] [-tag <tag>] [-rf <read_filter>] [-L <intervals>] [-XL <excludeIntervals>] [-isr
       <interval_set_rule>] [-im <interval_merging>] [-ip <interval_padding>] [-R <reference_sequence>] [-ndrs]
       [--disableRandomization] [-maxRuntime <maxRuntime>] [-maxRuntimeUnits <maxRuntimeUnits>] [-dt <downsampling_type>]
       [-dfrac <downsample_to_fraction>] [-dcov <downsample_to_coverage>] [-baq <baq>] [-baqGOP <baqGapOpenPenalty>] [-PF
       <performanceLog>] [-OQ] [-BQSR <BQSR>] [-DIQ] [-EOQ] [-preserveQ <preserve_qscores_less_than>] [-DBQ
       <defaultBaseQualities>] [-S <validation_strictness>] [-rpr] [-kpr] [-U <unsafe>] [-nt <num_threads>] [-nct
       <num_cpu_threads_per_data_thread>] [-mte] [-bfh <num_bam_file_handles>] [-rgbl <read_group_black_list>] [-ped
       <pedigree>] [-pedString <pedigreeString>] [-pedValidationType <pedigreeValidationType>] [-l <logging_level>] [-log
       <log_to_file>] [-h]
```

Example Command:

```
module load mutect/1.1.4
module load igenome-human/hg19

java -Djava.io.tmpdir=./tmp/ -Xmx8g -jar $mutect_dir/muTect-1.1.4.jar --analysis_type MuTect \
--enable_extended_output --fraction_contamination 0.02 -dt NONE -L Interval.bed --reference_sequence $REF \
--input_file:normal /path/to/normal/file.bam --input_file:tumor /path/to/tumour/file.bam \
--out /path/to/output.call_stats --vcf /path/to/output.vcf --coverage_file /path/to/output_coverage.wig.txt
```

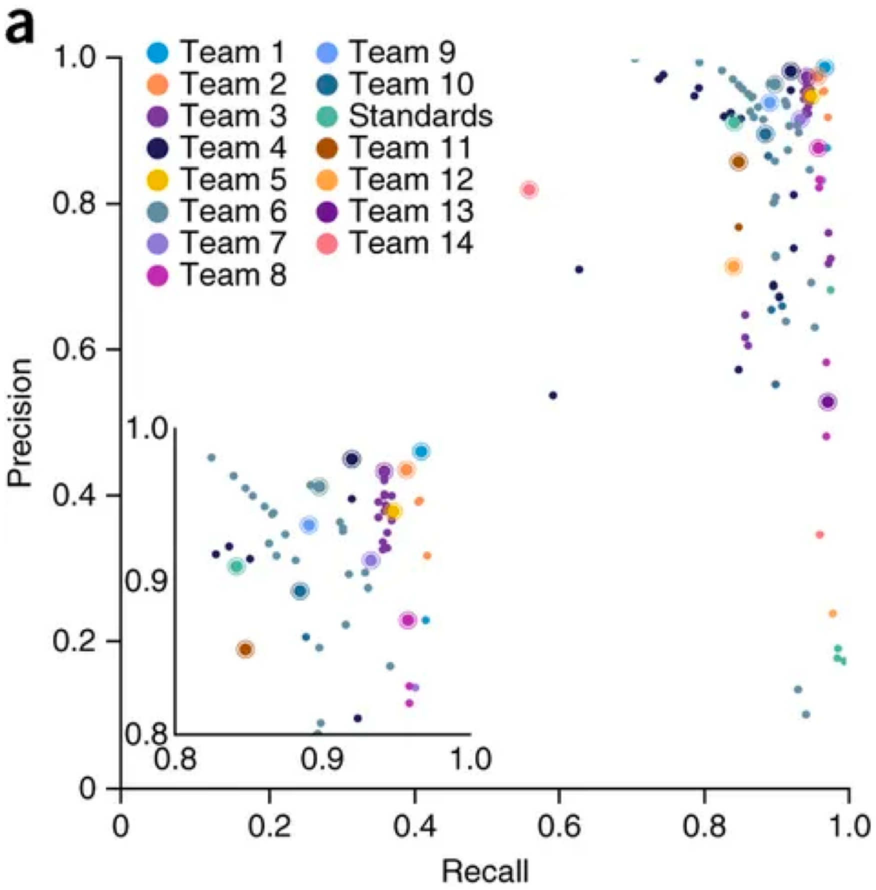
UHN Princess Margaret Cancer Centre

DREAM CHALLENGES

powered by Sage Bionetworks

Combining tumor genome simulation with crowdsourcing to benchmark somatic singlenucleotide-variant detection

Ewing ... Boutros *et al.,* ICGC-TCGA Network Nature Methods volume 12, pages 623–630 (2015)



Winner: MuTect – Broad Institute

DREAM CHALLENGES

powered by Sage Bionetworks

Combining tumor genome simulation with crowdsourcing to benchmark somatic singlenucleotide-variant detection

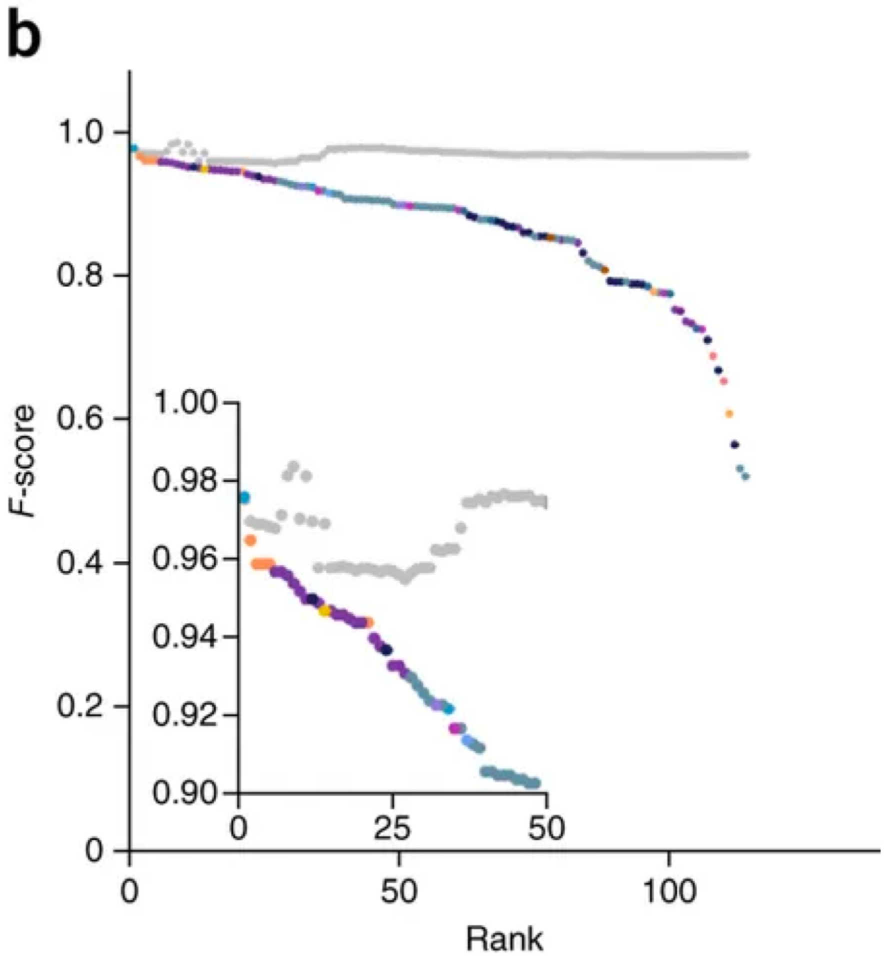
Ewing ... Boutros *et al.,* ICGC-TCGA Network Nature Methods volume 12, pages 623–630 (2015)

Name	Entity ID	Team	# True Positives	# False Positives	Recall	Precision	F-score
MuTect - L10	syn2343084	Broad SMC	**3421**	**57**	0.967204	**0.983611**	0.975339
MuTect - Stock	syn2343082	Broad SMC	**3431**	**547**	0.970031	**0.862494**	0.913107

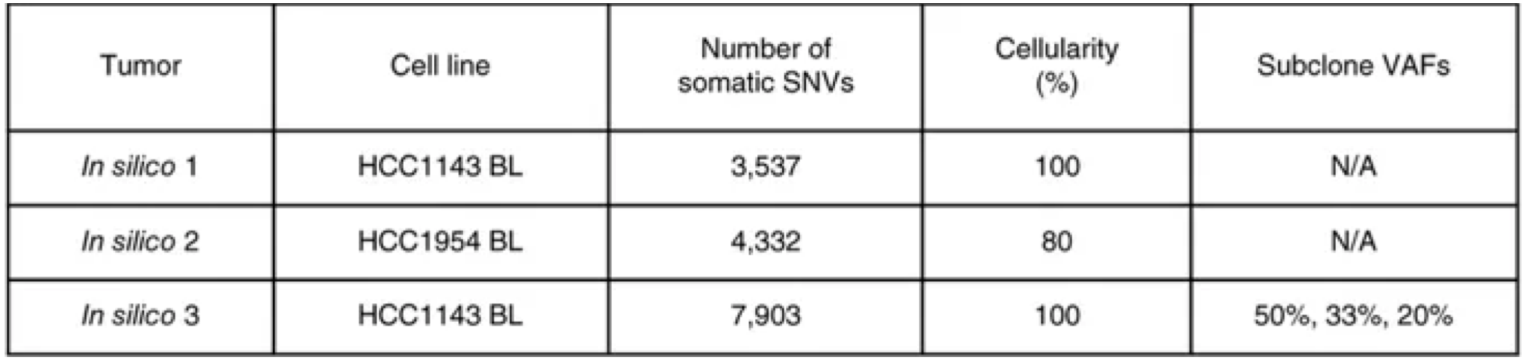
DREAM CHALLENGES
powered by Sage Bionetworks

Combining tumor genome simulation with crowdsourcing to benchmark somatic singlenucleotide-variant detection

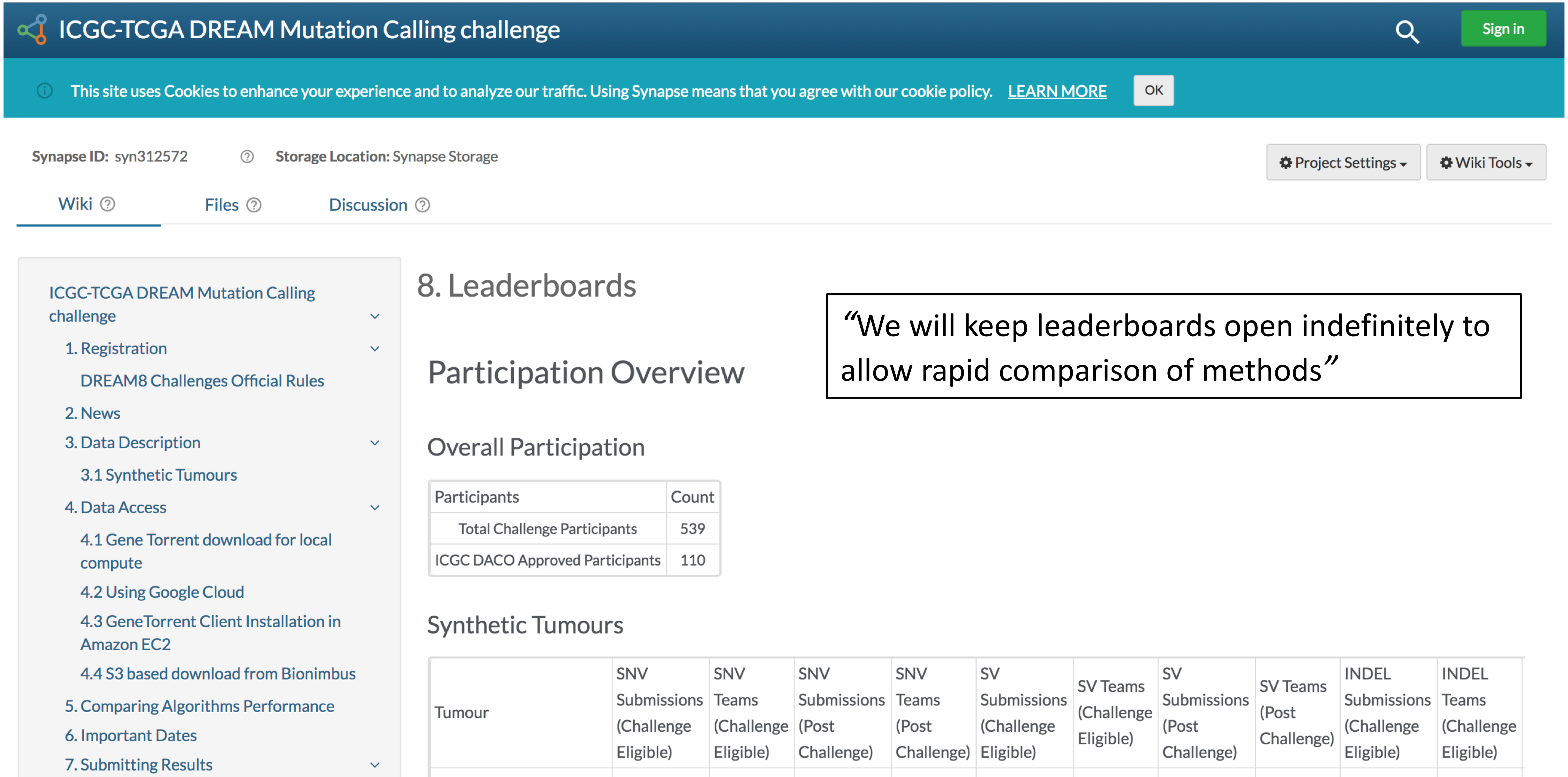
Ewing ... Boutros *et al.,* ICGC-TCGA Network Nature Methods volume 12, pages 623–630 (2015)



“…and in subclonality, an ensemble of pipelines outperforms the best individual pipeline in all cases”



Tumor	Cell line	Number of somatic SNVs	Cellularity (%)	Subclone VAFs
In silico 1	HCC1143 BL	3,537	100	N/A
In silico 2	HCC1954 BL	4,332	80	N/A
In silico 3	HCC1143 BL	7,903	100	50%, 33%, 20%



ICGC-TCGA DREAM Mutation Calling challenge
Sign in
This site uses Cookies to enhance your experience and to analyze our traffic. Using Synapse means that you agree with our cookie policy. LEARN MORE
OK
Project Settings
Wiki Tools
Wiki
Files
Discussion

Synapse ID: syn312572 **Storage Location:** Synapse Storage

- ICGC-TCGA DREAM Mutation Calling challenge
  - 1. Registration
    - DREAM8 Challenges Official Rules
  - 2. News
  - 3. Data Description
    - 3.1 Synthetic Tumours
  - 4. Data Access
    - 4.1 Gene Torrent download for local compute
    - 4.2 Using Google Cloud
    - 4.3 GeneTorrent Client Installation in Amazon EC2
    - 4.4 S3 based download from Bionimbus
  - 5. Comparing Algorithms Performance
  - 6. Important Dates
  - 7. Submitting Results

8. Leaderboards

Participation Overview

Overall Participation

Participants	Count
Total Challenge Participants	539
ICGC DACO Approved Participants	110

Synthetic Tumours

Tumour	SNV Submissions (Challenge Eligible)	SNV Teams (Challenge Eligible)	SNV Submissions (Post Challenge)	SNV Teams (Post Challenge)	SV Submissions (Challenge Eligible)	SV Teams (Challenge Eligible)	SV Submissions (Post Challenge)	SV Teams (Post Challenge)	INDEL Submissions (Challenge Eligible)	INDEL Teams (Challenge Eligible)

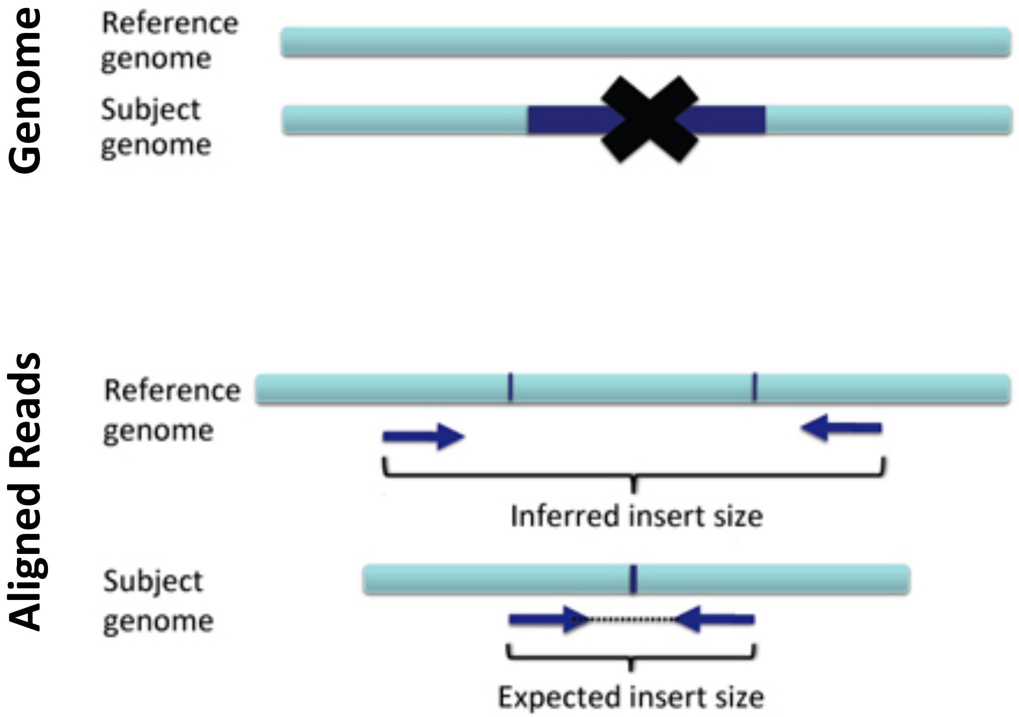
> “We will keep leaderboards open indefinitely to allow rapid comparison of methods”

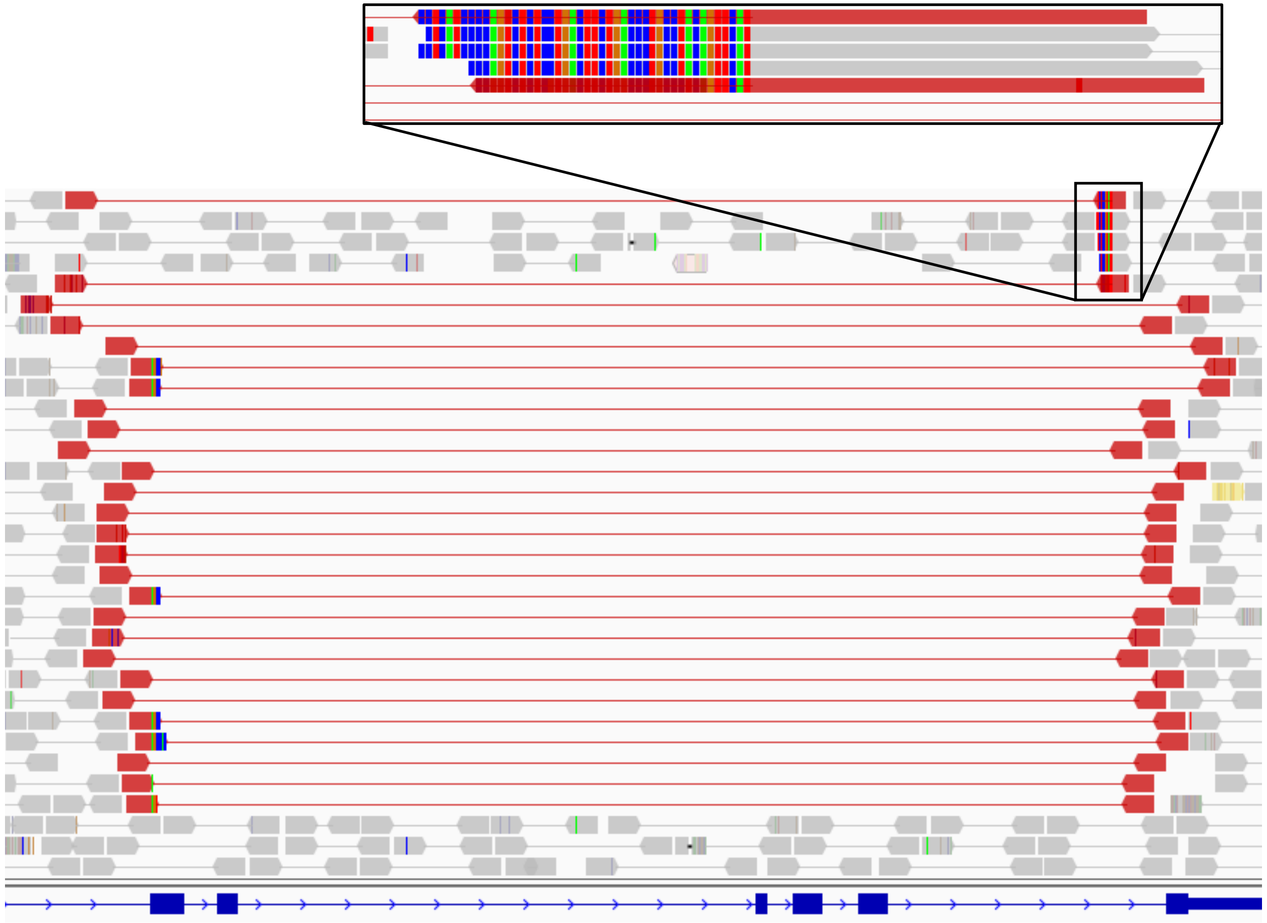
https://www.synapse.org/#!Synapse:syn312572/wiki/61509

Small (SNV/Indel) Variant Annotation

- **Purpose**
  - To aid in the interpretation of variants
- **Annotations**
  - Classification (Missense, frameshift etc.)
  - Predicted amino acid change
  - Predicted impact (ex. SIFT, Polyphen)
  - Occurrence in public databases (dbSNP, COSMIC, ExAC, Gnomad)
  - +++
- **Some available tools:**
  - Ensembl Variant Effect Predictor (VEP)
  - Annovar
  - Oncotator

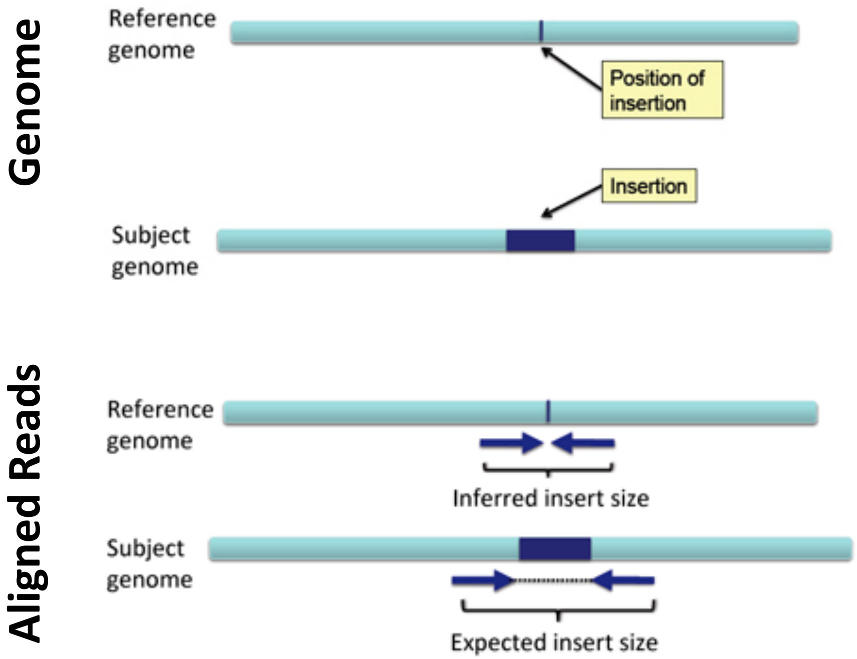
Structural variants affecting insert size: Deletions

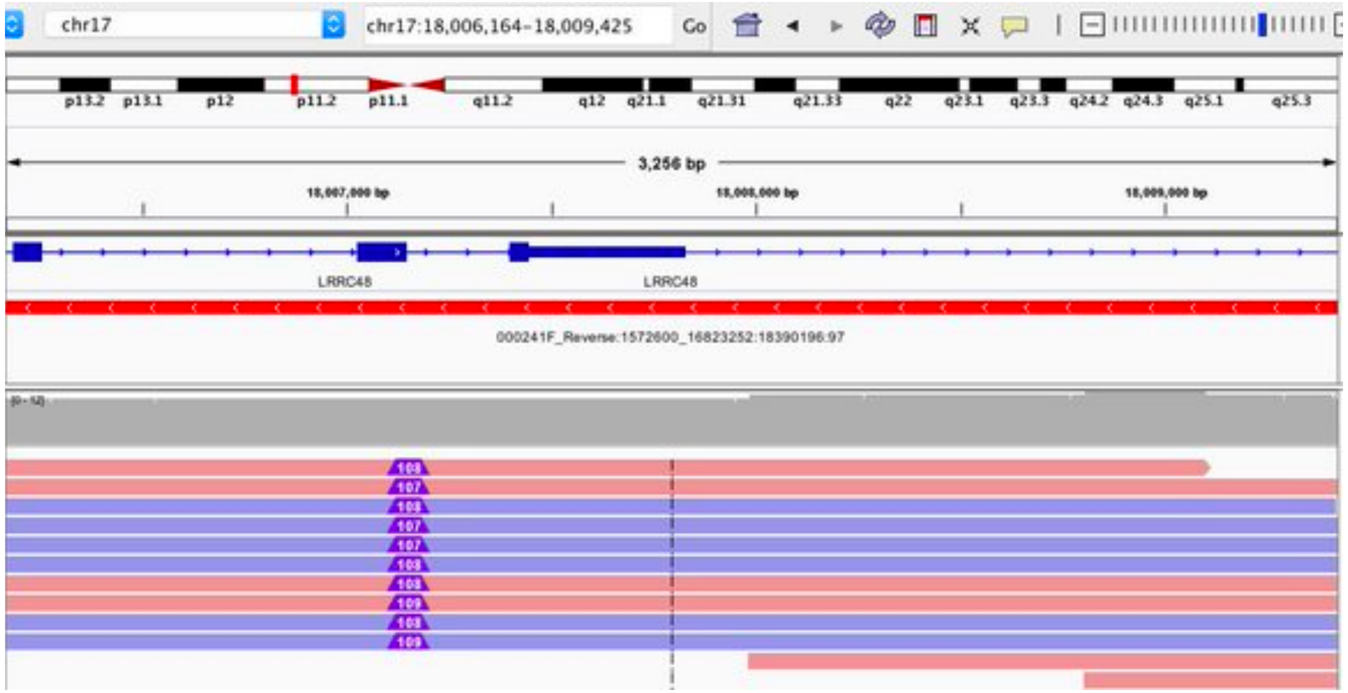




https://software.broadinstitute.org/software/igv/interpreting_insert_size

Structural variants affecting insert size: Insertions





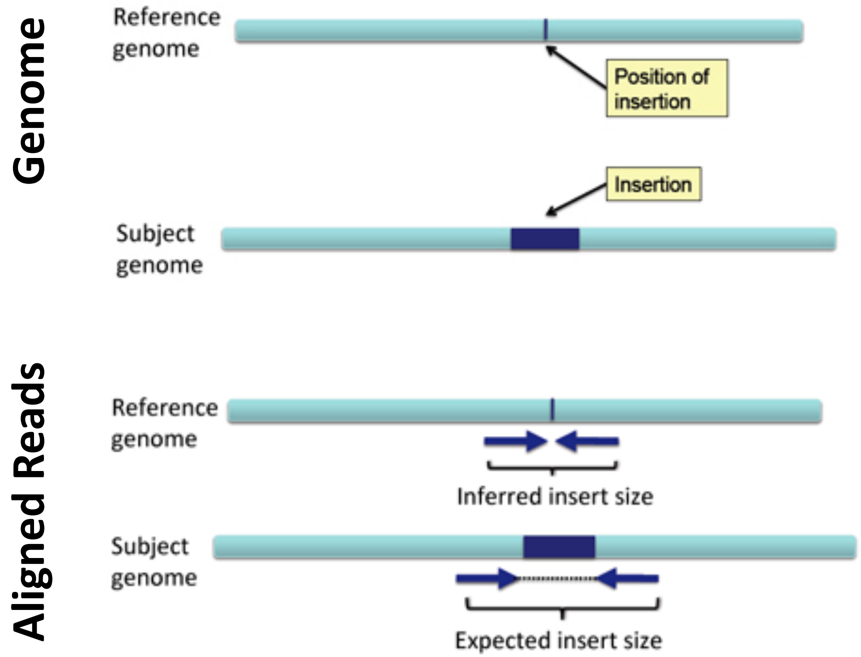
Pacbio long-reads https://twitter.com/infoecho/

Note: The maximum size of an insertion detectable by variant bases
Is limited by read length
The maximum detectable size is approximately equal to:

read length/2 <- pushing it

https://software.broadinstitute.org/software/igv/interpreting_insert_size

Structural Variants: Insertions



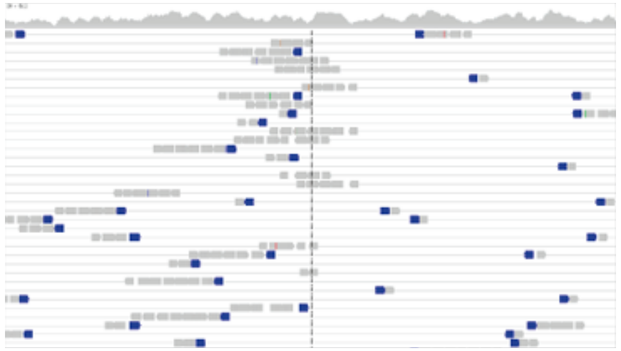
Note: The maximum size of an insertion detectable by insert size anomaly is limited by the size of the fragments.

They must be long enough to span the insertion and include sequences on both ends that are mapped to the reference.

The maximum detectable size is approximately equal to:

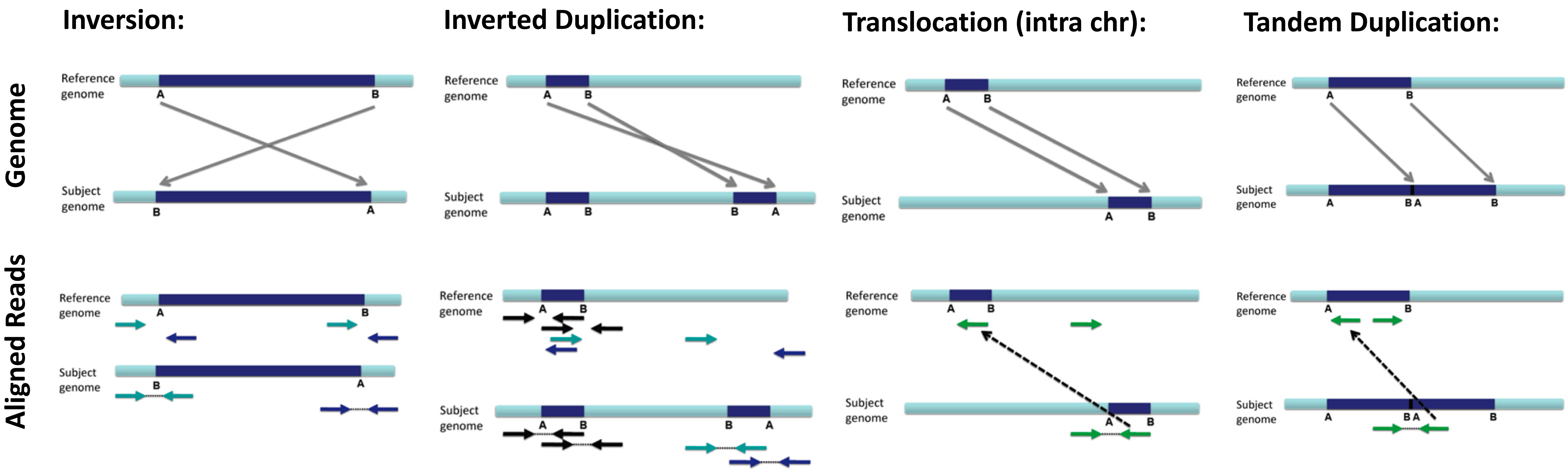
Fragment length - (2x read length)

Detection of this event is therefore more likely with larger fragment libraries, such as Illumina mate-pair (not paired-end) and SOLID.



https://software.broadinstitute.org/software/igv/interpreting_insert_size

Structural Variants affecting mapped-read direction:



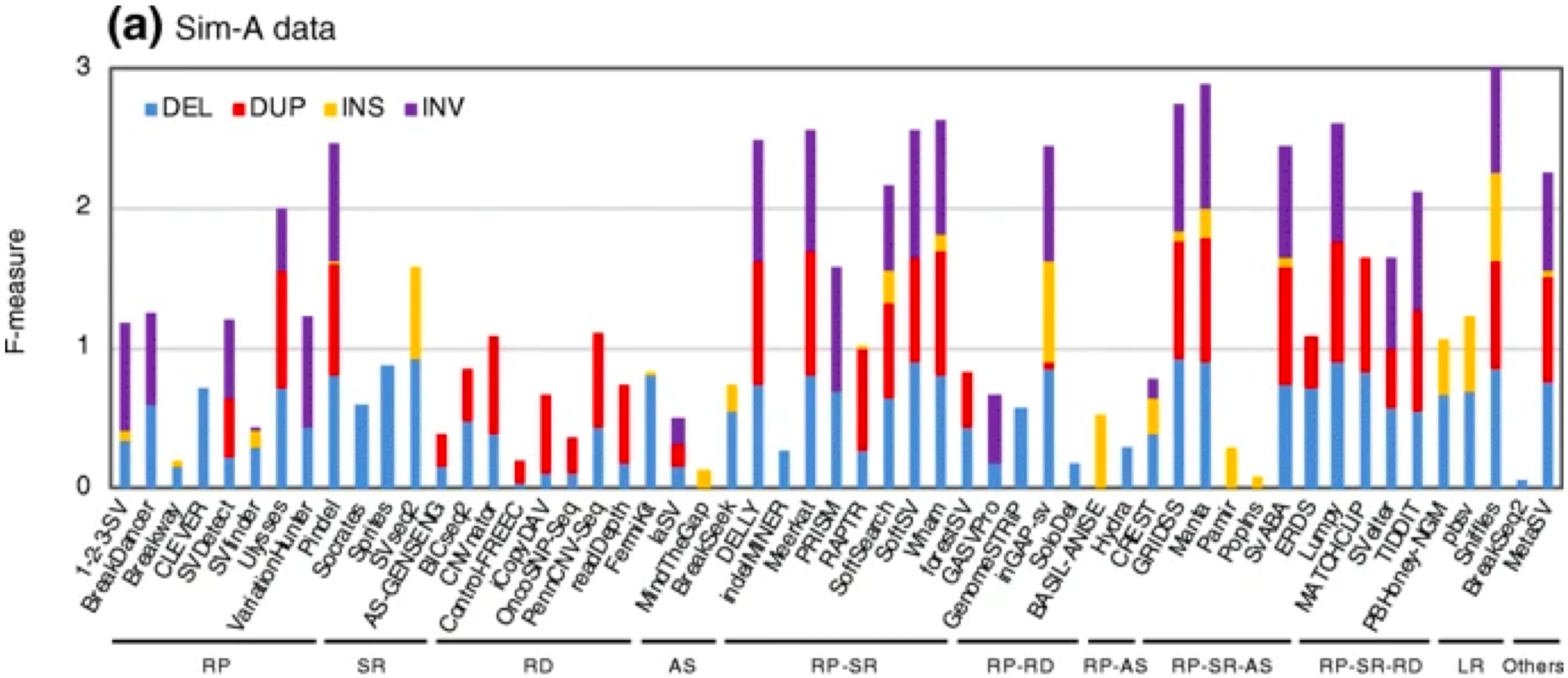
Inversion:

Inverted Duplication:

Translocation (intra chr):

Tandem Duplication:

Structural variant caller Performance varies depending on SV type

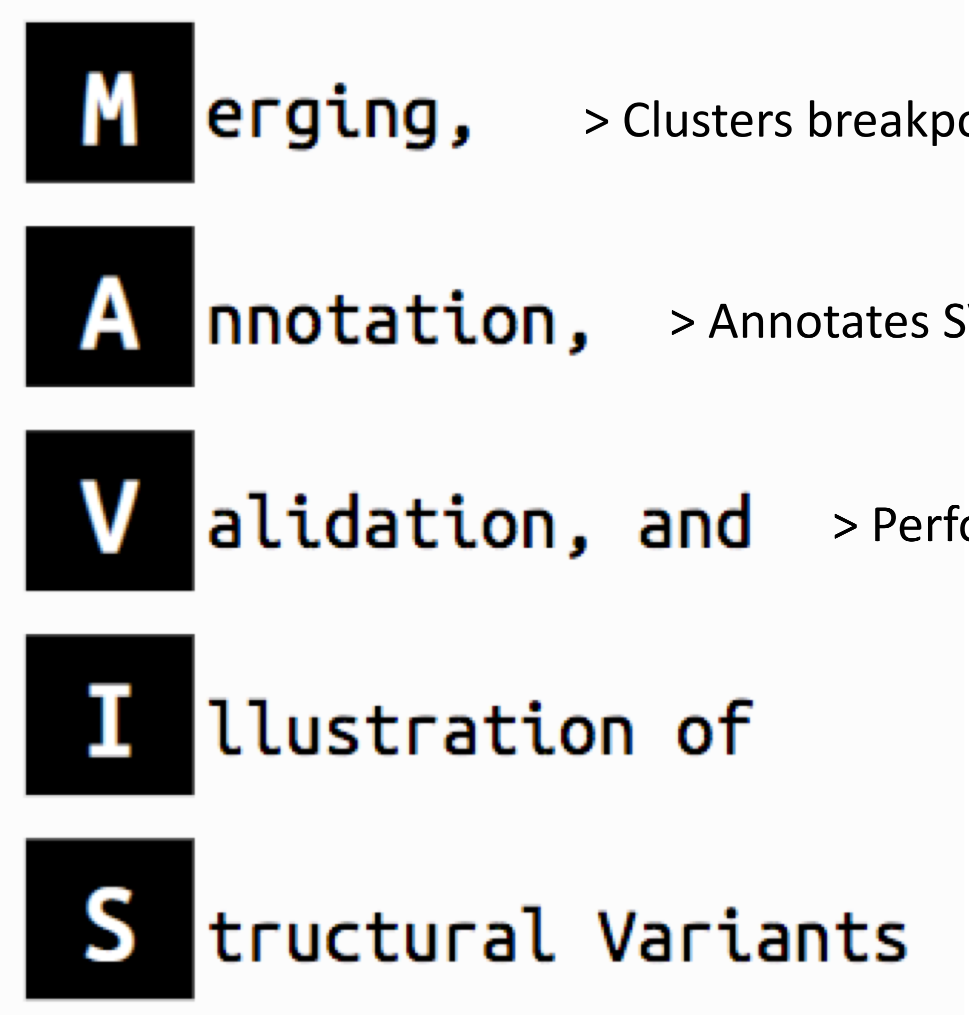


Kosugi *et al.* Genome Biology 2019

Structural Variant Annotation

- **Purpose**
  - To aid in the interpretation of variants
- **Annotations**
  - Classification (Deletion, translocation, inverted translocation etc.)
  - Genes/regions affected
  - Predicted amino acid change
  - +++
- **Some available tools:**
  - MAVIS
  - SVAnnotator

Structural Variant Annotator Example: MAVIS



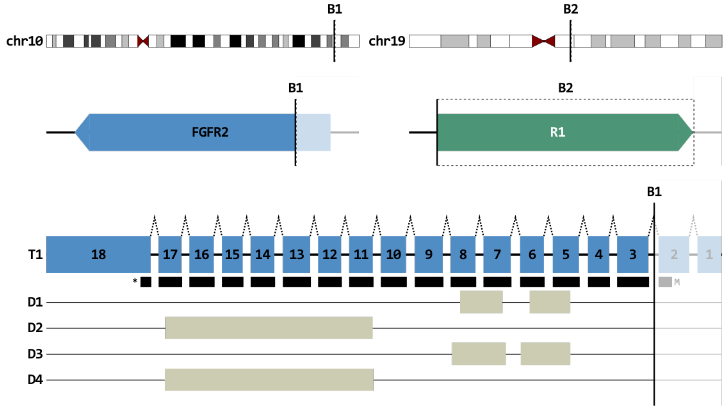
M erging, > Clusters breakpoints and SVs the same or multiple tools

A nnotation, > Annotates SV with genes, somatic/germline status, AA effect etc.

V alidation, and > Performs in-bam validation to further polish/filter results

I llustration of

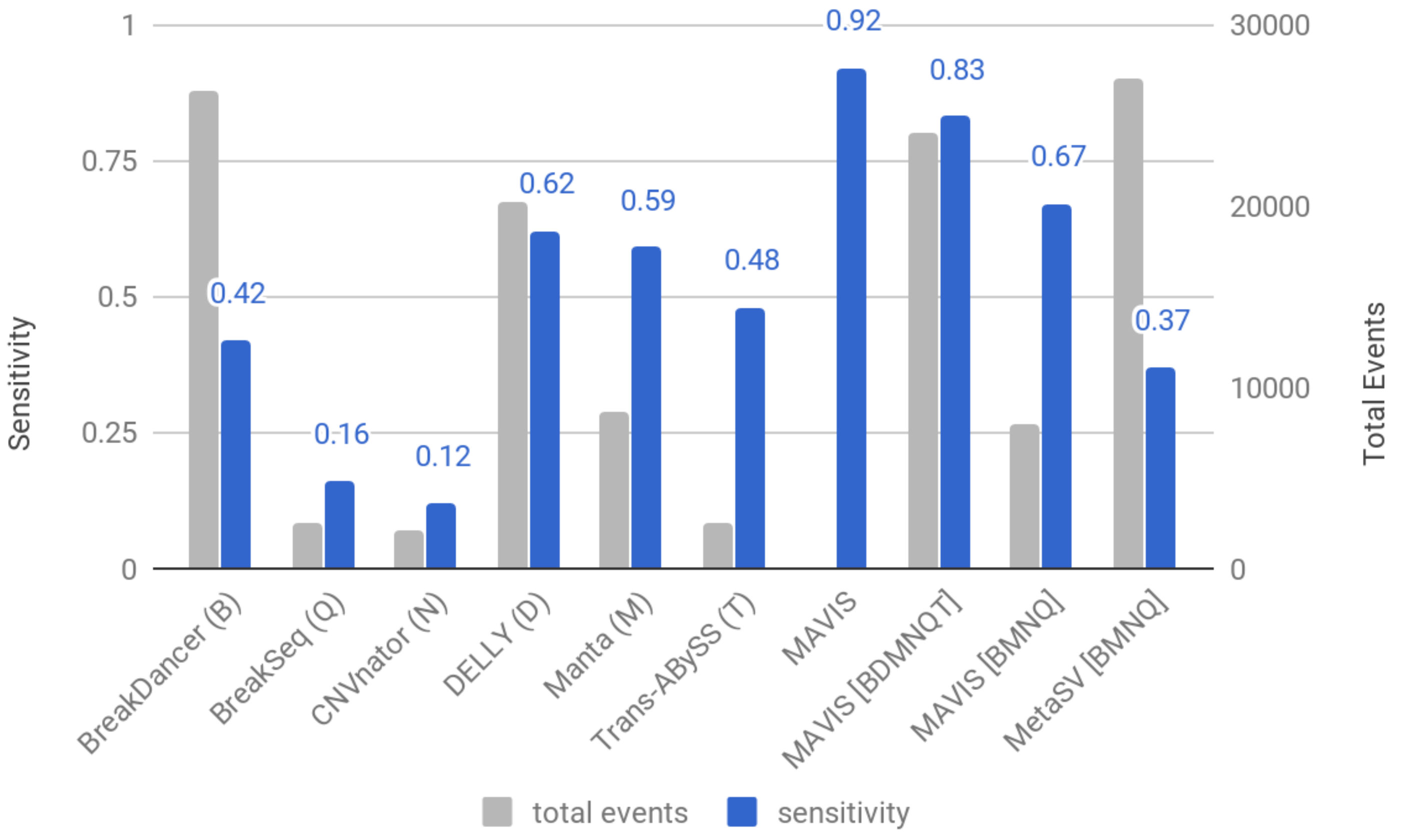
S tructural Variants



http://mavis.bcgsc.ca

Structural Variant Annotator Example: MAVIS

*Adding additional callers appears to improve structural variant detection as well

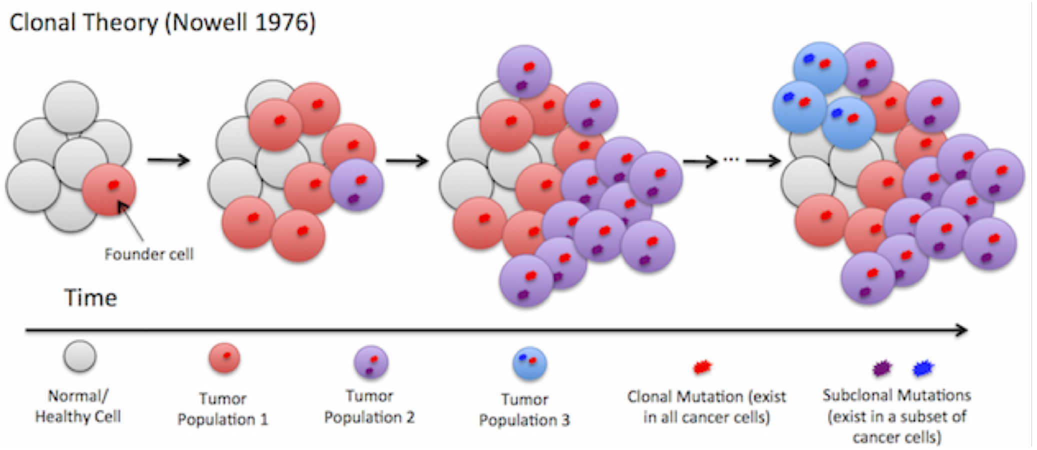


Reisle *et al.* Bioinformatics 2019

Challenges facing modern day small variant calling algorithms:

Low variant allele frequencies (VAF)

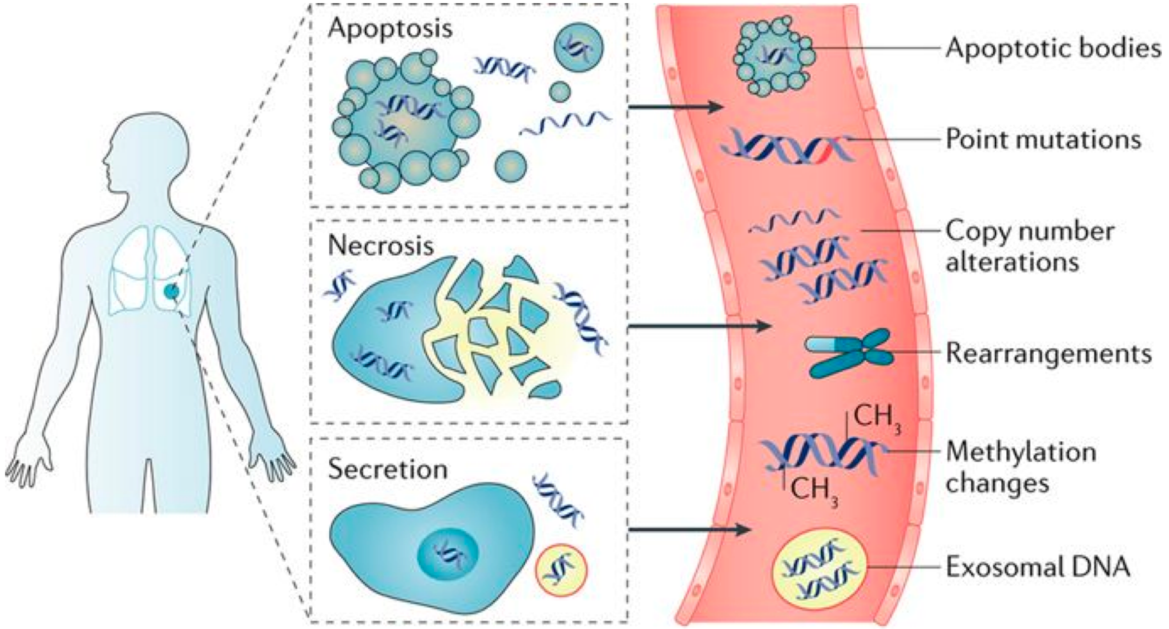
Subclonal Mutations



Clonal Theory (Nowell 1976)
Founder cell
Time
Normal/ Healthy Cell
Tumor Population 1
Tumor Population 2
Tumor Population 3
Clonal Mutation (exist in all cancer cells)
Subclonal Mutations (exist in a subset of cancer cells)

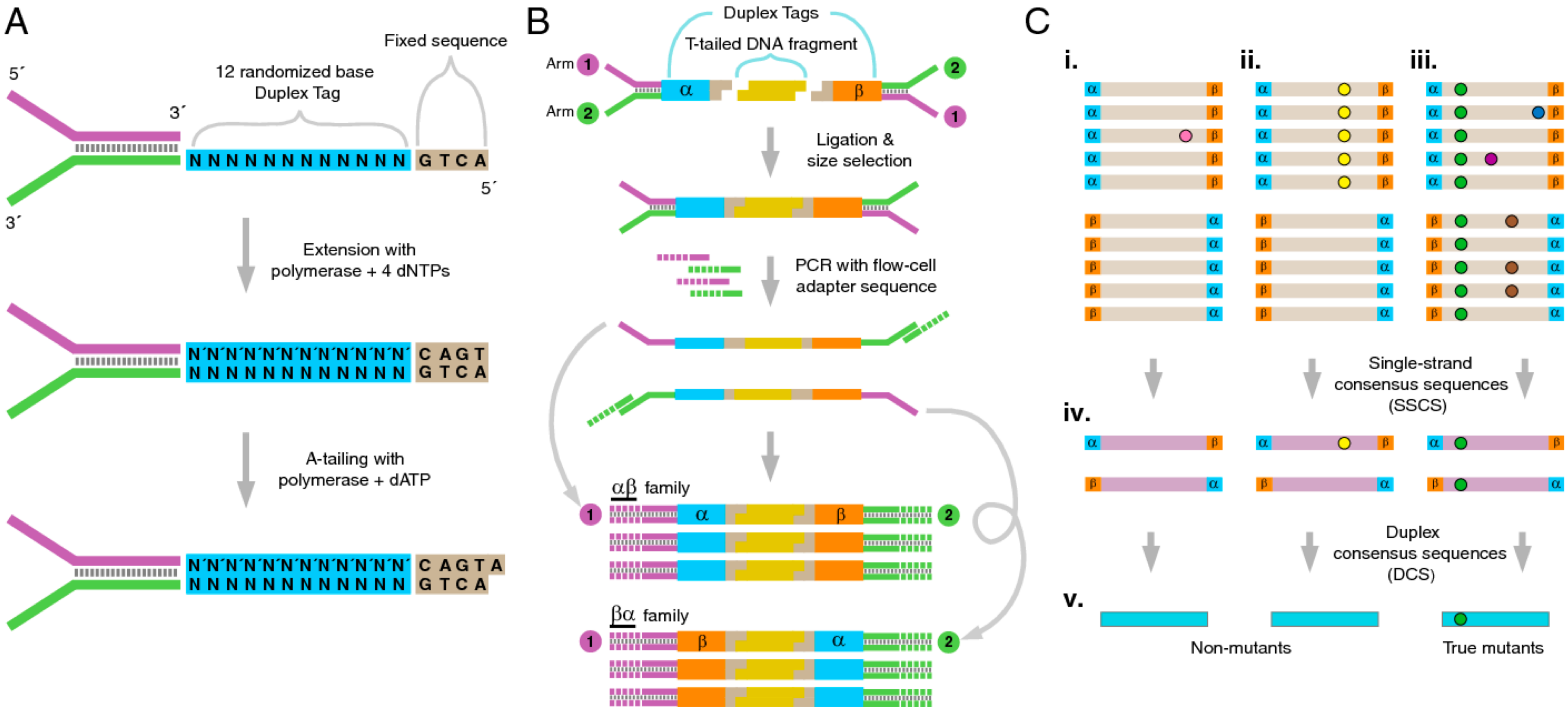
http://www.cs.carleton.edu/faculty/loesper/research.html

Cell-Free DNA



Nature Reviews Cancer **volume17**, pages223–238 (2017)

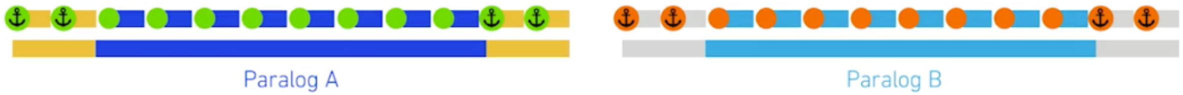
Detecting rare variants using unique molecular indices and duplex sequences



Challenges facing modern day small variant calling algorithms:
Difficult to align regions

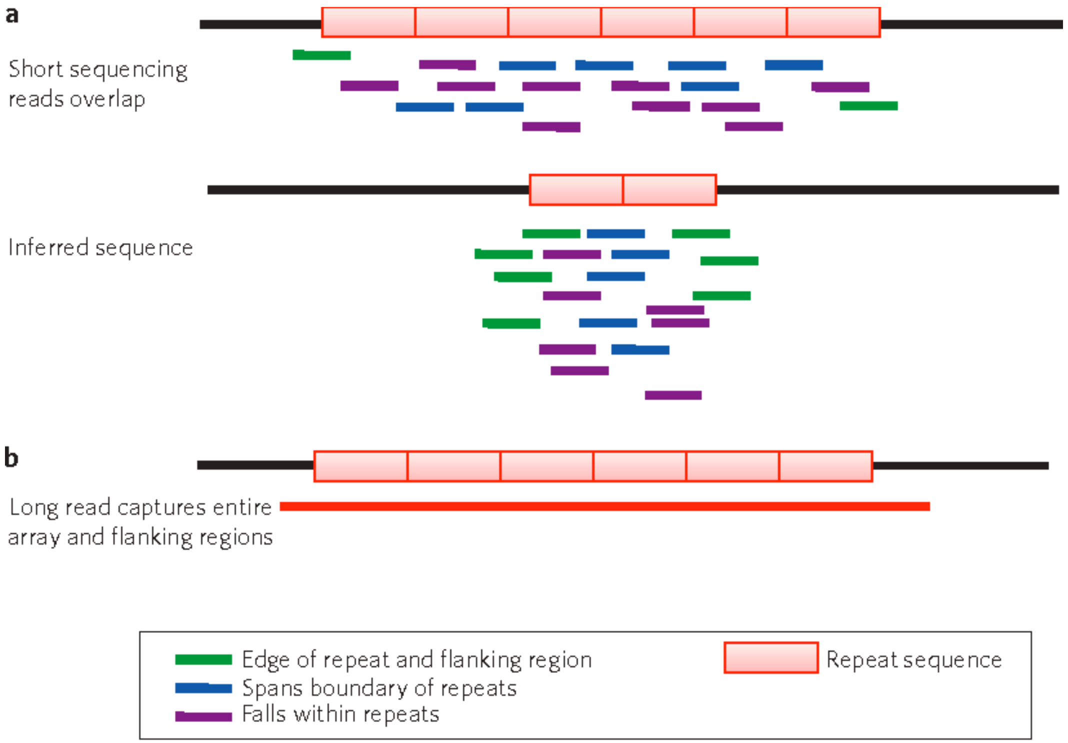
Linked-Reads from 10X genomics

Barcodes recruit short-reads into paralogous gene loci



https://www.10xgenomics.com/solutions/genome/

Long Reads from Pacbio



a
Short sequencing reads overlap
Inferred sequence
b
Long read captures entire array and flanking regions
Edge of repeat and flanking region
Spans boundary of repeats
Falls within repeats
Repeat sequence

Nature plants 2015 DOI:10.1038/nplants.2015.169

Questions?

Break

Hands on Exercise

Practical Exercise: Calling Somatic SNVs and Indels

- **Getting started:**
  - I have e-mailed the group a link with the necessary files: also Here
  - Those **with** access to Mordor: scp (copy) these data to a directory of your choosing
  - Those **without** access to Mordor: pair up with a) a user with access or b) another user without access. I will provide a laptop and log you in
- **Your task:**

1) Alter the two bash scripts in the `~/scripts/` directory to correspond with the location of the files you are processing
2) Submit these jobs to the cluster (ex. `qsub runMutect2_mordor.sh)`
3) Compare the resulting final output from each tool

   hint: final Mutect2 vcf has ‘filtered’ in the file name

   final Strelka vcfs have ‘passed.somatic’ in the file name
4) Find the two variants that were found by one tool and not by the other
5) Load the two provided BAM files into IGV and take a snapshot of these two loci

Bonus: Why do you think one tool did not report each of these?

Practical Exercise: Calling Somatic SNVs and Indels

- **Useful Linux Commands:**
  - Change directory: `cd` Ex: `cd ./a_subdirectory`
  - Go back one directory level: `cd ../`
  - List files in current directory: `ls` or `ll`
  - View the contents of a file: `cat`
  - Edit a file on the cluster: nano or vim Ex: `nano runMutect2_mordor.sh`

Thanks!