Genomic Data Analysis in R

Lecture 16

Tuesday, November 24, 2020 @ 1pm

Gavin Ha, Ph.D.
Assistant Professor
Computational Biology Program
Public Health Sciences

FRED HUTCH
CURES START HERE®

Genome Variant Analysis: Overview

1. Types of genomic variation

2. Visualization using IGV

3. File Formats for Variation Data

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

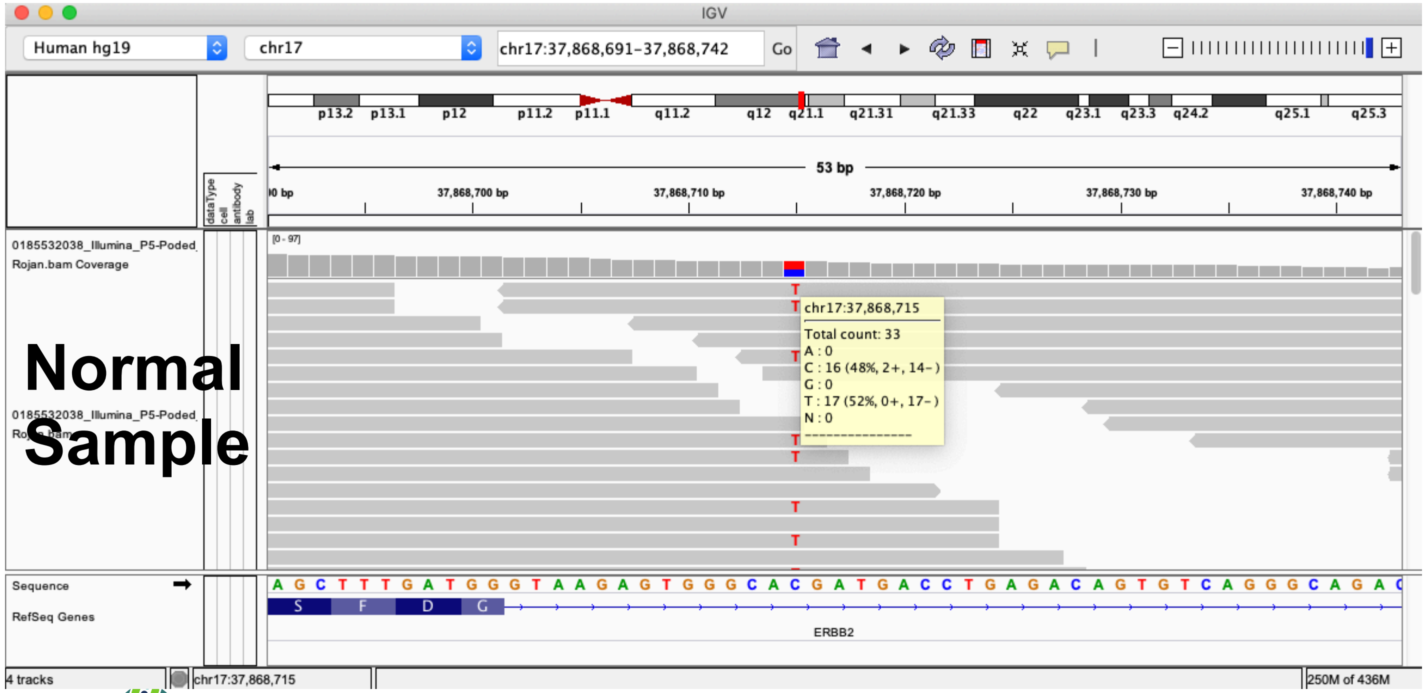
- Germline copy number variant (CNV) or polymorphism (CNP)
- Somatic copy number variant (CNV) or alterations (CNA)

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)



Normal Sample

chr17:37,868,715

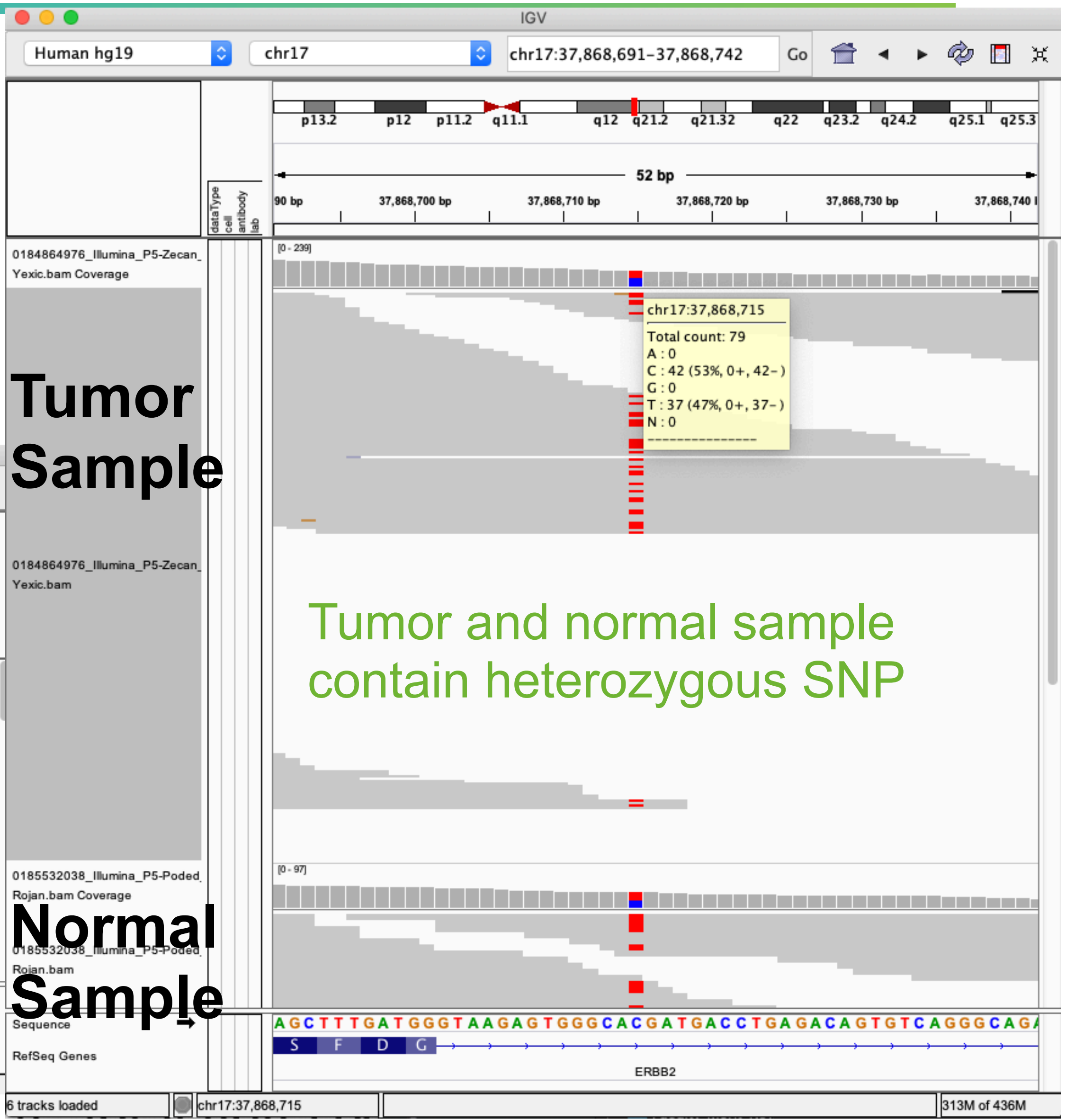
Total count: 33
A : 0
C : 16 (48%, 2+, 14-)
G : 0
T : 17 (52%, 0+, 17-)
N : 0

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)



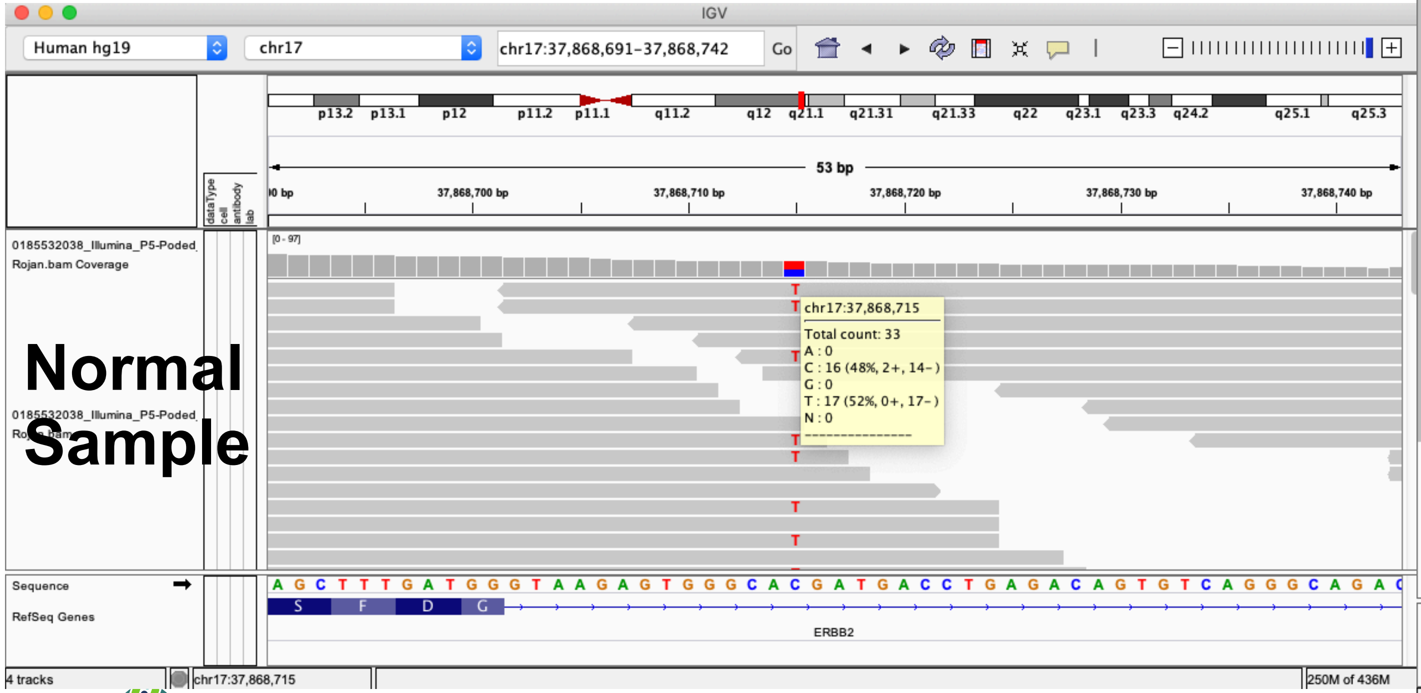
chr17:37,868,715

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

Tumor Sample

Tumor and normal sample contain heterozygous SNP

Normal Sample



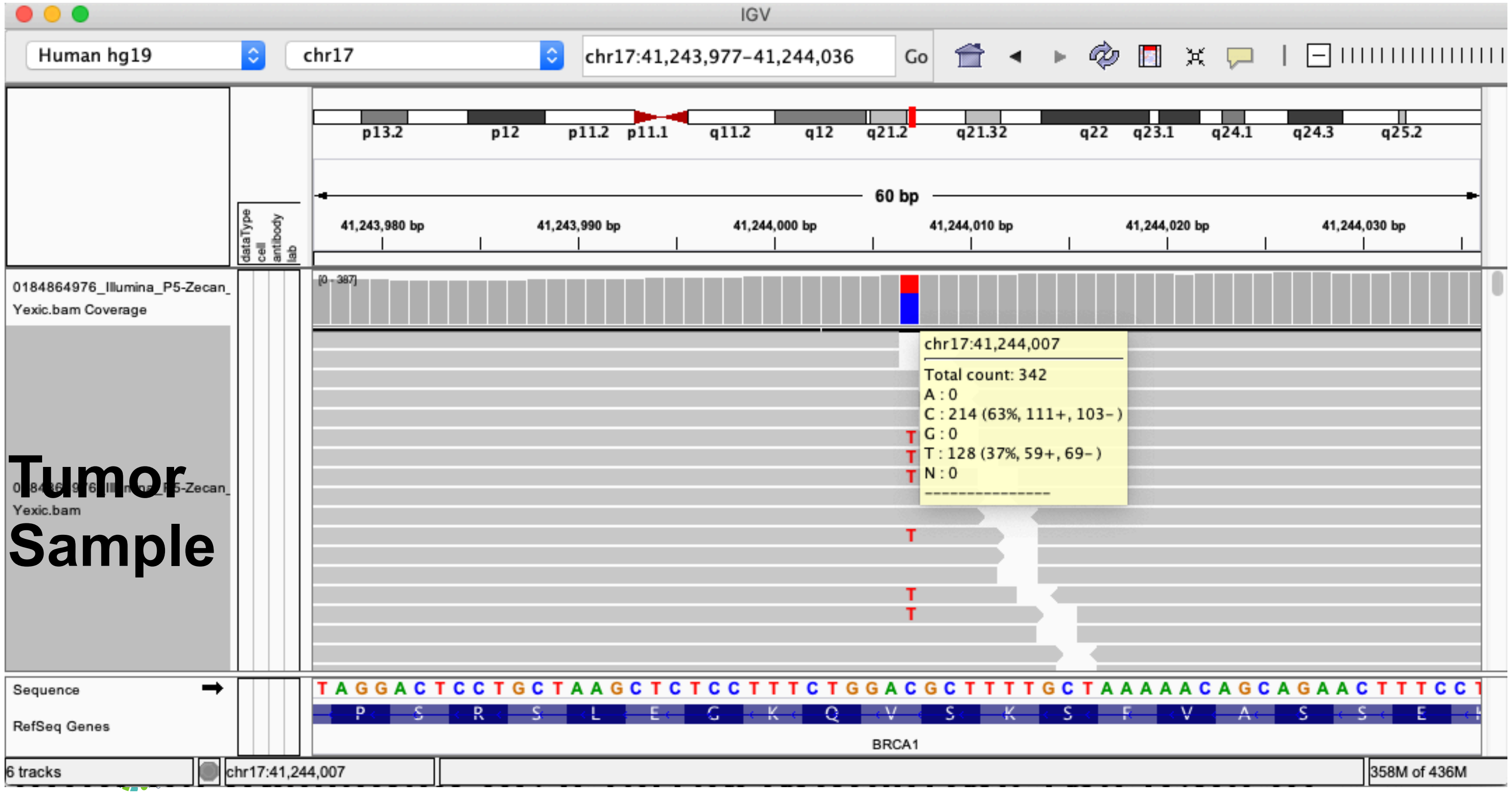
chr17:37,868,715

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

Normal Sample

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

0184864976_Illumina_P5-Zecan_Yexic.bam Coverage

0184864976_Illumina_P5-Zecan_Yexic.bam

```
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

Sequence: 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

RefSeq Genes: P S R S L E G K Q V S K S F V A S S E

BRCA1

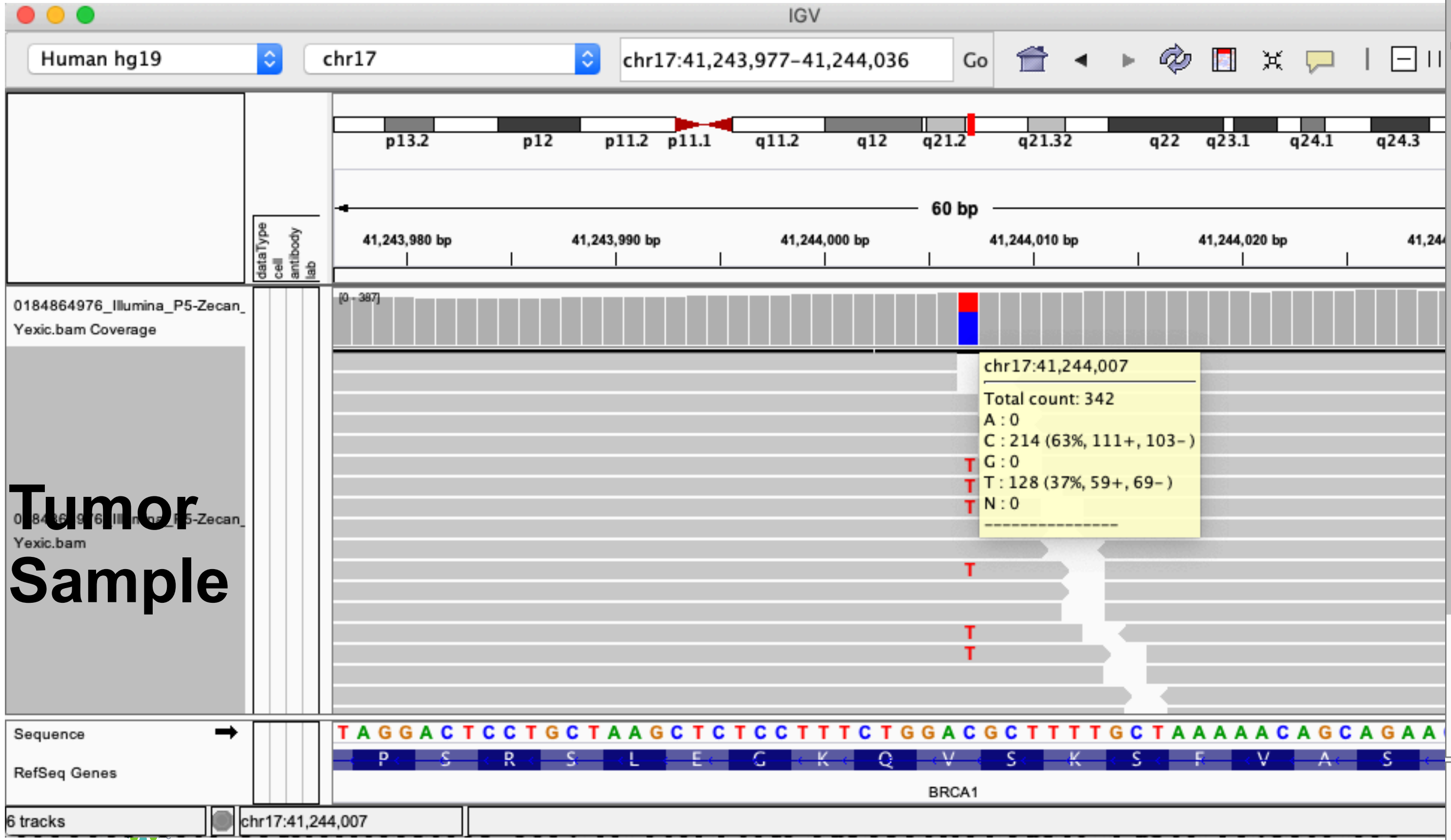
6 tracks | chr17:41,244,007 | 358M of 436M

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)



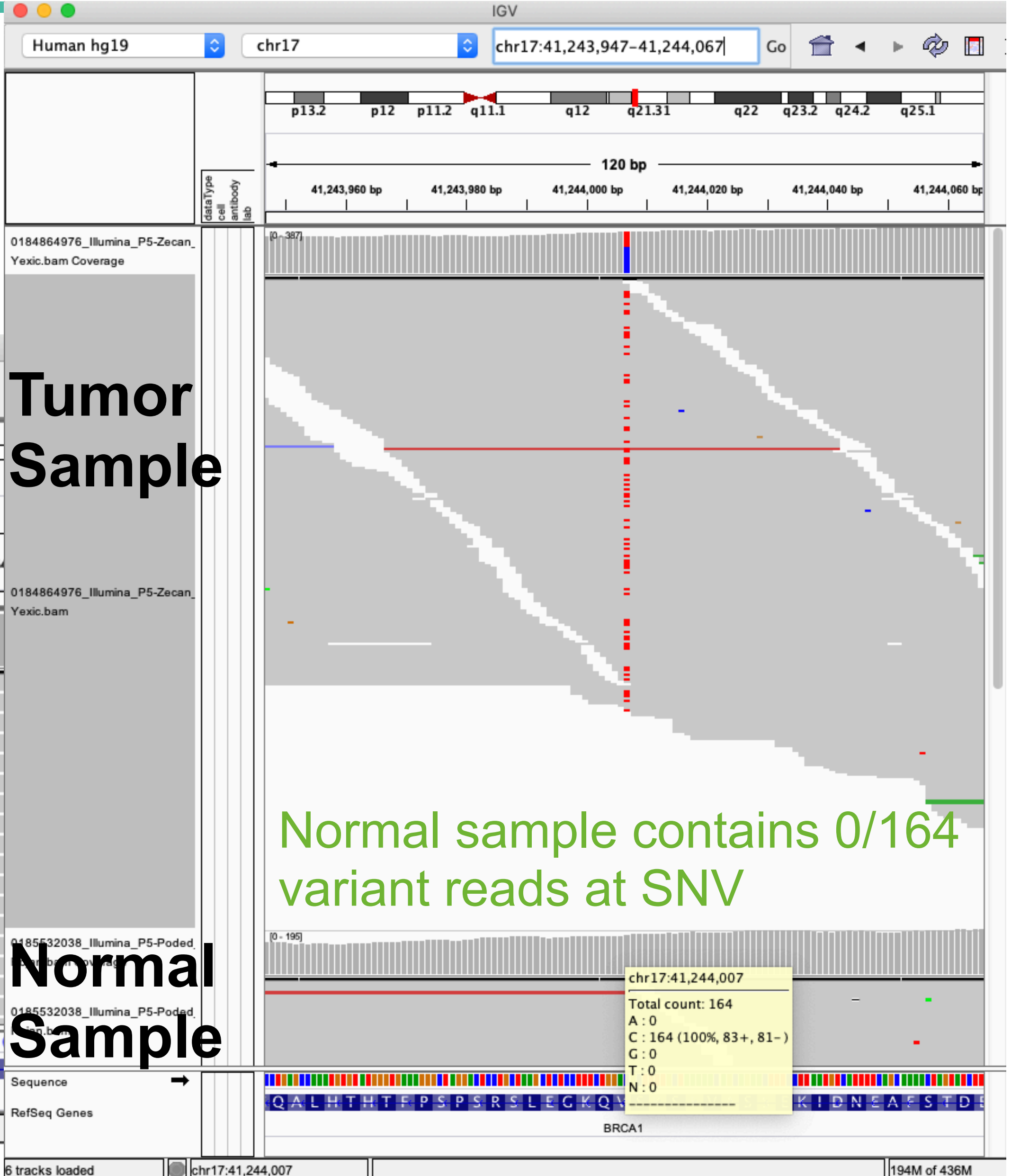
IGV

Human hg19 | chr17 | chr17:41,243,977-41,244,036 | Go

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

6 tracks | chr17:41,244,007

Tumor Sample



IGV

Human hg19 | chr17 | chr17:41,243,947-41,244,067 | Go

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

6 tracks loaded | chr17:41,244,007 | 194M of 436M

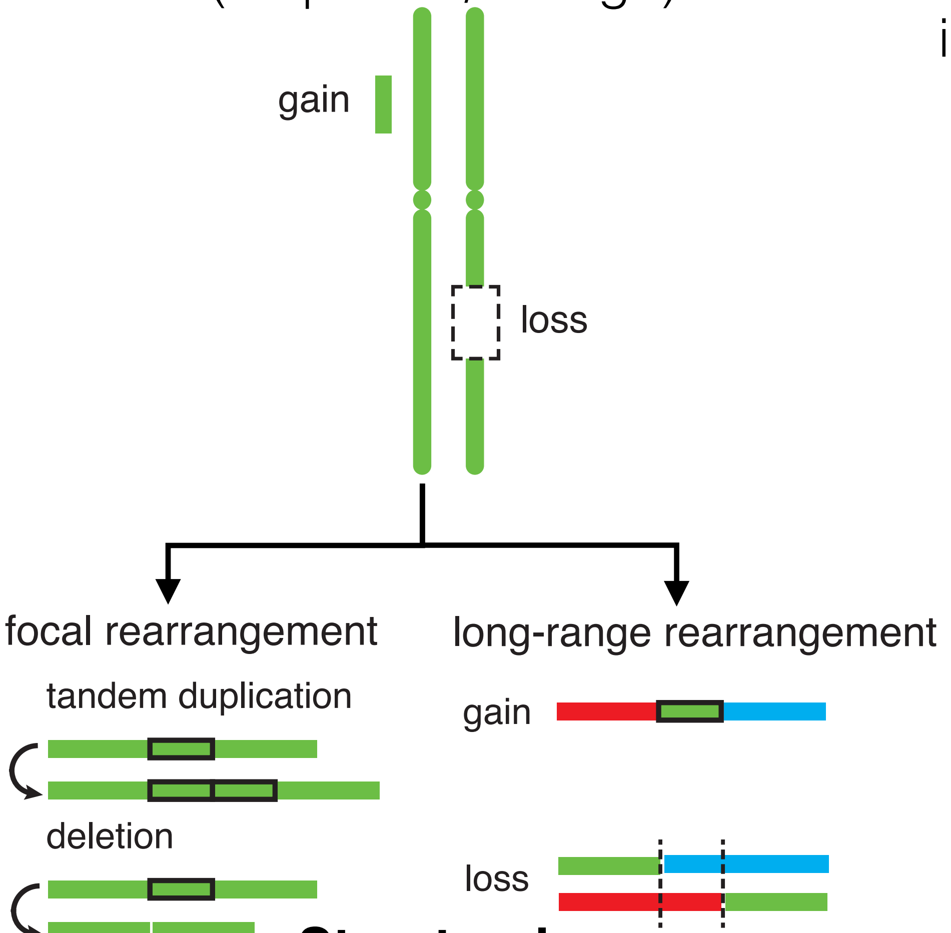
Tumor Sample

Normal sample contains 0/164 variant reads at SNV

Normal Sample

Genome Variant Analysis: Copy Number and Structural Variation

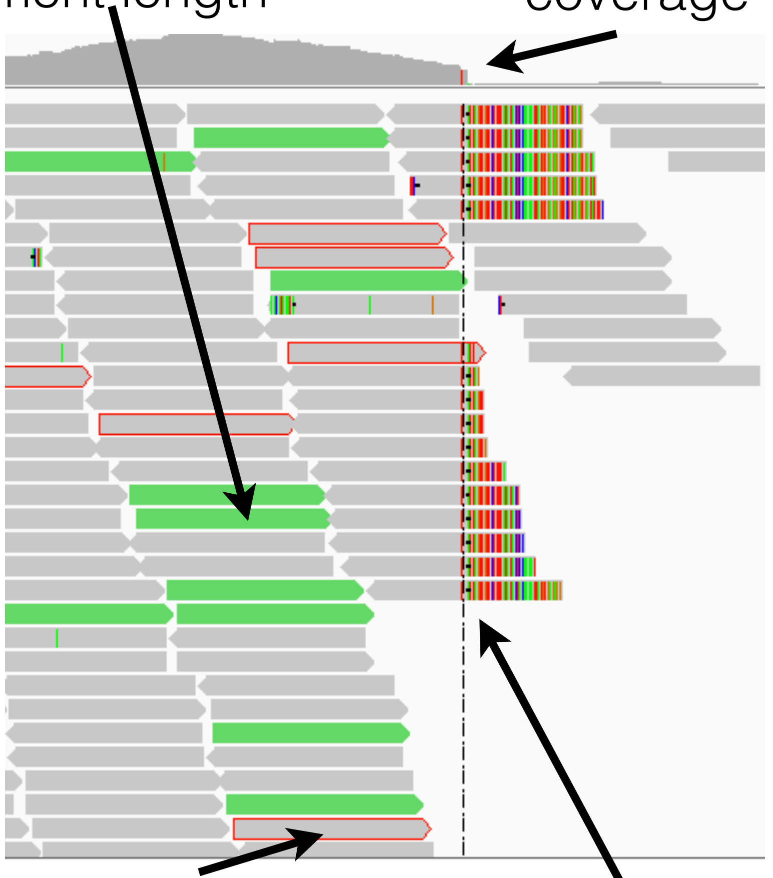
Copy number alterations
(amplitude/dosage)



Structural rearrangements
(location/configuration)

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

“copy number change”
abrupt change in read coverage

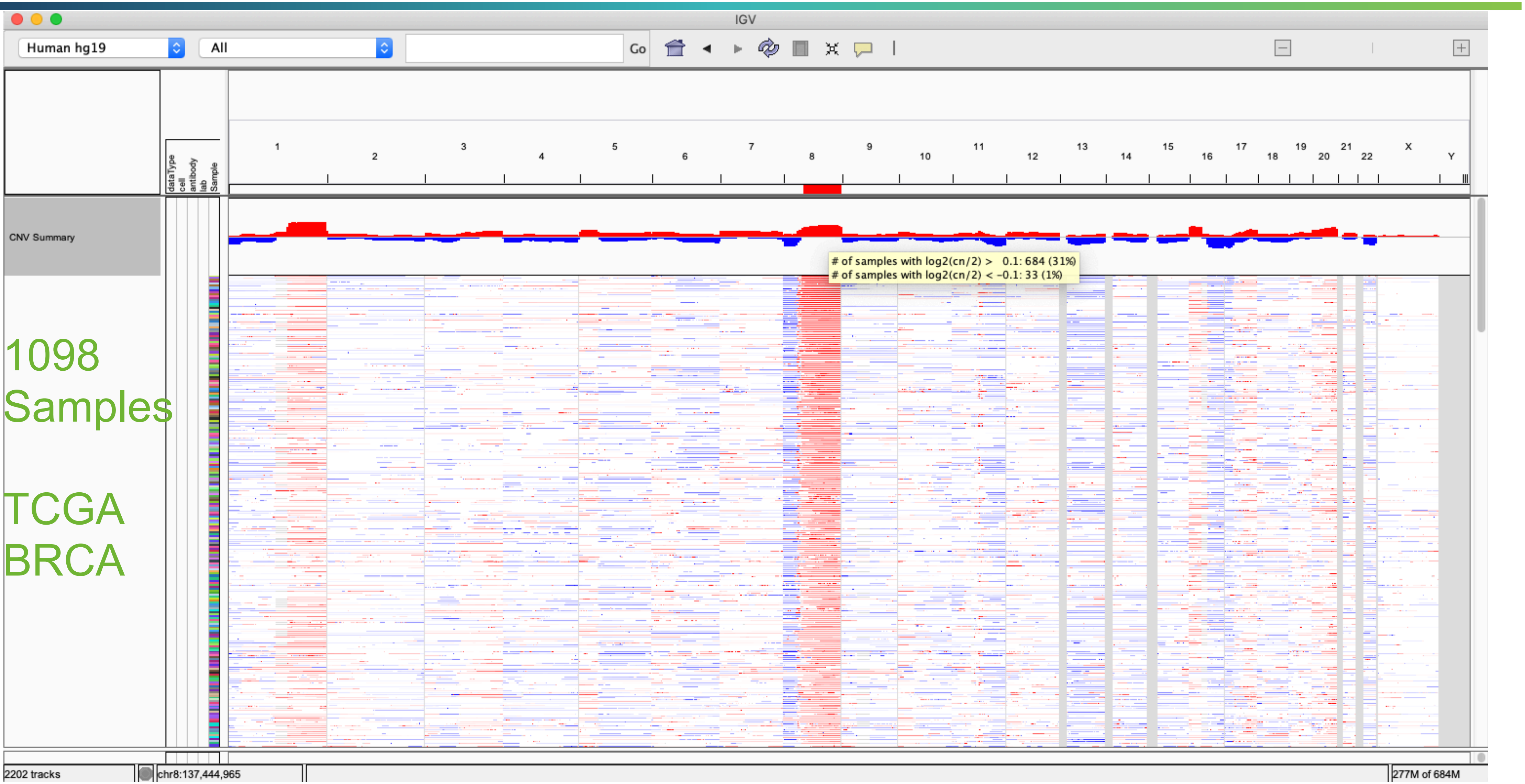


pair-mates unmapped

“split read”
split alignments

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

Genome Variant Analysis: Copy Number Variation



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

2202 tracks
chr8:137,444,965
277M of 684M

1098 Samples

TCGA BRCA

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

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

Genomic Data Analysis in R

Lecture 16

Tuesday, November 124, 2020 @ 1pm

Gavin Ha, Ph.D.
Assistant Professor
Computational Biology Program
Public Health Sciences

FRED HUTCH
CURES START HERE®

Overview

I. R and Bioconductor Packages for Genomic Data

II. Tutorials

- Analyzing Genomic Data
- Analyzing and Annotating Variants
- Analyzing Sequence Data

Overview: Learning Objectives

I. R Bioconductor Packages for Genomic Data

- `GenomicRanges, Rsamtools, VariantAnnotation`

II. Tutorials

1. Genomic Data Analysis (`GenomicRanges`)
   i. Load, inspect, query a BED/SEG file
   ii. Genomic regions overlap
2. Sequence Data Analysis (`Rsamtools`)
   i. Load, inspect, query a BAM alignment file
   ii. Extract sequences and qualities
   iii. Compute “pile-up” statistics at genomic loci
3. Genomic Variants and Annotations (`VariantAnnotation`)
   i. Load, inspect, query a VCF file

Tutorial #1: Genomic Data Analysis

1. Loading and querying BED/SEG text files
   a. Use packages `GenomicRanges`
2. Download the VCF and SEG files for this tutorial
   - https://www.dropbox.com/sh/zoitjnobgp7l7c2/AABBIpTQcNA4lWYOFnV5dlMKa?dl=0
   - BRCA.genome_wide_snp_6_broad_Level_3_scna.seg
   - GIAB_highconf_v.3.3.2.vcfgz, GIAB_highconf_v.3.3.2.vcf.gz
   - GIAB_highconf_v.3.3.2.vcfgz, GIAB_highconf_v.3.3.2.vcf.gz.tbi
3. R Markdown file for tutorial on GitHub: Lecture16_GenomicData.Rmd

Tutorial #2: Sequence Data Analysis

1. Loading and querying a BAM file using Rsamtools
    a. Define the genomic coordinates and components to query (`ScanBamParam`)
    b. Scanning the BAM file (`scanBam`)
2. We will use the example from Lecture 15: Slides 19-22.
3. Download the BAM file for this tutorial
    - https://www.dropbox.com/sh/zoitjnobgp7l7c2/AABBIpTQcNA4lWYOFnV5dlMKa?dl=0
    - BRCA_IDC_cfDNA.bam, BRCA_IDC_cfDNA.bam.bai
4. R Markdown file for tutorial on GitHub: Lecture16_Rsamtools.Rmd

Tutorial #3: Variant Call Format (VCF)

1. Loading and querying VCF files in R
    a. Use packages `VariantAnnotation`
    b. Download the VCF files for this tutorial
    - https://www.dropbox.com/sh/zoitjnobgp7l7c2/AABBIpTQcNA4lWYOFnV5dlMKa?dl=0
    - GIAB_highconf_v.3.3.2.vcfgz, GIAB_highconf_v.3.3.2.vcf.gz
    - GIAB_highconf_v.3.3.2.vcfgz, GIAB_highconf_v.3.3.2.vcf.gz.tbi
2. R Markdown file for tutorial on GitHub: Lecture16_VariantCalls.Rmd

Homework #7: Genomic Data Analysis in R

Problem Set in R Markdown file

- Contains 4 problems with some code to prepare you for the questions.
- Please complete the assignment within the markdown file
- You will be evaluated on
  1. the results and outputs
  2. your code and documentation
  3. Partial points awarded for code with correct logic/function even if the final answer may be incorrect