Moderní metody analýzy genomu

Bioinformatika II

Karol Pál

(Šárka Pospíšilová Research Group - Centre for Molecular Medicine CEITEC)

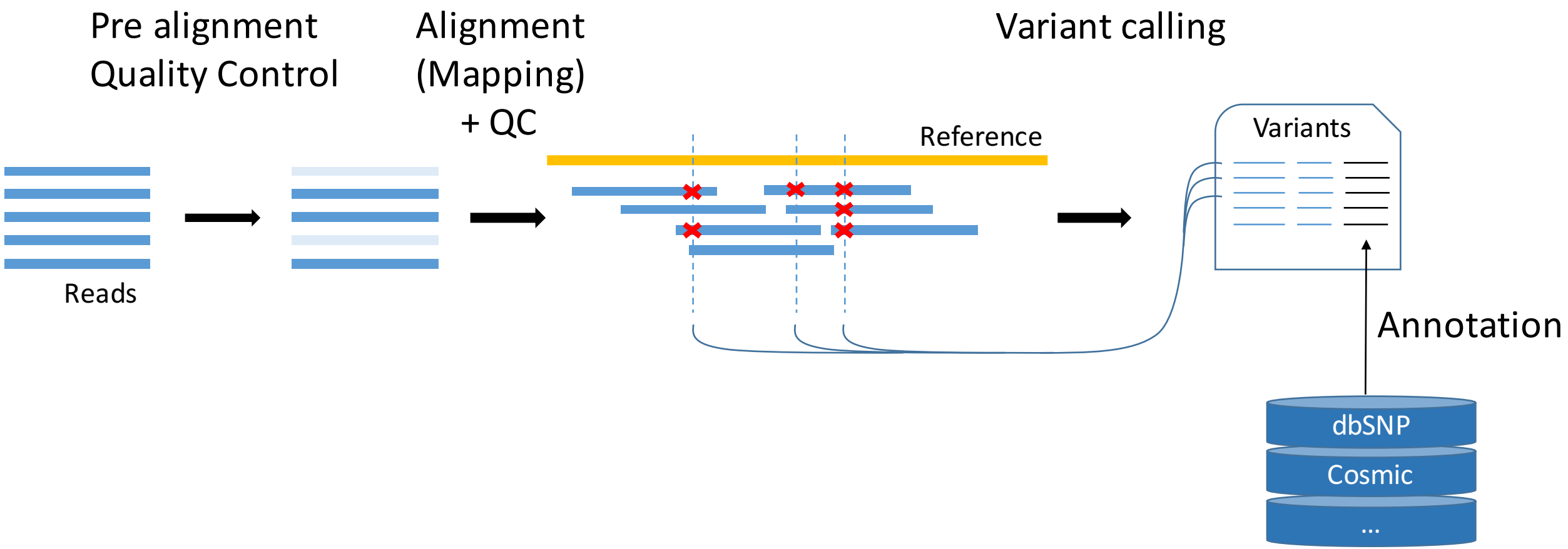
Content of this lecture

- Recap
  - Sequencing reads
  - Data analysis pipeline (workflow)
- Quality control
  - Raw Reads
  - Alignment
  - Variants
- Visualizing NGS Data
  - IGV

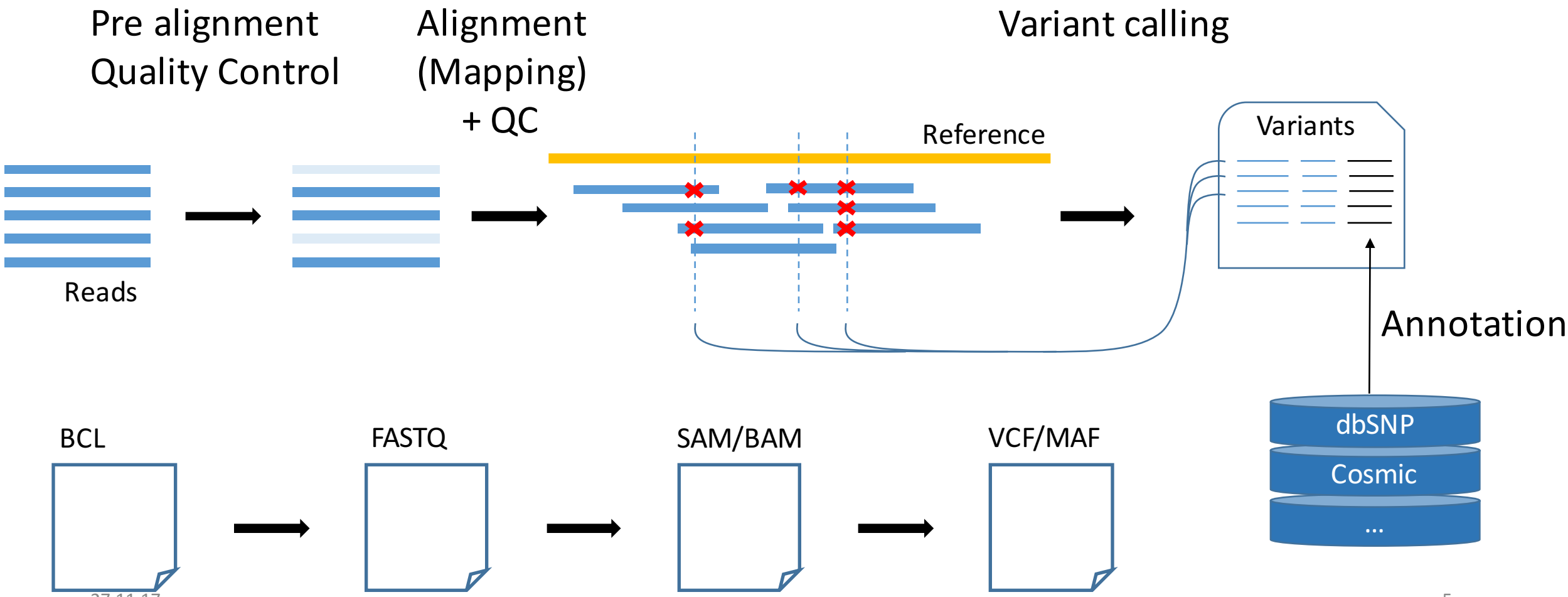
Recap – Sequencing reads

http://nextgen.mgh.harvard.edu/CustomPrimer.html

Recap – Data analysis pipeline

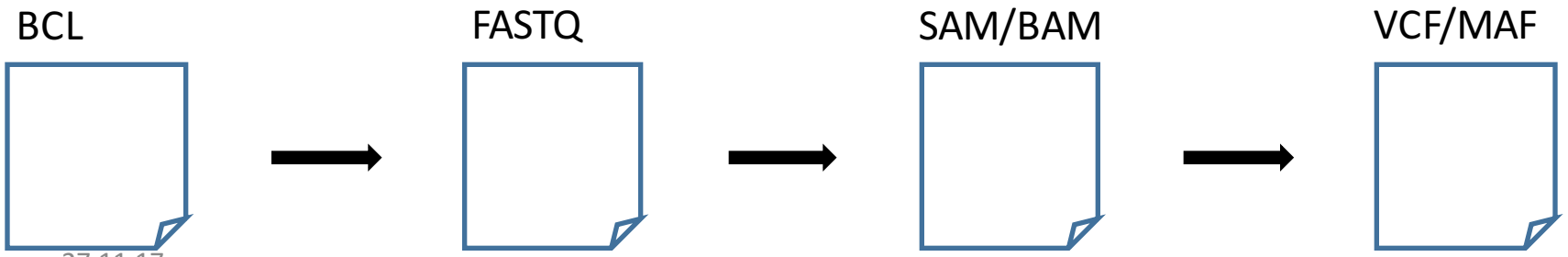


Recap – Data analysis pipeline



Quality Control

- Different steps
  - Reads summary statistics
  - Aligners
  - Post alignment statistics
- Different tools
  - Different kinds of outputs



Current version: v1.3 | Home | Docs | Plugins | Logo | Example Reports

MultiQC

Aggregate results from bioinformatics analyses across many samples into a single report

MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools.

- Introduction to MultiQC *(1:19)*
- Installing MultiQC *(4:33)*
- Running MultiQC *(5:21)*
- Using MultiQC Reports *(6:06)*

GitHub

Python Package Index

Documentation

56 supported tools

Publication / Citation

Get help on Gitter

Quick Install

```
pip install multiqc    # Install
multiqc .              # Run
```

pip | conda | manual

Need a little more help? See the full installation instructions.

Current version: v1.3 | Home | Docs | Plugins | Logo | Example Reports

MultiQC

Aggregate results from bioinformatics analyses across many samples into a single report

MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools.

- Introduction to MultiQC *(1:19)*
- Installing MultiQC *(4:33)*
- Running MultiQC *(5:21)*
- Using MultiQC Reports *(6:06)*

GitHub

Python Package Index

Documentation

56 supported tools

Publication / Citation

Get help on Gitter

Quick Install

```
pip install multiqc    # Install
multiqc .              # Run
```

pip | conda | manual

Need a little more help? See the full installation instructions.

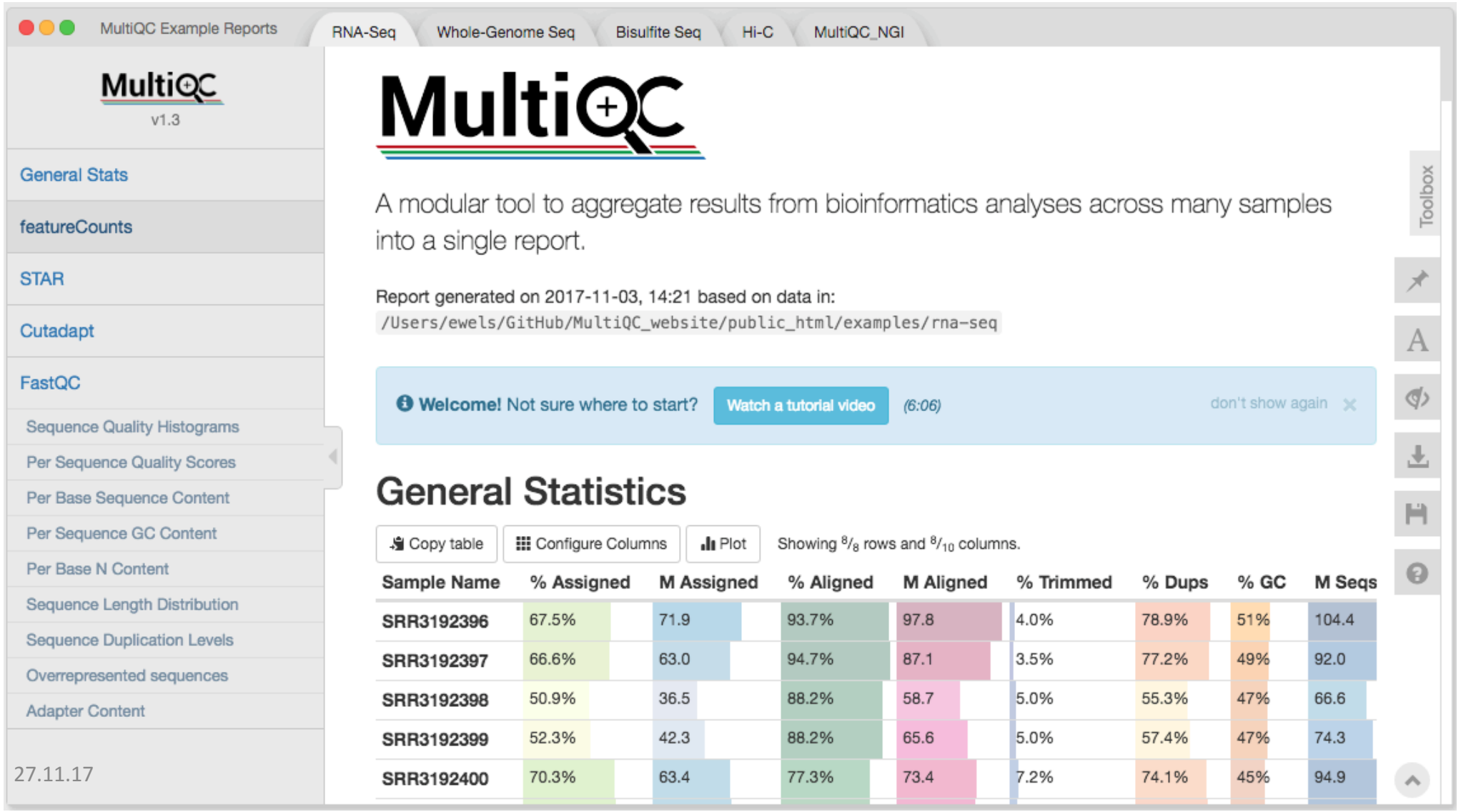
Quality Control

MultiQC: Supported Tools

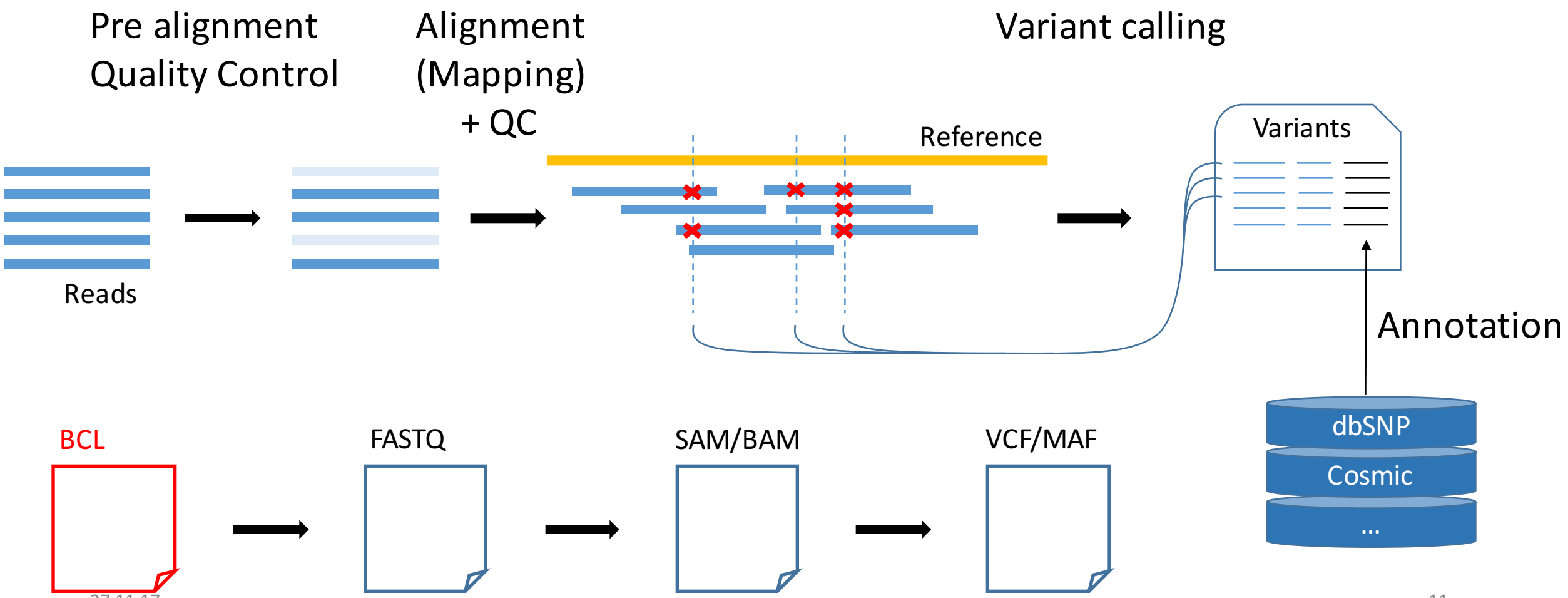
Pre-alignment tools | Alignment tools | Post-alignment tools

Skewer	Skewer is an adapter trimming tool specially designed for processing next-generation sequencing (NGS) paired-end sequences.
SortMeRNA	SortMeRNA is a program tool for filtering, mapping and OTU-picking NGS reads in metatranscriptomic and metagenomic data.
Trimmomatic	Trimmomatic is a flexible read trimming tool for Illumina NGS data
Bismark	Bismark is a tool to map bisulfite converted sequence reads and determine cytosine methylation states.
Bowtie 1	Bowtie 1 is an ultrafast, memory-efficient short read aligner.
Bowtie 2	Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.
BBMap	BBMap is a suite of pre-processing, assembly, alignment, and statistics tools for DNA/RNA sequencing reads.
HiCUP	HiCUP (Hi-C User Pipeline) is a tool for mapping and performing quality control on Hi-C data.
HISAT2	HISAT2 is a fast and sensitive alignment program for mapping NGS reads (both DNA and RNA) to reference genomes.
Kallisto	kallisto is a program for quantifying abundances of transcripts from RNA-Seq data.
Salmon	Salmon is a tool for quantifying the expression of transcripts using RNA-seq data.
STAR	STAR is an ultrafast universal RNA-seq aligner.
TopHat	TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes.
Bamtools	BamTools provides both a programmer's API and an end-user's toolkit for handling BAM files.
Bcftools	BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF.
BUSCO	BUSCO assesses genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs.
Conpair	Conpair estimates concordance and contamination for tumour–normal pairs
Disambiguate	Disambiguation algorithm for reads aligned to two species (e.g. human and mouse genomes) from Tophat, Hisat2, STAR or BWA mem.

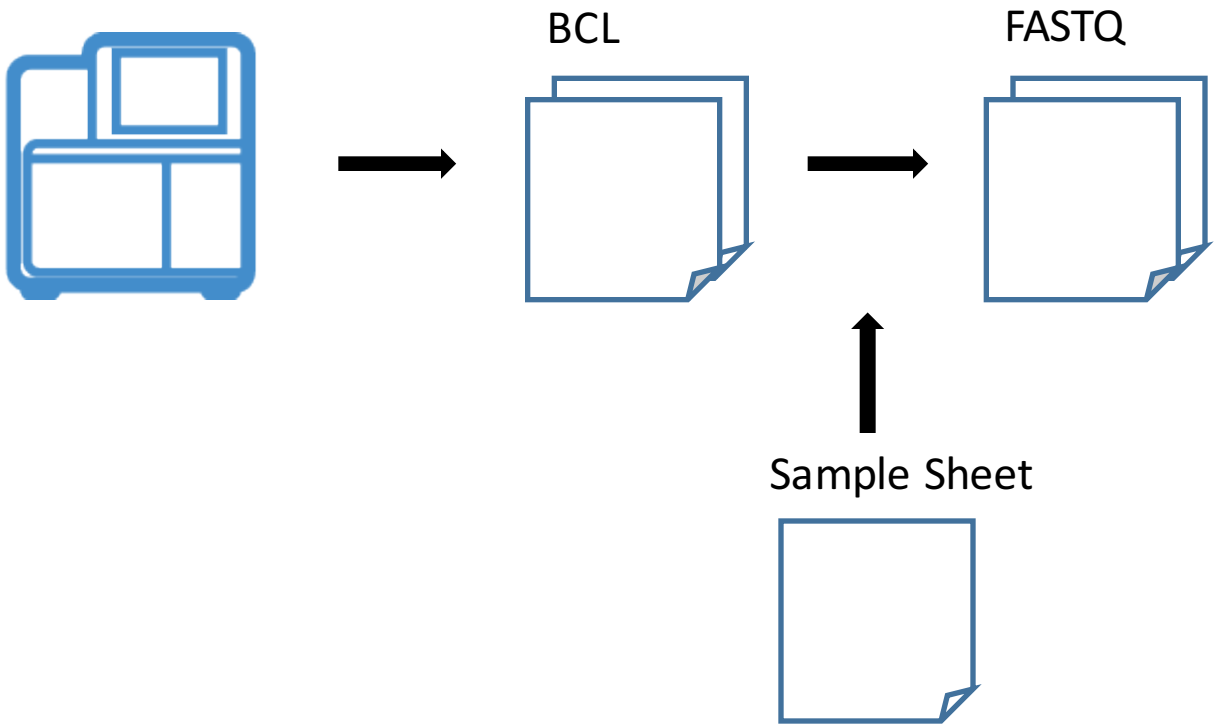
Quality Control



Data analysis pipeline



BCL to FASTQ



- BCL - raw sequencing output
- Convert to FASTQ format
- Split into sample files
- May be automated

Sample sheet

```
[Header]
IEMFileVersion,4
Experiment Name,Exom.20171013
Date,9.10.2017
Workflow,GenerateFASTQ
Application,FASTQ Only
Assay,TruSeq LT
Description,
Chemistry,Default

[Reads]


[Settings]
ReverseComplement,0

[Data]
Sample_ID,Sample_Name,Sample_Plate,Sample_Well,I7_Index_ID,index,Sample_Project,Description,
BRN01077_normal,,,,AD002,CGATGT,,
BRN01404_normal,,,,AD007,CAGATC,,
BRN01503_normal,,,,AD019,GTGAAA,,
```

Sample sheet

```
[Header]
IEMFileVersion,4
Experiment Name,Exom.20171013
Date,9.10.2017
Workflow,GenerateFASTQ
Application,FASTQ Only
Assay,TruSeq LT
Description,
Chemistry,Default

[Reads]


[Settings]
ReverseComplement,0

[Data]
Sample_ID,Sample_Name,Sample_Plate,Sample_Well,I7_Index_ID,index,Sample_Project,Description,
BRN01077_normal,,,,AD002,CGATGT,,
BRN01404_normal,,,,AD007,CAGATC,,
BRN01503_normal,,,,AD019,GTGAAA,,
```

Raw reads – bcl2fastq

MultiQC output

General Statistics

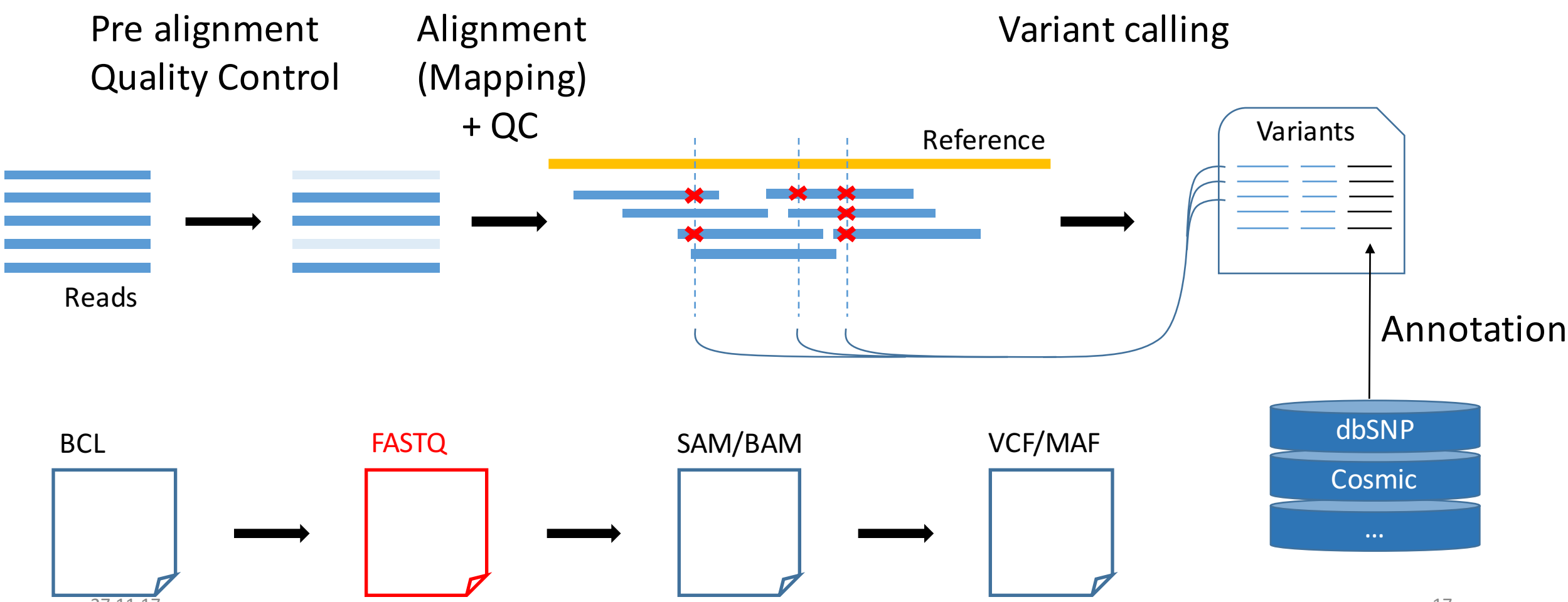
Copy table | Configure Columns | Plot | Showing 4/4 rows and 3/3 columns.

Sample Name	Total Reads	Mb Yield ≥ Q30	% Perfect Index
BRNO062_tumor	54 513 947.0	7 943.6	100.0%
BRNO0047_tumor	52 169 492.0	7 596.2	100.0%
BRNO1503_tumor	49 439 468.0	7 199.7	100.0%
undetermined	8 933 116.0	1 024.4	0.0%

Raw reads – bcl2fastq

MultiQC output

Data analysis pipeline

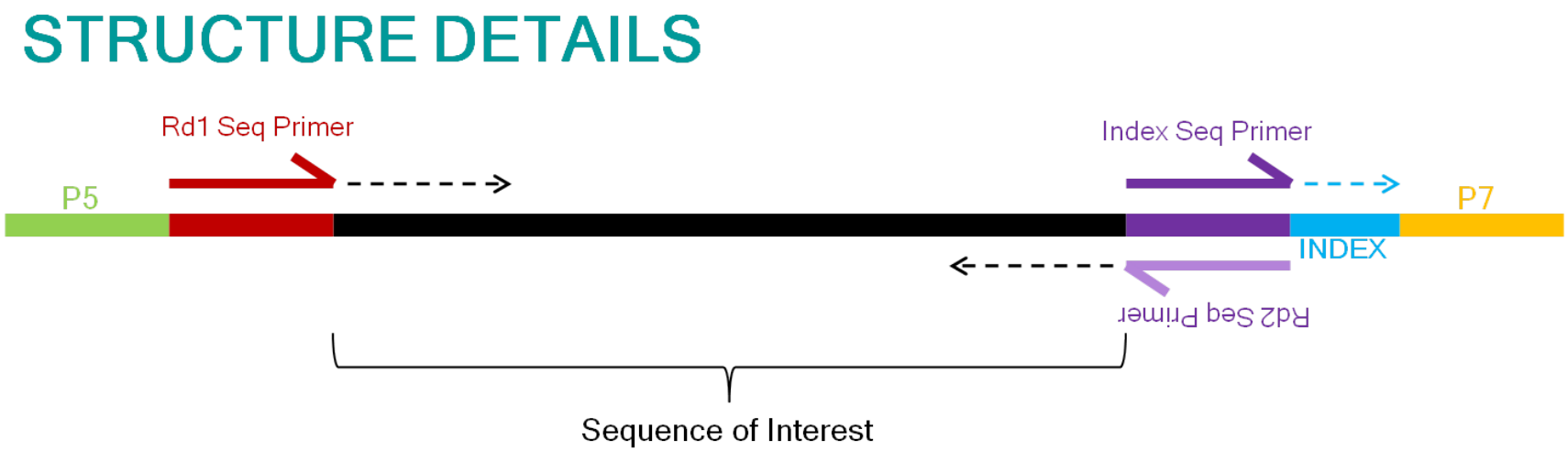


FastQC

- Summary statistics
- Two modes
  - Stand alone program
  - Command line (output can be integrated to MultiQC)
- Input: Fastq or BAM file
- [Demo]

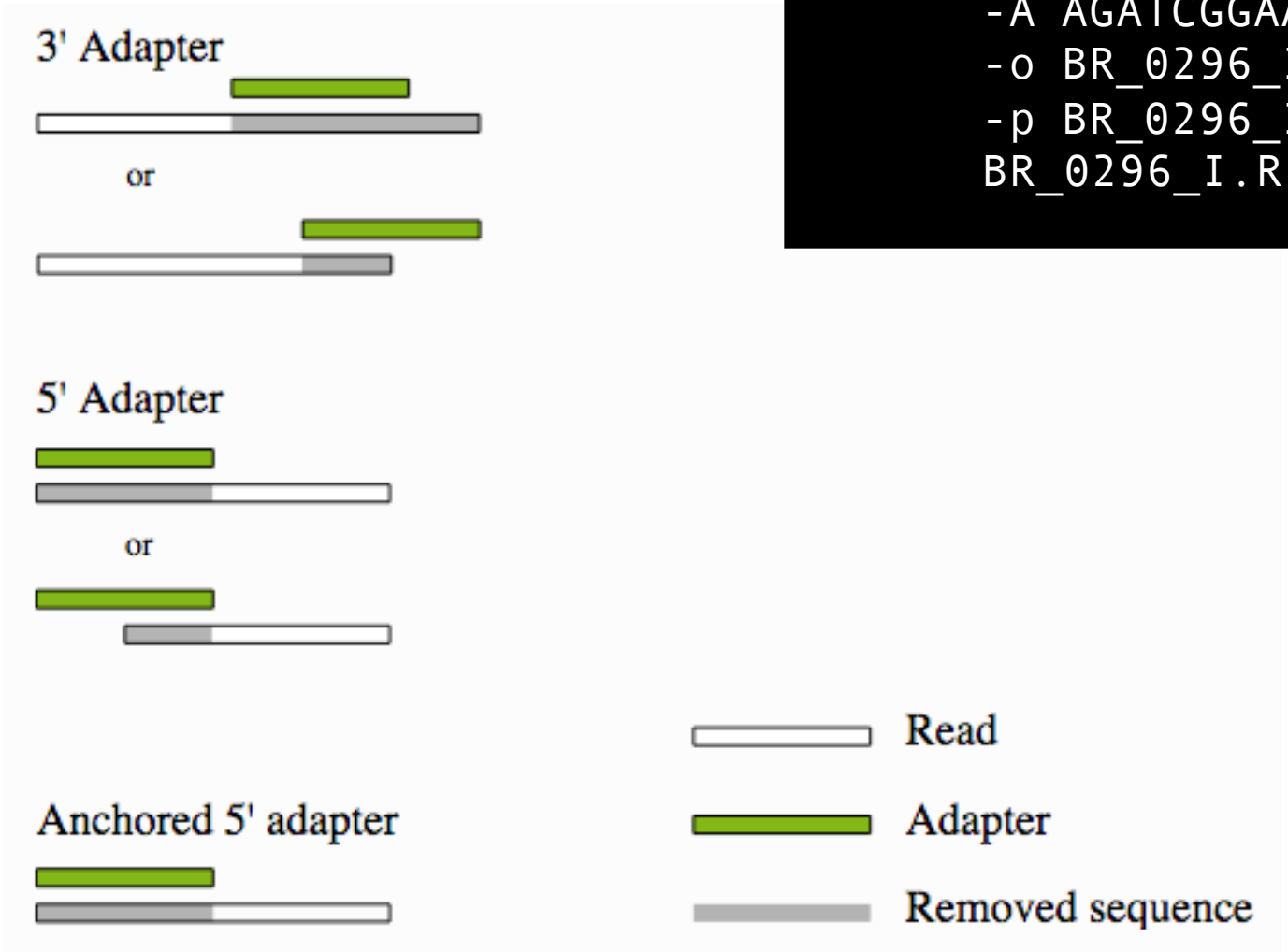
Trimming

- Adaptors
- Low quality ends of reads
- Tools:
  - Cutadapt
  - Trimmomatic



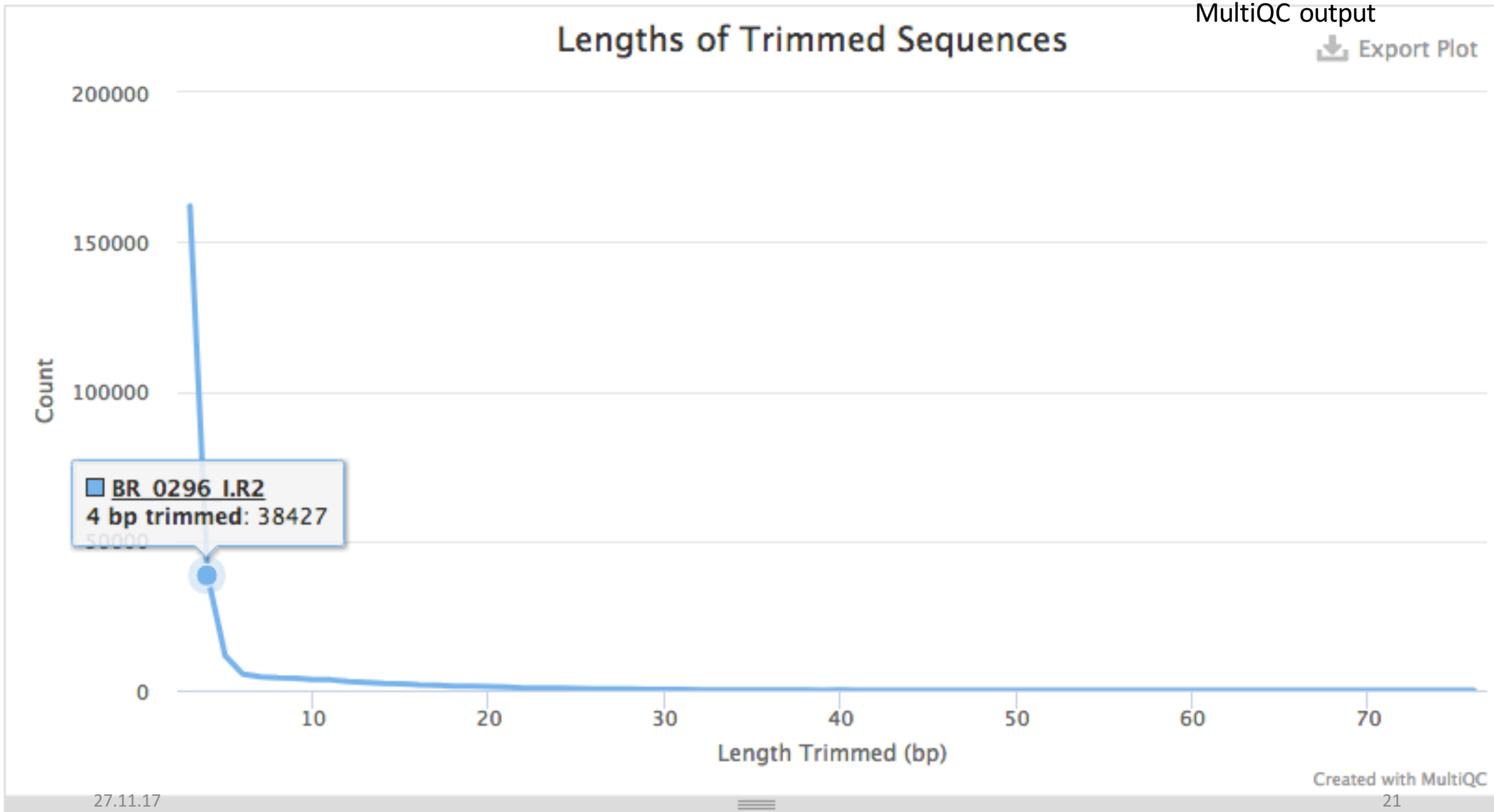
Trimming

```
> cutadapt \
      -a AGATCGGAAGAGC \
      -A AGATCGGAAGAGC \
      -o BR_0296_I.trimmed.1.fastq.gz \
      -p BR_0296_I.trimmed.2.fastq.gz \
      BR_0296_I.R1.fq.gz BR_0296_I.R2.fq.gz
```



http://tucf-genomics.tufts.edu/documents/protocols/TUCF_Understanding_Illumina_TruSeq_Adapters.pdf

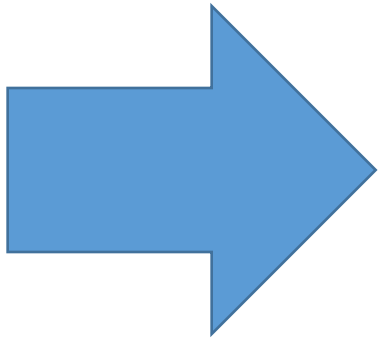
MultiQC output



FastQC II – after trimming

Summary

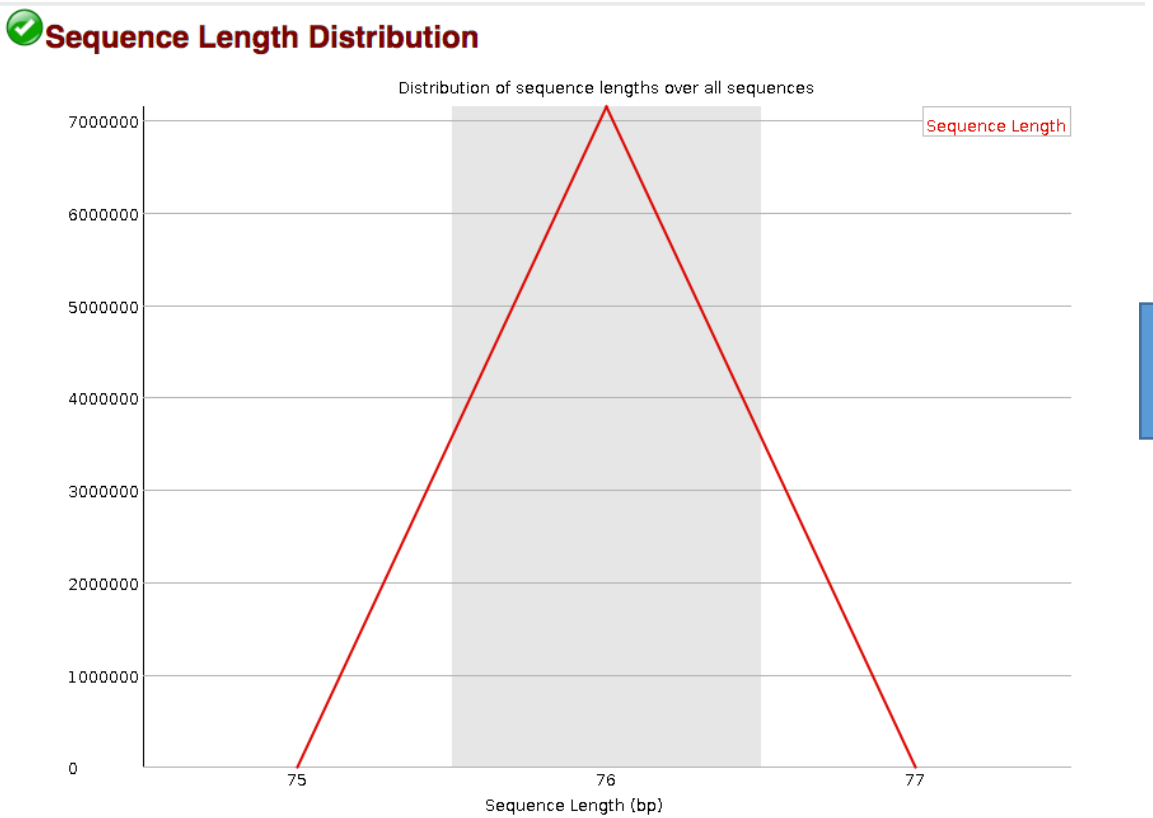
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

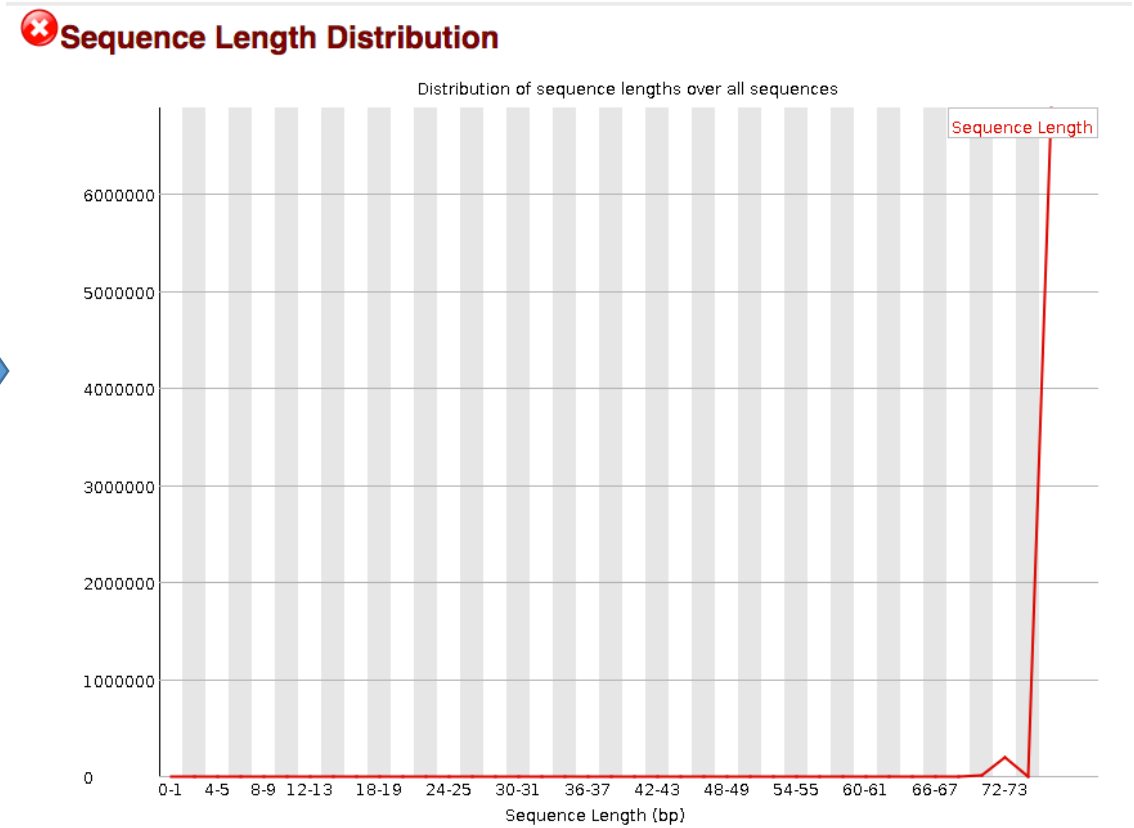


Summary

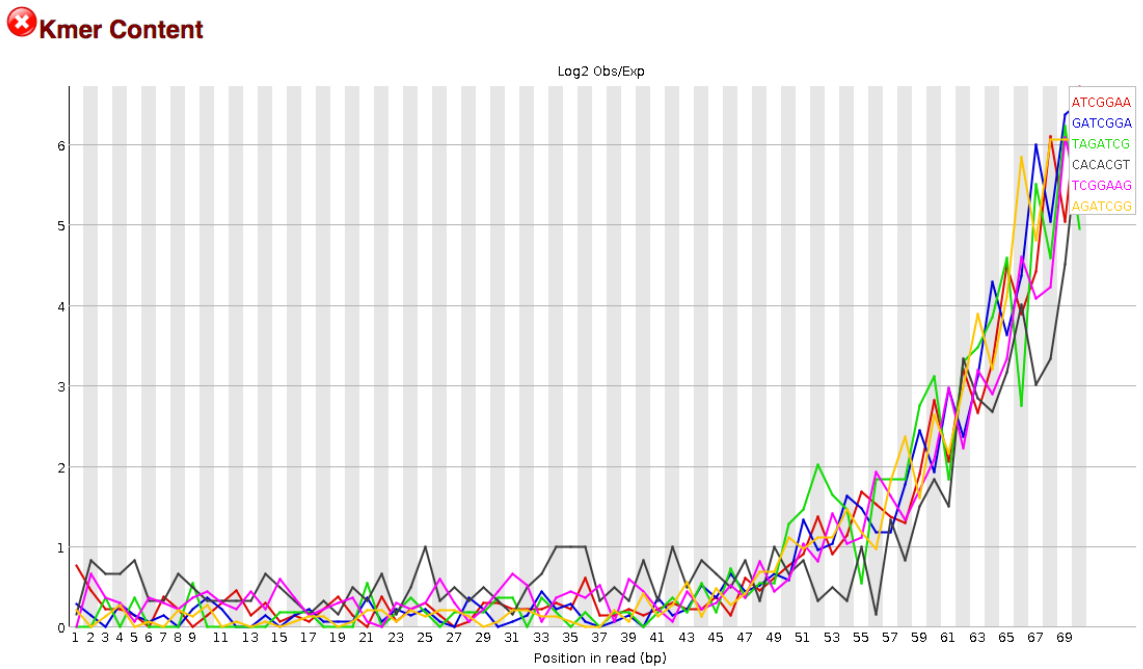
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

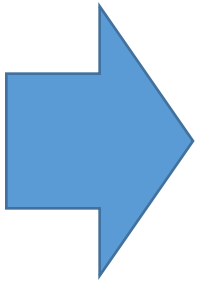
FastQC II – after trimming





FastQC II – after trimming



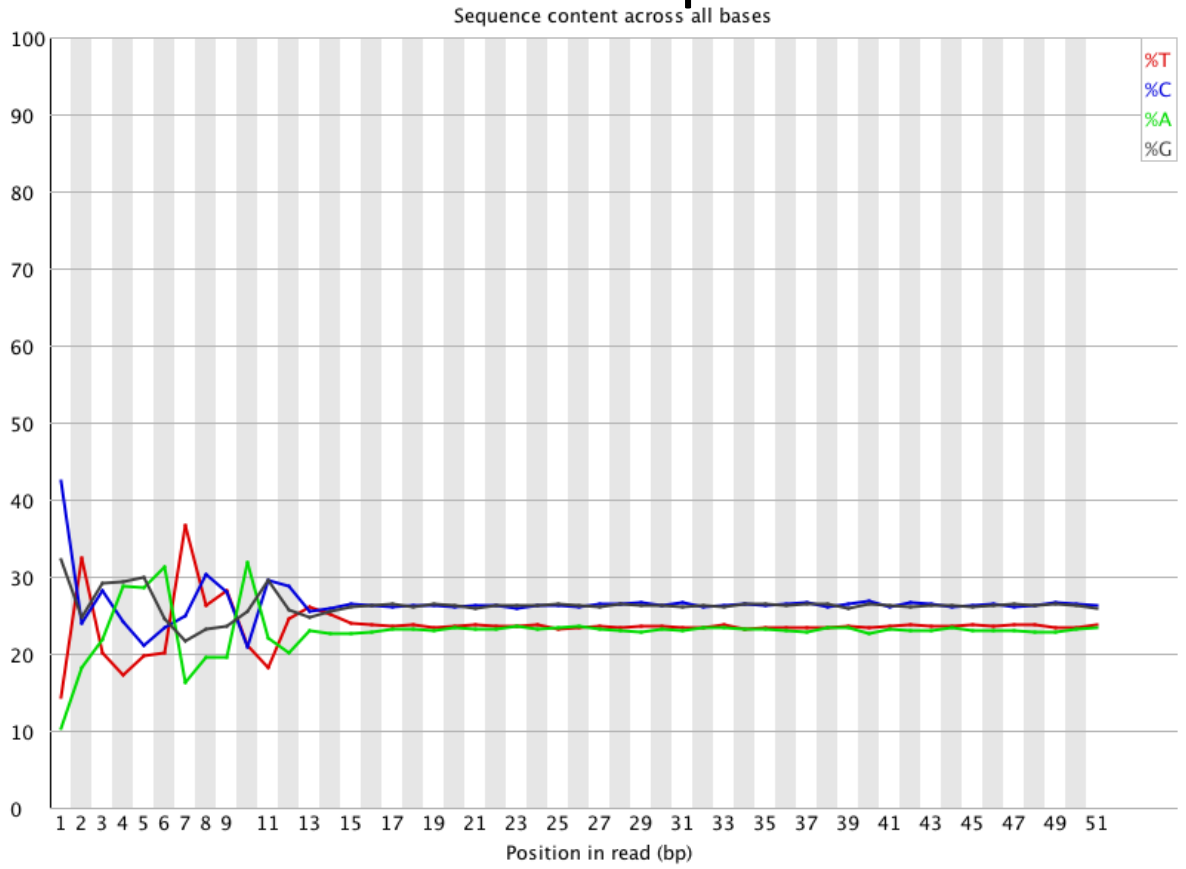


FastQC

- Overrepresented sequences are generally OK
  - Highly expressed genes?
  - cca 10 000
- If the number is too high (> 100 000) may indicate a problem
  - rRNA not depleated?

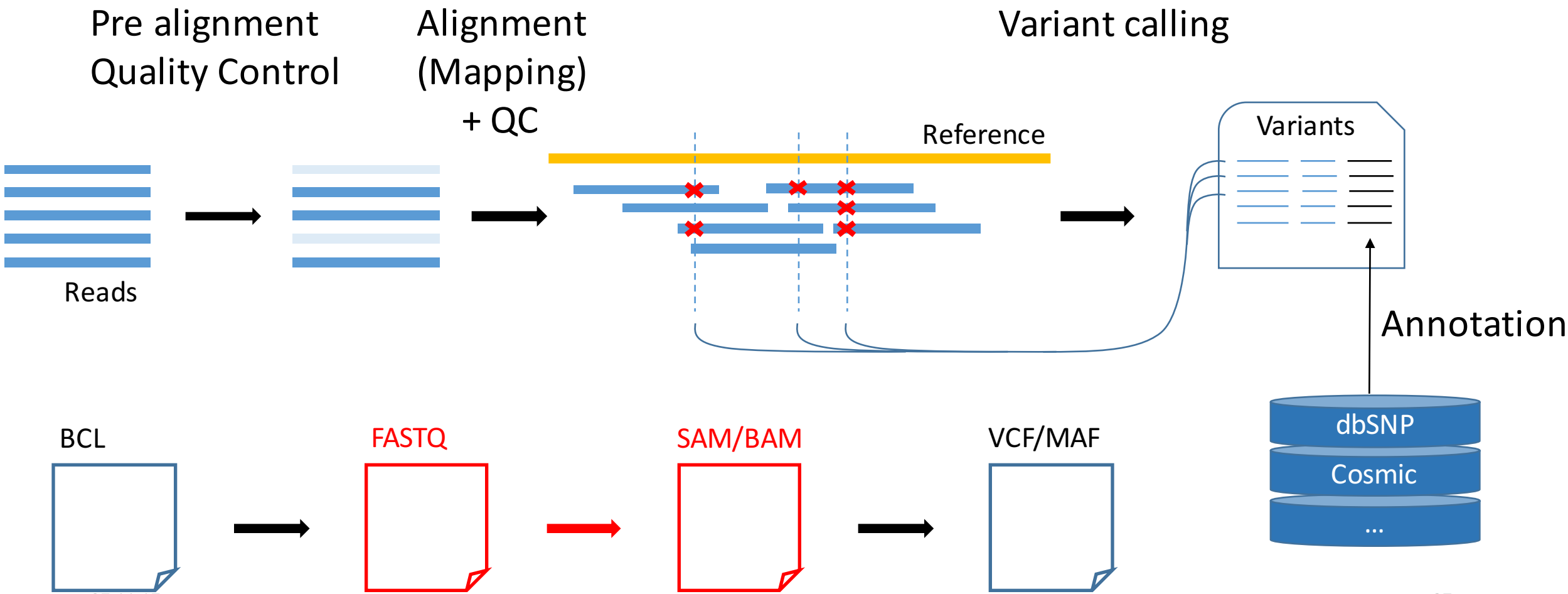
FastQC

- Nearly all RNA-Seq libraries inherit an intrinsic bias in the positions at which reads start.[1]



[1] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html

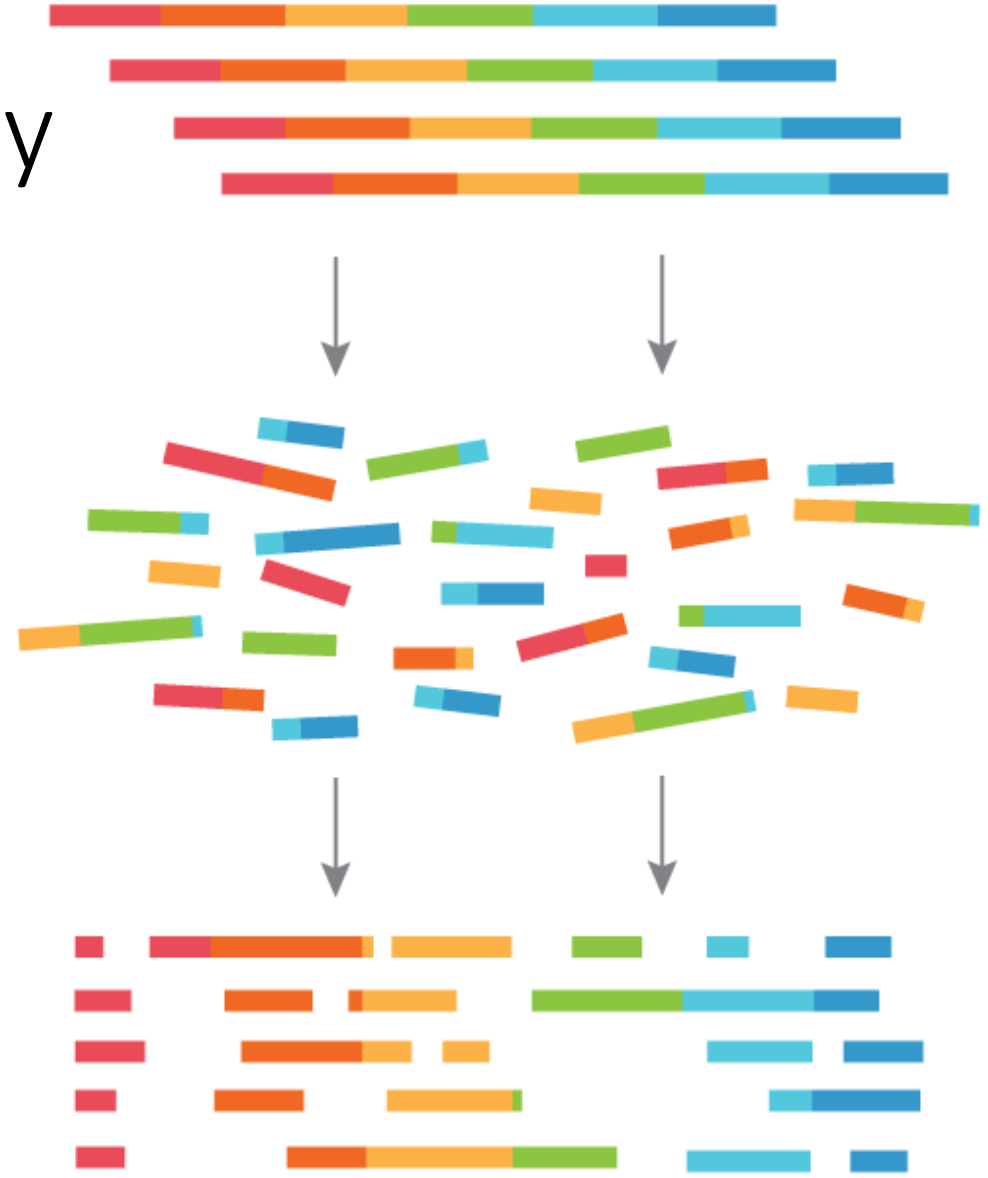
Recap – Data analysis pipeline



DNA

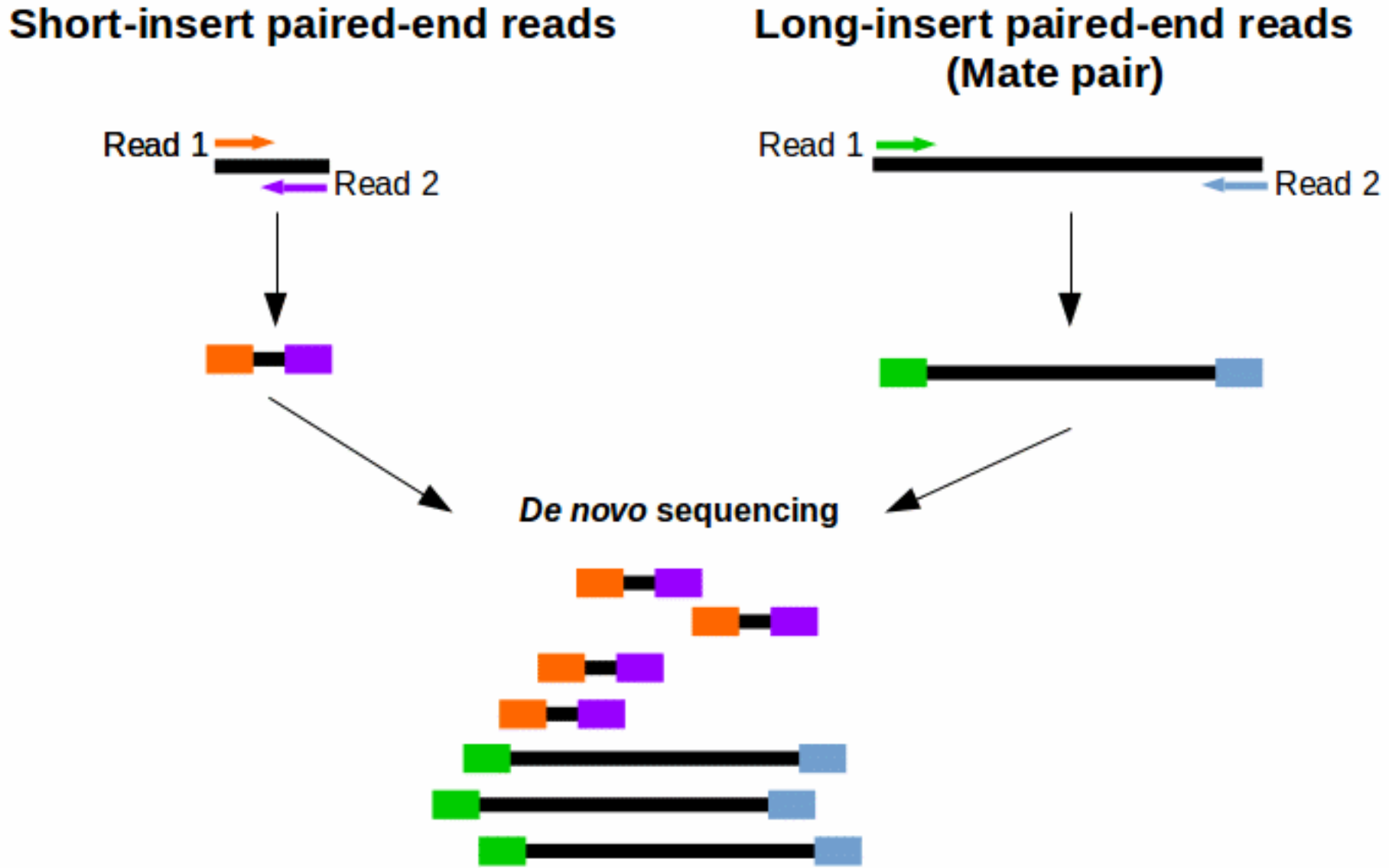
- De Novo Assembly
  - Create a new reference
  - Find structural variants
- Map to an existing reference
  - Alignment (BWA)
- Map against several references
  - Blast

De Novo Assembly



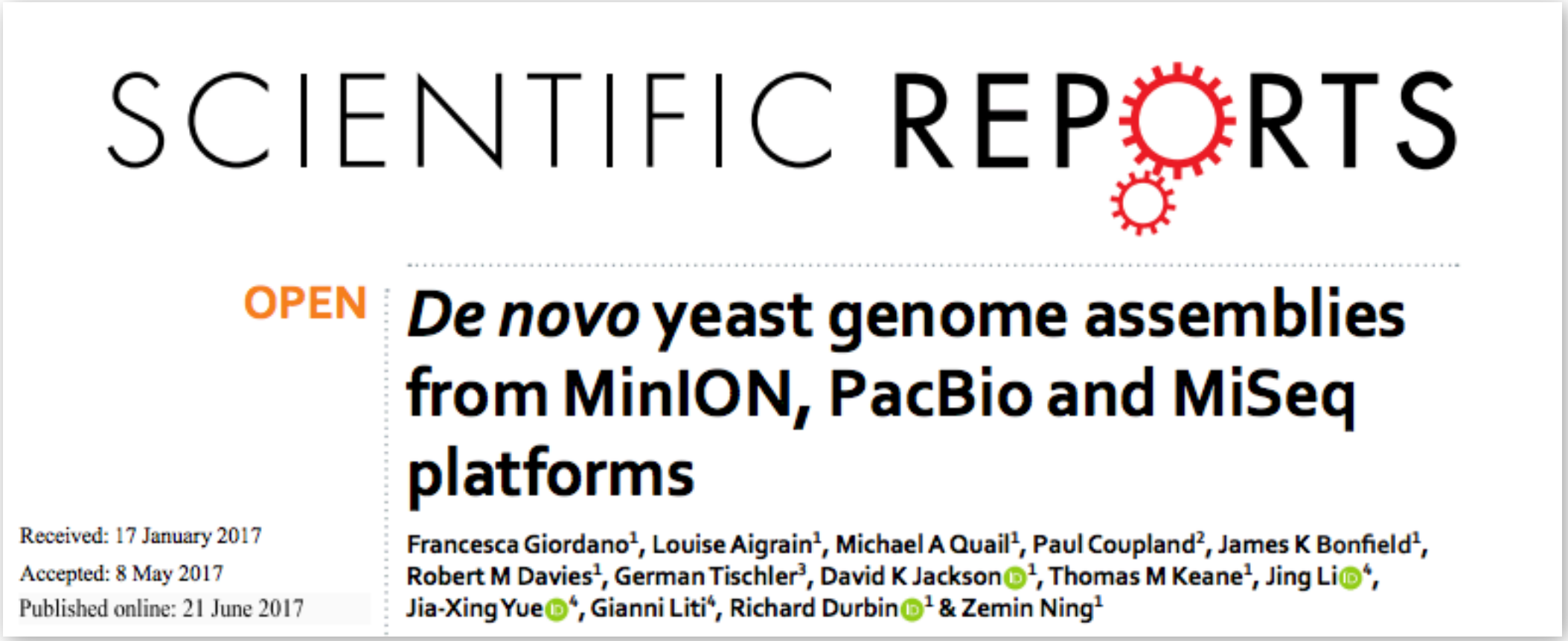
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAA

https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_data_analysis.php



https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for

De Novo Assembly



SCIENTIFIC REPORTS

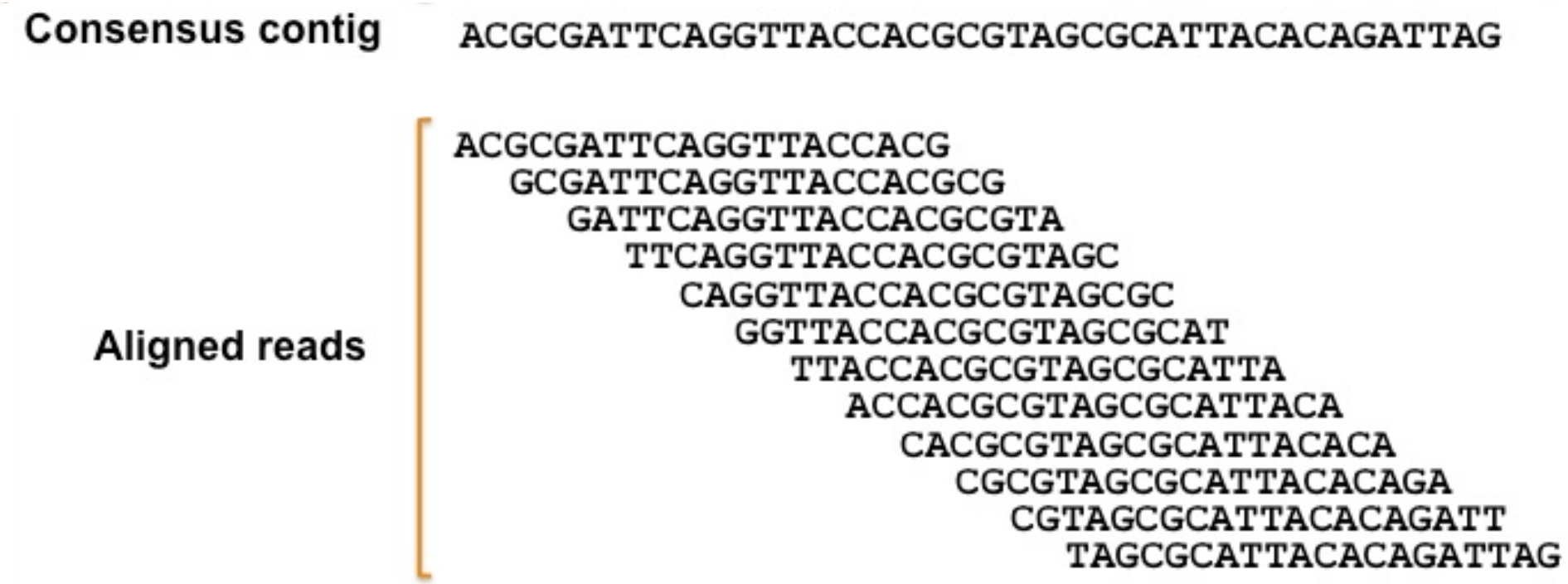
OPEN

De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms

Received: 17 January 2017
Accepted: 8 May 2017
Published online: 21 June 2017

Francesca Giordano[1], Louise Aigrain[1], Michael A Quail[1], Paul Coupland[2], James K Bonfield[1], Robert M Davies[1], German Tischler[3], David K Jackson[1], Thomas M Keane[1], Jing Li[4], Jia-Xing Yue[4], Gianni Liti[4], Richard Durbin[1] & Zemin Ning[1]

Alignment

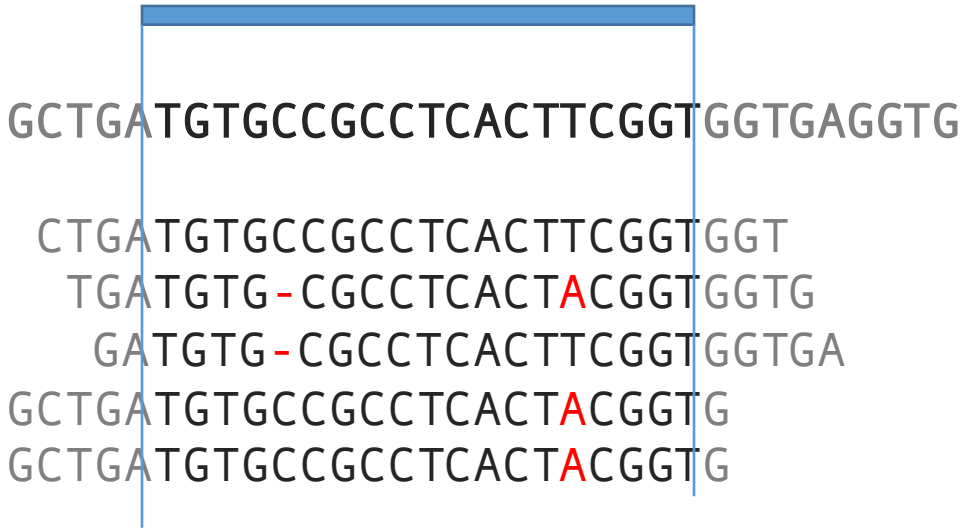

Consensus contig
ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
Aligned reads
ACGCGATTCAGGTTACCACG
GCGATTCAGGTTACCACGCG
GATTCAGGTTACCACGCGTA
TTCAGGTTACCACGCGTAGC
CAGGTTACCACGCGTAGCGC
GGTTACCACGCGTAGCGCAT
TTACCACGCGTAGCGCATTA
ACCACGCGTAGCGCATTACA
CACGCGTAGCGCATTACACA
CGCGTAGCGCATTACACAGA
CGTAGCGCATTACACAGATT
TAGCGCATTACACAGATTAG

Alignment

```
GCTGATGTGCCGCCTCACTTCGGTGGTGAGGTG   Reference sequence

 CTGATGTGCCGCCTCACTTCGGTGGT         Short read 1
  TGATGTG-CGCCTCACTACGGTGGTG        Short read 2
   GATGTG-CGCCTCACTTCGGTGGTGA       Short read 3
GCTGATGTGCCGCCTCACTACGGTG           Short read 4
GCTGATGTGCCGCCTCACTACGGTG           Short read 5
```

Alignment



BED file

```
Chr7   127471196   127472363   Pos1   0   +
```

Reference sequence

Short read 1

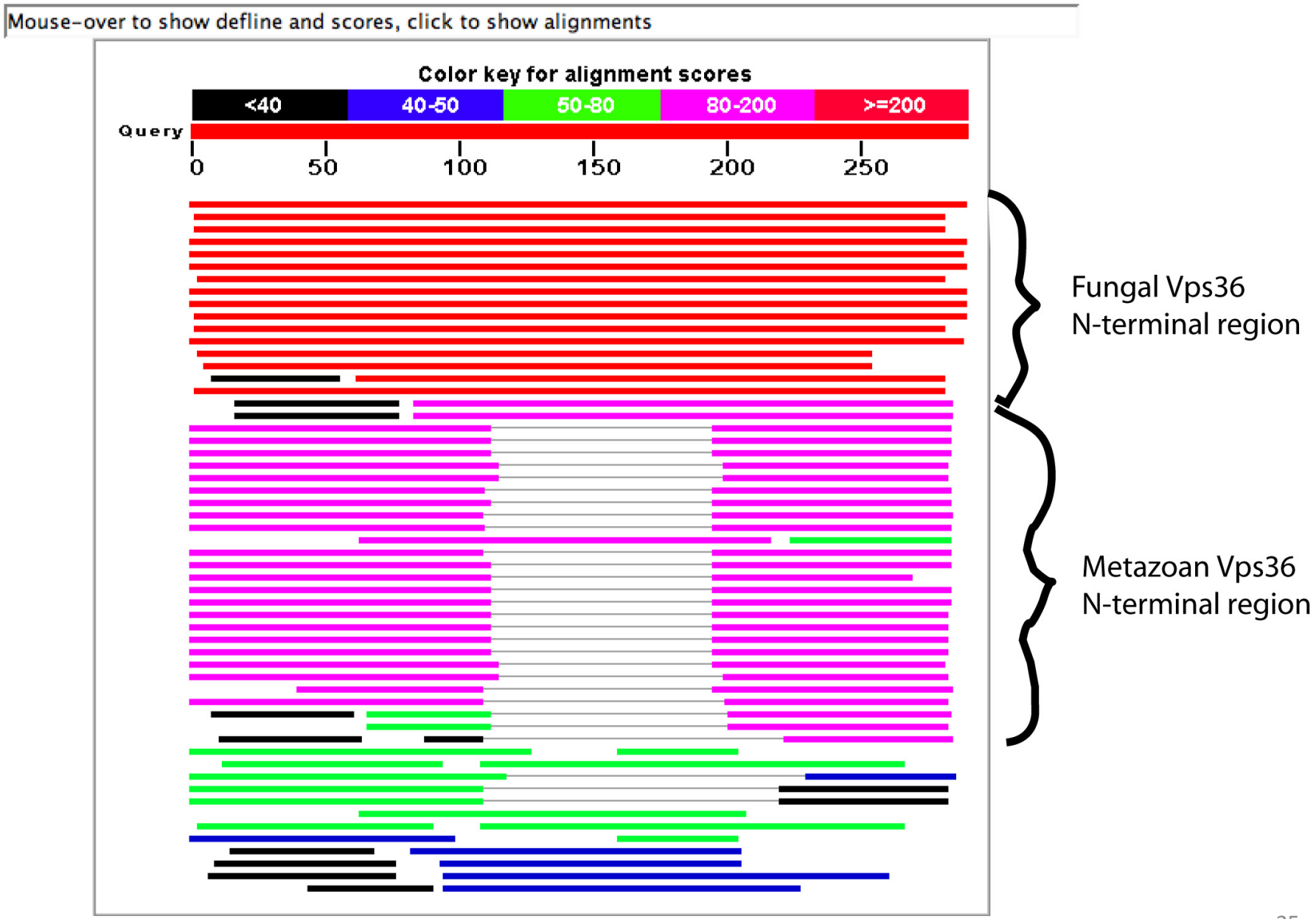
Short read 2

Short read 3

Short read 4

Short read 5

Blast



http://www.mrc-lmb.cam.ac.uk/rlw/text/bioinfo_tuto/sequence.html

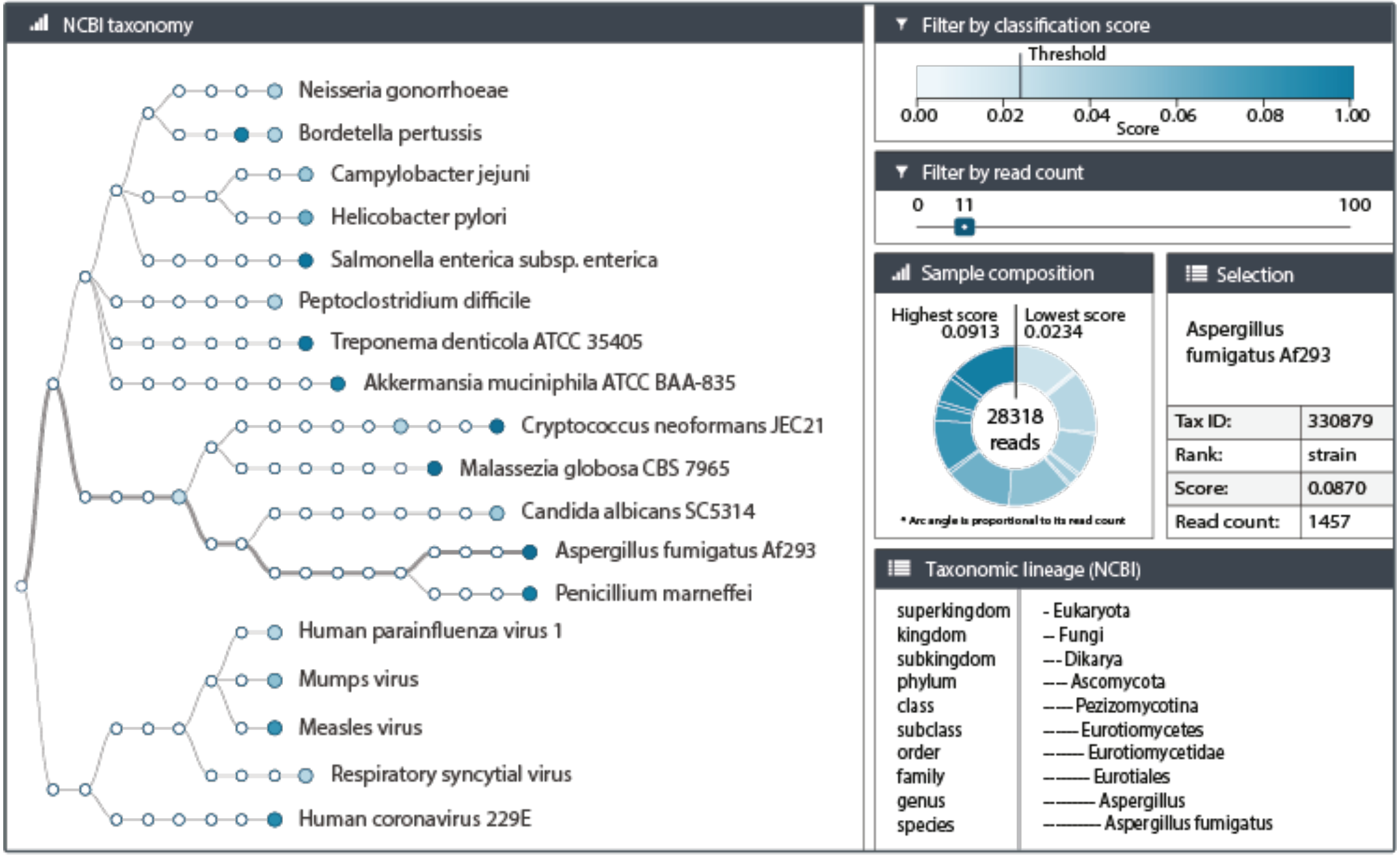


Fig. 1 Metrichor WIMP report, shown for a sample containing bacteria, viruses and fungi

No alignment?

Nucleic Acids Res. 2017 Jan 9; 45(1): 39–53.
Published online 2016 Nov 28. doi: 10.1093/nar/gkw1002
PMCID: PMC5224470

Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences

Nathan A. Ahlgren,[1,*†] Jie Ren,[2,†] Yang Young Lu,[2] Jed A. Fuhrman,[1] and Fengzhu Sun[1,2,3]

Author information ▶ Article notes ▶ Copyright and License information ▶

Alignment to human genome

- GRCh37(NCBI) vs hg19(UCSC) released 2009

Fasta format:

Unique sequence name

```
>seq1
ACGTCGTG
>seq2 additional info
TCGCAGCG
```

Alignment to human genome

- GRCh37(NCBI) vs hg19(UCSC) released 2009

```
173390@BioDA-server /m/n/s/0/r/G/seq> cat GRCh37.fa | grep ">"
>1 dna:chromosome chromosome:GRCh37:1:1:249250621:1
>2 dna:chromosome chromosome:GRCh37:2:1:243199373:1
>3 dna:chromosome chromosome:GRCh37:3:1:198022430:1
>4 dna:chromosome chromosome:GRCh37:4:1:191154276:1
>5 dna:chromosome chromosome:GRCh37:5:1:180915260:1
```

…

```
>22 dna:chromosome chromosome:GRCh37:22:1:51304566:1
>X dna:chromosome chromosome:GRCh37:X:1:155270560:1
>Y dna:chromosome chromosome:GRCh37:Y:2649521:59034049:1
>MT gi|251831106|ref|NC_012920.1| Homo sapiens mitochondrion, complete genome
>GL000207.1 dna:supercontig supercontig::GL000207.1:1:4262:1
>GL000226.1 dna:supercontig supercontig::GL000226.1:1:15008:1
>GL000229.1 dna:supercontig supercontig::GL000229.1:1:19913:1
```

Alignment to human genome

- GRCh37(NCBI) vs hg19(UCSC) released 2009

```
173390@BioDA-server /m/n/s/0/r/G/seq>
>1 dna:chromosome chromosome:GRCh37:1
>2 dna:chromosome chromosome:GRCh37:2
>3 dna:chromosome chromosome:GRCh37:3
>4 dna:chromosome chromosome:GRCh37:4
>5 dna:chromosome chromosome:GRCh37:5
```

```
173390@BioDA-server /m/n/s/0/r/h/seq> grep ">" hg19.fa
>chrM
>chr1
>chr2
>chr3
>chr4
```

…

…

```
>22 dna:chromosome chromosome:GRCh37:
>X dna:chromosome chromosome:GRCh37:X
>Y dna:chromosome chromosome:GRCh37:Y
>MT gi|251831106|ref|NC_012920.1| Hom
>GL000207.1 dna:supercontig supercont
>GL000226.1 dna:supercontig supercont
>GL000229.1 dna:supercontig supercont
```

```
>chr21
>chr22
>chrX
>chrY
>chr1_gl000191_random
>chr1_gl000192_random
>chr4_ctg9_hap1
>chr4_gl000193_random
```

Alignment to human genome

GRCh37(NCBI) vs hg19(UCSC) released Feb 2009

VS

GRCh38(NCBI) or hg38(UCSC) released Dec 2013

Alignment to human genome

GR

GR

Heng Li's blog Archive Categories Pages Tags

Which human reference genome to use?

13 November 2017

TL;DR: If you map reads to GRCh37 or hg19, use `hs37-1kg`:

```
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/human_g1k_v37.fasta.gz
```

If you map to GRCh37 and believe decoy sequences help with better variant calling, use `hs37d5`:

Alignment to genome

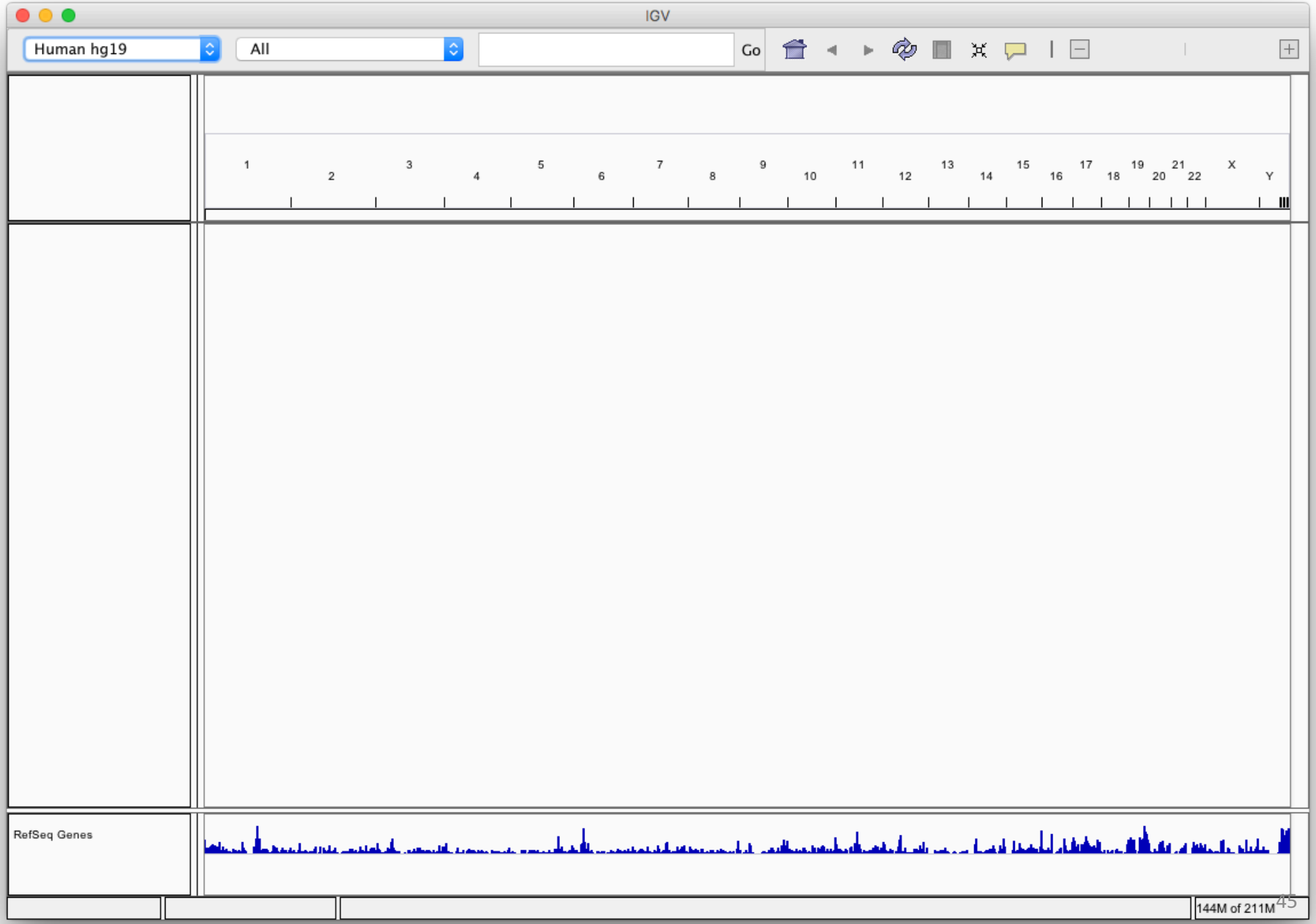
NCBI/UCSC applies also to the mouse genome

GRCm38/mm10

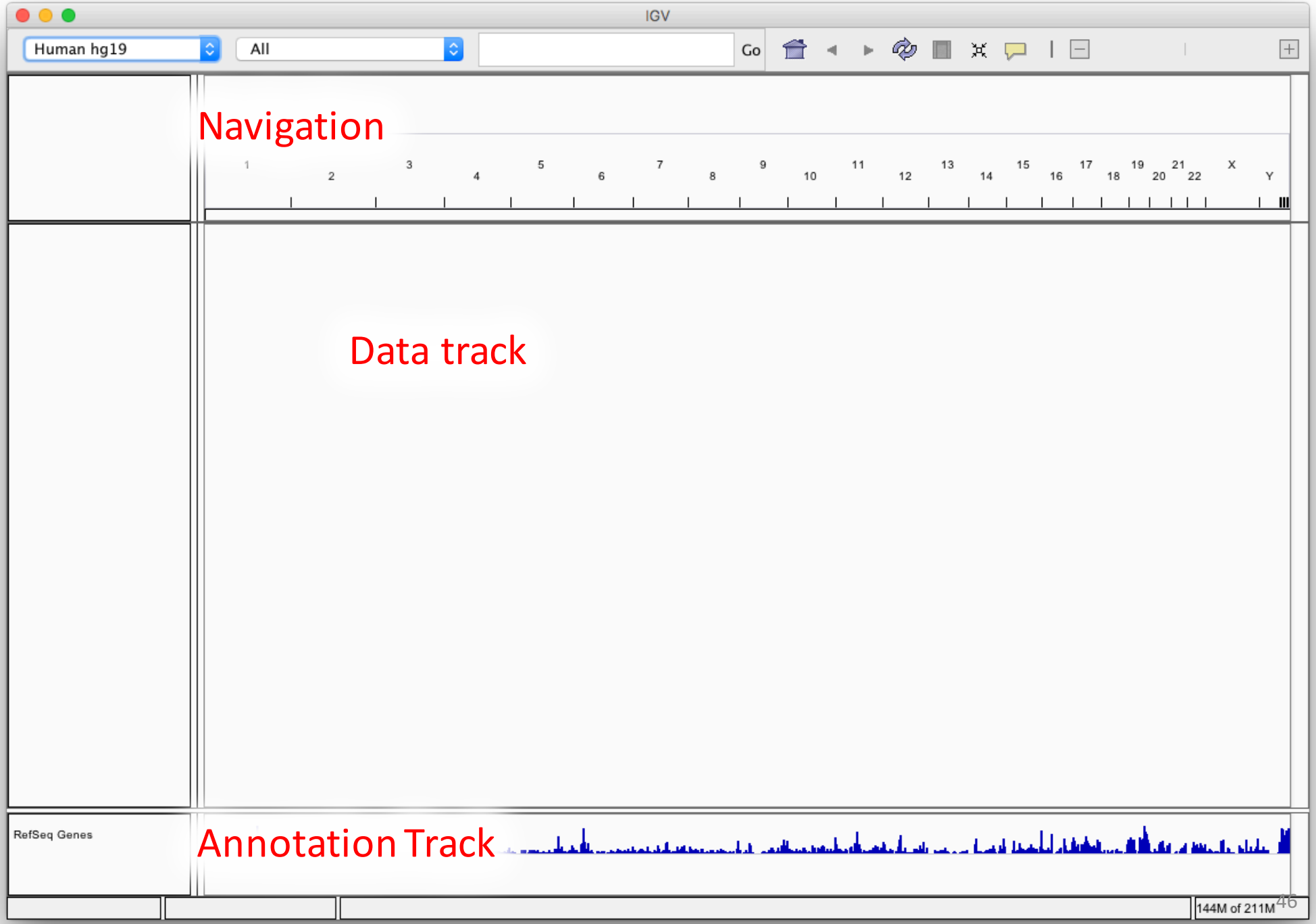
Alignment to genome

- Ungapped
  - BWA
  - Novoalign
- Gapped
  - Bowtie
  - Star

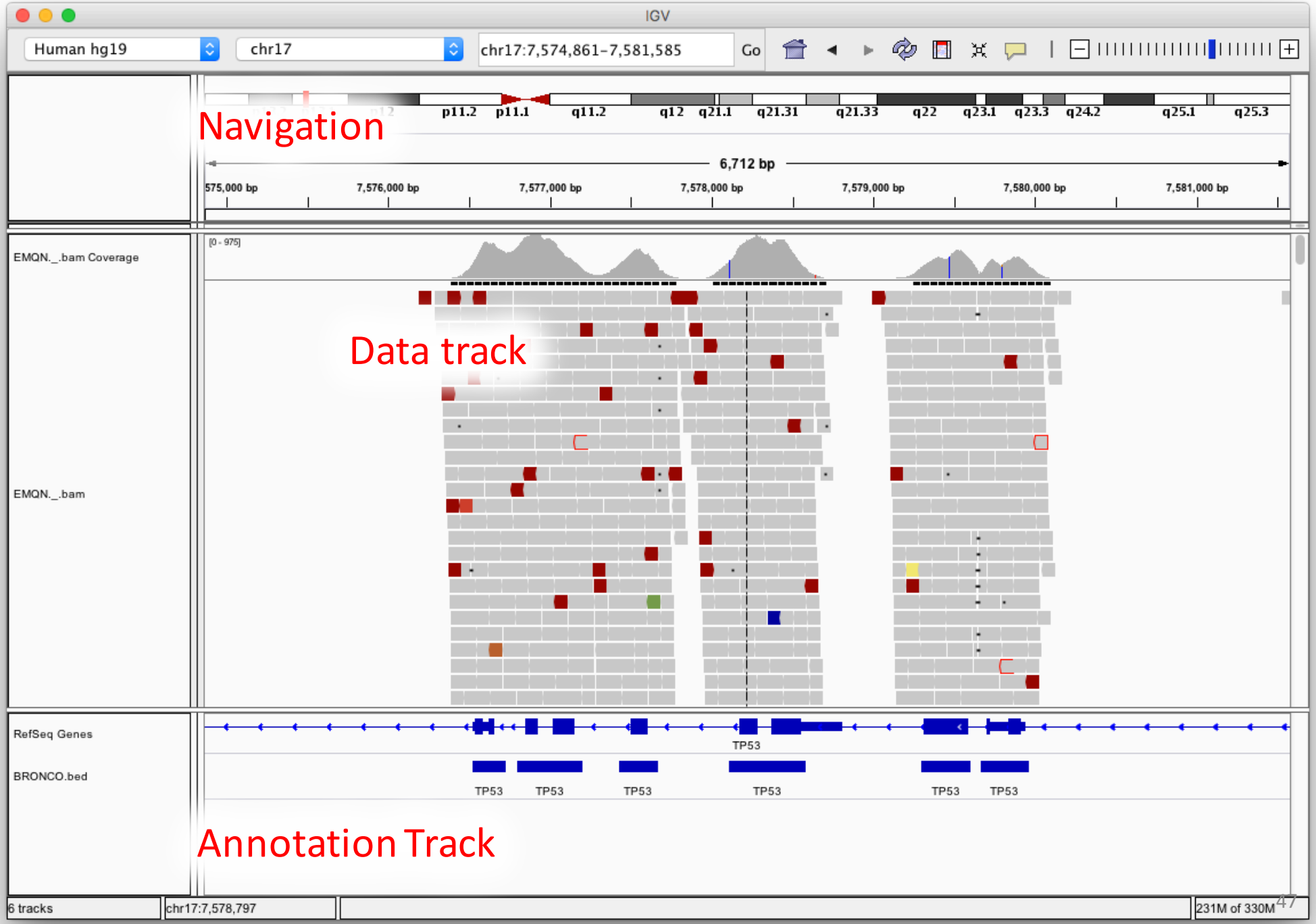
IGV


IGV
Human hg19
All
Go
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
RefSeq Genes
144M of 211M

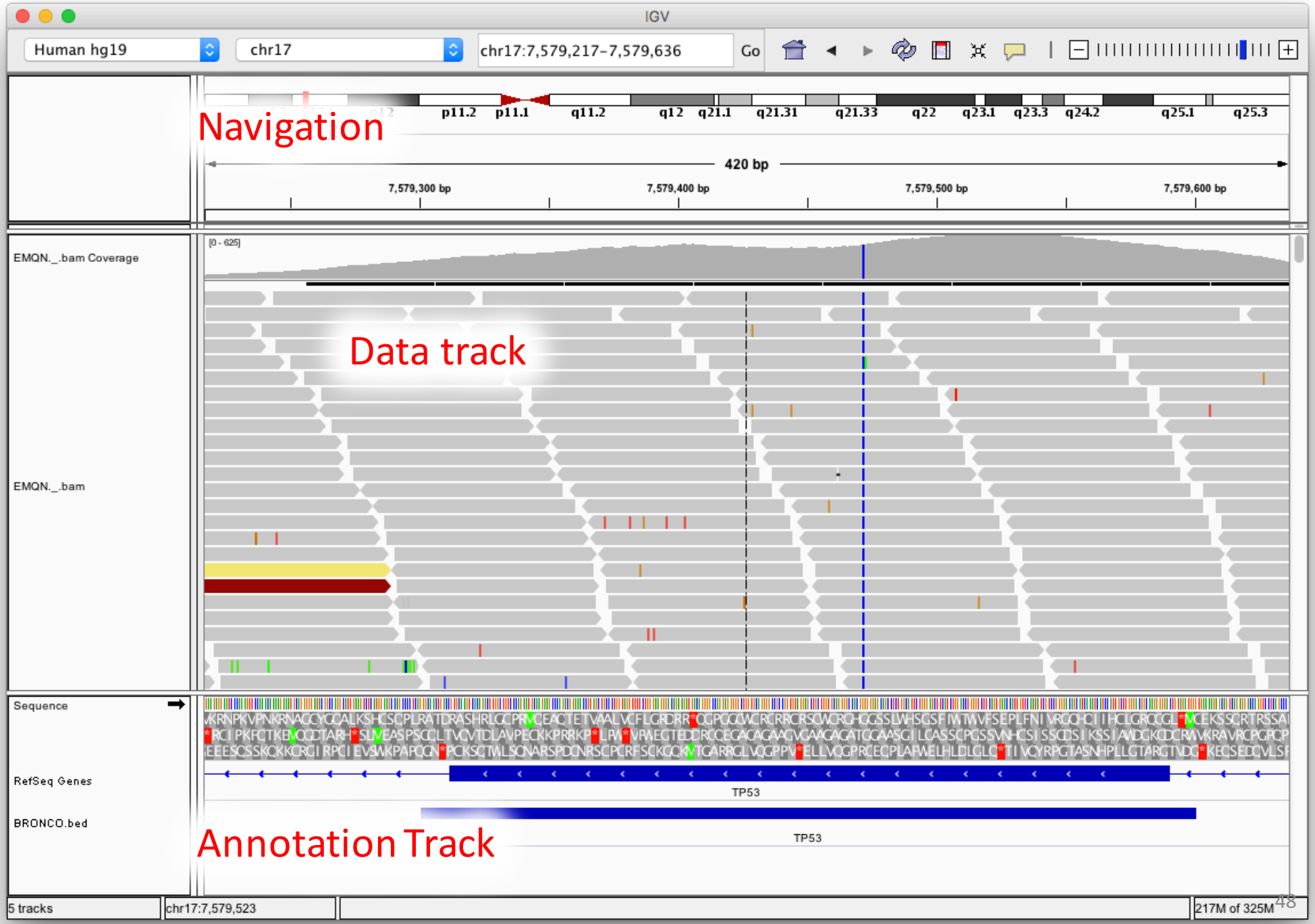
IGV


IGV
Human hg19
All
Go
Navigation
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Data track
RefSeq Genes
Annotation Track
144M of 211M

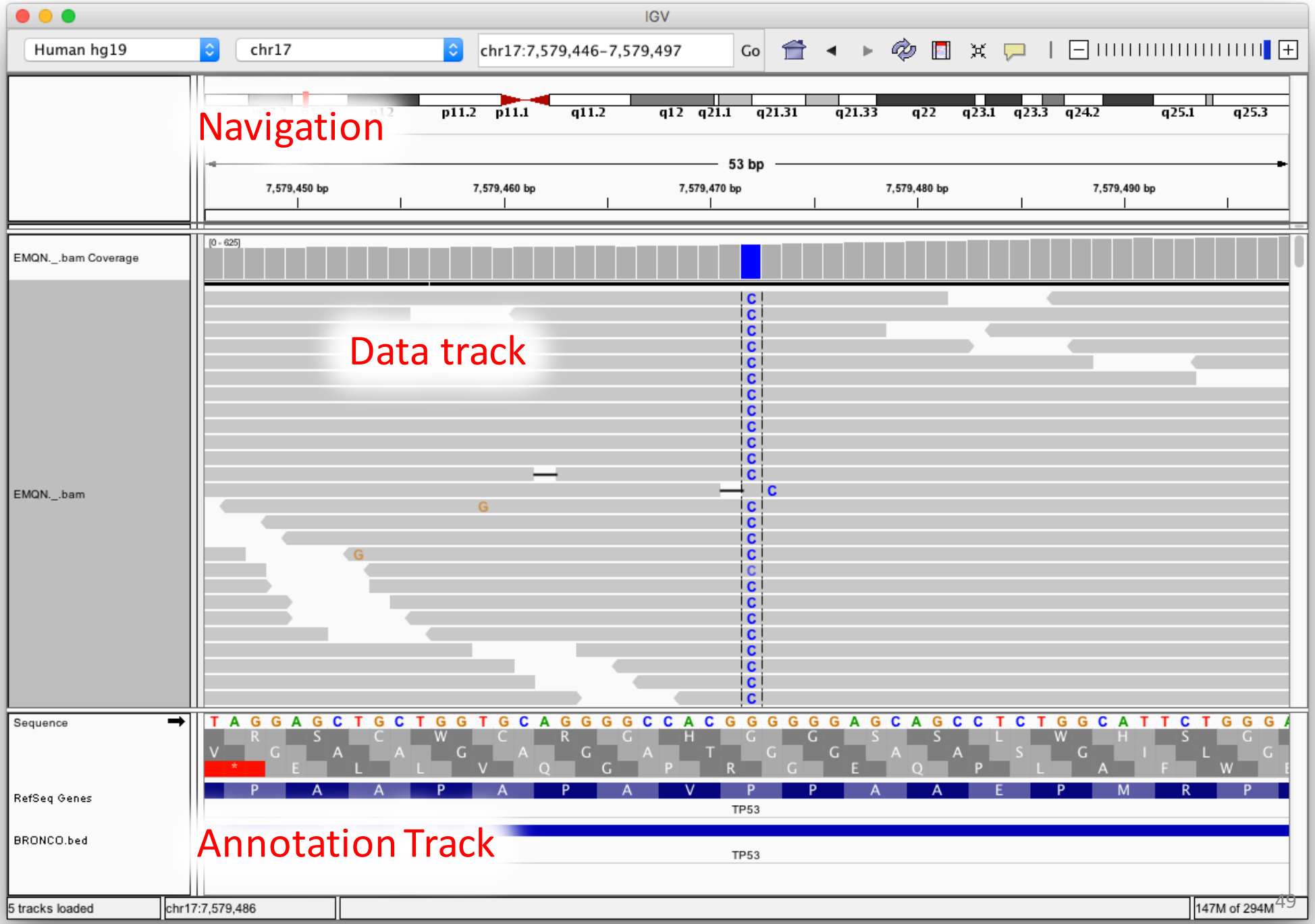
IGV


IGV
Human hg19
chr17
chr17:7,574,861-7,581,585
Go
Navigation
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.33
q22
q23.1
q23.3
q24.2
q25.1
q25.3
6,712 bp
575,000 bp
7,576,000 bp
7,577,000 bp
7,578,000 bp
7,579,000 bp
7,580,000 bp
7,581,000 bp
EMQN._.bam Coverage
[0 - 975]
Data track
EMQN._.bam
RefSeq Genes
TP53
BRONCO.bed
TP53
TP53
TP53
TP53
TP53
TP53
Annotation Track
6 tracks
chr17:7,578,797
231M of 330M

IGV


IGV
Human hg19
chr17
chr17:7,579,217-7,579,636
Go
Navigation
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.33
q22
q23.1
q23.3
q24.2
q25.1
q25.3
420 bp
7,579,300 bp
7,579,400 bp
7,579,500 bp
7,579,600 bp
EMQN._.bam Coverage
[0 - 625]
Data track
EMQN._.bam
Sequence
RefSeq Genes
TP53
BRONCO.bed
TP53
Annotation Track
5 tracks
chr17:7,579,523
217M of 325M

IGV


IGV
Human hg19
chr17
chr17:7,579,446-7,579,497
Go
Navigation
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.33
q22
q23.1
q23.3
q24.2
q25.1
q25.3
53 bp
7,579,450 bp
7,579,460 bp
7,579,470 bp
7,579,480 bp
7,579,490 bp
EMQN._.bam Coverage
[0 - 625]
Data track
EMQN._.bam
Sequence
RefSeq Genes
TP53
BRONCO.bed
Annotation Track
TP53
5 tracks loaded
chr17:7,579,486
147M of 294M

Alignment QC - Coverage statistics

- How many reads are aligned?
- How even is the overall coverage
- Average insert size
- How many reads come from the region of interest
  - On/Off target reads
  - Bed file – defines region of interest
- What is the average coverage
- How many % of target bases have at least X coverage

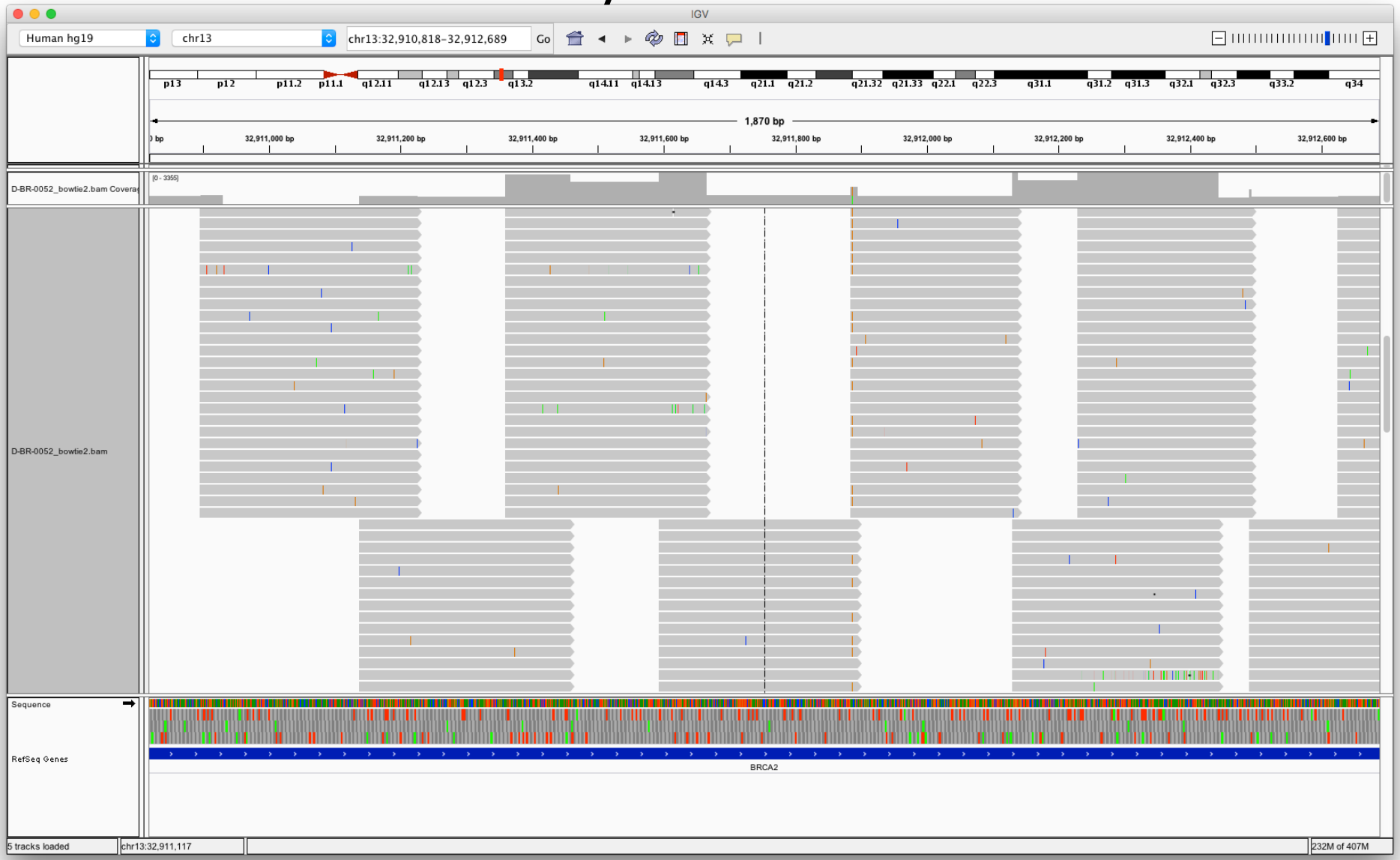
Alignment – Coverage statistics

- [Multiqc demo]

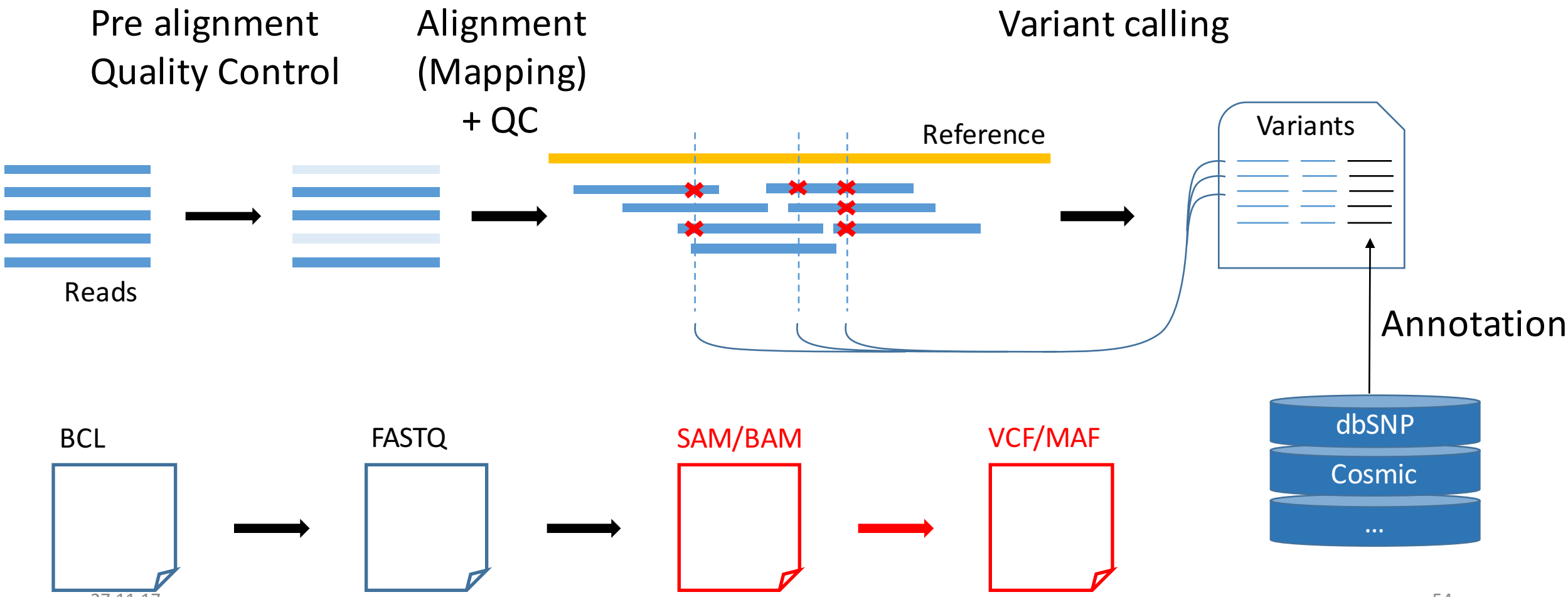
Alignment – Coverage statistics Picard-Tools

BAIT_SET		BRNO1077norm	BRNO1404norm	BRNO1503norm
TOTAL_READS		79842182	98157468	106336660
PF_READS		79842182	98157468	106336660
PF_UNIQUE_READS		69127998	87214990	95287834
PCT_PF_UQ_READS		0.865808	0.888521	0.896096
PF_UQ_READS_ALIGNED		68554192	86493216	94875098
PCT_PF_UQ_READS_ALIGNED		0.991699	0.991724	0.995669
ON_BAIT_BASES		4259279846	5132027582	5495198420
NEAR_BAIT_BASES		921266922	1284426047	1344776758
OFF_BAIT_BASES		1112205142	1324505871	1573330165
ON_TARGET_BASES		2765311894	3544263597	3773976192
PCT_SELECTED_BASES		0.823256	0.828896	0.812995
PCT_OFF_BAIT		0.176744	0.171104	0.187005
ON_BAIT_VS_SELECTED		0.822168	0.799823	0.803394
MEAN_BAIT_COVERAGE		93.968208	113.222763	121.235036
MEAN_TARGET_COVERAGE		61.008295	78.193523	83.261441
MEDIAN_TARGET_COVERAGE		52	68	72
PCT_TARGET_BASES_1X		0.971254	0.973571	0.974168
PCT_TARGET_BASES_2X		0.965912	0.969276	0.970293
PCT_TARGET_BASES_10X		0.928313	0.941296	0.945291
PCT_TARGET_BASES_20X		**0.858879**	**0.896376**	**0.90591**
PCT_TARGET_BASES_30X		**0.760404**	**0.835876**	**0.852135**
PCT_TARGET_BASES_40X		0.646449	0.760458	0.783717
PCT_TARGET_BASES_50X		0.530314	0.674041	0.702404

Alignment PCR library



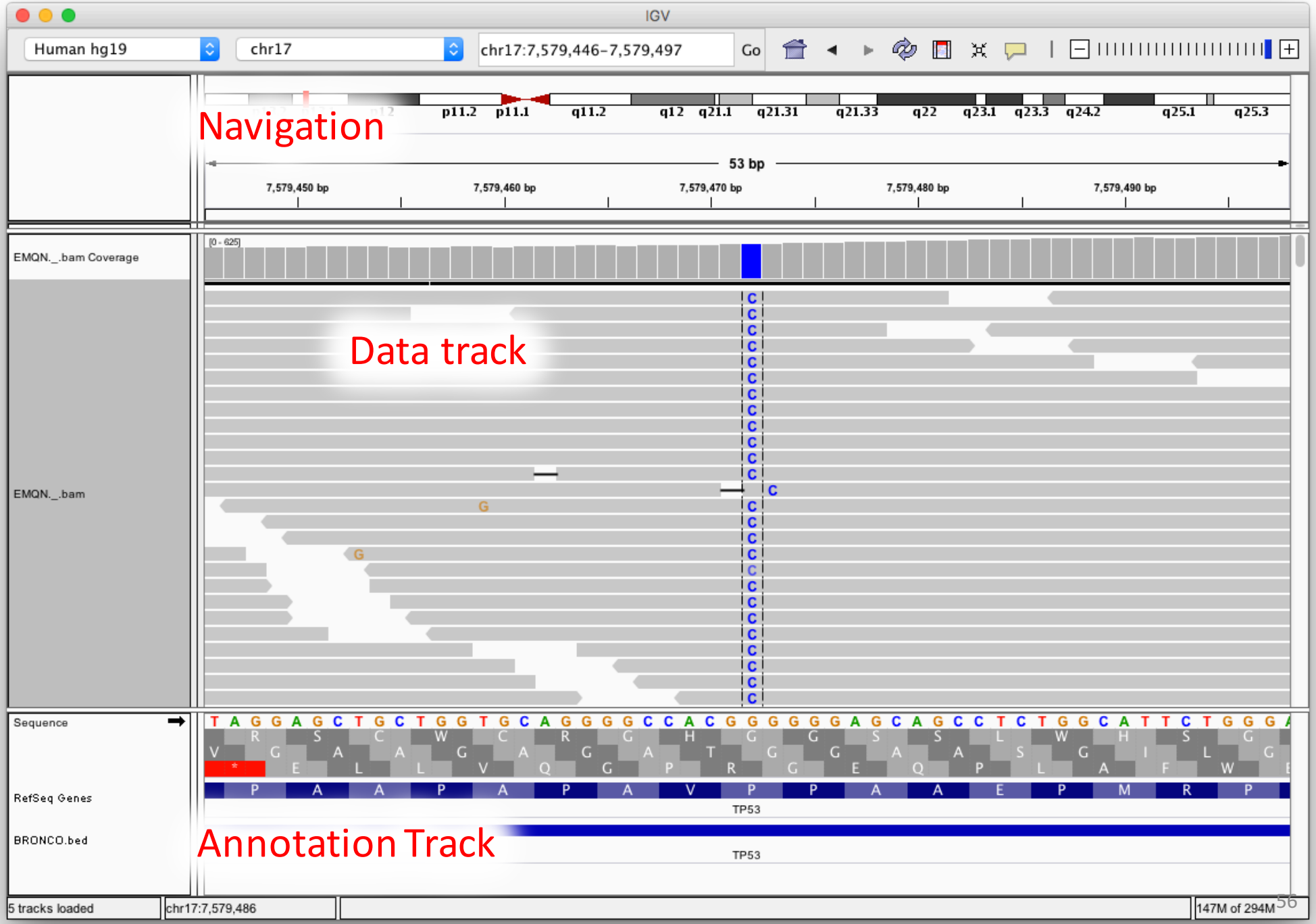
Recap – Data analysis pipeline



DNA Variant calling

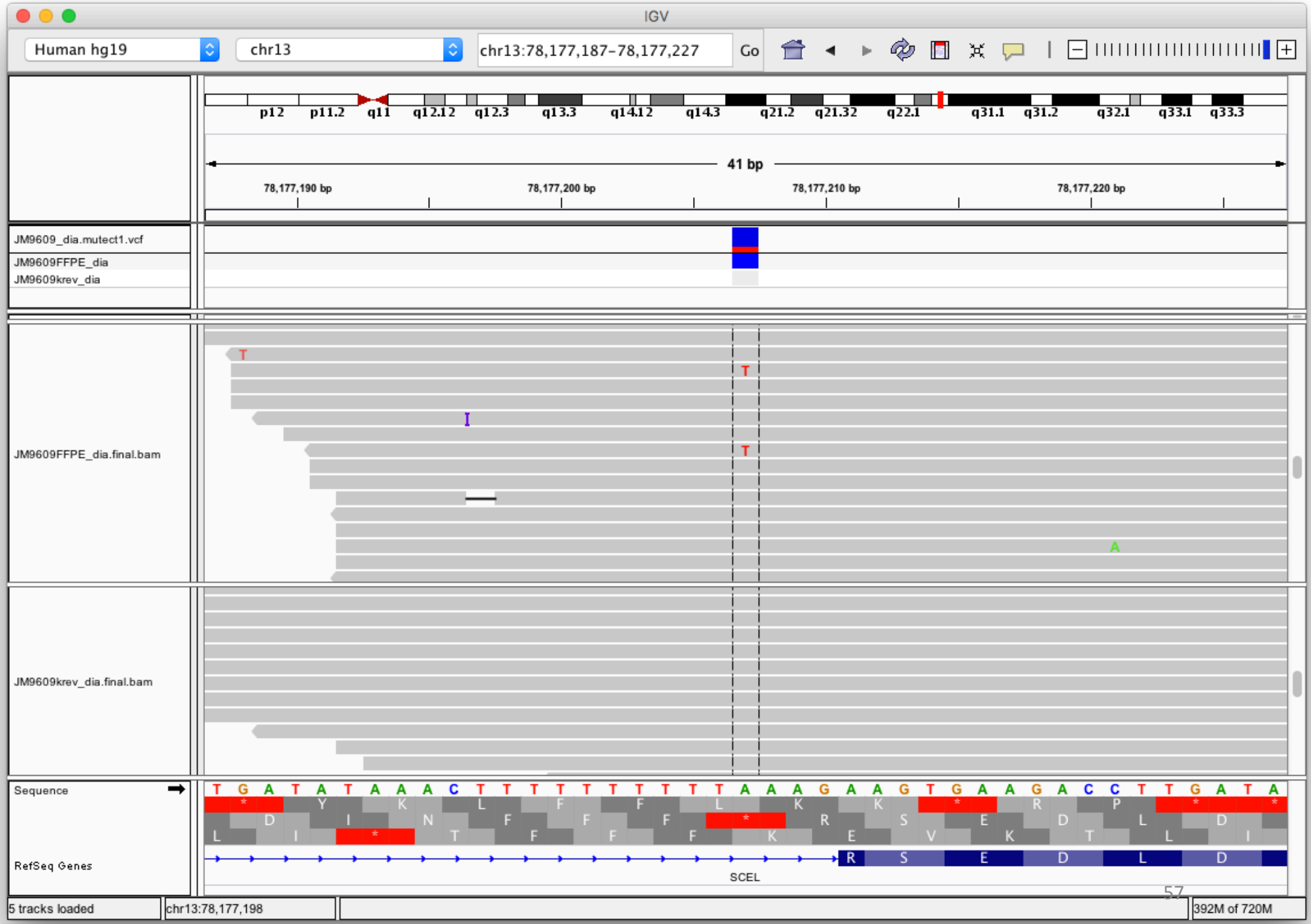
- Single Nucleotide Variants (SNV's) + short indels
  - Somatic/Germline
- Copy Number variants (CNV)
- Structural Variants

IGV

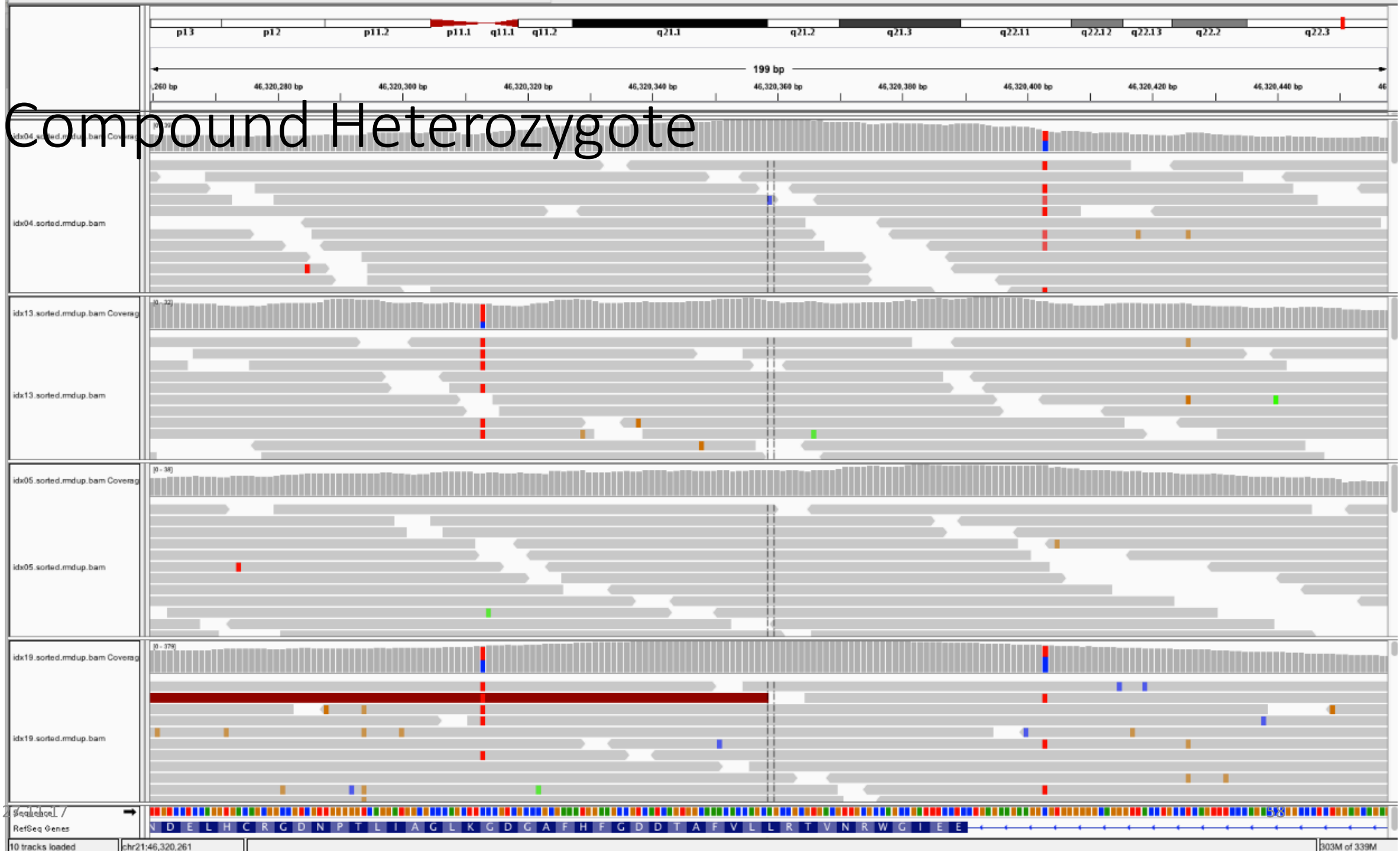

IGV
Human hg19
chr17
chr17:7,579,446-7,579,497
Go
Navigation
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.33
q22
q23.1
q23.3
q24.2
q25.1
q25.3
53 bp
7,579,450 bp
7,579,460 bp
7,579,470 bp
7,579,480 bp
7,579,490 bp
EMQN._.bam Coverage
[0 - 625]
Data track
EMQN._.bam
Sequence
RefSeq Genes
TP53
BRONCO.bed
Annotation Track
TP53
5 tracks loaded
chr17:7,579,486
147M of 294M

IGV inspect variants

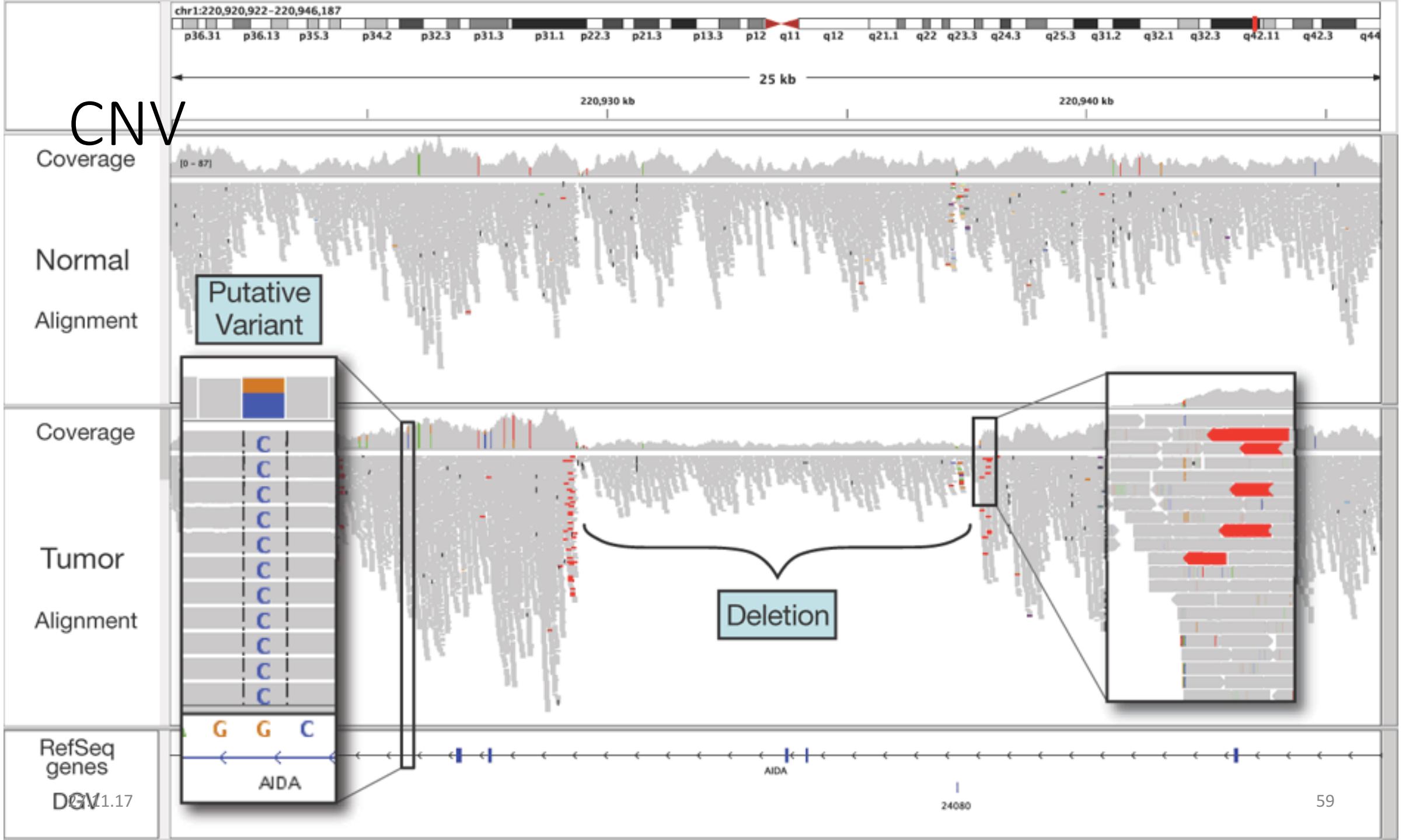
Somatic



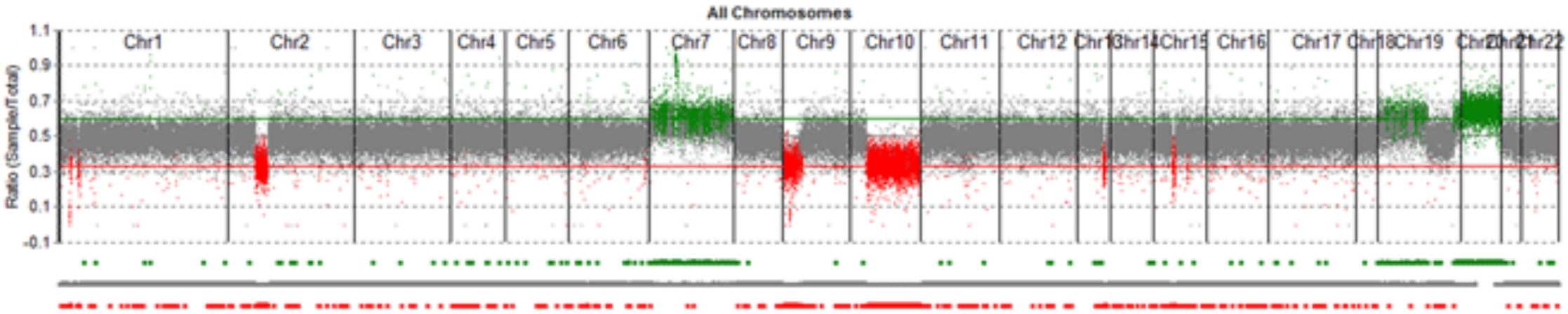
Compound Heterozygote



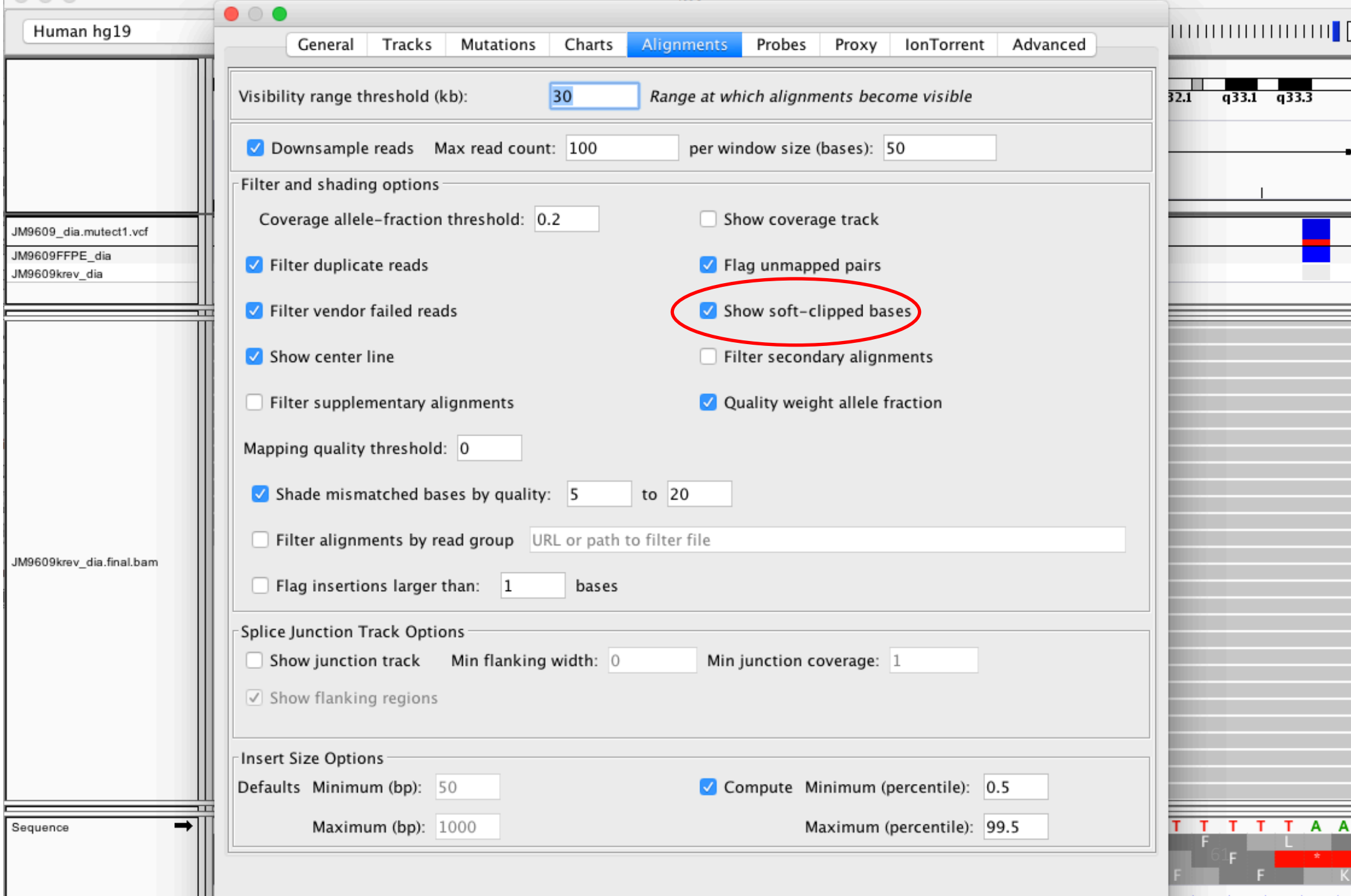
CNV


chr1:220,920,922-220,946,187
25 kb
220,930 kb
220,940 kb
Coverage
Normal
Alignment
Putative Variant
Coverage
Tumor
Alignment
Deletion
RefSeq genes
DGV
AIDA
24080

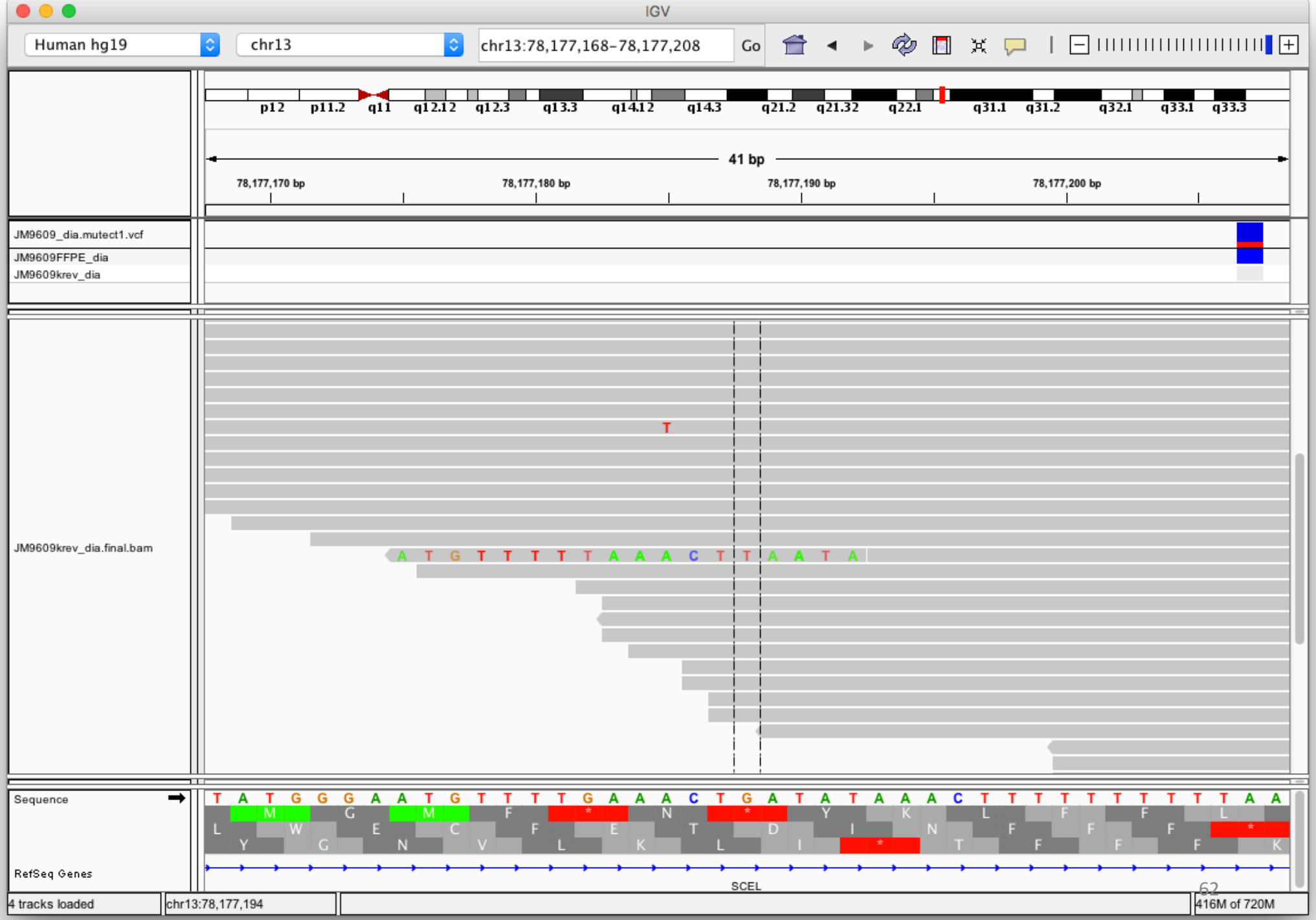
CNV

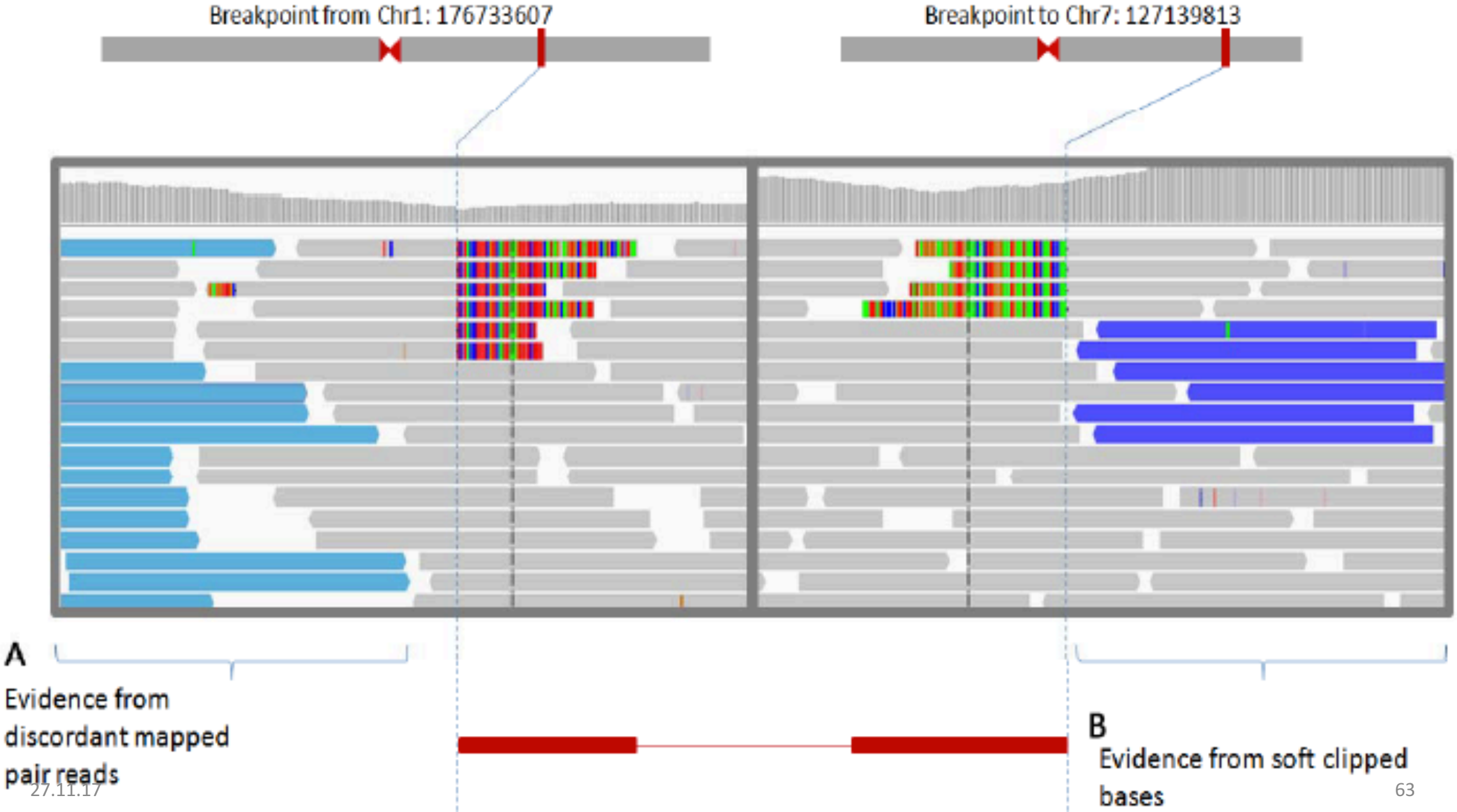

All Chromosomes
Ratio (Sample/Total)
1.1
0.9
0.7
0.5
0.3
0.1
-0.1
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr14
Chr15
Chr16
Chr17
Chr19
Chr22

IGV soft clips

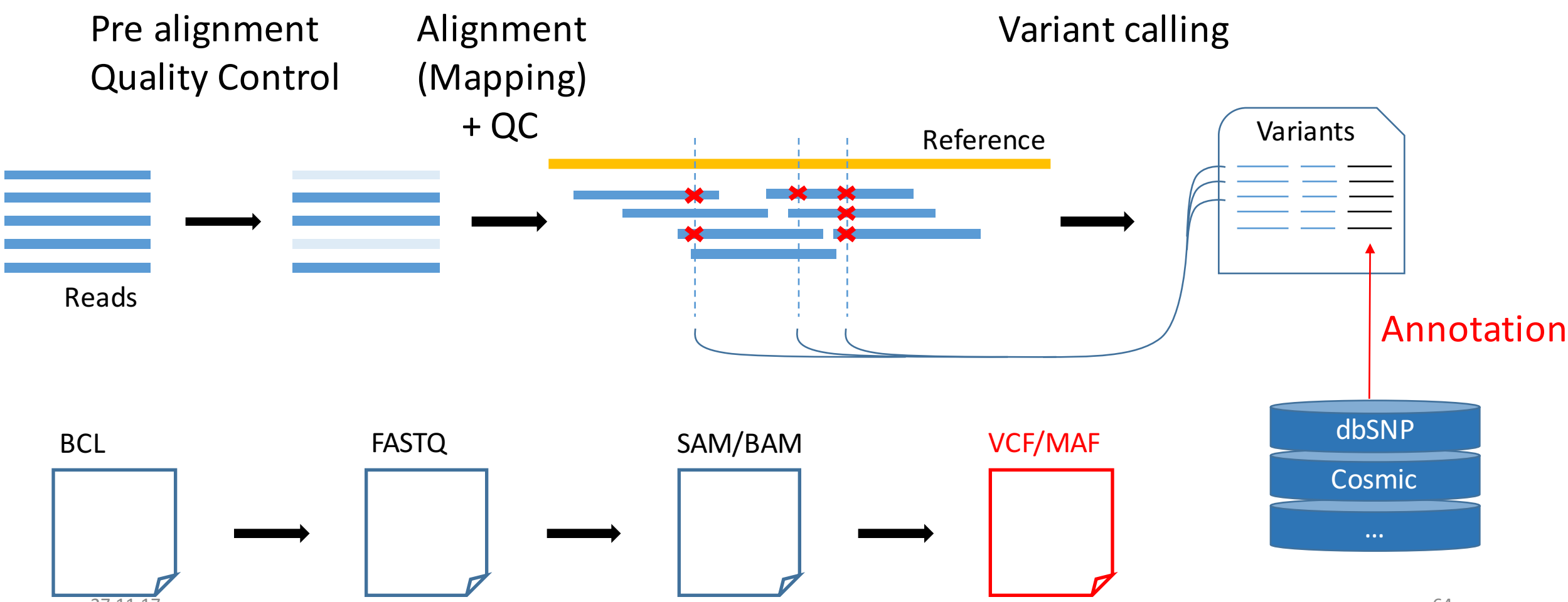


IGV soft clips



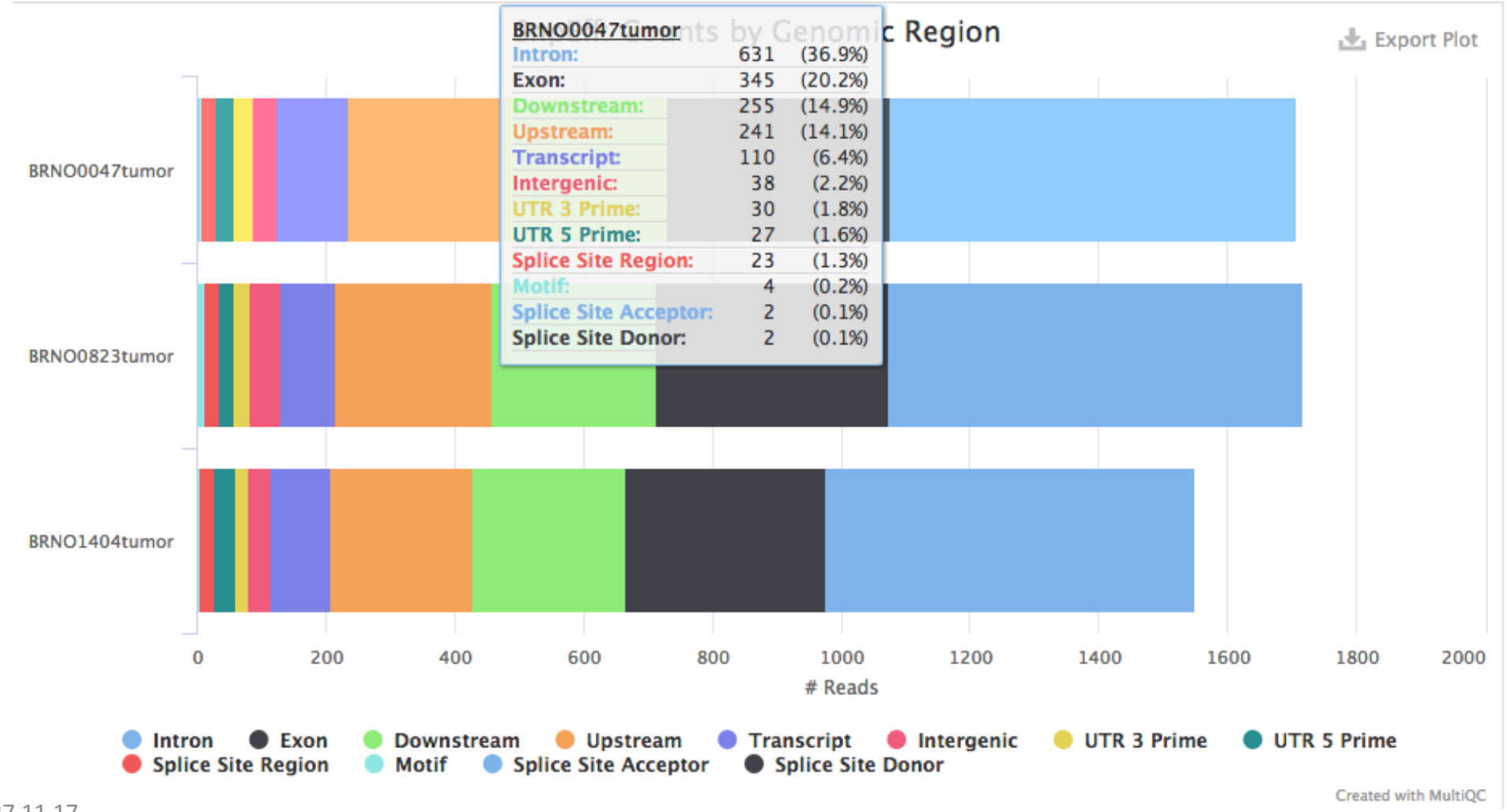

Breakpoint from Chr1: 176733607
Breakpoint to Chr7: 127139813
A
Evidence from discordant mapped pair reads
B
Evidence from soft clipped bases

Recap – Data analysis pipeline

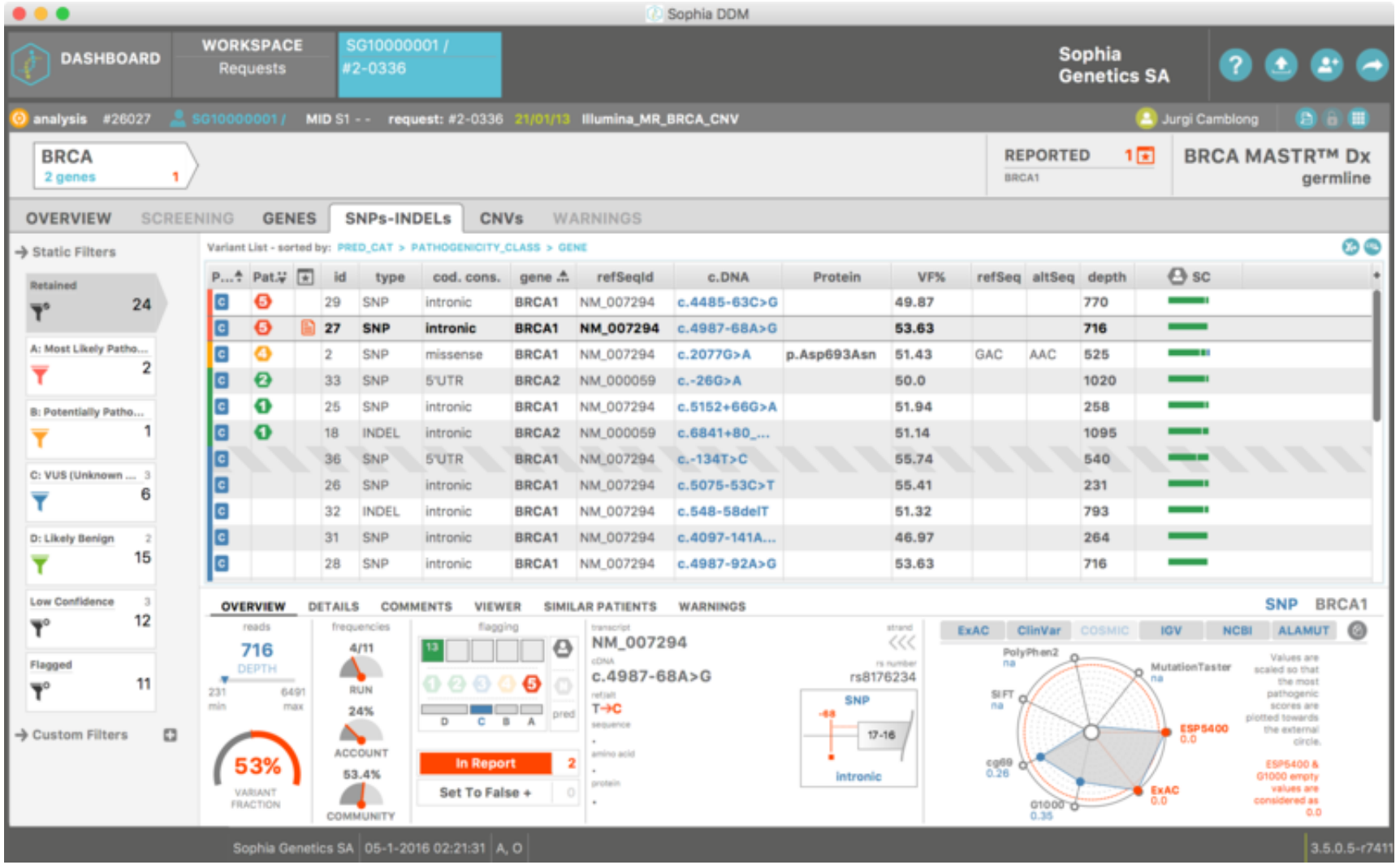


Variant annotation

MultiQC output

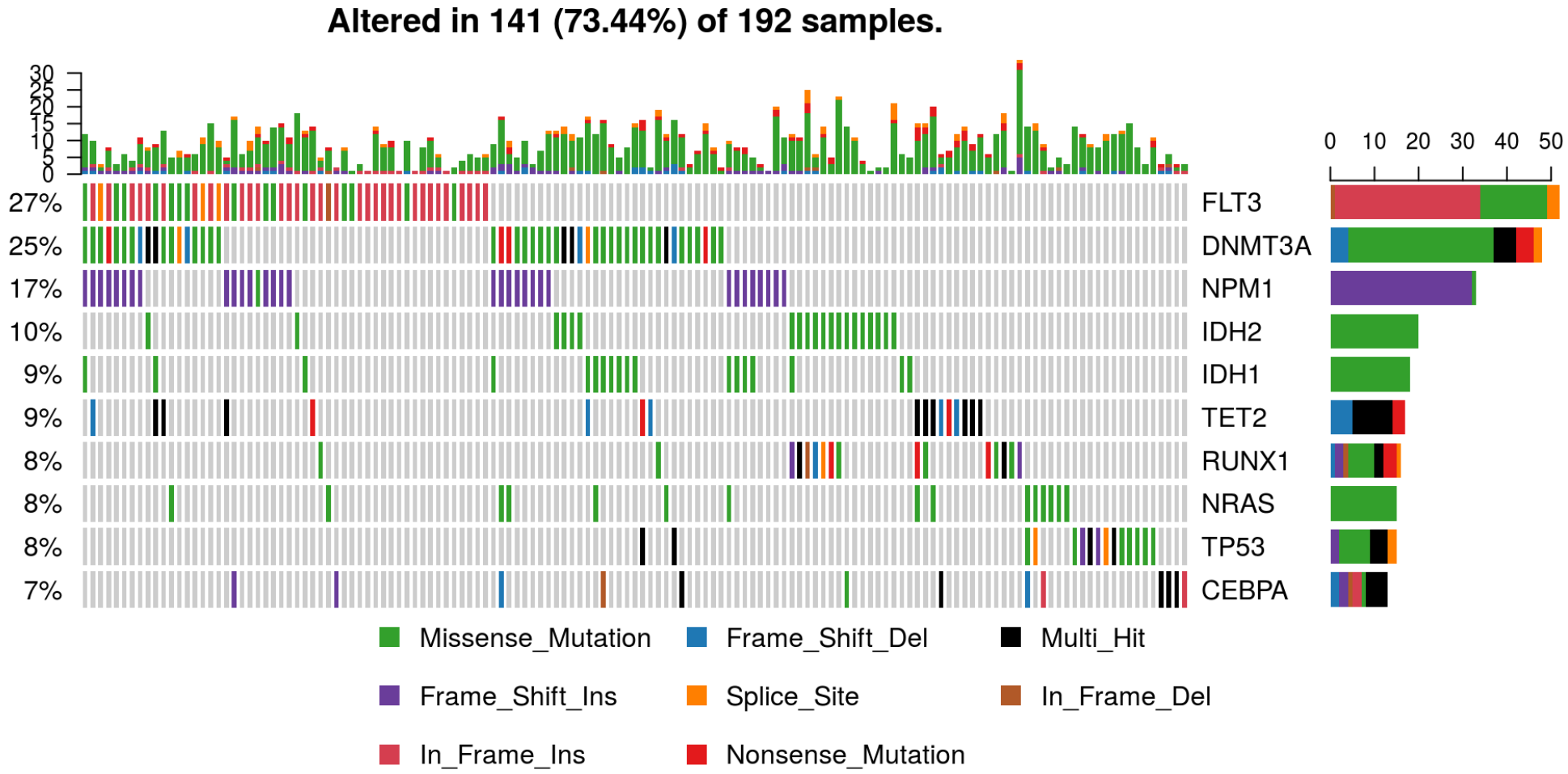


Variant annotation + report



https://www.kti.admin.ch/kti/en/home/ueber-uns/nsb-news/weitere-news/sophiagenetics.html

Variant annotation + report

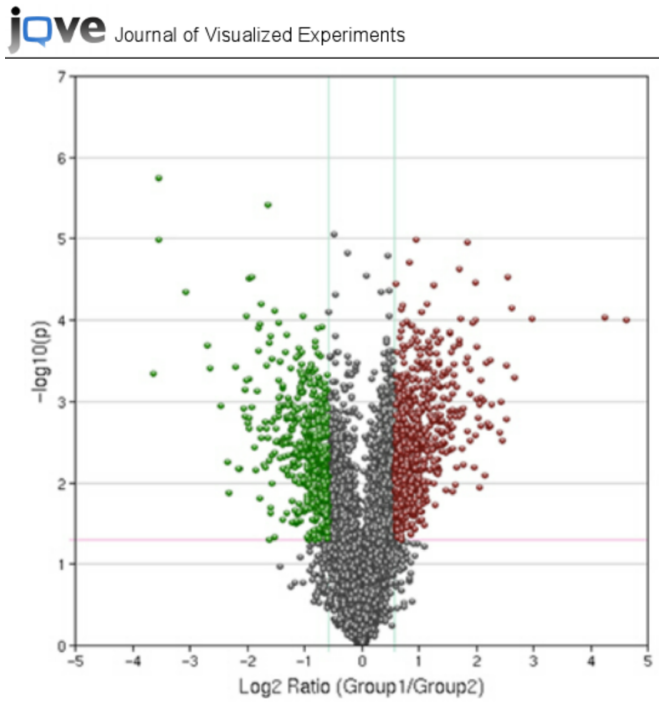


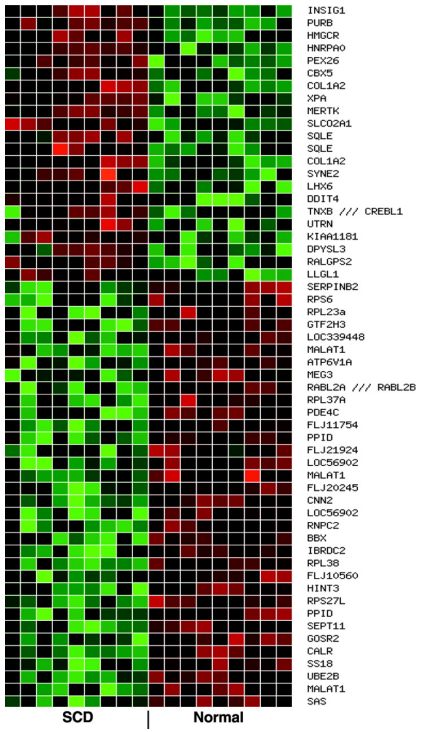
https://bioconductor.org/packages/devel/bioc/vignettes/maftools/inst/doc/maftools.html

*RNASeq

RNAseq pipeline

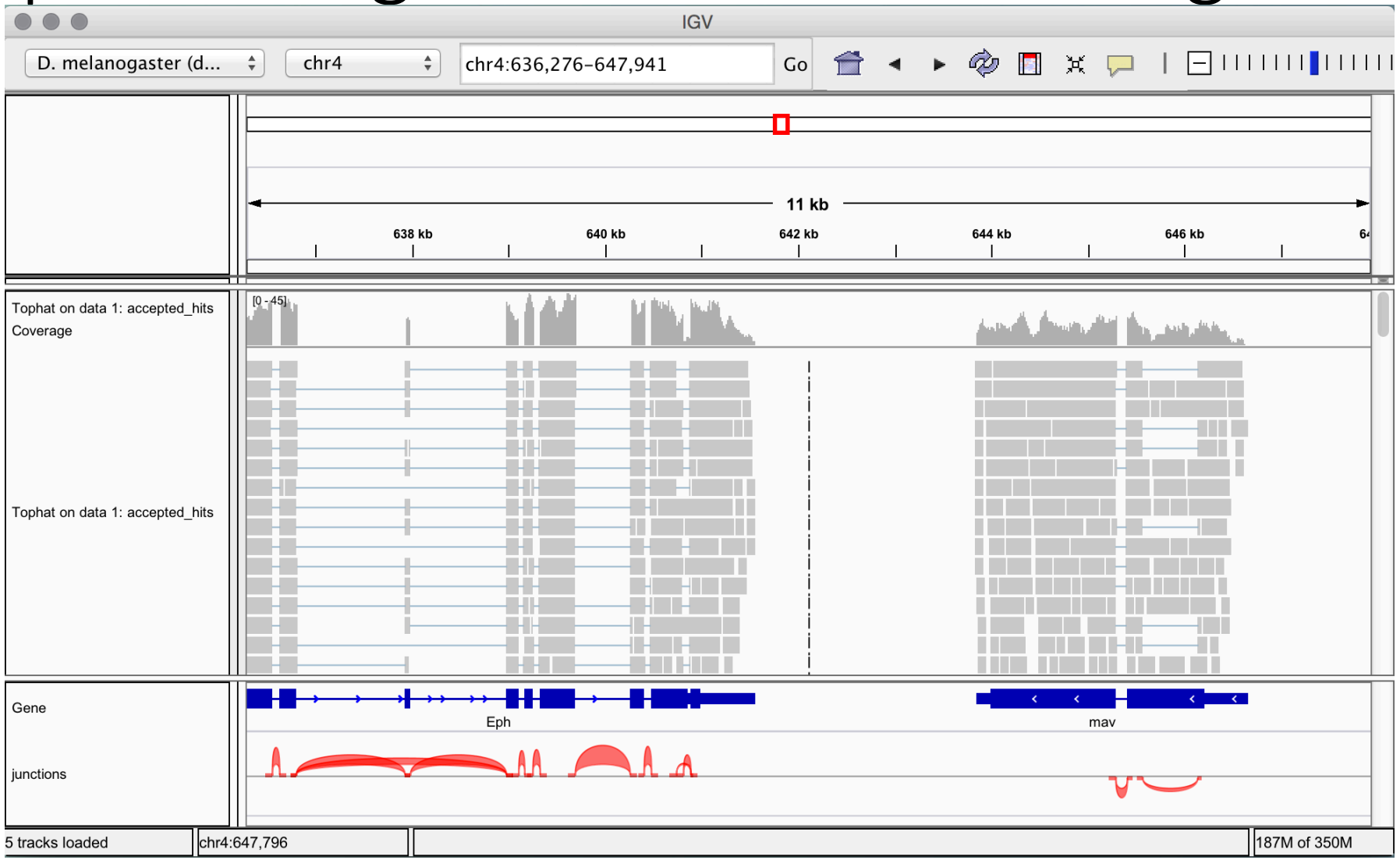
#	G1:MeOH	G1:MeOH	G1:MeOH	G2:R3G	G2:R3G	G2:R3G
#Feature	MeOH_Rep1	MeOH_Rep2	MeOH_Rep3	R3G_Rep1	R3G_Rep2	R3G_Rep3
LOC100288778	38	48	47	51	46	47
IQSEC3	0	0	0	0	0	0
CCDC77	51	51	51	40	40	39
B4GALNT3	4	4	3	6	6	11
WNK1	264	293	268	281	256	272
ERC1	55	55	68	83	57	49
LOC100292680	0	0	0	2	1	0
WNT5B	3	1	0	1	0	1
ADIPOR2	96	83	109	79	65	81
LRTM2	0	0	0	1	0	0
CACNA1C	5	1	2	7	3	4
CACNA1C-IT3	0	0	0	0	0	0
FKBP4	466	472	466	257	229	257
ITFG2	51	63	64	46	41	44
LOC100507424	5	1	2	0	1	4
RHNO1	73	82	74	61	58	66
TULP3	32	19	32	18	19	27
TEAD4	1	0	0	0	1	0
TSPAN9	0	0	1	1	1	0
PRMT8	1	0	1	0	0	0
CCND2	4440	4496	4694	2743	2739	2726





Feature counts (+normalization) → Volcano Plot → Heat map

RNASeq reads aligned to the reference genome



Annotation

https://www.melbournebioinformatics.org.au/tutorials/tutorials/rna_seq_dge_basic/rna_seq_basic_tutorial/

Taka away

- Terminology
- Interpreting Different QC metrics
- Interpreting NGS data visually
- Basic intuition (reads, alignments, references, variants)

Thank you for your attention