

# Introduction to Sequencing Data Analysis

## Lecture 15

Thursday, November 15, 2022

**Gavin Ha, Ph.D.**

Assistant Professor

Herbold Computational Biology Program

Divisions of Public Health Sciences and Human Biology

 [@GavinHa](https://twitter.com/GavinHa)  
[GavinHaLab.org](https://GavinHaLab.org)

# Overview

**I. Introduction to sequence data and resources**

**II. Tools for analyzing and visualizing sequencing data**

# 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 (Background; Next Lecture)

- 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/>)

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

The Cancer Genome Atlas (TCGA) (<https://portal.gdc.cancer.gov/>)

International Cancer Genome Consortium (ICGC) (<https://icgc.org/>)

**IGSR: The International Genome Sample Resource**

Providing ongoing support for the 1000 Genomes Project data



**UK10K**

*Rare Genetic Variants in Health and Disease*



#100kThankYous



International  
Cancer Genome  
Consortium

# Sequence Data: Databases and Online Resources

## Common Repositories/Databases for human 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

### 4. ICGC Data Portal

- <https://dcc.icgc.org/>

# Sequence Data: Databases and Online Resources

Harmonized Cancer Datasets

## Genomic Data Commons Data Portal

Get Started by Exploring:



Projects



Exploration



Analysis



Repository

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

### Data Portal Summary

[Data Release 19.0 - September 17, 2019](#)

PROJECTS



53

PRIMARY SITES



67

CASES



37,075

FILES



427,407

GENES

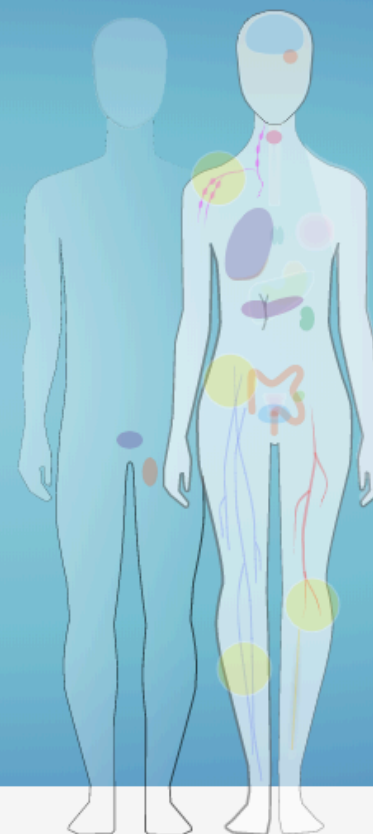


22,872

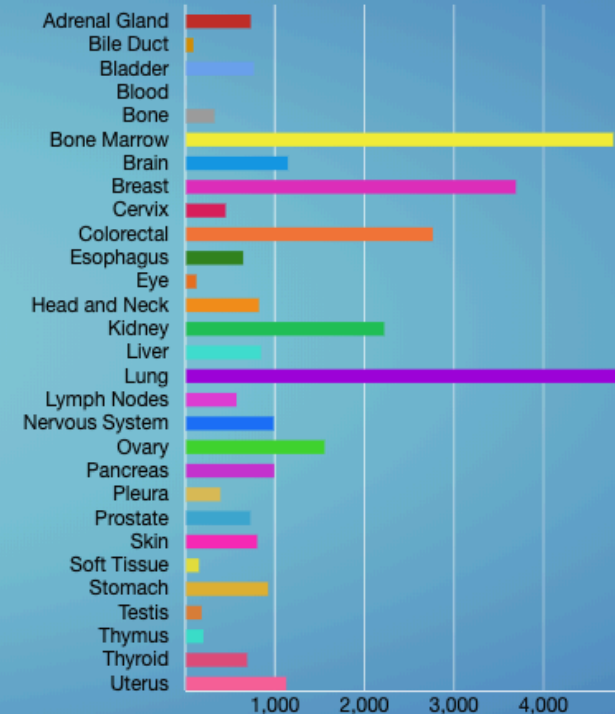
MUTATIONS



3,142,246




### Cases by Major Primary Site



### 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

 **ICGC Data Portal**

[Cancer Projects](#) [Advanced Search](#) [Data Analysis](#) [DCC Data Releases](#) [Data Repositories](#)

Cancer genomics data sets visualization, analysis and download.

[Search](#)

e.g. BRAF, KRAS G12D, DO35100, MU7870, FI998, apoptosis, Cancer Gene Census, imatinib, GO:0016049

Advanced Search

[By donors](#) [By genes](#) [By mutations](#)

**Data Release 28** March 27th, 2019

|                                  |            |
|----------------------------------|------------|
| Cancer projects                  | 86         |
| Cancer primary sites             | 22         |
| Donor with molecular data in DCC | 22,330     |
| Total Donors                     | 24,289     |
| Simple somatic mutations         | 81,782,588 |

[Download Release](#)

# Sequence Data: Databases and Online Resources

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


The screenshot shows the NCBI GEO website interface. At the top left is the NCBI logo. To its right is the GEO logo with the text "Gene Expression Omnibus". Below these are navigation links: HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, and Email GEO. A breadcrumb trail reads "NCBI > GEO > Accession Display". On the right, it says "Not logged in | Login". A help message states: "GEO help: Mouse over screen elements for information." Below this is a search bar with fields for Scope (Self), Format (HTML), Amount (Quick), and GEO accession (GSE71378), followed by a GO button. The main content area is titled "Series GSE71378" and includes a link "Query DataSets for GSE71378". The series details are as follows:

|                 |   |
|-----------------|---|
| 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        | <a href="#">Homo sapiens</a>  |
| 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



HOME | SEARCH | SITE MAP

NCBI > GEO > Accession

[GEO help: Mouse over screen](#)

Scope:

Series **GSE71378**

Status

Title

Organism

Experiment type

Summary

Pub

Cell info

Hon

Gen

Nuc

Contributor(s)  
Citation(s)

Shendure J  
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  
Last update date  
Contact name  
Organization name  
Department  
Lab  
Street address  
City  
State/province  
ZIP/Postal code  
Country

Jul 27, 2015  
May 15, 2019  
Jay Shendure  
University of Washington  
Genome Sciences  
Shendure  
3720 15th Ave NE  
Seattle  
WA  
98195-5065  
USA

Platforms (1)  
Samples (60)  
[More...](#)

[GPL11154](#) Illumina HiSeq 2000 (Homo sapiens)  
[GSM1833219](#) BH01  
[GSM1833220](#) IA01  
[GSM1833221](#) IA02

Relations

BioProject  
SRA

[PRJNA291063](#)  
[SRP061633](#)

Download family

[SOFT formatted family file\(s\)](#)  
[MINiML formatted family file\(s\)](#)  
[Series Matrix File\(s\)](#)

Format  
SOFT [?](#)  
MINiML [?](#)  
TXT [?](#)

| Supplementary file | Size     | Download                                     | File type/resource |
|--------------------|----------|--|--------------------|
| GSE71378_BH01.bb   | 311.8 Mb | <a href="#">(ftp)</a> <a href="#">(http)</a> | BB                 |
| GSE71378_CA01.bb   | 325.0 Mb | <a href="#">(ftp)</a> <a href="#">(http)</a> | BB                 |
| GSE71378_CH01.bb   | 319.7 Mb | <a href="#">(ftp)</a> <a href="#">(http)</a> | BB                 |
| GSE71378_IH01.bb   | 296.6 Mb | <a href="#">(ftp)</a> <a href="#">(http)</a> | BB                 |
| GSE71378_IH02.bb   | 248.3 Mb | <a href="#">(ftp)</a> <a href="#">(http)</a> | BB                 |

[SRA Run Selector](#) [?](#)

1378)

NAME | Email GEO

Not logged in | [Login](#) [?](#)

78

that

ind

# Sequence Data: Databases and Online Resources

Sequence Read

1378)

NCBI

SRA Run Selector

?

⚙

🔍

Log in

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 | <div>RunInfo Table or Accession List</div> |
| Selected | 0    | 0         | 0      | <div>RunInfo Table or Accession List</div> |

Found 60 Items

Search...

🔍

Clear

<

1

1

2

>

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

# 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) or **CRAM** (.cram)

# 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_cfdDNA.bam
```

For your reference.

# 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](#)]

For your reference.

# 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>)

## 3. Sequencing metrics and Processing

- SAMtools
- Genomic Analysis Toolkit (GATK) and Picard Tools

# 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>)

## 3. Sequencing metrics and Processing

- SAMtools
- Genomic Analysis Toolkit (GATK) and Picard Tools



# 1. Inspecting and Reading BAM Files

**SAMtools** (<http://www.htslib.org/>)

**Demo & Exercise**



# Sequence Data: Inspecting and Reading BAM Files

## SAMtools (<http://www.htslib.org/>)

- Indexing

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

- File operations

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

- Statistics

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

- Viewing

```
# view header information  
samtools view -H BRCA_IDC_cfdDNA.bam  
  
# view aligned reads at chr17:37844393  
samtools view BRCA_IDC_cfdDNA.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
...
```

# 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 17:37844393-37844393
```

```
...
```

```
41976152      163      17      37844359      60      39M      =      37844477
157
ACTCTCCGCTGAAGTCCACACAGTTTAAATTAAAGTTCC .AAAAFFFFFFFFFFFFFF)FAFFFFFFFFFFFFFFFFFFFF
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 17:37844393-37844393
```

Query (Read)

Read

Mate's

... Name

Reference and Position

Reference and Position

41976152

163

17

37844359

60

39M

=

37844477

157

ACTCTCCGCTGAAGTCCACACAGTTTAAATTAAAGTTCC).AAAAFFFFFFFFFFFFFF)FAFFFFFFFFFFFFFFFFFFFFFF

RG:Z:P12.17.7\_Breast NH:i:1 NM:i:0

Read Sequence

# Sequence Data: SAM Format

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

## B. Alignment information

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

| Template Length<br>(Insert Size or<br>Fragment Size) | Flag |    |          | Mapping<br>Quality | CIGAR<br>string |   |          |
|--|------|----|----------|--------------------|-----------------|---|----------|
| 41976152   | 163  | 17 | 37844359 | 60                 | 39M             | = | 37844477 |
| 157  |      |    |          |                    |                 |   |          |

```
ACTCTCCGCTGAAGTCCACACAGTTTAAATTAAAGTTCC .AAAAFFFFFFFFFFFFFF)FAFFFFFFFFFFFFFFFFFFFF  
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 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)

# Exercise: SAMtools

```
# While in dev container
conda activate samtools
# Go to directory where class data has been downloaded
cd myDataDirectory
```

1. Run samtools view header command on BRCA\_IDC\_cfDNA.bam
  - a. What is the read group (@RG) ID?
  
2. Run samtools view at 17:7579472-7579472
  - a. What is the insert size?



# 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>)

## 3. Sequencing metrics and Processing

- SAMtools
- Genomic Analysis Toolkit (GATK) and Picard Tools



## 2. Interactive Visualization

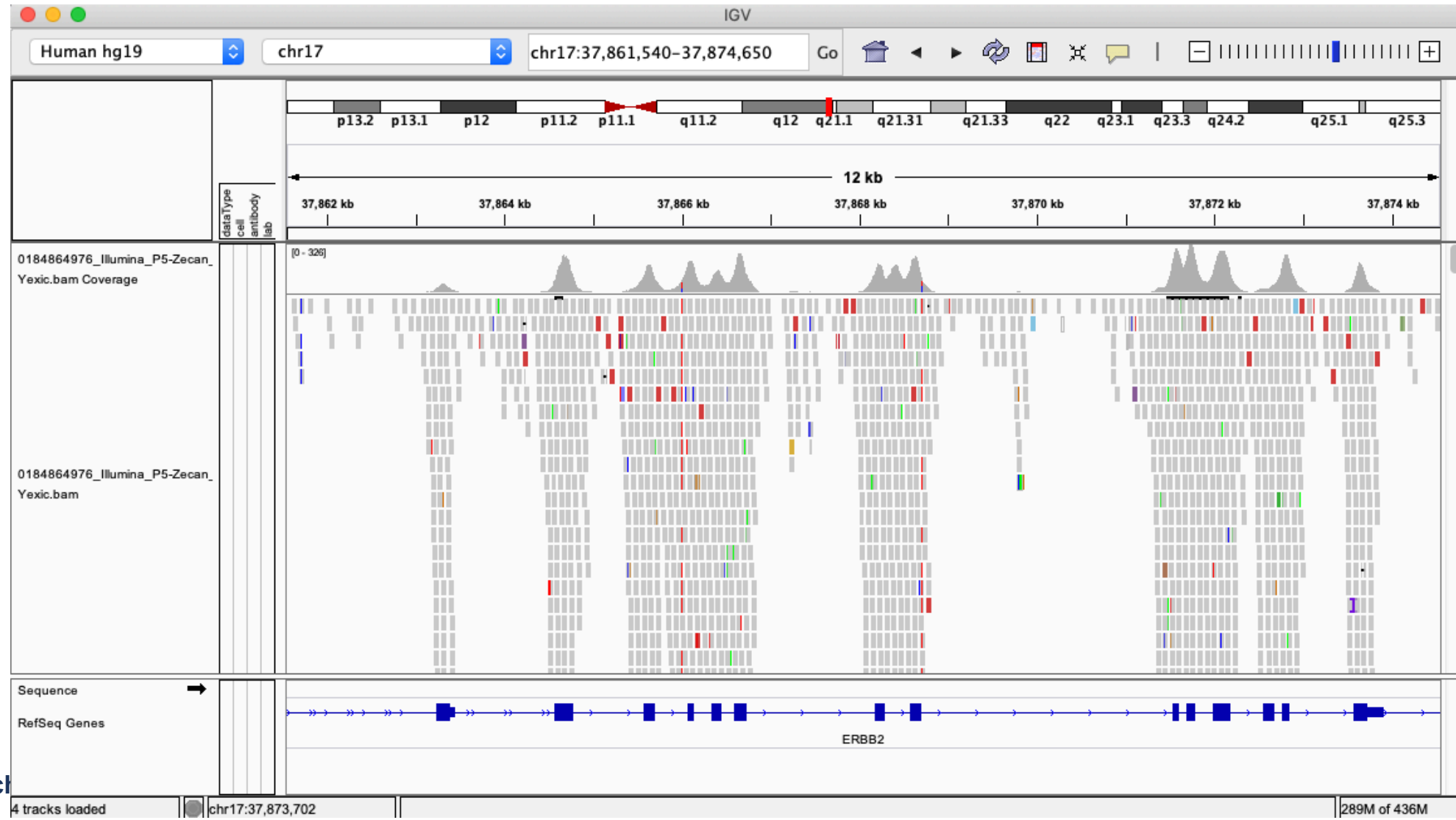
### Integrative Genomics Viewer

(<https://software.broadinstitute.org/software/igv>)

### Demo + Exercise

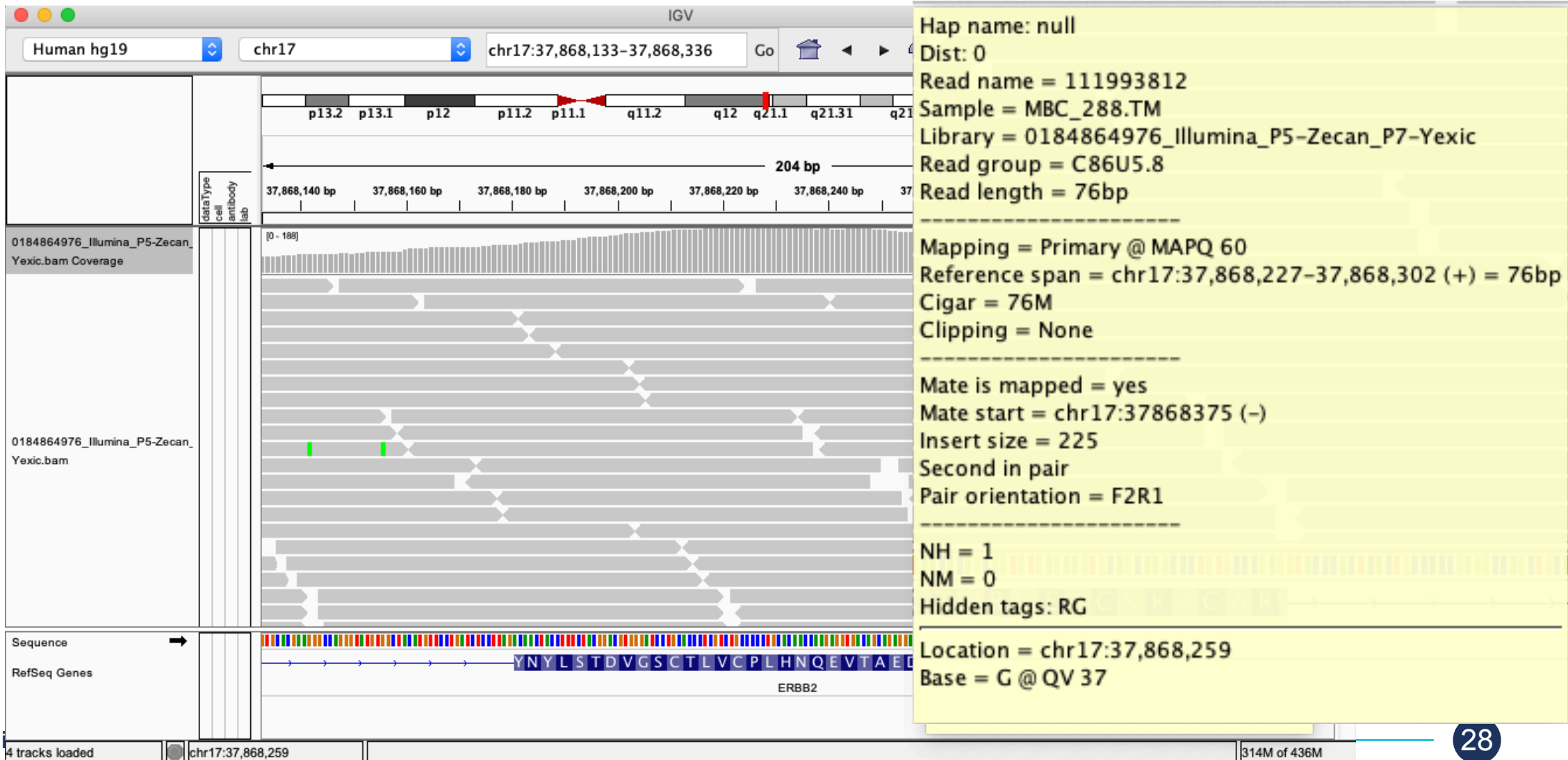
# Tools for Sequencing Data: Interactive Visualization

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



# Tools for Sequencing Data: Interactive Visualization

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



# Exercise: IGV

## Instructions:

Launch IGV-Web (<https://igv.org/app/>).

Tracks > Local File > select both BRCA\_IDC\_cfDNA.bam and BRCA\_IDC\_cfDNA.bam.bai

## Questions:

1. Go to location chr17:7,579,517
  - a. Which gene and exon # is at this location?
  - b. How many reads match the reference? How many don't? What are the nucleotides bases?
2. Go to location chr13:32,912,062
  - a. Which gene and exon # is at this location?
  - b. What is the “Read length”, “Insert size”, and “CIGAR” for the read found here?
  - c. File > Load from Server > Annotations > Variation and Repeats > *check* dbSNP
    - i. What is the “Name” (rs ID) and “Class” of the SNP located at this position?

# 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>)

## 3. Sequencing metrics and Processing

- SAMtools
- Genomic Analysis Toolkit (GATK) and Picard Tools



# 3. Tools for Sequence Data Processing

## PICARD and GATK

<https://broadinstitute.github.io/picard/>

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

## Demo + Exercise

# Tools for Sequencing Data: Processing

## Picard Tools & GATK4: Best practices

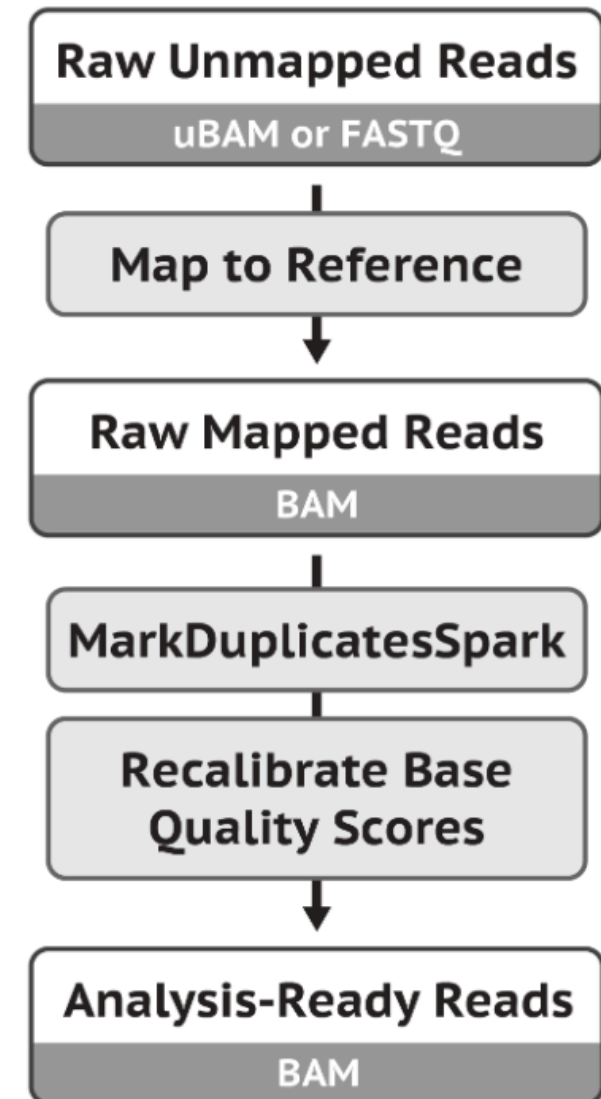
### 1. Mark Duplicates

1. MarkDuplicates + SortSam (Picard)

### 2. Base Quality Score Recalibration (BQSR)

1. BaseRecalibrator (GATK4)
2. ApplyBQSR (GATK4)

```
picard MarkDuplicates \  
INPUT=BRCA_IDC_cfdNA.bam \  
REMOVE_DUPLICATES=false \  
OUTPUT=BRCA_IDC_cfdNA.marked_duplicates.bam \  
METRIC_FILE=BRCA_IDC_cfdNA.markDupMetrics.txt
```





# 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) CollectWgsMetrics

```
picard CollectWgsMetrics \  
INPUT=BRCA_IDC_cfdNA.bam \  
OUTPUT=BRCA_IDC_cfdNA.alignMetrics.txt \  
REFERENCE_SEQUENCE=hs37d5.fa \  
VALIDATION_STRINGENCY=LENIENT
```

| 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 |

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

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

# Exercise: PICARD

Run `CollectAlignmentSummaryMetrics` for `BRCA_IDC_cfDNA.bam`

```
#While in Dev container
conda activate Picard
# Go to directory where class data has been downloaded
cd myDataDirectory
# Run Picard command
picard CollectAlignmentSummaryMetrics \
. . .
```

How many `PF_READS_ALIGNED` for PAIR Category?

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

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

## Python

- PySam

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

## R and Bioconductor

- 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>

# Lecture 15

R Bioconductor packages:

- VariantAnnotation
- GenomicRanges
- plyranges

Download data:

[https://drive.google.com/drive/folders/  
13jM29nhzELyThKQXI27MrRXxbziVunQr?usp=sharing](https://drive.google.com/drive/folders/13jM29nhzELyThKQXI27MrRXxbziVunQr?usp=sharing)