Fred Hutch
Cancer Center

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
GavinHaLab.org

Overview

I. Introduction to sequence data and resources

II. Tools for analyzing and visualizing sequencing data

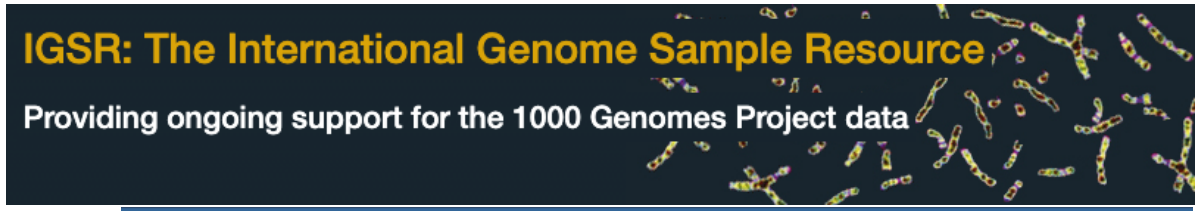
Overview: Learning Objectives

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/)

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

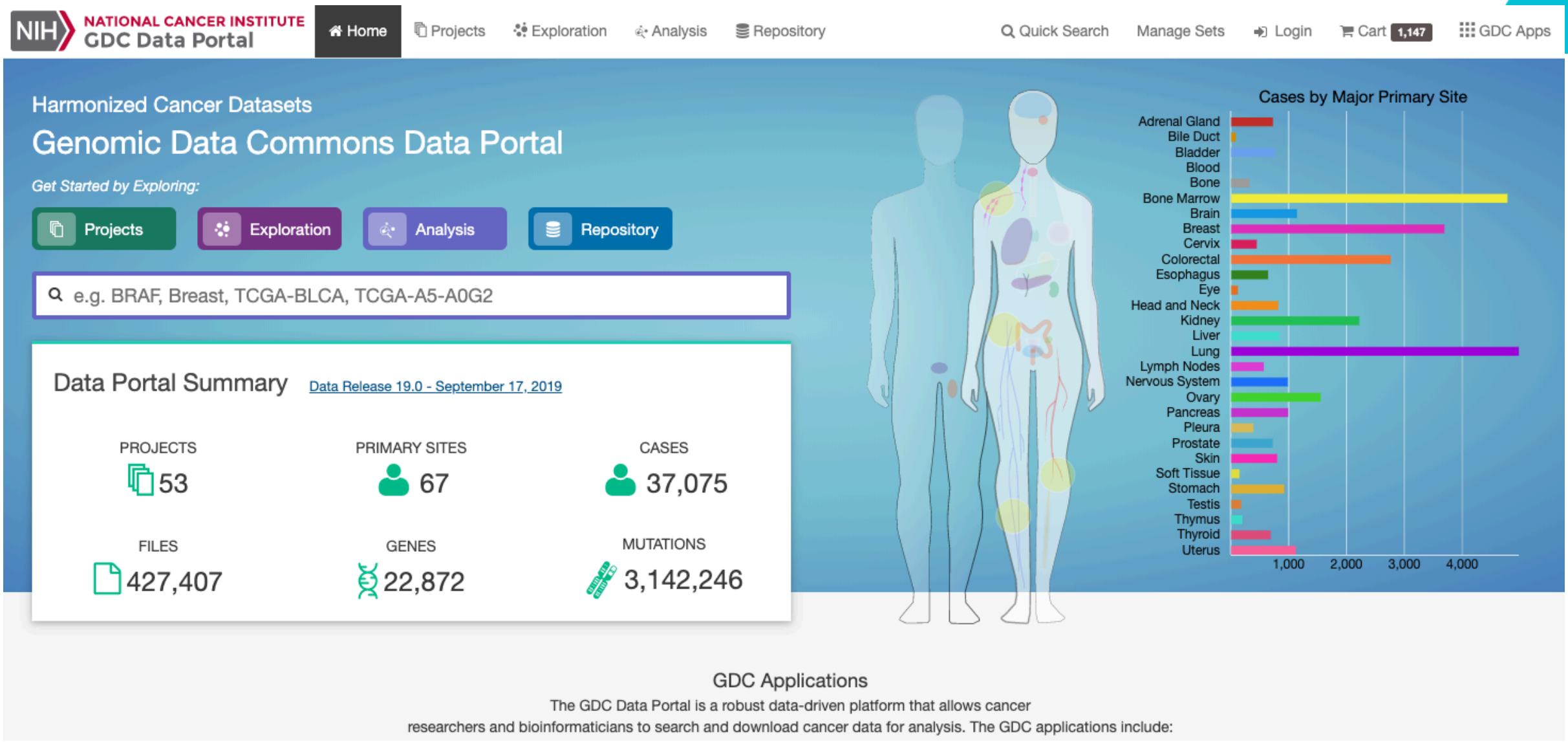
3. NIH NCI Genomic Data Commons (GDC) Data Portal

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

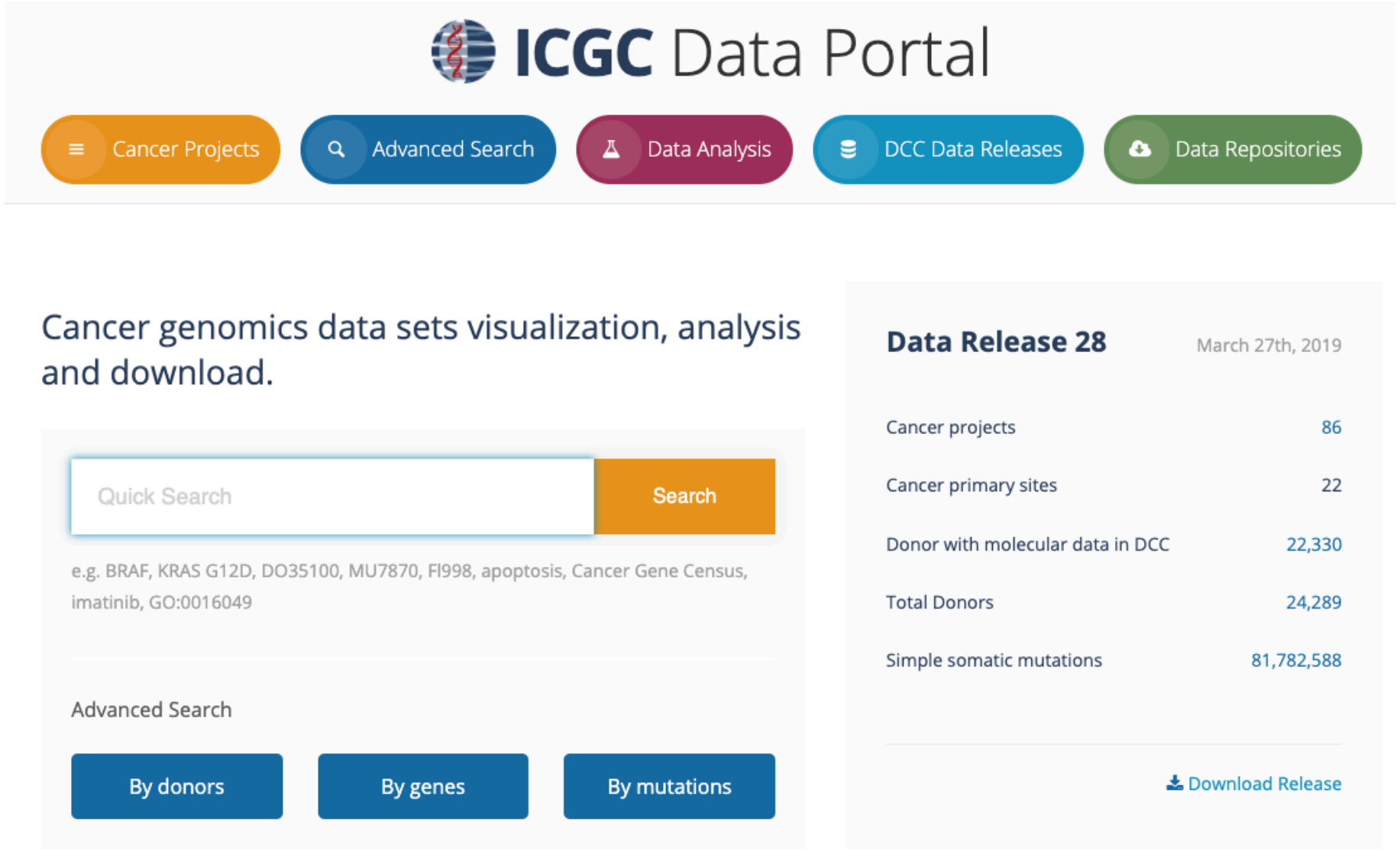
4. ICGC Data Portal

- https://dcc.icgc.org/

Sequence Data: Databases and Online Resources

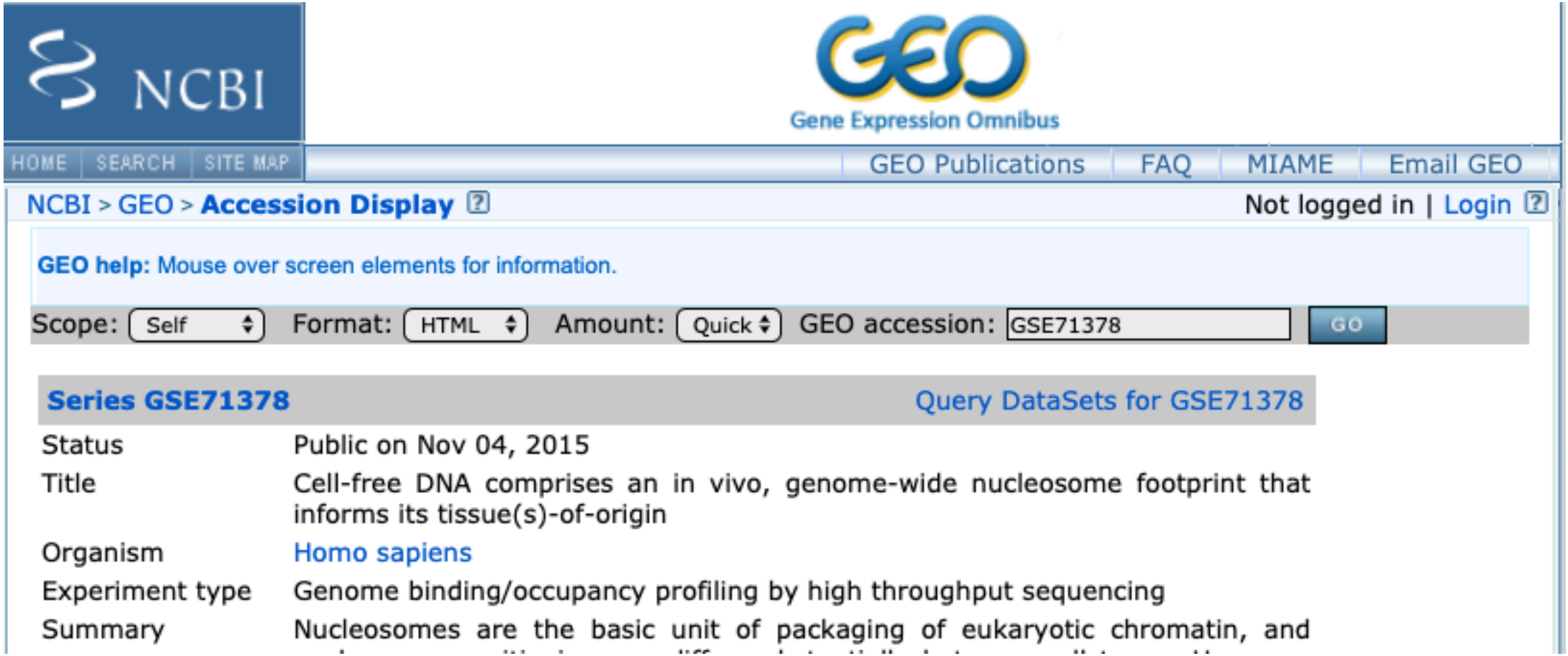


Sequence Data: Databases and Online Resources



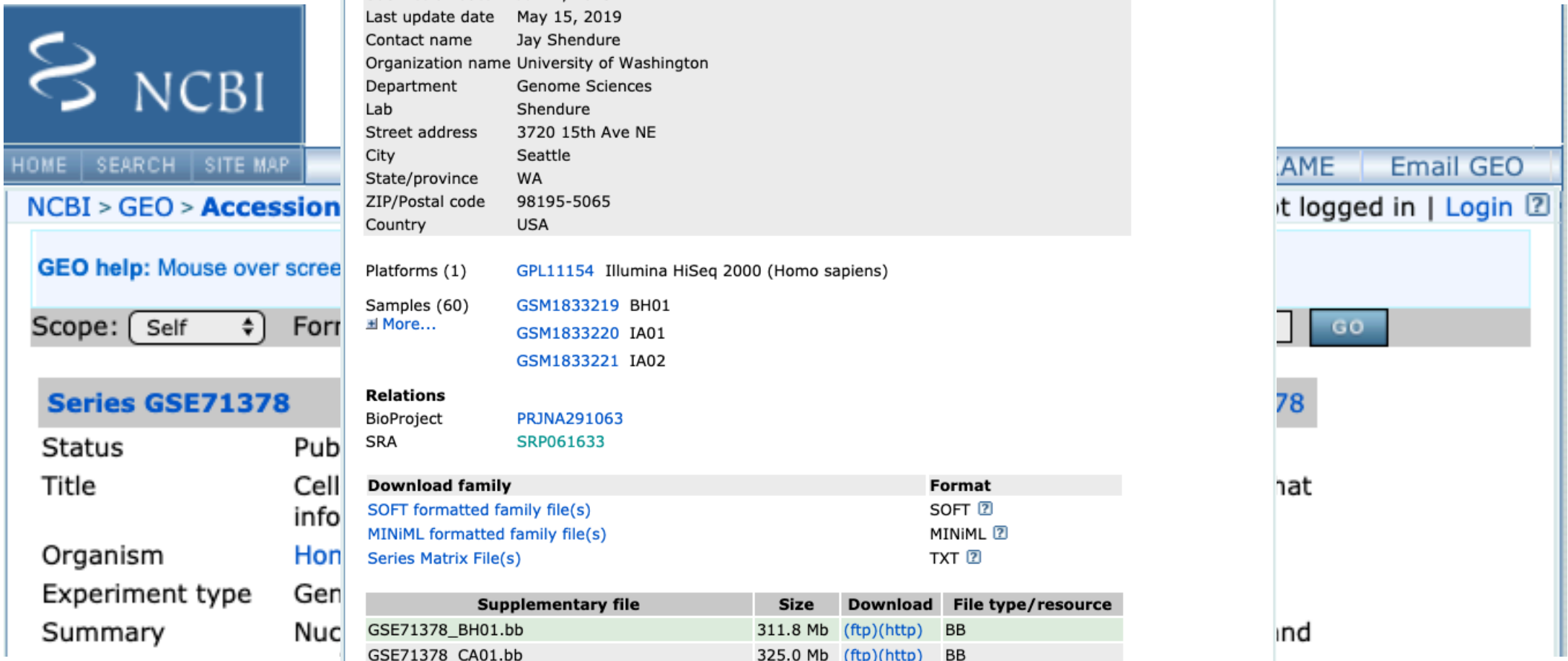
Sequence Data: Databases and Online Resources

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



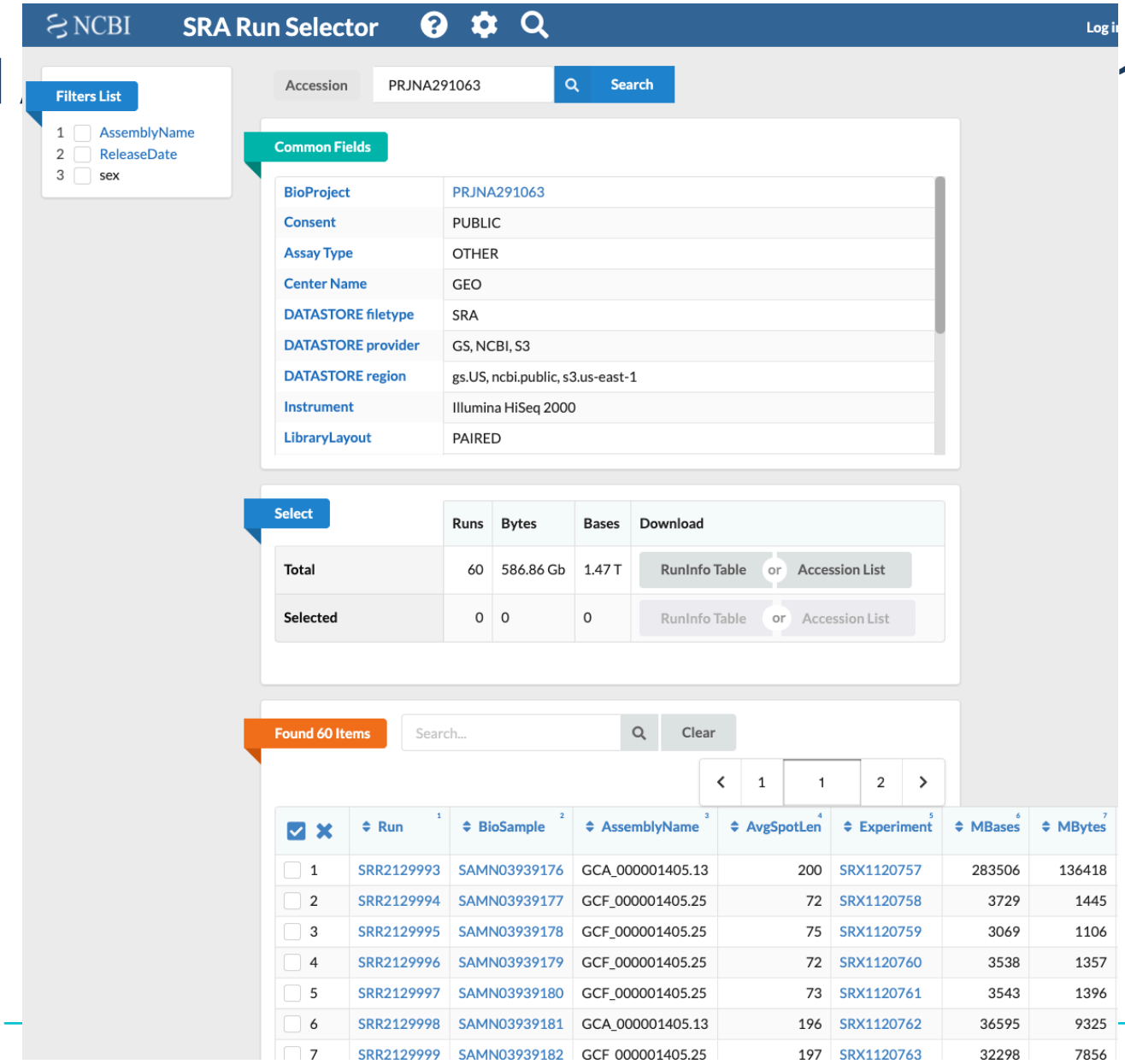
Sequence Data: Databases and Online Resources

Sequence Read … 1378)



Sequence Data: Databases and Online Resources

Sequence Read A 1378)



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_cfDNA.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_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 chr17:37844393
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
...
```

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 .AAAAFFFFFFFFFFF)FAFFFFFFFFFFFFFFFFFF
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) ··· Name | Read Reference and Position | Mate's Reference and Position

```
41976152        163     17      37844359        60      39M     =       37844477
157
ACTCTCCGCTGAAGTCCACACAGTTTAAATTAAAGTTCC .AAAAFFFFFFFFFFF)FAFFFFFFFFFFFFFFFFF
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
--
41976152        163     17      37844359        60      39M     =       37844477
157
ACTCTCCGCTGAAGTCCACACAGTTTAAATTAAAGTTCC .AAAAFFFFFFFFFFF)FAFFFFFFFFFFFFFFFFF
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)

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

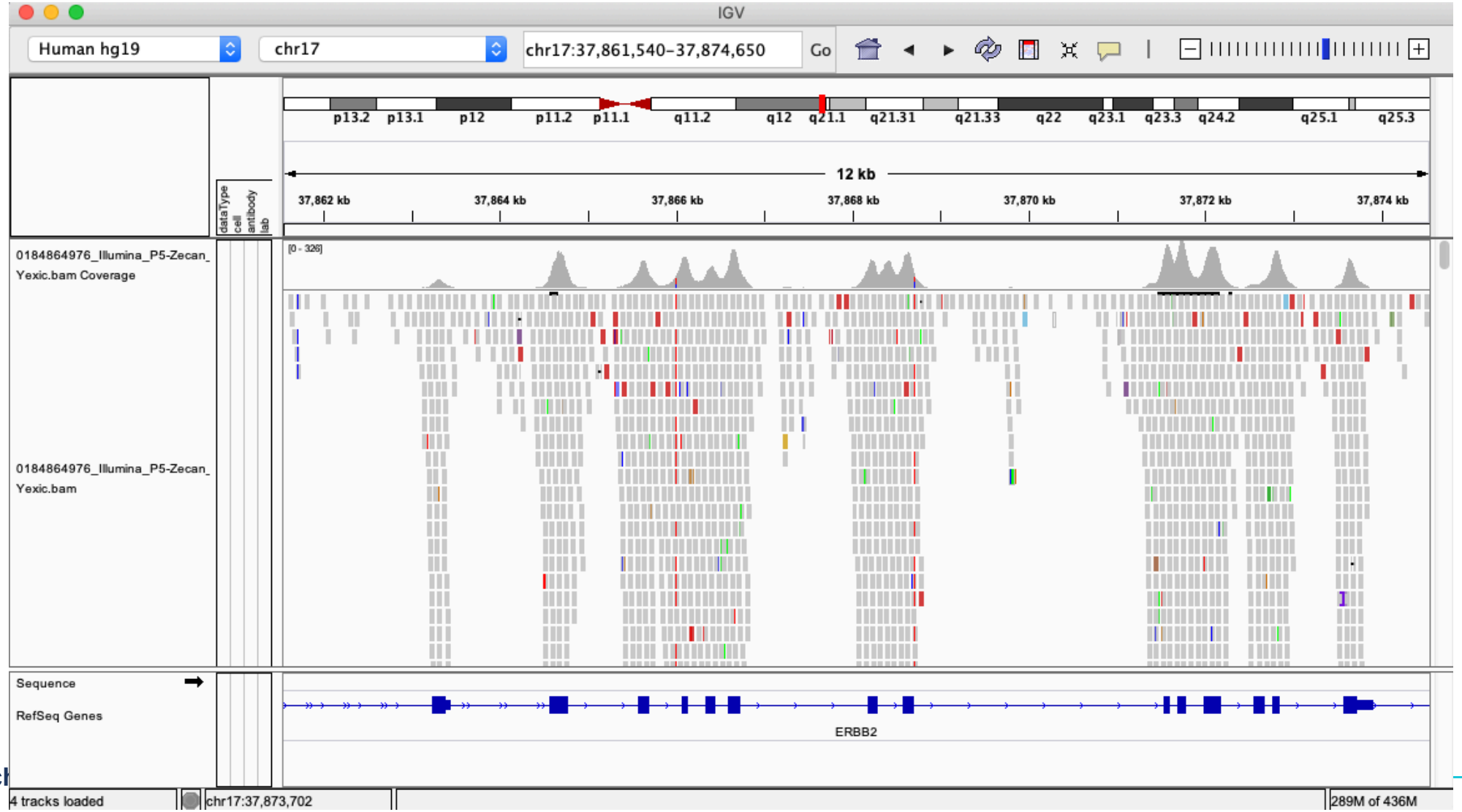
Integrative Genomics Viewer

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

Demo + Exercise

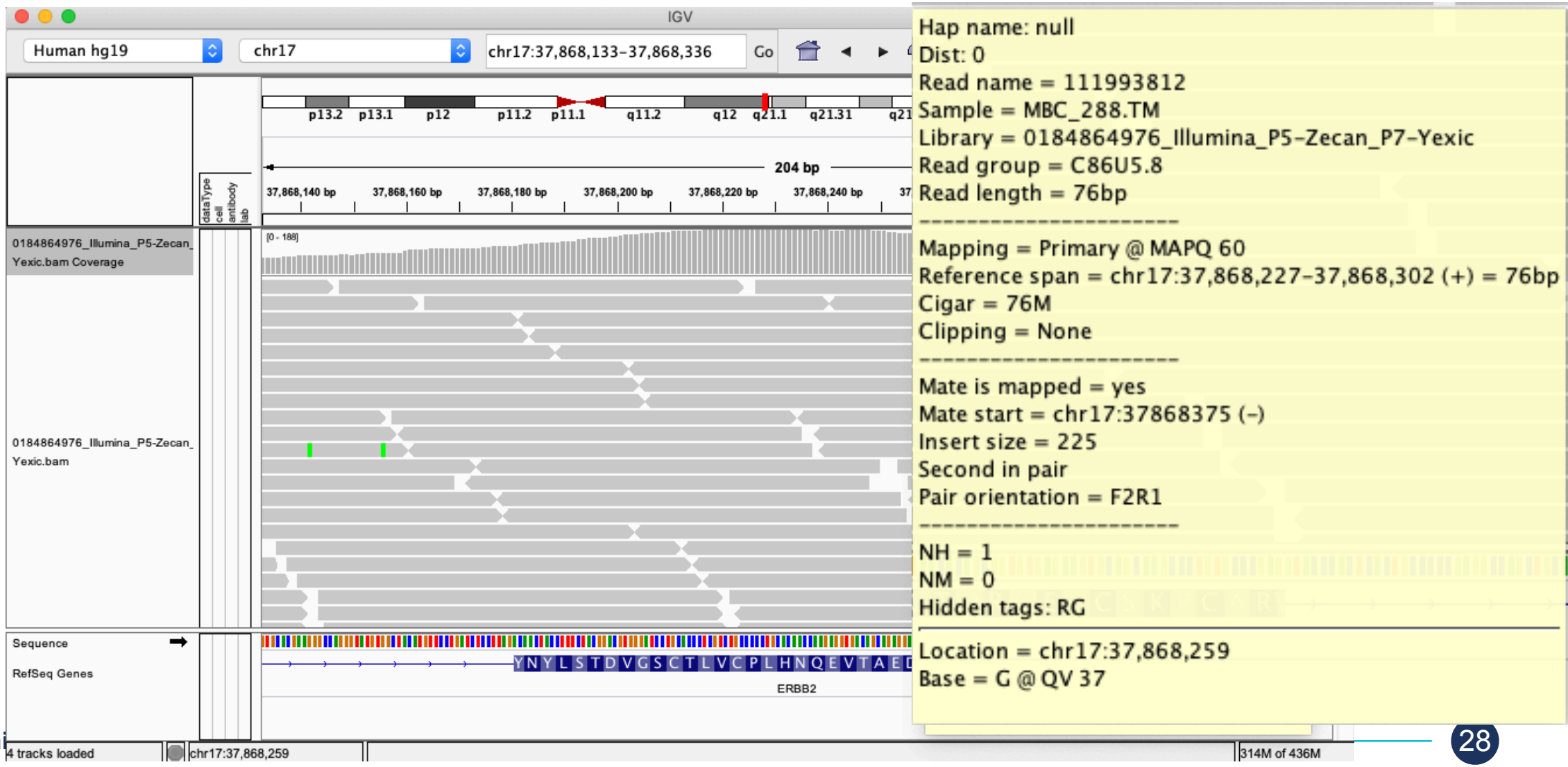
Tools for Sequencing Data: Interactive Visualization

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/

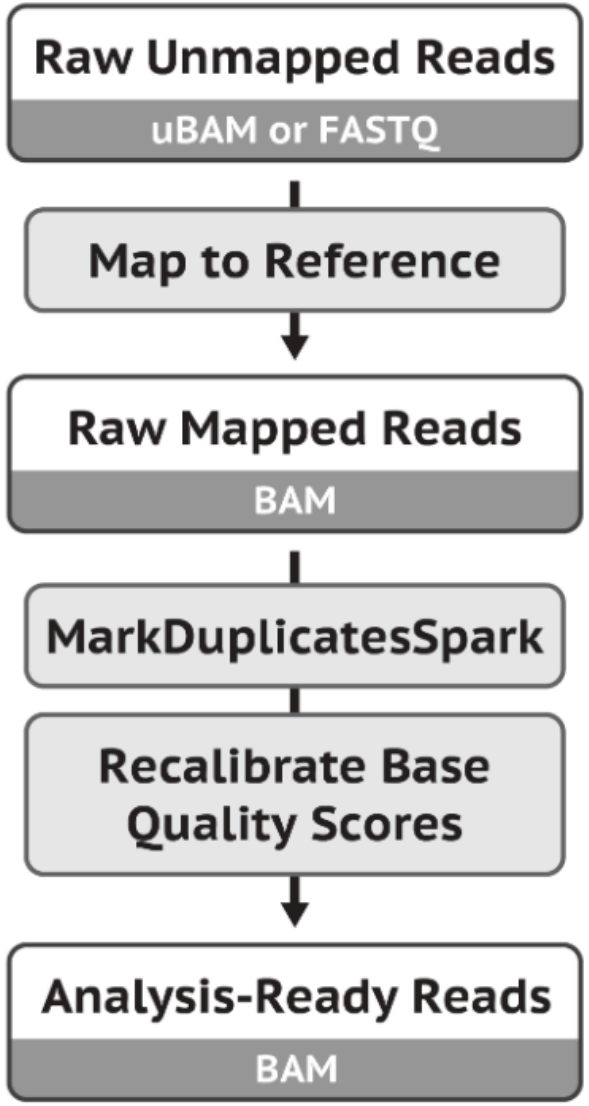
Demo + Exercise

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
```



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) `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

For your reference.

Lecture 15

R Bioconductor packages:

- VariantAnnotation
- GenomicRanges
- plyranges

Download data:

https://drive.google.com/drive/folders/13jM29nhzELyThKQXI27MrRXxbziVunQr?usp=sharing