Introduction to Sequencing Data Analysis

Lecture 7

October 17, 2019

Gavin Ha, Ph.D.
Assistant Member
Public Health Sciences

FRED HUTCH
CURES START HERE®

Overview

I. Sequence data

II. Tools for analyzing and visualizing sequencing data

III. Genome variant analysis

Overview: Learning Objectives

1. Sequence data
   - Databases and online resources for sequence data
   - Learn the common sequence data file formats
2. Tools for sequencing data
   - Tools to query, inspect, visualize an aligned sequence file
   - Learn the contents of sequence data files
   - Learn to generate sequencing metrics and to process sequence data
   - Learn about Python and R libraries/packages to read sequence data
3. Genome variant analysis
   - Types of genomic variation
   - Tools to predict genomic variations
   - Learn the common file formats for variation data
   - Databases and online resources for human variation data

Sequence Data: International Consortia and Projects

1000 Genomes Project (https://www.internationalgenome.org/)

UK10K (https://www.uk10k.org/)

The 100,000 Genomes Project (https://www.genomicsengland.co.uk/)

- Rare disease, cancer, infectious disease

Genome 10K Project (https://genome10k.soe.ucsc.edu/)

- Genomic “zoo” of 16,000 vertebrate species

Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org/)

- Lek et al. Nature, 536, 285-91 (2016)

Genome Aggregation Database (gnomAD) (https://gnomad.broadinstitute.org/)

- Karczewski et al. bioRxiv (2019)

IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

UK10K

Rare Genetic Variants in Health and Disease

Genomics england

#100kThankYous

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

GENOME 10K

Sequence Data: Databases and Online Resources

Repositories/Databases for sequence data

1. NCBI Sequence Read Archive (SRA)

- Publicly available data submitted from studies (e.g. Gene Expression Omnibus [GEO])
- https://www.ncbi.nlm.nih.gov/gds/
- Controlled access (e.g. dbGaP)

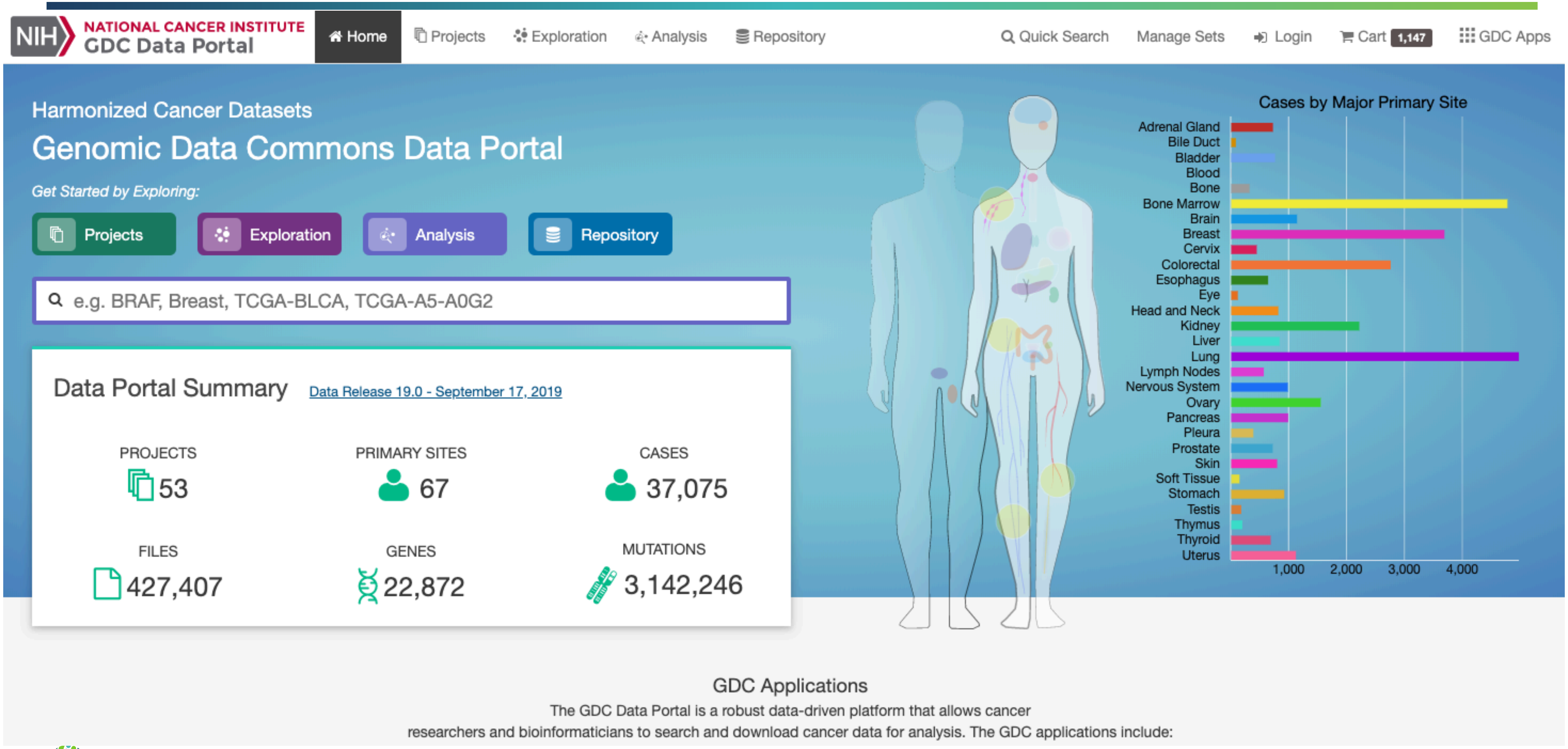
2. European Genome Phenome Archive (EGA)

- https://www.ebi.ac.uk/ega/home

3. NIH NCI Genomic Data Commons (GDC) Data Portal

- https://portal.gdc.cancer.gov/
- Harmonized Cancer Datasets

Sequence Data: Databases and Online Resources



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary Data Release 19.0 - September 17, 2019

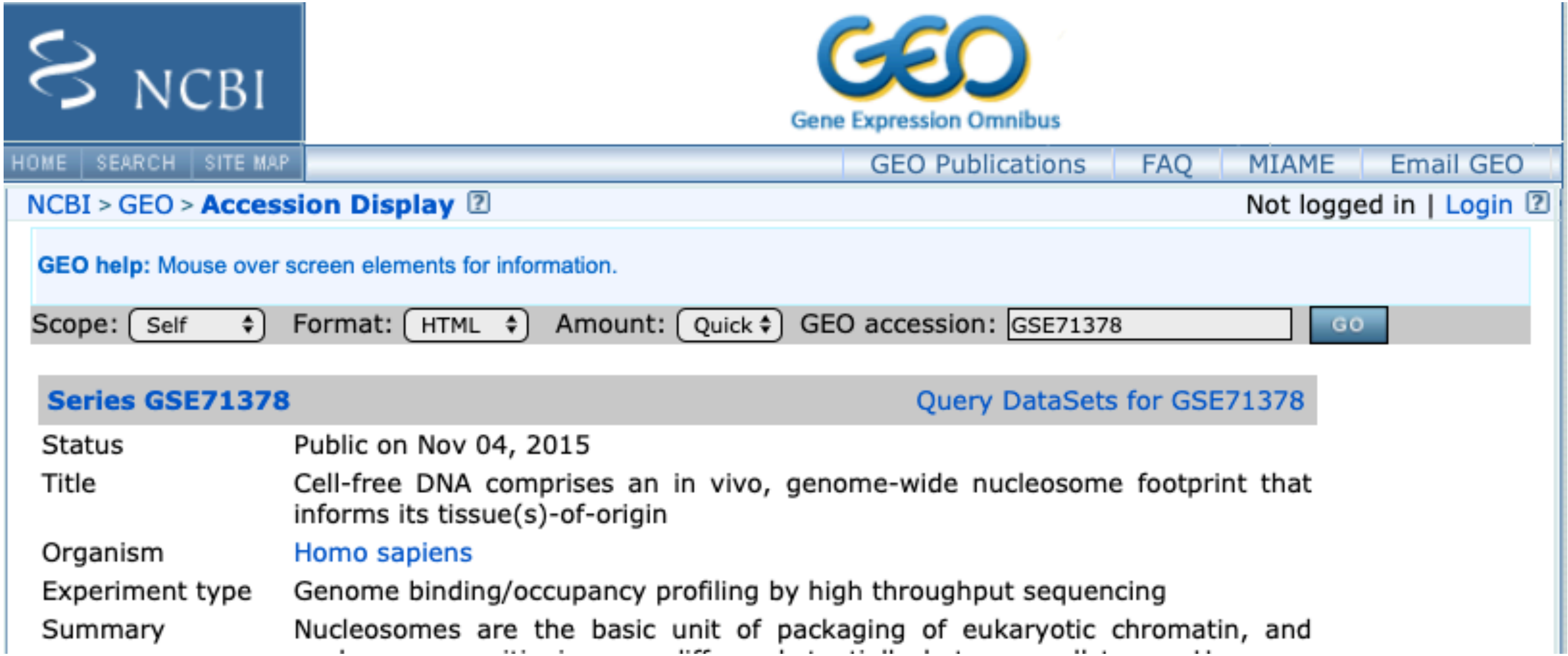
PROJECTS	PRIMARY SITES	CASES
53	67	37,075
FILES	**GENES**	**MUTATIONS**
427,407	22,872	3,142,246

GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

Sequence Data: Databases and Online Resources

Sequence Read Archive (SRA) & GEO example (GSE71378)



GEO help: Mouse over screen elements for information.

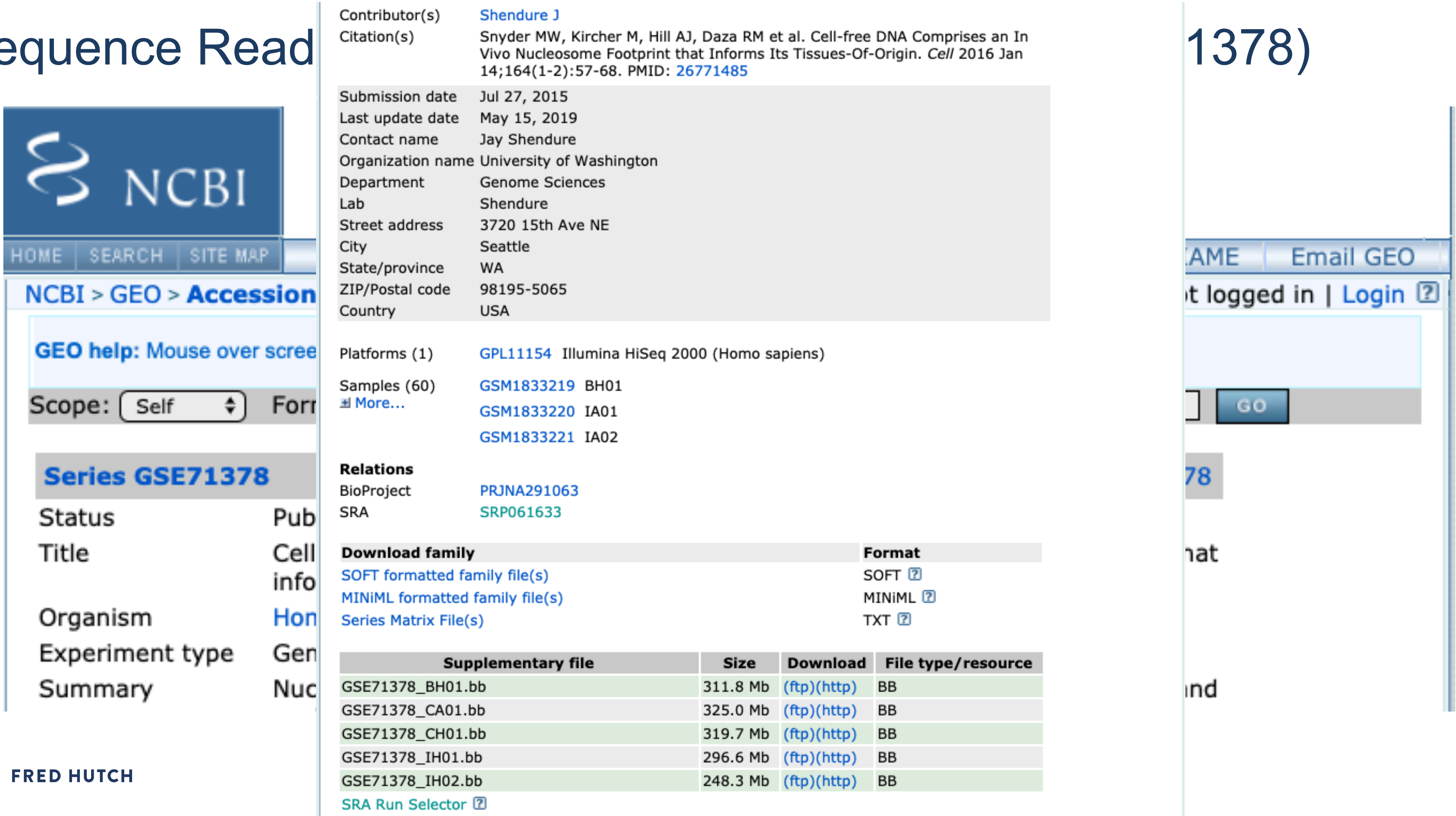
Scope: Self Format: HTML Amount: Quick GEO accession: GSE71378

Series GSE71378

Status	Public on Nov 04, 2015
Title	Cell-free DNA comprises an in vivo, genome-wide nucleosome footprint that informs its tissue(s)-of-origin
Organism	Homo sapiens
Experiment type	Genome binding/occupancy profiling by high throughput sequencing
Summary	Nucleosomes are the basic unit of packaging of eukaryotic chromatin, and

Sequence Data: Databases and Online Resources

Sequence Read … 1378)



Series GSE71378 … 78

Status Pub

Title Cell info

Organism Hon

Experiment type Gen

Summary Nuc

nat

nd

Contributor(s) Shendure J

Citation(s) Snyder MW, Kircher M, Hill AJ, Daza RM et al. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* 2016 Jan 14;164(1-2):57-68. PMID: 26771485

Submission date	Jul 27, 2015
Last update date	May 15, 2019
Contact name	Jay Shendure
Organization name	University of Washington
Department	Genome Sciences
Lab	Shendure
Street address	3720 15th Ave NE
City	Seattle
State/province	WA
ZIP/Postal code	98195-5065
Country	USA

Platforms (1) GPL11154 Illumina HiSeq 2000 (Homo sapiens)

Samples (60) GSM1833219 BH01
More... GSM1833220 IA01
GSM1833221 IA02

Relations
BioProject PRJNA291063
SRA SRP061633

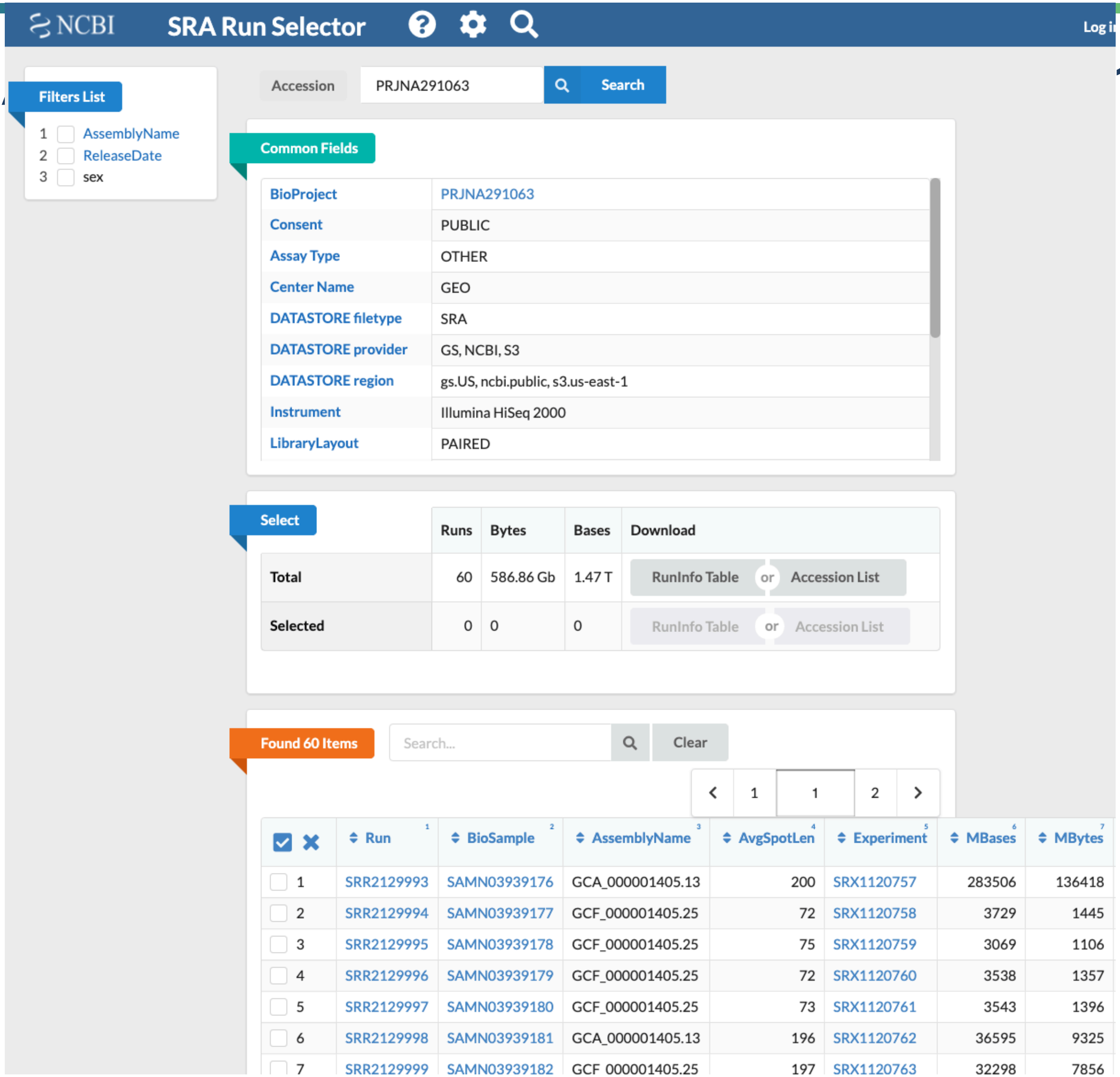
Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

Supplementary file	Size	Download	File type/resource
GSE71378_BH01.bb	311.8 Mb	(ftp)(http)	BB
GSE71378_CA01.bb	325.0 Mb	(ftp)(http)	BB
GSE71378_CH01.bb	319.7 Mb	(ftp)(http)	BB
GSE71378_IH01.bb	296.6 Mb	(ftp)(http)	BB
GSE71378_IH02.bb	248.3 Mb	(ftp)(http)	BB

SRA Run Selector ?

Sequence Data: Databases and Online Resources

Sequence Read 1378)



Accession: PRJNA291063 Search

Filters List
1 □ AssemblyName
2 □ ReleaseDate
3 □ sex

Common Fields

BioProject	PRJNA291063
Consent	PUBLIC
Assay Type	OTHER
Center Name	GEO
DATASTORE filetype	SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.US, ncbi.public, s3.us-east-1
Instrument	Illumina HiSeq 2000
LibraryLayout	PAIRED

Select

	Runs	Bytes	Bases	Download
Total	60	586.86 Gb	1.47 T	RunInfo Table or Accession List
Selected	0	0	0	RunInfo Table or Accession List

Found 60 Items

Search... Clear

‹ 1 1 2 ›

☑ ✖	Run	BioSample	AssemblyName	AvgSpotLen	Experiment	MBases	MBytes
□ 1	SRR2129993	SAMN03939176	GCA_000001405.13	200	SRX1120757	283506	136418
□ 2	SRR2129994	SAMN03939177	GCF_000001405.25	72	SRX1120758	3729	1445
□ 3	SRR2129995	SAMN03939178	GCF_000001405.25	75	SRX1120759	3069	1106
□ 4	SRR2129996	SAMN03939179	GCF_000001405.25	72	SRX1120760	3538	1357
□ 5	SRR2129997	SAMN03939180	GCF_000001405.25	73	SRX1120761	3543	1396
□ 6	SRR2129998	SAMN03939181	GCA_000001405.13	196	SRX1120762	36595	9325
□ 7	SRR2129999	SAMN03939182	GCF_000001405.25	197	SRX1120763	32298	7856

Sequence Data: Databases and Online Resources

Sequence Read Archive (SRA) & GEO example (GSE71378)

SRA Toolkit required to download and extract **`.sra`** files

- Download .sra file

```
prefetch SRR2130004
```

- Convert `.sra` file to fastq

```
fastq-dump SRR2130004 # use accession
fastq-dump SRR2130004.sra # use file if already downloaded
```

- Convert `.sra` file to SAM/BAM file

```
# will write data to a SAM file
sam-dump --header SRR2130004.sra > SAMN03160688.sam
# will write data to a BAM file
sam-dump --header SRR2130004.sra | samtools view -bS - > BRCA_IDC_cfDNA.bam
```

Sequence Data: File formats

Sequences

- Genome sequences - **FASTA** (.fasta or .fa)
- Sequenced reads - **FASTQ** (.fastq or .fq)

Sequence Alignment/Map Format

- https://samtools.github.io/hts-specs/SAMv1.pdf
- Sequence Alignment - **SAM** (.sam)
- Binary Alignment - **BAM** (.bam)

Sequence Data: Sequence alignment

Burrows-Wheeler Aligner, bwa (http://bio-bwa.sourceforge.net/)

- aln - for 35bp to 100bp reads
- mem - for reads with length 70bp to 1Mb (Recommended for most)

```
# If two fastq files, one for each mate of paired-end reads
bwa mem -M reference.fa BRCA_IDC_cfDNA_R1.fq BRCA_IDC_cfDNA_R2.fq > BRCA_IDC_cfDNA.bam


# If single fastq file with paired-end reads interleaved
bwa mem -M -p reference.fa BRCA_IDC_cfDNA.fq > BRCA_IDC_cfDNA.bam
```

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]

FRED HUTCH

Sequence Data: Inspecting and Reading SAM/BAM Files

SAMtools (http://www.htslib.org/)

- Indexing

```
samtools index BRCA_IDC_cfDNA.bam #required for all BAM files
```

- File operations

```
samtools sort BRCA_IDC_cfDNA.bam #sort by coordinate
```

- Statistics

```
samtools flagstat BRCA_IDC_cfDNA.bam #get general alignment metrics
```

- Viewing

```
# view header information
samtools view -H BRCA_IDC_cfDNA.bam

# view aligned reads at chr17:25,000,000
samtools view BRCA_IDC_cfDNA.bam 17:37844393
```

Sequence Data: SAM Format

https://samtools.github.io/hts-specs/SAMv1.pdf

A. Header information

```
samtools view -H BRCA_IDC_cfDNA.bam

@HD     VN:1.2  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
...
@RG     ID:P12.17.7_Breast
```

Sequence Data: SAM Format

https://samtools.github.io/hts-specs/SAMv1.pdf

A. Header information

- `@HD`: Header line
  - `SO`: Sorting order of alignments (`unknown, unsorted, coordinate, queryname`)
- `@SD`: Reference sequence dictionary
  - `SN`: Reference sequence name - typically, one row for each chromosome
  - `LN`: Length of reference sequence
- `@RG`: Read group
  - `ID`: Read group identifier (must be unique)
  - `PL`: Platform or technology used (e.g. ILLUMINA)
  - `SM`: Sample ID and/or pool being sequenced
- `@PG`: Program/tool information
  - `ID`: Unique name, `PN`: Program name; `CL`: Command line

Sequence Data: SAM Format

https://samtools.github.io/hts-specs/SAMv1.pdf

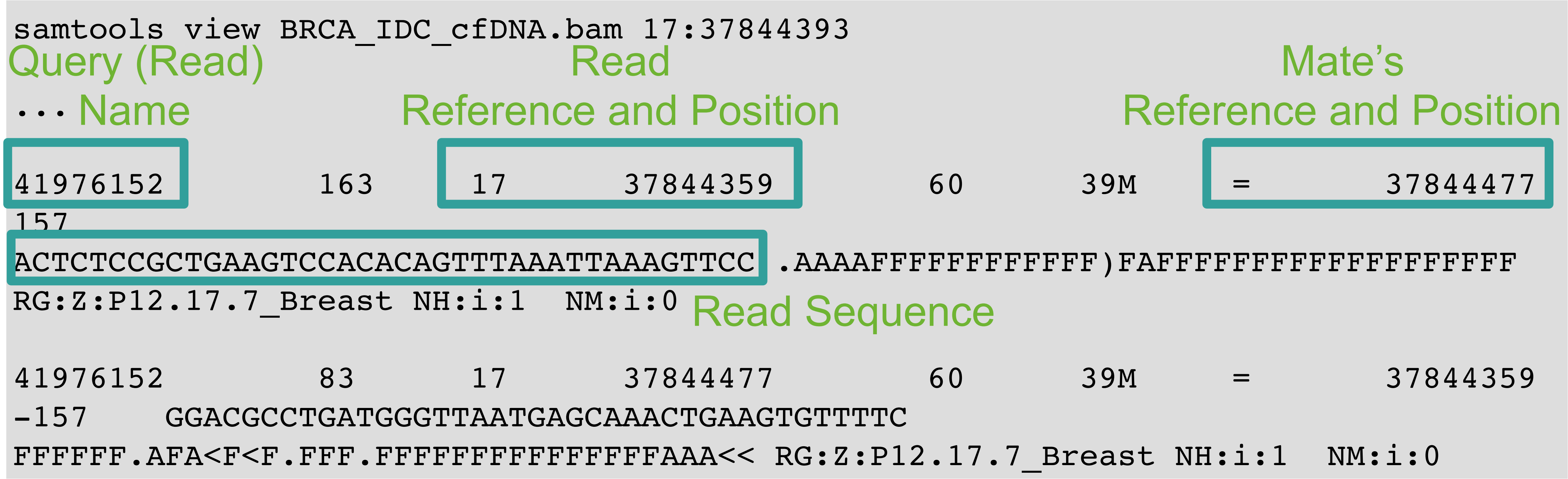
B. Alignment information

```
samtools view BRCA_IDC_cfDNA.bam 17:37844393

...

41976152        163     17      37844359        60      39M     =       37844477
157
ACTCTCCGCTGAAGTCCACACAGTTTAAATTAAAGTTCC .AAAAFFFFFFFFFFFF)FAFFFFFFFFFFFFFFFFFFF
RG:Z:P12.17.7_Breast NH:i:1  NM:i:0

41976152        83      17      37844477        60      39M     =       37844359
-157    GGACGCCTGATGGGTTAATGAGCAAACTGAAGTGTTTTC
FFFFFF.AFA<F<F.FFF.FFFFFFFFFFFFFFFAAA<< RG:Z:P12.17.7_Breast NH:i:1  NM:i:0
```

Sequence Data: SAM Format

https://samtools.github.io/hts-specs/SAMv1.pdf

B. Alignment information

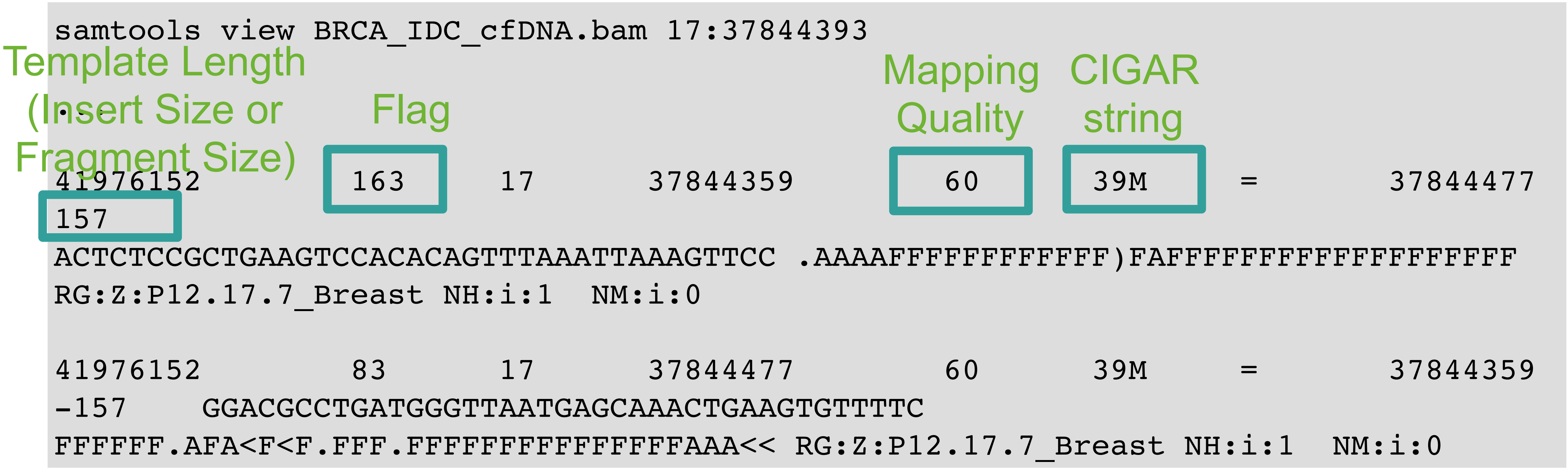


```
samtools view BRCA_IDC_cfDNA.bam 17:37844393
…

41976152       163     17      37844359        60      39M     =       37844477
157
ACTCTCCGCTGAAGTCCACACAGTTTAAATTAAAGTTCC .AAAAFFFFFFFFFFFF)FAFFFFFFFFFFFFFFFFFFF
RG:Z:P12.17.7_Breast NH:i:1  NM:i:0

41976152       83      17      37844477        60      39M     =       37844359
-157    GGACGCCTGATGGGTTAATGAGCAAACTGAAGTGTTTTC
FFFFFF.AFA<F<F.FFF.FFFFFFFFFFFFFFFAAA<< RG:Z:P12.17.7_Breast NH:i:1  NM:i:0
```

Sequence Data: SAM Format

https://samtools.github.io/hts-specs/SAMv1.pdf

B. Alignment information



```
samtools view BRCA_IDC_cfDNA.bam 17:37844393
...
41976152        163      17       37844359        60       39M      =        37844477
157
ACTCTCCGCTGAAGTCCACACAGTTTAAATTAAAGTTCC .AAAAFFFFFFFFFFFF)FAFFFFFFFFFFFFFFFFFF
RG:Z:P12.17.7_Breast NH:i:1  NM:i:0

41976152        83       17       37844477        60       39M      =        37844359
-157    GGACGCCTGATGGGTTAATGAGCAAACTGAAGTGTTTTC
FFFFFF.AFA<F<F.FFF.FFFFFFFFFFFFFFAAA<< RG:Z:P12.17.7_Breast NH:i:1  NM:i:0
```

Template Length (Insert Size or Fragment Size)

Flag

Mapping Quality

CIGAR string

Sequence Data: SAM Format

https://samtools.github.io/hts-specs/SAMv1.pdf

B. Alignment Format

1. QNAME: query (read) template name
2. FLAG: bitwise value describing the alignment
   - e.g. 4 - read is unmapped; 2 - proper pair; 1024 - PCR duplicate
   - https://www.samformat.info/sam-format-flag
3. RNAME: reference sequence name (i.e. chr1 or 1)
4. POS: position of aligned read (leftmost; 1-based)
5. MAPQ: Mapping quality
6. CIGAR: Code string to describe read alignment sequence match to reference
7. RNEXT: reference sequence name of mate read
8. PNEXT: position of mate read
9. TLEN: template (read) length; 0 if mates on different chromosomes
10. SEQ: sequence of mapped reads on forward genomic strand
11. QUAL: base qualities (Phred-scale)

Sequence Data: SAM Format

https://samtools.github.io/hts-specs/SAMv1.pdf

B. Alignment Format: CIGAR string (common operators)

M	alignment match (sequence match or mismatch)
I	insertion relative to reference
D	deletion relative to reference
S	soft clipping (mismatch bases included in SEQ)
H	hard clipping (mismatch bases excluded in SEQ)
N	skipped sequence from reference
=	sequence match
X	sequence mismatch

```
Reference: GACCTTACTTCATCTTGTG--CTTACTATCAAGTGATTA
     Read:     TTACTT----TTCTGAACTTACTGCTCCTA
```

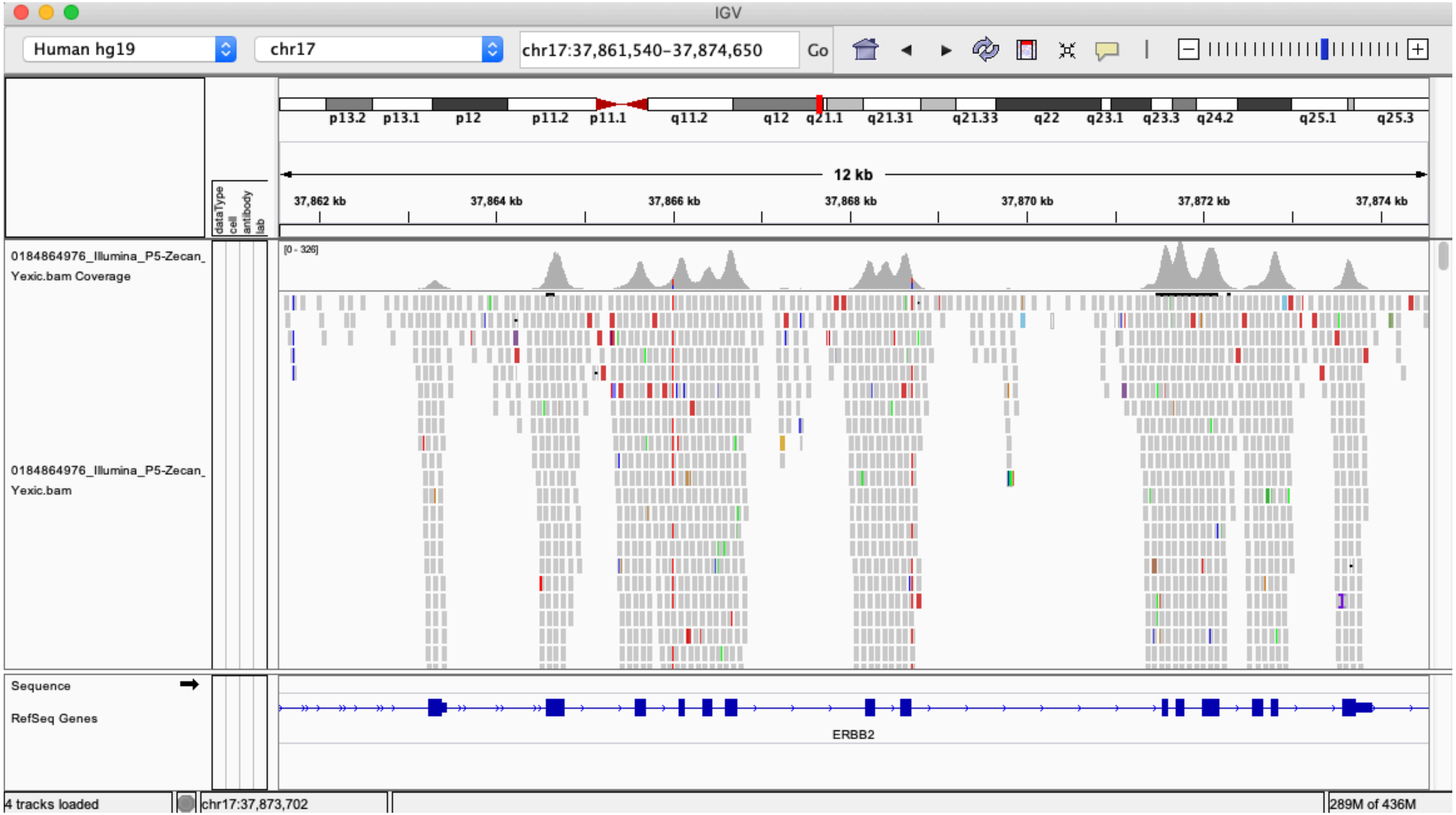
What is the CIGAR?

Tools for Sequencing Data: Overview

1. **Inspecting and Reading SAM/BAM files**
   - SAMtools
2. **Interactive Visualization**
   - Integrative Genomics Viewer (https://software.broadinstitute.org/software/igv)
   - BioViz (https://bioviz.org/)
   - Table (https://ics.hutton.ac.uk/tablet/)
3. **Sequencing metrics and Processing**
   - SAMtools
   - Picard Tools
   - Genomic Analysis Toolkit (GATK)
4. **Genome Variation Analysis**

Tools for Sequencing Data: Interactive Visualization

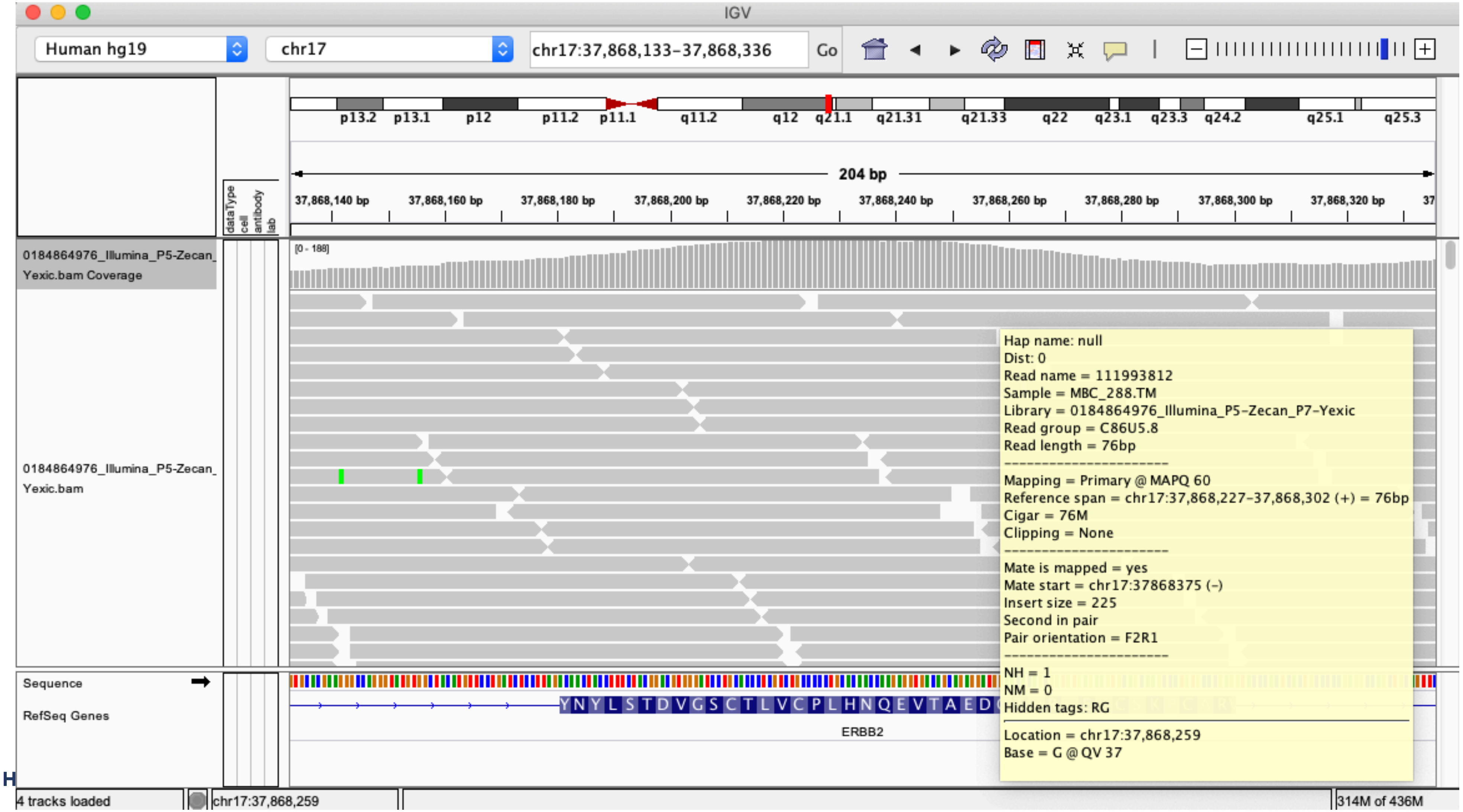
Integrative Genomics Viewer (https://software.broadinstitute.org/software/igv)



FRED

Tools for Sequencing Data: Interactive Visualization

Integrative Genomics Viewer (https://software.broadinstitute.org/software/igv)



Hap name: null
Dist: 0
Read name = 111993812
Sample = MBC_288.TM
Library = 0184864976_Illumina_P5-Zecan_P7-Yexic
Read group = C86U5.8
Read length = 76bp

Mapping = Primary @ MAPQ 60
Reference span = chr17:37,868,227-37,868,302 (+) = 76bp
Cigar = 76M
Clipping = None

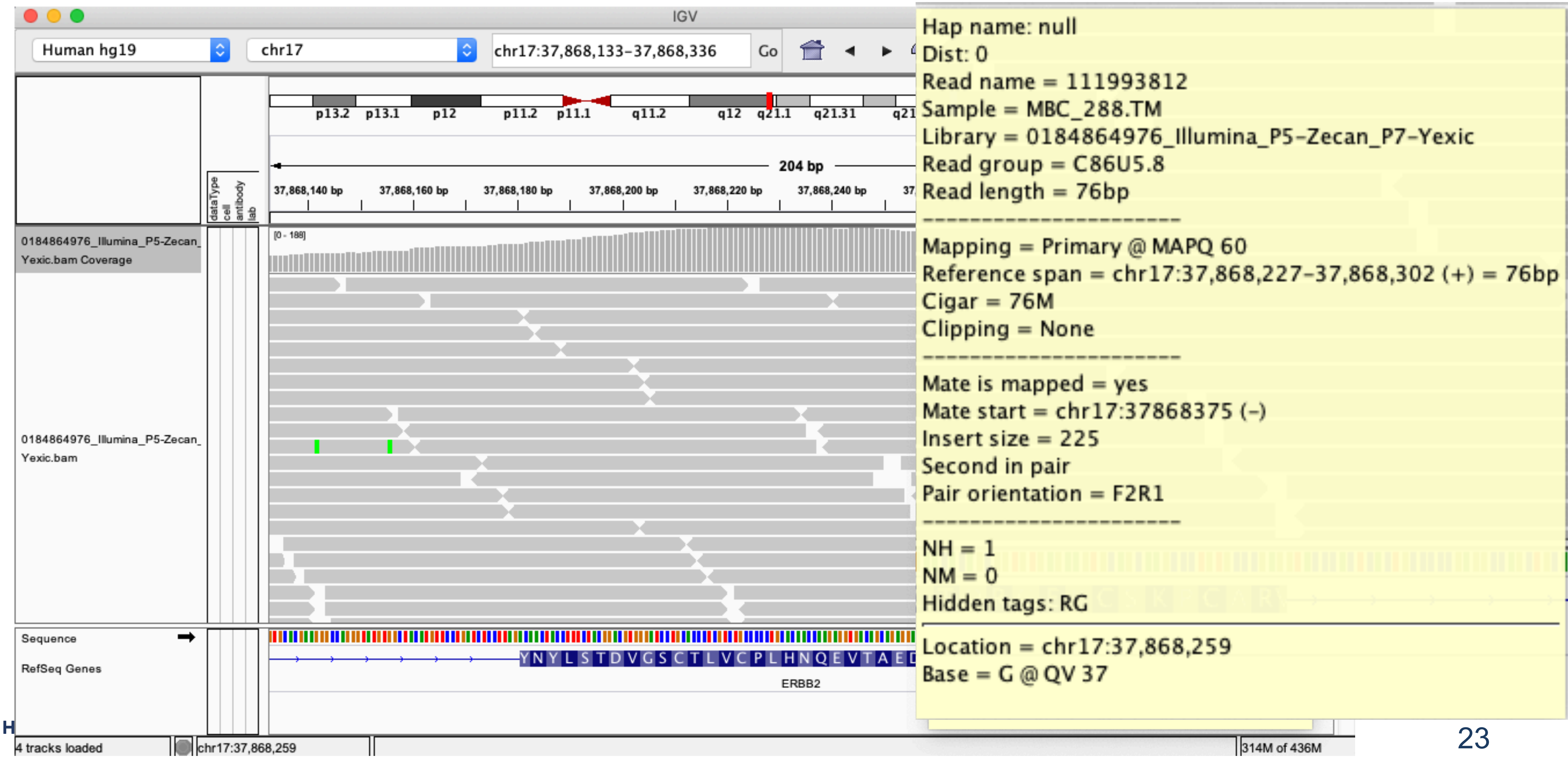
Mate is mapped = yes
Mate start = chr17:37868375 (-)
Insert size = 225
Second in pair
Pair orientation = F2R1

NH = 1
NM = 0
Hidden tags: RG

Location = chr17:37,868,259
Base = G @ QV 37

Tools for Sequencing Data: Interactive Visualization

Integrative Genomics Viewer (https://software.broadinstitute.org/software/igv)



Human hg19

chr17

chr17:37,868,133-37,868,336

Hap name: null
Dist: 0
Read name = 111993812
Sample = MBC_288.TM
Library = 0184864976_Illumina_P5-Zecan_P7-Yexic
Read group = C86U5.8
Read length = 76bp

Mapping = Primary @ MAPQ 60
Reference span = chr17:37,868,227-37,868,302 (+) = 76bp
Cigar = 76M
Clipping = None

Mate is mapped = yes
Mate start = chr17:37868375 (-)
Insert size = 225
Second in pair
Pair orientation = F2R1

NH = 1
NM = 0
Hidden tags: RG

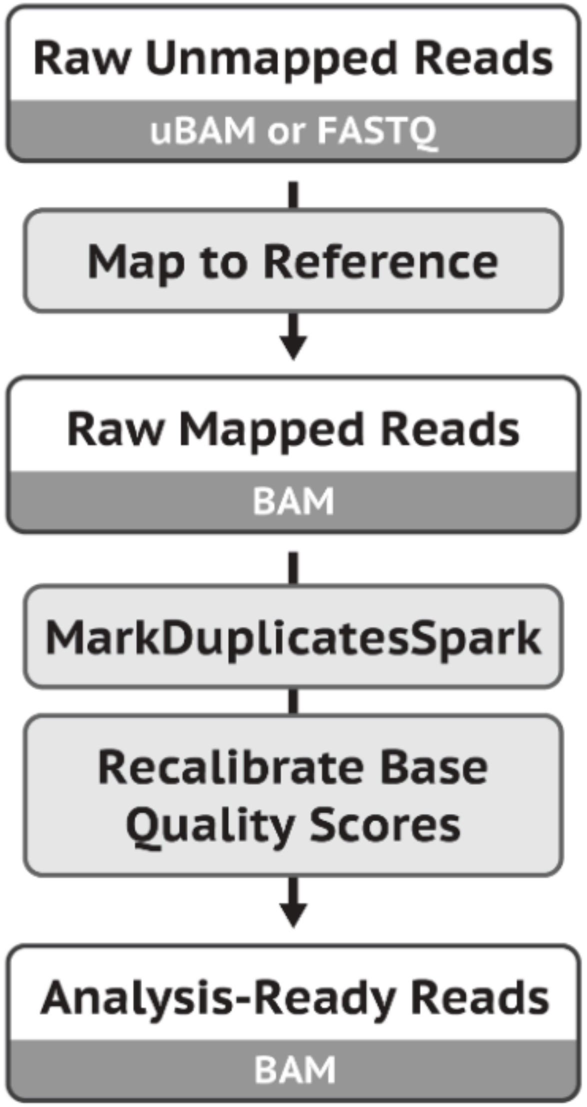
Location = chr17:37,868,259
Base = G @ QV 37

Tools for Sequencing Data: Processing

Picard Tools & GATK4: Best practices

1. Mark Duplicates
    1. MarkDuplicates + SortSam (Picard)
2. Base Quality Score Recalibration (BQSR)
    1. BaseRecalibrator (GATK4)
    2. ApplyBQSR (GATK4)

```
java -jar picard.jar \
MarkDuplicates \
INPUT=BRCA_IDC_cfDNA.bam \
REMOVE_DUPLICATES=false \
OUTPUT=BRCA_IDC_cfDNA.marked_duplicates.bam \
METRIC_FILE=BRCA_IDC_cfDNA.markDupMetrics.txt
```



Raw Unmapped Reads

uBAM or FASTQ

Map to Reference

Raw Mapped Reads

BAM

MarkDuplicatesSpark

Recalibrate Base Quality Scores

Analysis-Ready Reads

BAM

FRED HUTCH

https://software.broadinstitute.org/gatk/best-practices/

Tools for Sequencing Data: Sequencing Metrics

Picard Tools & GATK4: Best practices

3. Generate alignment metrics
   a. `CollectMultipleMetrics`
      - `CollectAlignmentSummaryMetrics`
      - `CollectInsertSizeMetrics`
   b. Collect assay-specific metrics
      - `CollectWgsMetrics` - Whole genome sequencing
      - `CollectHsMetrics` - Hybrid Selection (i.e. whole exome)
      - `CollectRnaSeqMetrics` - RNA-seq
      - `CollectTargetedPcrMetrics` - Targeted PCR amplicon sequencing
   c. `EstimateLibraryComplexity`
      a. Estimates the number of unique molecules in the library

https://broadinstitute.github.io/picard/command-line-overview.html

http://broadinstitute.github.io/picard/picard-metric-definitions.html

Tools for Sequencing Data: Sequencing Metrics

Picard Tools & GATK4: Best practices

3. Generate alignment metrics: (a) `CollectAlignmentSummaryMetrics`

```
java -jar picard.jar CollectAlignmentSummaryMetrics \
INPUT=BRCA_IDC_cfDNA.bam \
OUTPUT=BRCA_IDC_cfDNA.alignMetrics.txt \
REFERENCE_SEQUENCE=hs37d5.fa \
```

CATEGORY	TOTAL_READS	PF_READS	PCT_PF_READS	PF_READS_ALIGNED	PCT_PF_READS_ALIGNED	PF_ALIGNED_BASES	PF_HQ_ALIGNED_READS	PF_HQ_ALIGNED_BASES	MEAN_READ_LENGTH	STRAND_BALANCE	PCT_CHIMERAS
FIRST_OF_PAIR	49333221	49333218	1	49333218	1	1920603633	42832799	1667902478	39	0.50011	0.019043
SECOND_OF_PAIR	49333221	49333218	1	49333218	1	1918882942	42822426	1665977301	39	0.500715	0.019904
PAIR	98666442	98666436	1	98666436	1	3839486575	85655225	3333879779	39	0.500412	0.019474
UNPAIRED	3371706	3349869	0.993523	3349869	1	106596546	2050309	69312874	31.872292	0.501895	0

http://broadinstitute.github.io/picard/picard-metric-definitions.html#AlignmentSummaryMetrics

Tools for Sequencing Data: Sequencing Metrics

Picard Tools & GATK4: Best practices

3. Generate alignment metrics: (a) `CollectWgsMetrics`

```
java -jar picard.jar CollectWgsMetrics \
INPUT=BRCA_IDC_cfDNA.bam \
OUTPUT=BRCA_IDC_cfDNA.alignMetrics.txt \
REFERENCE_SEQUENCE=hs37d5.fa \
```

GENOME_TERRITORY	MEAN_COVERAGE	SD_COVERAGE	MEDIAN_COVERAGE	PCT_EXC_MAPQ	PCT_EXC_DUPE	PCT_1X	PCT_5X
2900340137	1.053882	1.383867	1	0.137741	0	0.578236	0.015963

```
coverage        high_quality_coverage_count
0       1223257622
1       854276028
2       475072046
3       215728575
4       85708030
5       30916117
6       10376403
7       3318514
8       1041100
9       329830
10      111513
```

https://broadinstitute.github.io/picard/picard-metric-definitions.html#CollectWgsMetrics.WgsMetrics

Tools for Sequencing Data: Accessing BAM files in R & Python

Python

- PySam

  https://pysam.readthedocs.io/en/latest/api.html

R and Bioconductor (more in next lecture)

- **Rsamtools**
  - Import BAM files into R
  - View the header information
  - Accessing read sequences, aligned positions, CIGAR, read names, etc
  - Large BAM files can be read in chunks to optimize memory
  - Create new BAM files using “Views” of a subset of reads

https://bioconductor.org/packages/release/bioc/vignettes/Rsamtools/inst/doc/Rsamtools-Overview.pdf

Genome Variant Analysis: Overview

1. Types of genomic variation

2. Visualization using IGV

3. Tools for Predicting Genome Variation

4. File Formats for Variation Data

5. Variant Annotation Tools

6. Variant databases

Genome Variant Analysis: Types of Genomic Variation

Variant or Mutation or Alteration or Polymorphism

- Changes in the genome sequence of a sample compared to a reference sequence
- Chromosomes: 22 autosomal pairs + 1 sex pair
  - Each set inherited from maternal and paternal germline cells

Germline Variant

- Variant inherited from one or both parental chromosomes
- Source of genetic differences between ancestral populations and individuals
- Polymorphism: >1% frequency in a population

Somatic Variant

- Mutation acquired during individual's lifetime
- Important to identify in sporadic cancers and other non-familial diseases

Genome Variant Analysis: Types of Genomic Variation

a. **Single nucleotide base substitutions**

- Germline single nucleotide polymorphism (SNP)
- Somatic single nucleotide variant (SNV)

b. **Small insertions or deletions**

- Germline or somatic insertion or deletion (INDEL)

c. **Copy number changes**

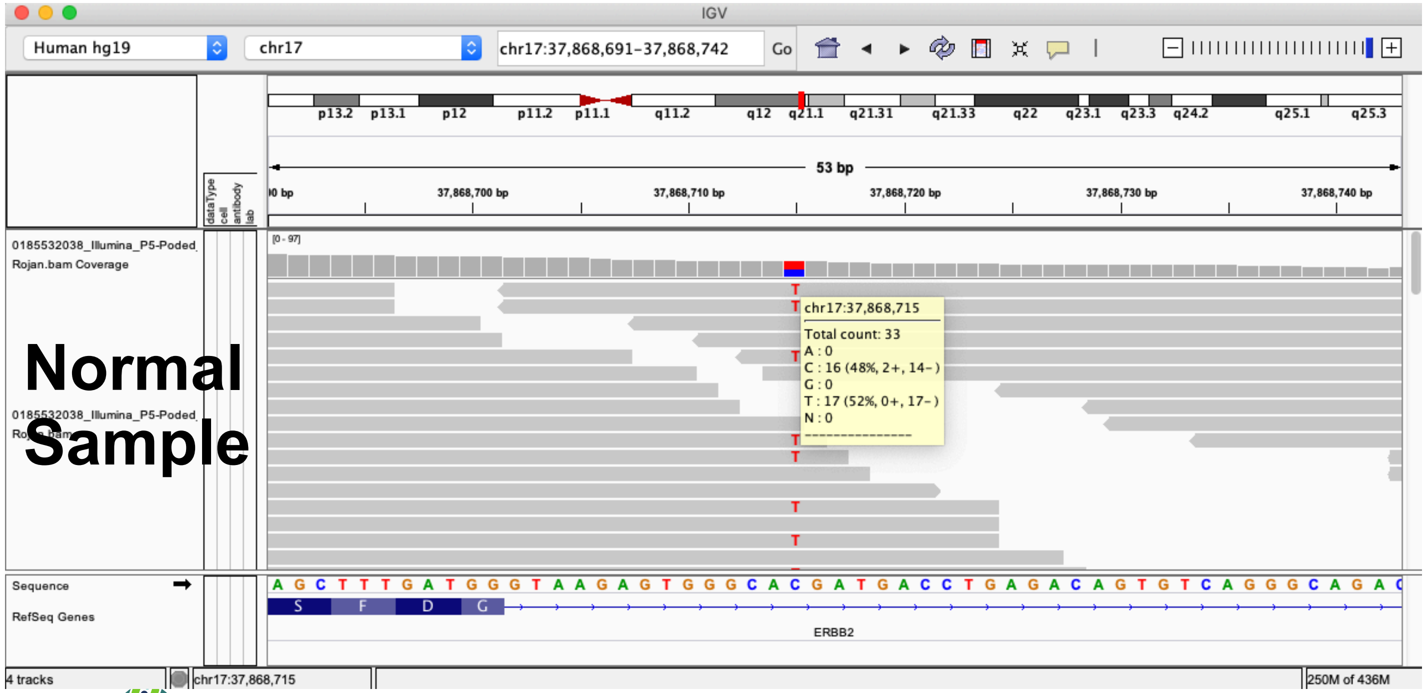
- Germline copy number variant (CNV) or polymorphism (CNP)
- Somatic copy number variant (CNV) or alterations (CNA)

d. **Structural rearrangements**

- Germline or Somatic structural variant (SV)

Genome Variant Analysis: Single Nucleotide Polymorphism

- ~1.5 to 2 million **SNPs** per individual
- Identify SNPs from normal peripheral blood mononuclear cells (PBMC)



chr17:37,868,715

Total count: 33
A : 0
C : 16 (48%, 2+, 14-)
G : 0
T : 17 (52%, 0+, 17-)
N : 0

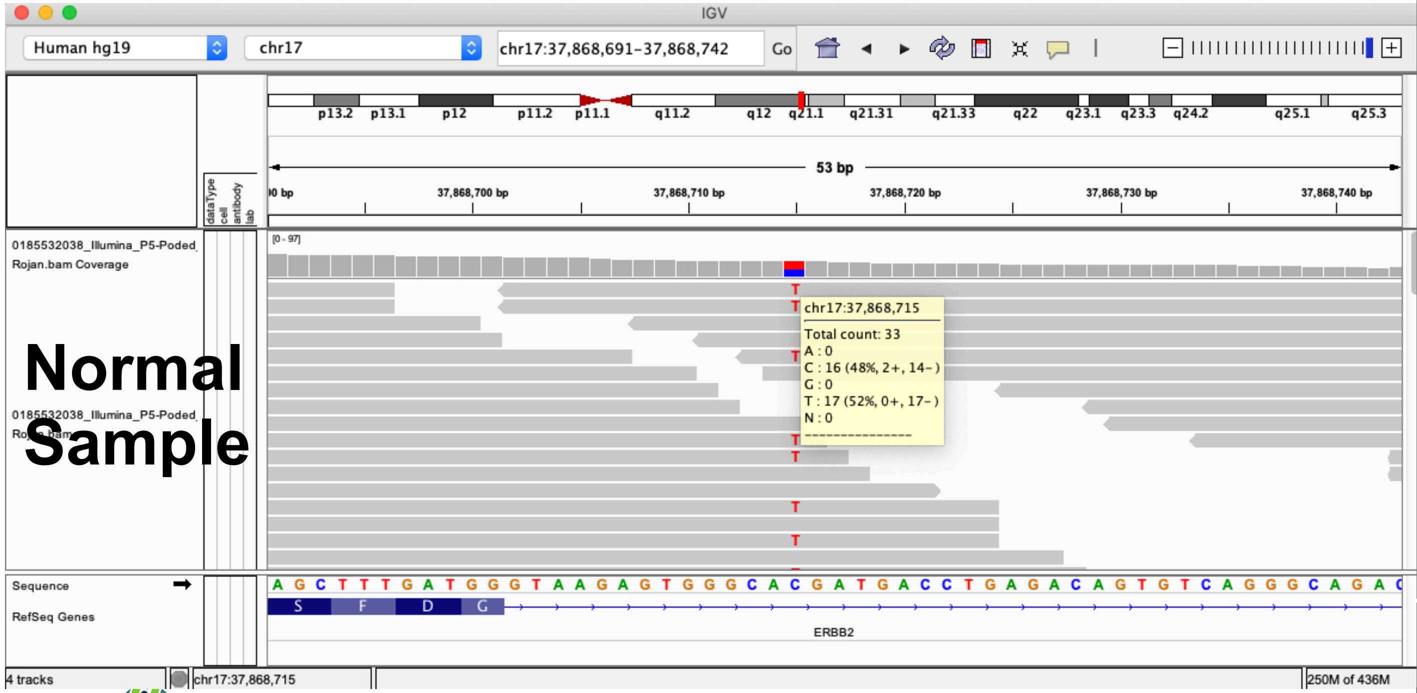
Normal Sample

Heterozygous SNP with 37 reads containing the variant and having depth 79 reads

37/79 (47%) variant allele fraction (VAF)

Genome Variant Analysis: Single Nucleotide Polymorphism

- ~1.5 to 2 million **SNPs** per individual
- Identify SNPs from normal peripheral blood mononuclear cells (PBMC)



0185532038_Illumina_P5-Poded_Rojan.bam Coverage

[0 - 97]

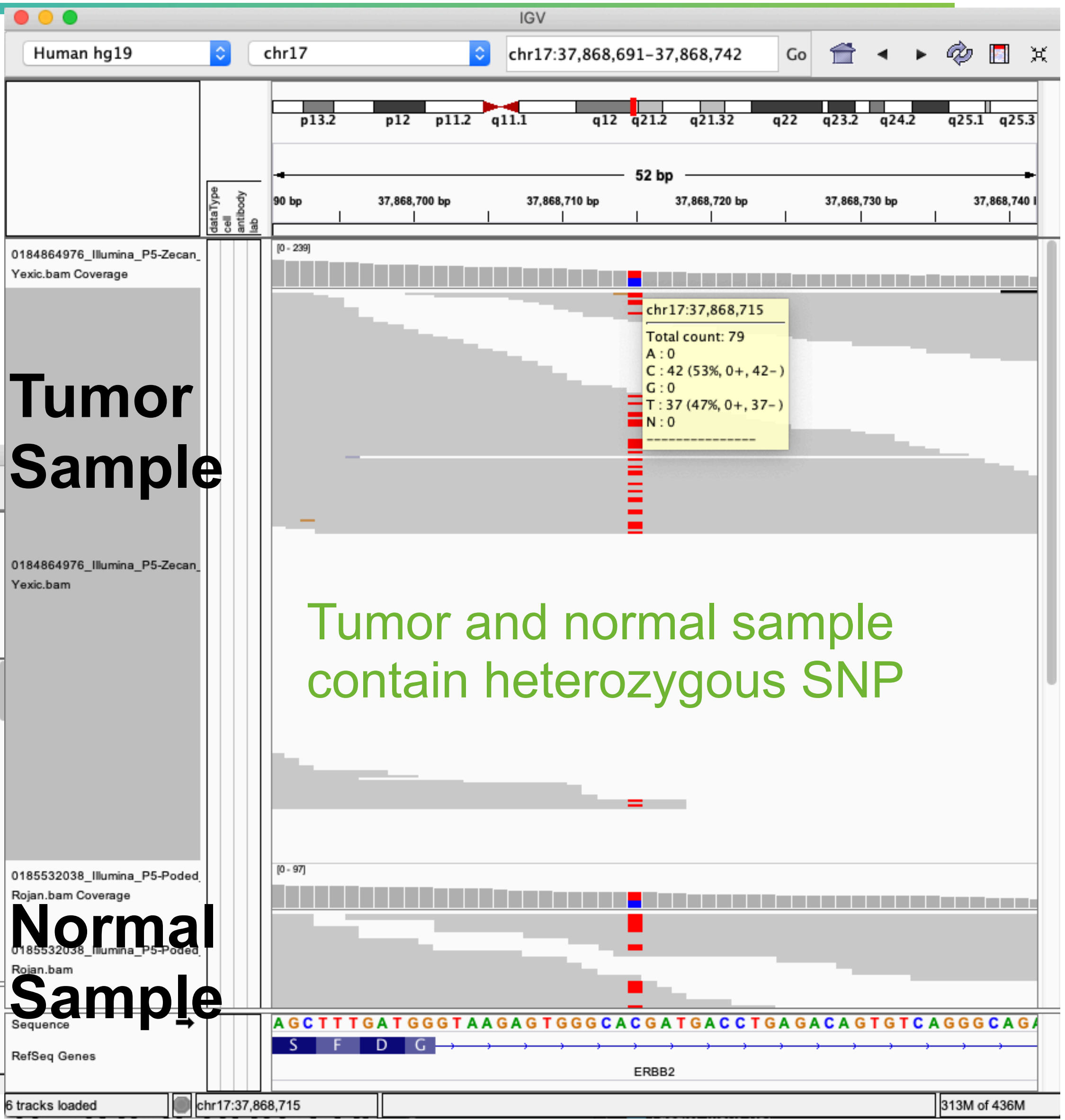
chr17:37,868,715
Total count: 33
A : 0
C : 16 (48%, 2+, 14-)
G : 0
T : 17 (52%, 0+, 17-)
N : 0

0185532038_Illumina_P5-Poded_Rojan.bam

A G C T T T G A T G G G T A A G A G T G G G C A C G A T G A C C T G A G A C A G T G T C A G G G C A G A C

4 tracks | chr17:37,868,715 | 250M of 436M

Normal Sample



0184864976_Illumina_P5-Zecan_Yexic.bam Coverage

[0 - 239]

chr17:37,868,715
Total count: 79
A : 0
C : 42 (53%, 0+, 42-)
G : 0
T : 37 (47%, 0+, 37-)
N : 0

0184864976_Illumina_P5-Zecan_Yexic.bam

0185532038_Illumina_P5-Poded_Rojan.bam Coverage

[0 - 97]

0185532038_Illumina_P5-Poded_Rojan.bam

A G C T T T G A T G G G T A A G A G T G G G C A C G A T G A C C T G A G A C A G T G T C A G G G C A G A

6 tracks loaded | chr17:37,868,715 | 313M of 436M

Tumor Sample

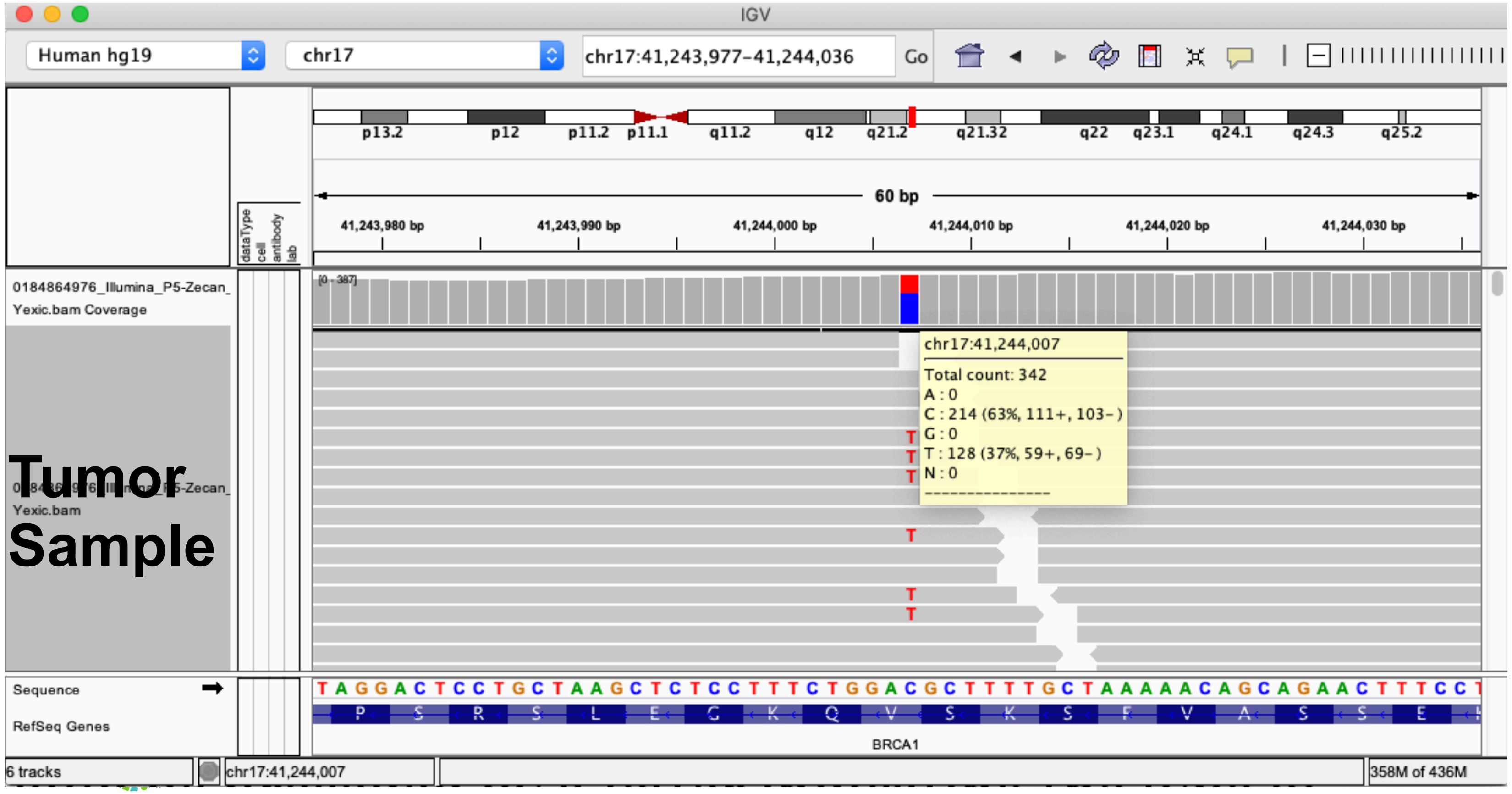
Tumor and normal sample contain heterozygous SNP

Normal Sample

FRED HUTCH

Genome Variant Analysis: Single Nucleotide Variant (SNV)

- Somatic **SNV** requires comparing case (tumor) with control (PBMC)


IGV
Human hg19
chr17
chr17:41,243,977-41,244,036
Go
p13.2
p12
p11.2
p11.1
q11.2
q12
q21.2
q21.32
q22
q23.1
q24.1
q24.3
q25.2
60 bp
41,243,980 bp
41,243,990 bp
41,244,000 bp
41,244,010 bp
41,244,020 bp
41,244,030 bp
dataType
cell
antibody
lab
0184864976_Illumina_P5-Zecan_Yexic.bam Coverage
[0 - 387]
0184864976_Illumina_P5-Zecan_Yexic.bam
T
T
T
T
T
T
Sequence
RefSeq Genes
T A G G A C T C C T G C T A A G C T C T C C T T T C T G G A C G C T T T T G C T A A A A A C A G C A G A A C T T T C C T
P S R S L E G K Q V S K S F V A S S E
BRCA1
6 tracks
chr17:41,244,007
358M of 436M

chr17:41,244,007

Total count: 342
A : 0
C : 214 (63%, 111+, 103-)
G : 0
T : 128 (37%, 59+, 69-)
N : 0

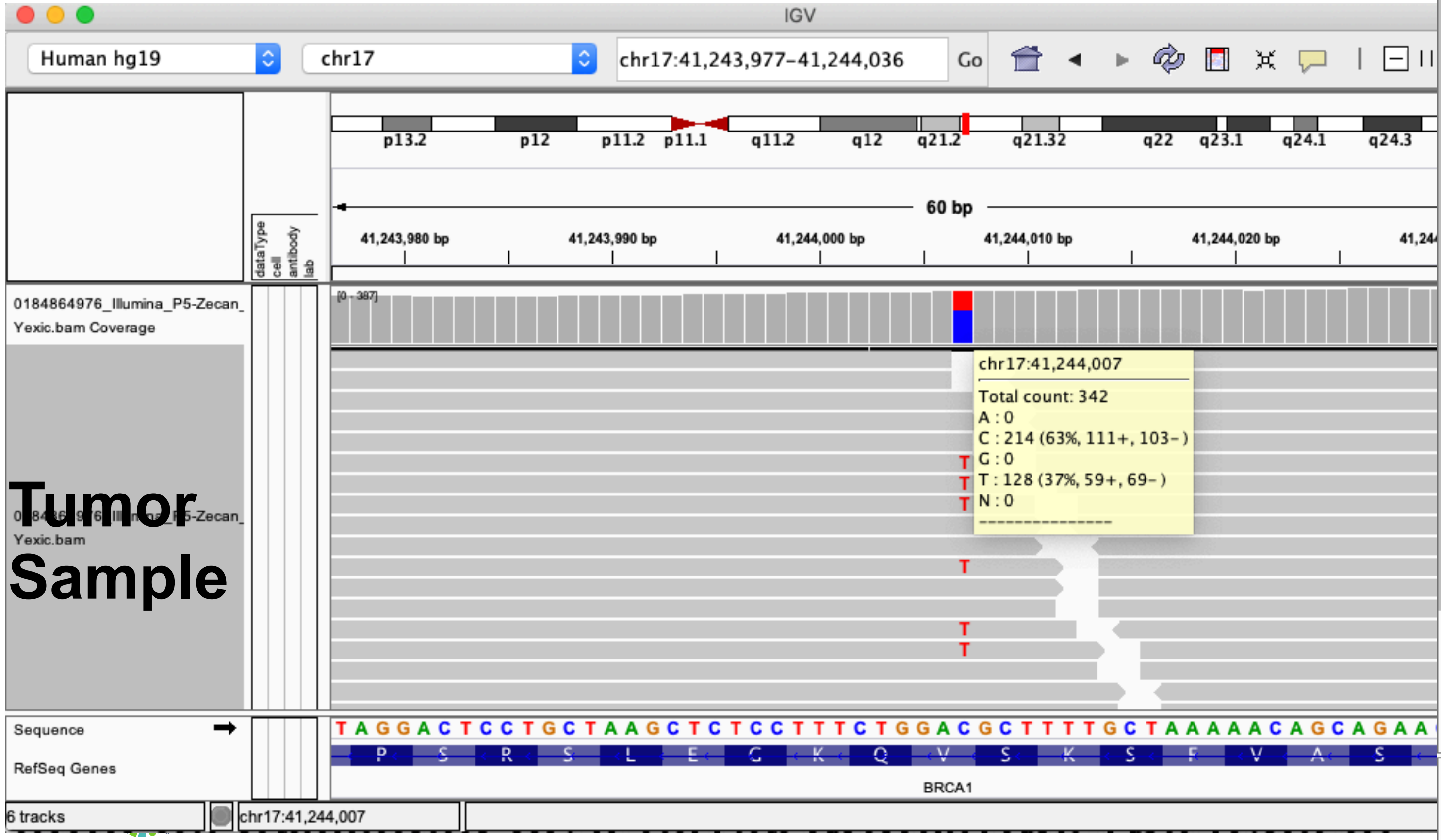
Tumor Sample

Potential SNV with 128/342 (37%) VAF

p.V1181I

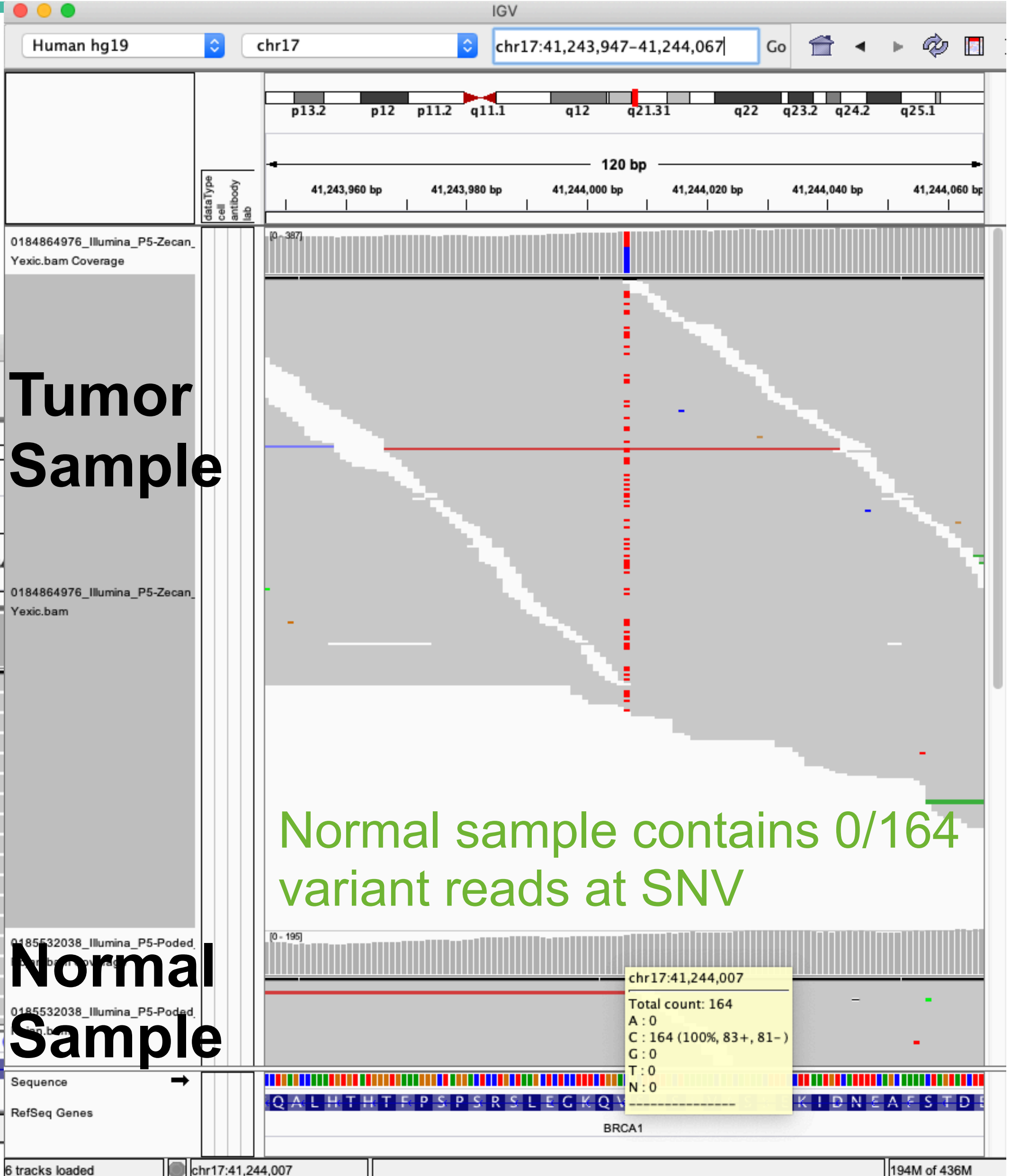
Genome Variant Analysis: Single Nucleotide Variant (SNV)

- Somatic **SNV** requires comparing case (tumor) with control (PBMC)



chr17:41,244,007
Total count: 342
A : 0
C : 214 (63%, 111+, 103-)
G : 0
T : 128 (37%, 59+, 69-)
N : 0

Tumor Sample



Tumor Sample

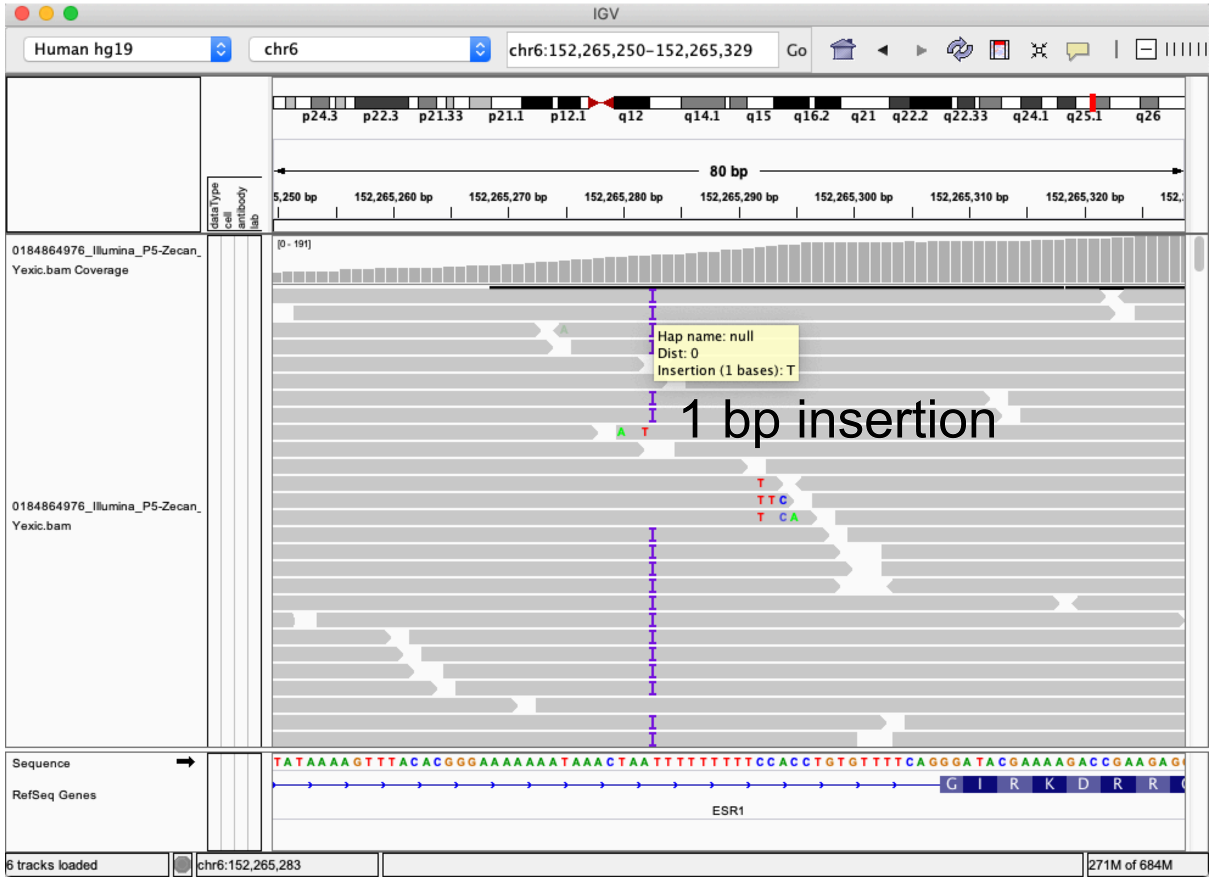
Normal sample contains 0/164 variant reads at SNV

Normal Sample

chr17:41,244,007
Total count: 164
A : 0
C : 164 (100%, 83+, 81-)
G : 0
T : 0
N : 0

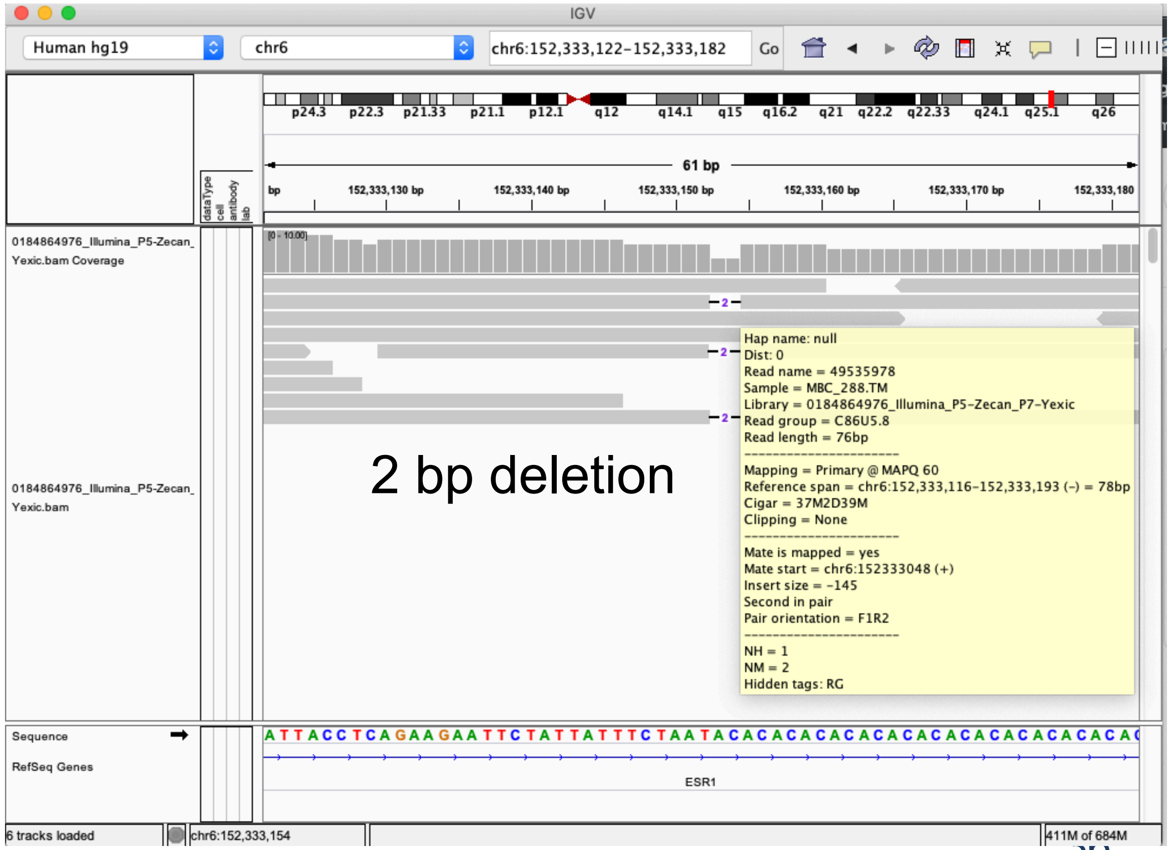
Genome Variant Analysis: Insertion & Deletion (INDEL)

- 1 to 10,000 bps size range
- Can lead to in-frame or frame-shift mutations
- Recall: CIGAR strings


IGV
Human hg19
chr6
chr6:152,265,250-152,265,329
Go
p24.3 p22.3 p21.33 p21.1 p12.1 q12 q14.1 q15 q16.2 q21 q22.2 q22.33 q24.1 q25.1 q26
80 bp
5,250 bp
152,265,260 bp
152,265,270 bp
152,265,280 bp
152,265,290 bp
152,265,300 bp
152,265,310 bp
152,265,320 bp
dataType
cell
antibody
lab
0184864976_Illumina_P5-Zecan_Yexic.bam Coverage
[0 - 191]
0184864976_Illumina_P5-Zecan_Yexic.bam
Sequence
RefSeq Genes
TATAAAAGTTTACACGGGAAAAAAATAAACTAATTTTTTTTTTCCACCTGTGTTTTCAGGGATACGAAAAGACCGAAGAG
G I R K D R R
ESR1
6 tracks loaded
chr6:152,265,283
271M of 684M

Hap name: null
Dist: 0
Insertion (1 bases): T

1 bp insertion


IGV
Human hg19
chr6
chr6:152,333,122-152,333,182
Go
p24.3 p22.3 p21.33 p21.1 p12.1 q12 q14.1 q15 q16.2 q21 q22.2 q22.33 q24.1 q25.1 q26
61 bp
bp
152,333,130 bp
152,333,140 bp
152,333,150 bp
152,333,160 bp
152,333,170 bp
152,333,180
dataType
cell
antibody
lab
0184864976_Illumina_P5-Zecan_Yexic.bam Coverage
[0 - 10.00]
0184864976_Illumina_P5-Zecan_Yexic.bam
2
Sequence
RefSeq Genes
ATTACCTCAGAAGAATTCTATTATTTCTAATACACACACACACACACACACACACACACA
ESR1
6 tracks loaded
chr6:152,333,154
411M of 684M

Hap name: null
Dist: 0
Read name = 49535978
Sample = MBC_288.TM
Library = 0184864976_Illumina_P5-Zecan_P7-Yexic
Read group = C86U5.8
Read length = 76bp

Mapping = Primary @ MAPQ 60
Reference span = chr6:152,333,116-152,333,193 (-) = 78bp
Cigar = 37M2D39M
Clipping = None

Mate is mapped = yes
Mate start = chr6:152333048 (+)
Insert size = -145
Second in pair
Pair orientation = F1R2

NH = 1
NM = 2
Hidden tags: RG

2 bp deletion

Genome Variant Analysis: Tools to Predict SNP/SNV/INDEL

1. GATK4 (https://software.broadinstitute.org/gatk/)

   a. `HaplotypeCaller`

   - Call germline SNPs and INDELs using local reassembly of haplotypes
   - Variant Quality Score Recalibration (VQSR)
     - VariantRecalibrator + ApplyVQSR

   b. `Mutect2`

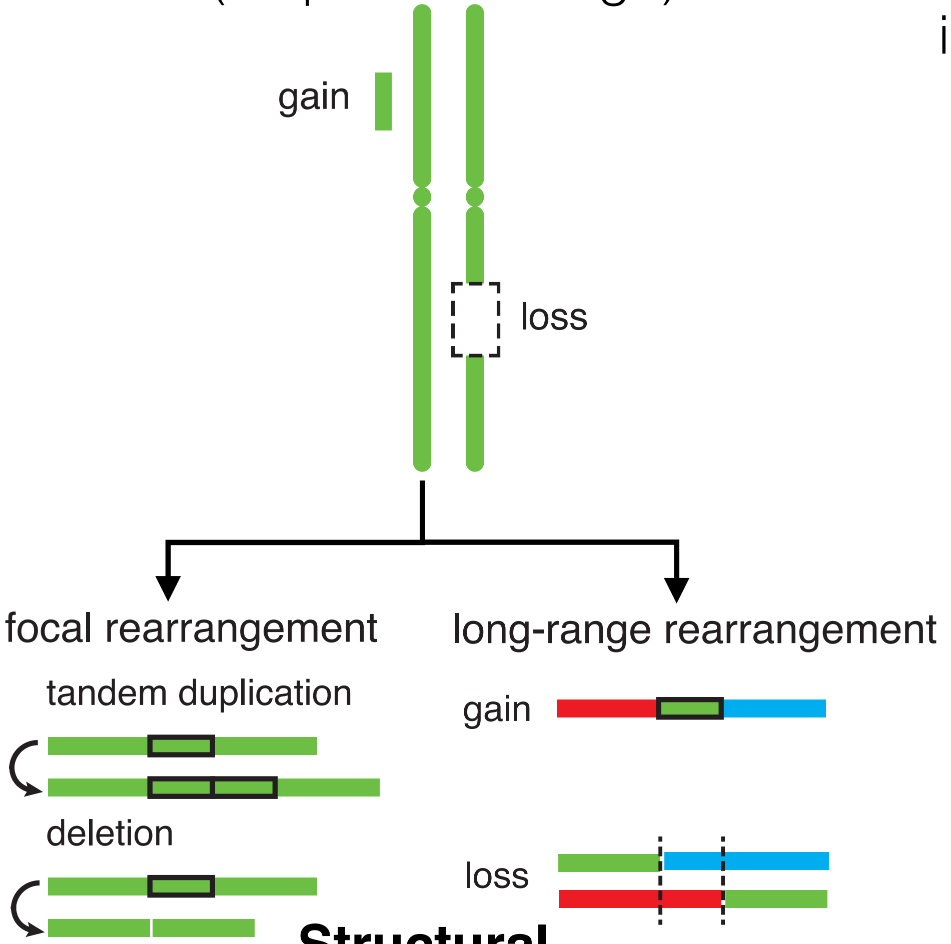
   - Call somatic SNVs using with tumor and normal pairing
   - https://software.broadinstitute.org/gatk/documentation/tooldocs/4.beta.5/org_broadinstitute_hellbender_tools_walkers_mutect_Mutect2.php

2. Strelka (https://github.com/Illumina/strelka, Kim et al. Nature Methods, 2018)

3. Others: VarScan2, SomaticSniper, MuSE, LoLoPicker, deepSNV, FreeBayes, Platypus, CaVEMan, DeepVariant, JointSNVMix2, ShearWater,

Genome Variant Analysis: Copy Number and Structural Variation

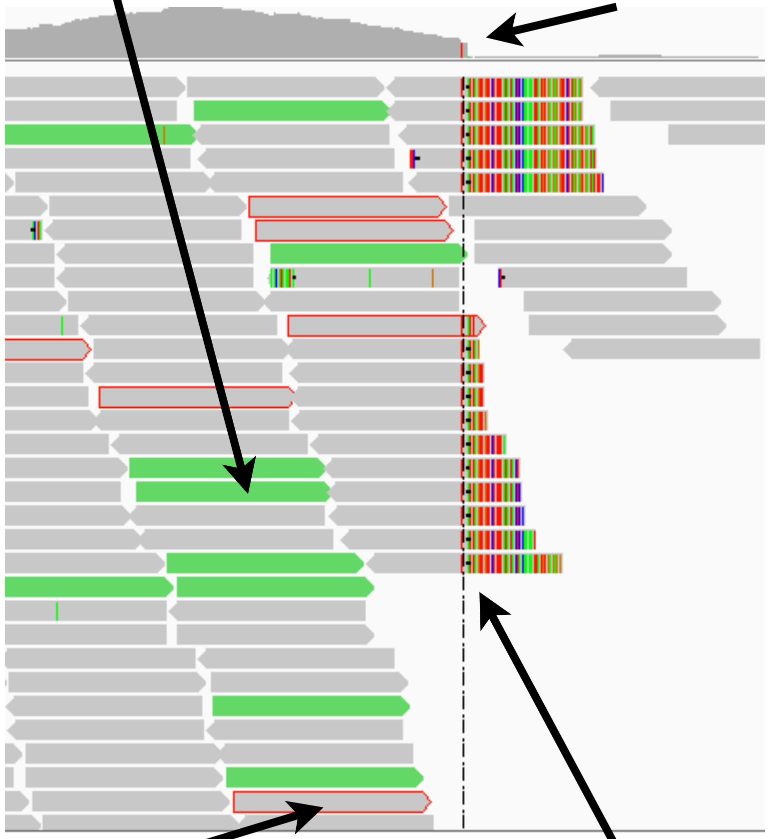
Copy number alterations
(amplitude/dosage)



Structural rearrangements
(location/configuration)

“discordant read pair”
read pairs with aberrant inferred fragment length

“copy number change”
abrupt change in read coverage



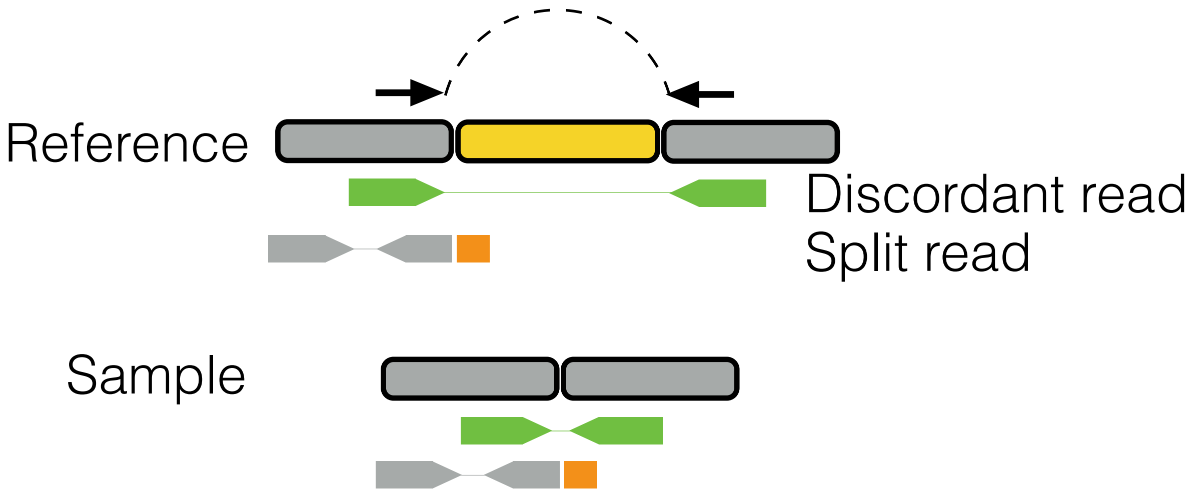
pair-mates unmapped

“split read”
split alignments

Zhang and Pellman. *CSH Symp Quant Biol.* **80**:117-37 (2016)

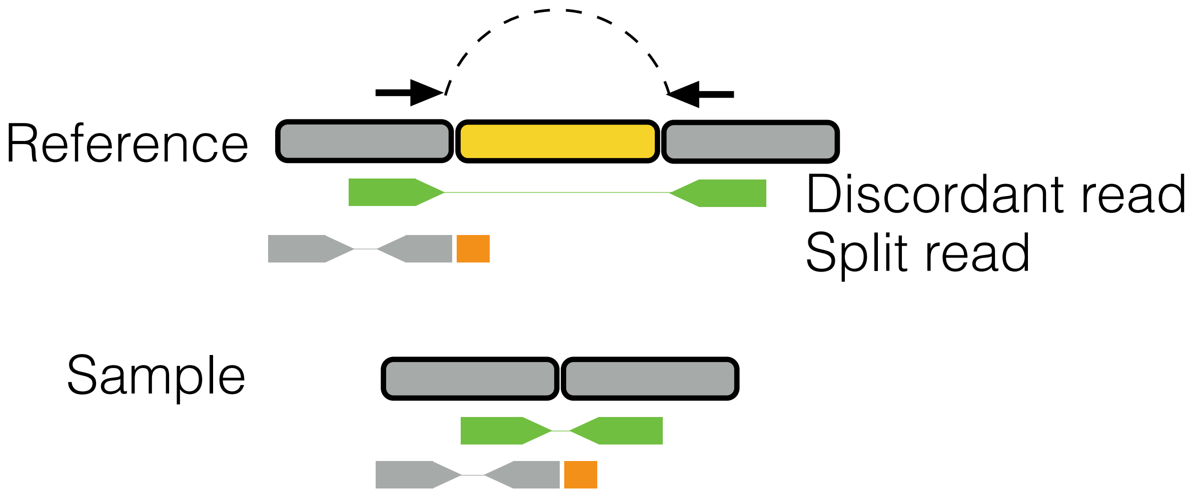
Genome Variant Analysis: Structural Variation

Deletion



Genome Variant Analysis: Structural Variation

Deletion



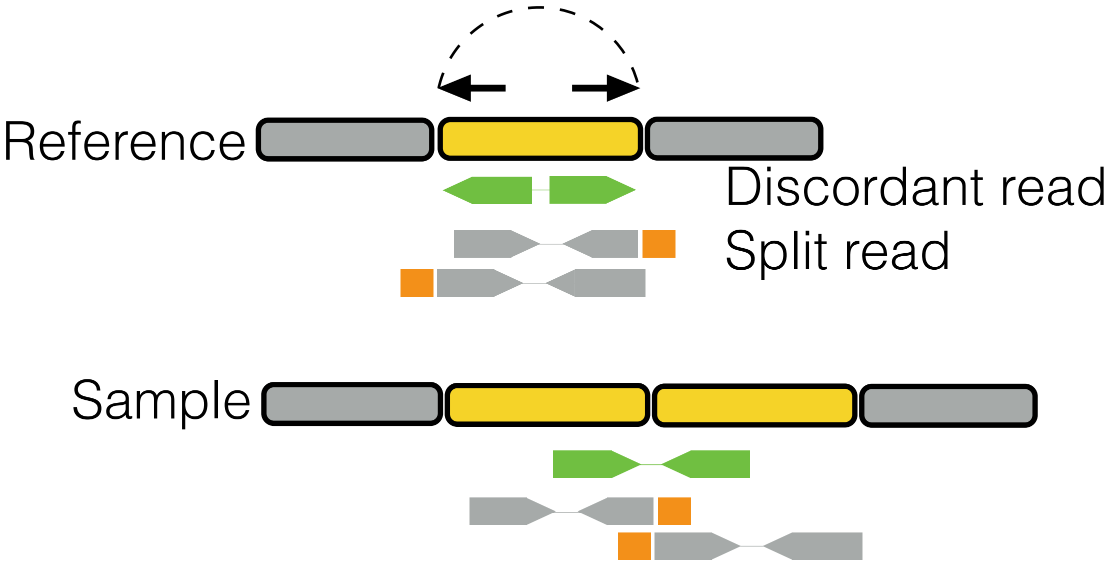
Reference

Discordant read

Split read

Sample

Tandem Duplication



Reference

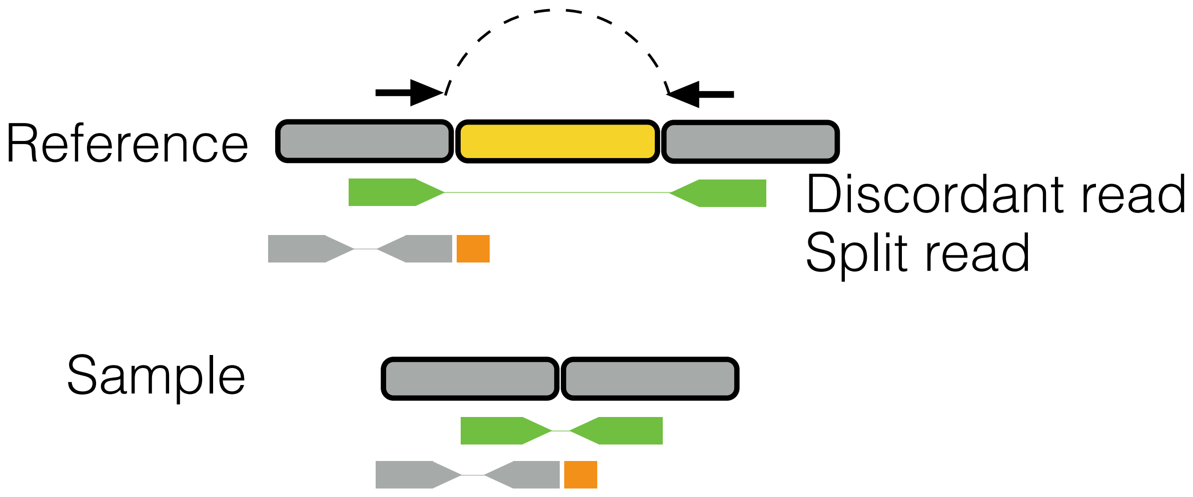
Discordant read

Split read

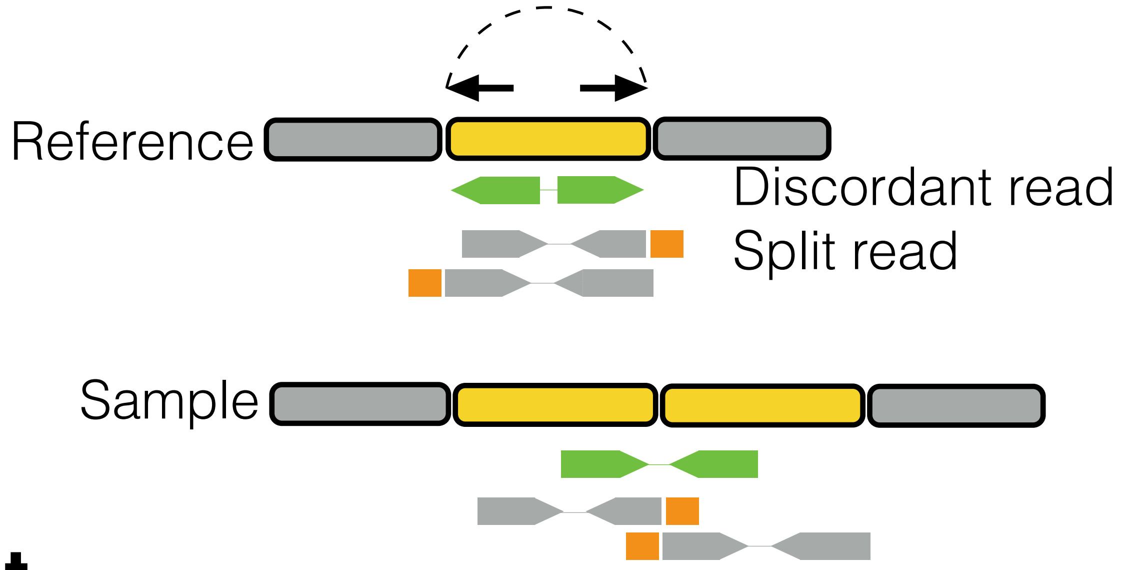
Sample

Genome Variant Analysis: Structural Variation

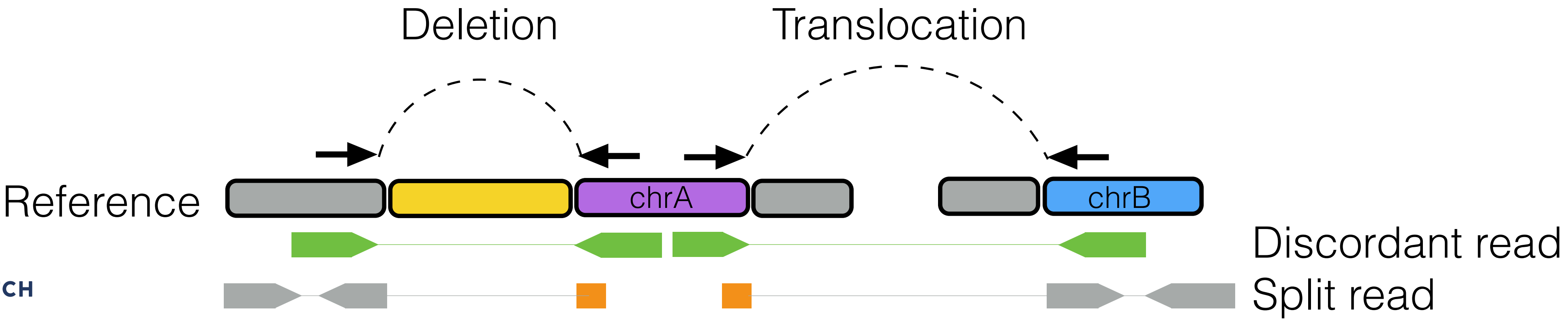
Deletion



Tandem Duplication



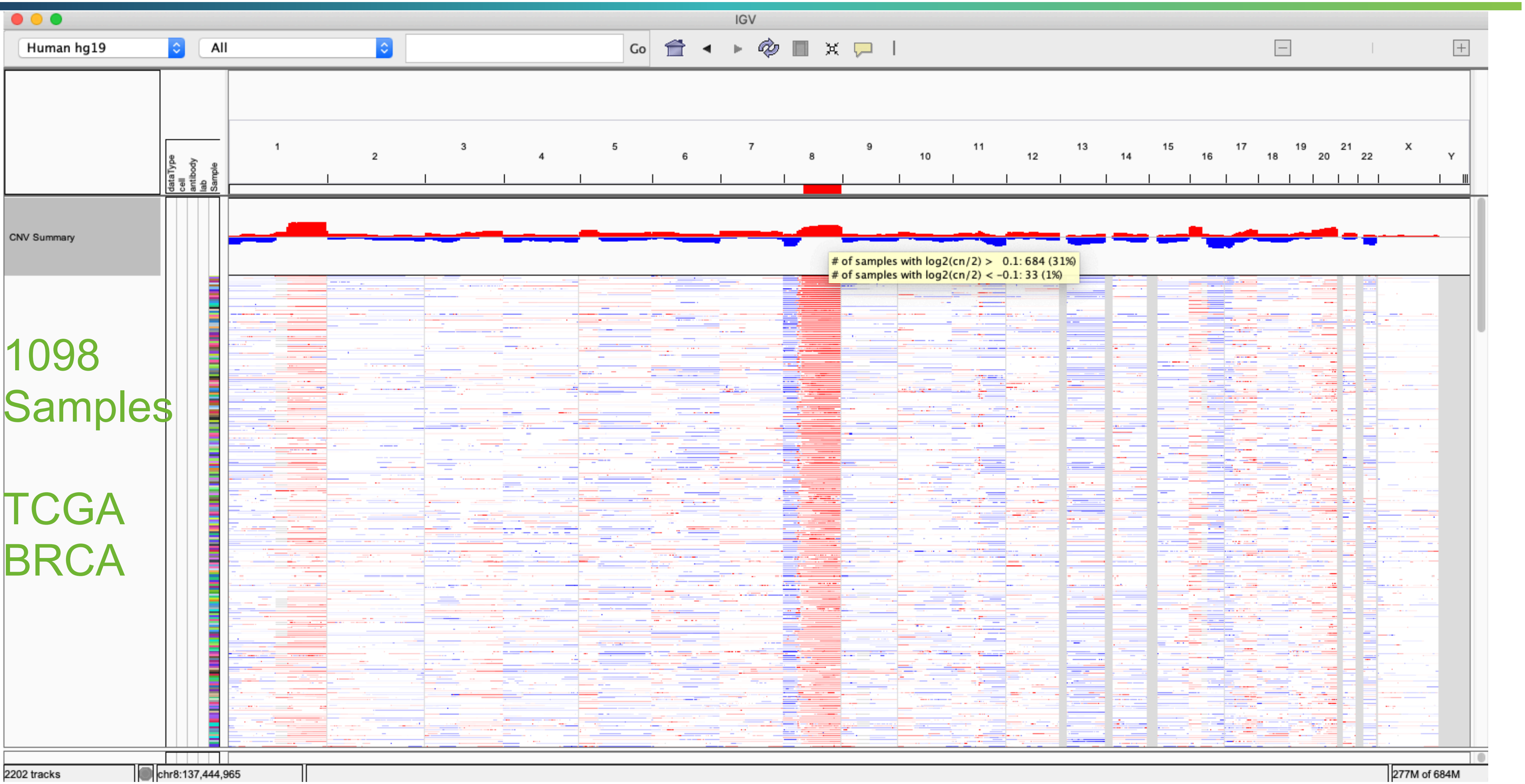
Complex Event



Genome Variant Analysis: Tools to Predict SVs

1. Germline SV
   - GATK4
   - LUMPY (https://github.com/arq5x/lumpy-sv)
   - DELLY (https://github.com/dellytools/delly)
   - Manta (https://github.com/Illumina/manta)
2. Somatic SV
   - BreakDancer (https://github.com/genome/breakdancer)
   - SvABA (https://github.com/walaj/svaba)
3. Others: Comparison of 69 SV tools (Kosugi et al. *Genome Biol*, 2019)

Genome Variant Analysis: Copy Number Variation



# of samples with log2(cn/2) > 0.1: 684 (31%)
# of samples with log2(cn/2) < -0.1: 33 (1%)

1098 Samples

TCGA BRCA

http://firebrowse.org/?cohort=BRCA
https://portal.gdc.cancer.gov/projects/TCGA-BRCA

Genome Variant Analysis: Tools to Predict CNVs

1. Germline CNV
    - GATK4
    - DNAcopy (https://github.com/veseshan/DNAcopy)
    - Others: cn.MOPS, VarScan2
2. Somatic CNV for Cancer
    - ASCAT (https://github.com/Crick-CancerGenomics/ascat)
    - ABSOLUTE (https://software.broadinstitute.org/cancer/cga/absolute)
    - TITAN (https://github.com/gavinha/TitanCNA)
    - Battenberg (https://github.com/cancerit/cgpBattenberg)
    - Others: CNVkit, Sequenza, ichorCNA, HMMcopy

Genome Variant Analysis: Common Variant File Formats

a. Variant Call Format (VCF)

- http://samtools.github.io/hts-specs/VCFv4.2.pdf
- Used mostly for SNV/SNP, INDEL, and SV

b. Mutation Annotation Format (MAF)

- https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/
- http://software.broadinstitute.org/software/igv/MutationData
- Tab-delimited format containing columns for mutation information and annotations
- Used primarily for SNV/SNP and INDEL data

c. Browser Embedded Data (BED)

  a. https://bedtools.readthedocs.io/

  b. Used for any genomic features/region and annotations, including CNV and SV (BEDPE)

d. Others

  a. http://genome.ucsc.edu/FAQ/FAQformat

  b. GFF, WIG/bigWIG, etc.

Genome Variant Analysis: Variant Call Format (VCF)

http://samtools.github.io/hts-specs/VCFv4.2.pdf

a. Header information

```
##fileformat=VCFv4.2
##GATKCommandLine=<ID=HaplotypeCaller,CommandLine=“HaplotypeCaller">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in
the VCF specification">
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="ID of Phase Set for Variant">
##FILTER=<ID=PASS,Description="All filters passed">
##FILTER=<ID=LowQual,Description="Low quality”>
```

b. Variant record

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample_1
chr1	11542	.	A	T	49.77	PASS	AC=1;AF=0.5;AN=2;DP=4	GT:AD:DP:GQ:PL:PS	0|1:2,2:4:78:78,0,78

Genome Variant Analysis: Variant Call Format (VCF)

http://samtools.github.io/hts-specs/VCFv4.2.pdf

c. Genotype Field (GT)

  a. 0=Reference allele, 1=Alternate allele

  b. 0/1=heterozygous, 0/0 or 1/1=homozygous

  c. 0|1 or 1|0 = heterozygous (phased)

SNP	S1	S2	S3
Reference	A	T	G
Haplotype 1	A	C	G
Haplotype 2	C	T	A
GT (unphased)	0/1	0/1	0/1
GT (phased)	0|1	1|0	??

Haplotype 1 | Haplotype 2

A(0)|C(1)

C(1)|T(0)

?? | ??

Genome Variant Analysis: Variant Annotation Tools

ANNOVAR (http://annovar.openbioinformatics.org)

SnpEff (http://snpeff.sourceforge.net)

SIFT (https://sift.bii.a-star.edu.sg/) - predict amino acid substitution effects on protein function

GATK VariantAnnotator

VariantAnnotation R Package (https://bioconductor.org/packages/release/bioc/html/VariantAnnotation.html)

Variant Annotation Integrator (UCSC, https://genome.ucsc.edu/cgi-bin/hgVai)

BioMart (http://www.biomart.org/)

Genome Variant Analysis: Variant Databases

1000 Genomes Project (https://www.internationalgenome.org/)

dbSNP (https://www.ncbi.nlm.nih.gov/snp/)

dbVar (https://www.ncbi.nlm.nih.gov/dbvar/)

ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/)

Exome Aggregation Consortium (ExAC, http://exac.broadinstitute.org/)

- Lek et al. Nature, 536, 285-91 (2016)

Genome Aggregation Database (gnomAD, https://gnomad.broadinstitute.org/)

- Karczewski et al. bioRxiv (2019)

Genome Data Commons (https://portal.gdc.cancer.gov/)

Preparation for Lecture 8: Genomic Data Analysis in R

1. Lecture 7 and 8 Data can be downloaded here:

https://www.dropbox.com/sh/zoitjnobgp7l7c2/AABBIpTQcNA4lWYOFnV5dlMKa?dl=0

2. Required Installation

   a. R Studio (R version 3.6.1)

   b. R Bioconductor packages

   - VariantAnnotation_1.31.3
   - GenomicRanges_1.37.7
   - Rsamtools_2.1.6
   - BSgenome.Hsapiens.UCSC.hg19_1.4.0
   - GenomicFeatures_1.37.1
   - TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
   - biomaRt_2.41.0

   c. R Cran packages

   - data.table_1.12.2