Genomic data formats and conversions

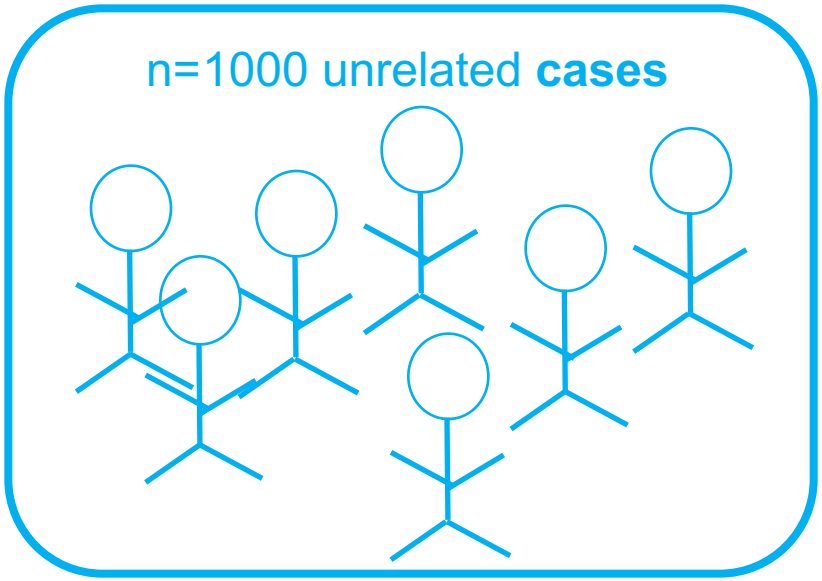
Biocomputing Boot Camp
Day 3 – Session 2
Instructor: Jacob Kitzman
kitzmanj@umich.edu

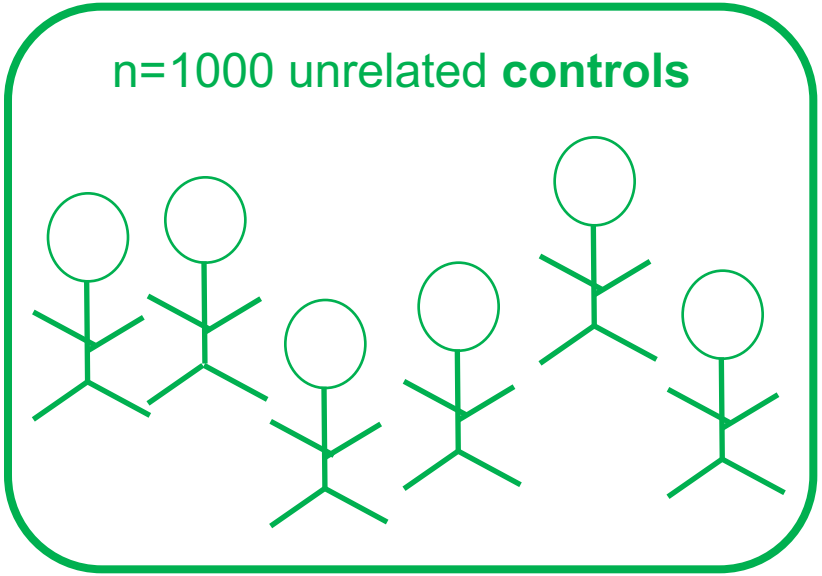
Why standardize data formats?

- Facilitates reliable exchange of data
- Interoperable tools can be used to reanalyze same data in new ways
- Analysis pipelines can be assembled from individual components

Example: rare variant association study

Study goal: identify genes which,
when mutated, increase risk for disease X





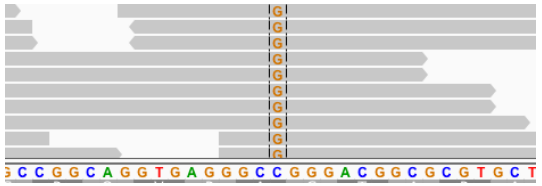
Study design: which genes are more
frequently mutated in cases vs controls?

Data flow for a variation discovery project

Reads

CTAGAAGGCGGGATGTCGCG
AGGCGGGATGTCGCGTGGGA
ATGTCGCGTGGGAAGGGAGC

Alignment to reference genome



Variants

Chr2 12390479 C > G

Joe	Homozygous ref C/C
Mary	Heterozygous C/G

Mutated genes

TP53 (5/20 individuals)
BRCA1 (3/20 individuals)
…

Common formats for genomic data

Machine-readable

Sequence alignments

cram

bam

Variant/genotype calls

bcf

Human-readable

sam

vcf

Sequence data

fastg

fasta

fastq

Track data (location + value)

bed

wig

gff

fastq.gz

gtf

bigBed

2bit

genePred

bigWig

tdf

Gene/functional annotations

https://genome.ucsc.edu/FAQ/FAQformat.html (bed, wig, genePred)

https://samtools.github.io/hts-specs/ (SAM, BAM, VCF, BCF)

Common formats for genomic data

Machine-readable

Sequence alignments

cram

bam

Variant/genotype calls

bcf

vcf

Human-readable

sam

Sequence data

fastg

fasta

fastq

fastq.gz

2bit

Track data (location + value)

bed

wig

bigBed

bigWig

tdf

gff

gtf

genePred

Gene/functional annotations

https://genome.ucsc.edu/FAQ/FAQformat.html (bed, wig, genePred)

https://samtools.github.io/hts-specs/ (SAM, BAM, VCF, BCF)

Sequence data: FASTA and FASTQ

- FASTA: text-based sequence format
- Can be DNA, RNA, protein sequences
- Nearly free-form, can be written by hand
- Extension is usually ".fa" or ".fasta"

"**>**" denotes start of a record

Rest of the line contains the sequence name (spaces are OK)

Sequence (one line or many)

my_file.fa

```
>a_palindrome
AATTAA
>sequence#1
ACGTACGATCGATCAGCATCACACACGTACGTACTGAACAACTACACT
CGCCGC
>chromosome 20
ACTACGTCAGTCAGCATCGA
```

FASTA format in one tweet

Tim Yates
@tim_yates

Follow

@pathogenomenick haha, isn't the standard:

> ANYTHING
ANYTHINGANYTHINGANYTHING
(repeat as much as you like)

4:39 PM - 24 Jan 2015

2 5

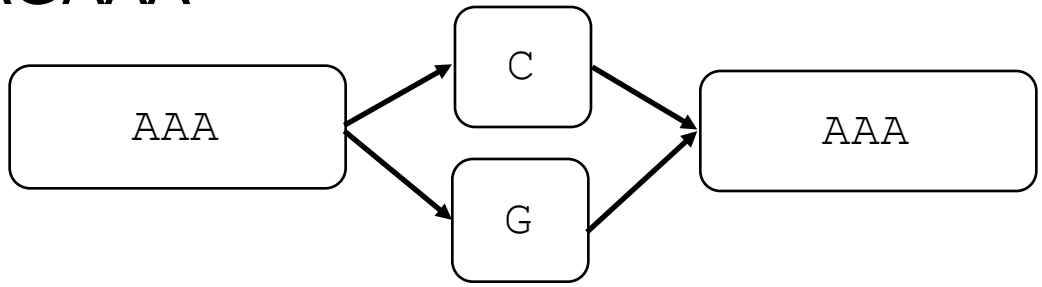
- Lack of a formal spec sometimes causes confusion
  - Is name everything after “>”, or just to the first space?
  - Are symbols (-_|$@&!></) OK? Emoji?
  - Does every line of sequence have to be the same length?
  - Should sequence be just on one line?
- Programs/scripts may have differing expectations…

Other sequence file formats

- FASTQ: text-based, with quality scores
- Typically for primary sequence data - short reads

vs

- FASTG: encode sequence as a graph
  - Haplotypes in diploid genomes
  - Uncertainty
  - "AAA**C**AAA or AAA**G**AAA"



Common formats for genomic data

Machine-readable

Sequence alignments

cram

bam

Variant/genotype calls

bcf

Human-readable

sam

vcf

Sequence data

fastg

fasta

fastq

Track data (location + value)

bed

wig

gff

fastq.gz

gtf

bigBed

2bit

genePred

bigWig

tdf

Gene/functional annotations

https://genome.ucsc.edu/FAQ/FAQformat.html (bed, wig, genePred)

https://samtools.github.io/hts-specs/ (SAM, BAM, VCF, BCF)

Tabular data formats

- Up to now: free form data, 1 record : 2-4 rows
- Next, consider tabular data formats
- Each row has same # of columns
- Values separated by a delimiter (usually tab or comma)

Rank	Name	Number
1	Mary	161,508
2	Helen	69,429
3	Margaret	57,923
4	Anna	54,917
5	Ruth	51,011
6	Elizabeth	41,708
7	Dorothy	39,112
8	Marie	37,089
9	Florence	36,191

```
Rank ⟶ name ⟶ number⏎ 1 ⟶
Mary ⟶ 161,508 ⏎ 2 ⟶ Helen
⟶ 69,429 ⏎ 3 ⟶ Margaret ⟶
57,923 . . .
```

Genomic tabular data

- Many genomic datasets are reference-based: a value associated with a genomic location
- Chrom: name of a chromosome (or a sequence in the reference)
- Start: start coordinate for this record, within that chromosome
- End: ending coordinate for this record

			Other fields (data type/application-dependent)			
Chrom	**Start**	**End**	**Name**	**Strand**	**Score**	**Exons**
1	46860010	46860225	gene1	+	100	5
1	46867752	46867886	gene2	-	49	3
…						
2	154410960	154411313	geneN	-	10	1

Common formats for genomic data

Machine-readable

Human-readable

Sequence alignments

cram

bam

sam

Variant/genotype calls

bcf

vcf

Sequence data

fastg

fasta

fastq

fastq.gz

2bit

Track data (location + value)

bed

wig

bigBed

bigWig

tdf

Gene/functional annotations

gff

gtf

genePred

https://genome.ucsc.edu/FAQ/FAQformat.html (bed, wig, genePred)

https://samtools.github.io/hts-specs/ (SAM, BAM, VCF, BCF)

Sequence alignments

- The optimal match between two possibly related sequences

```
Sequence 1: ACTACTATCACATGGATACTTT
                      | |||| ||
Sequence 2:           C-CATGCAT
```

- Examples:
  - A human protein and its related ortholog in mouse
  - A human protein and a related paralog
  - The reference genome and a sequencing read

- Includes any differences between the two sequences.

Sequence alignments

- Basic information we might want to have:
  - For each query seq, which reference sequence is matched
  - Where? (start, end)
  - What strand?
  - Within alignment, where & what are the gaps (indels) and the mismatches?

SAM file format

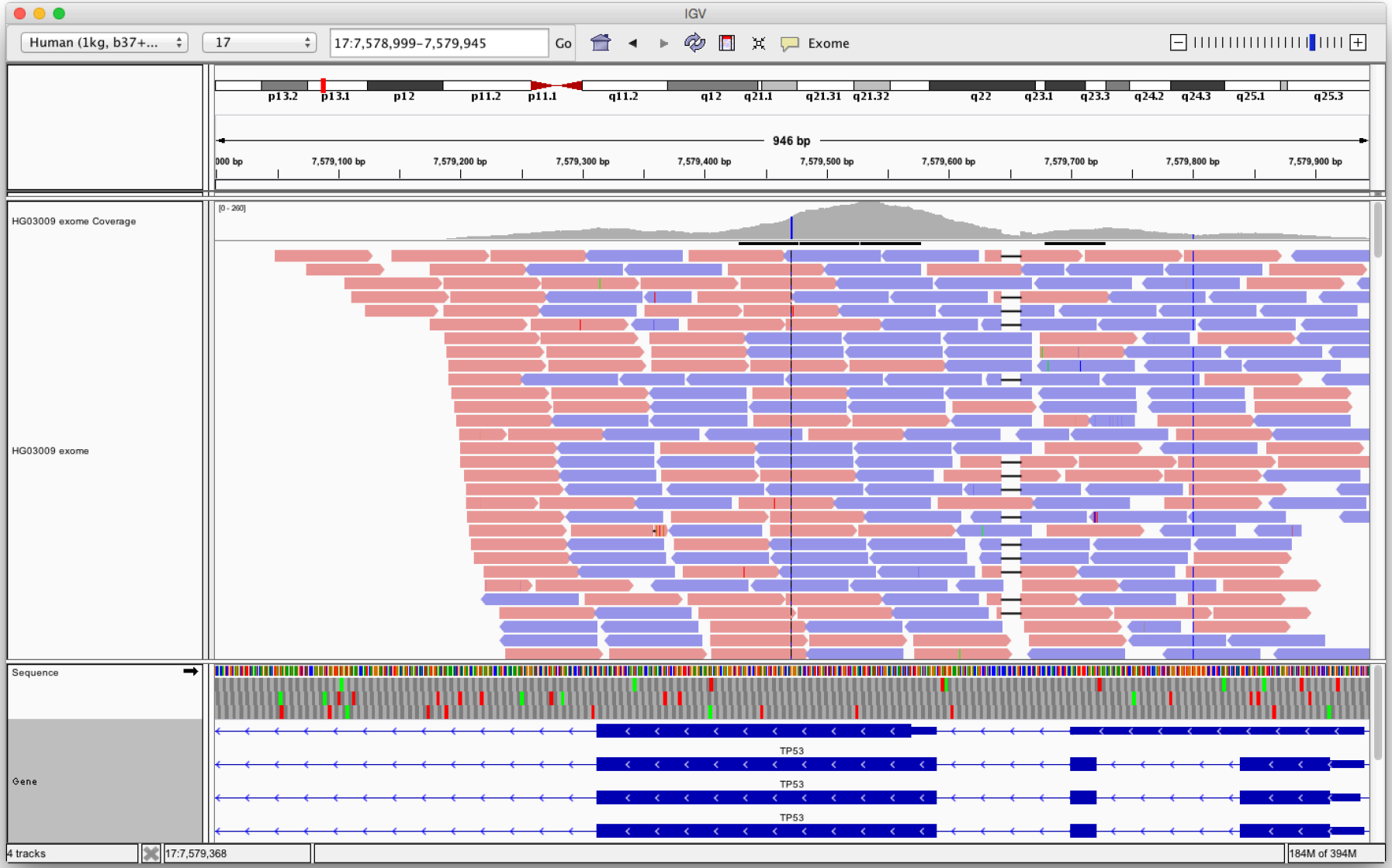
```
@HD     VN:1.4  SO:coordinate  GO:none
@SQ     SN:1    LN:249250621
@SQ     SN:2    LN:243199373
@SQ     SN:3    LN:198022430
@SQ     SN:4    LN:191154276
@SQ     SN:5    LN:180915260@SQ    SN:6    LN:171115067
@SQ     SN:7    LN:159138663@SQ    SN:8    LN:146364022
@SQ     SN:9    LN:141213431
. . . .
.
.
read1  99  1  2160213  60  75M    =  2160288  113  GGAGCCGGAGCGC.....GGGCTGCAGAAGAC  CCCCCGGGGGGGGGGG...GGGGGGGGGGGGGGGGCE  YC:i:0
read2  99  1  2160213  60  75M    =  2160288  113  GGAGCCGGAGCGC.....GGGCTGCAGAAGAC  CCCCCGGGGGGGGGGG...GGGGGGGGGGGGGGGGGG  YC:i:0
read3  99  1  2160213  60  70M5S  =  2160288  113  GGAGCCGGAGCGC.....GGGCTGCAGAAGAC  CCCCCGGGGGGGGGGG...GG7FGGGGGGGGGGGGGG  YC:i:0
read4  99  1  2160213  60  70M1D5 =  2160288  113  GGAGCCGGAGCGC.....GGGCTGCAGAAGAC  CCCCCGGGGGGGGGGG...GGGGGGGGGGGGGGGGGG  YC:i:0
read5  99  1  2160213  60  75M     =  2160288  113  GGAGCCGGAGCGC.....GGGCTGCAGAAGAC  CCCCCGGGGGGGGGGG...GGGGGGGGGGGGGGGGGG  YC:i:0
```

Header – metadata, e.g., names and lengths of reference sequences

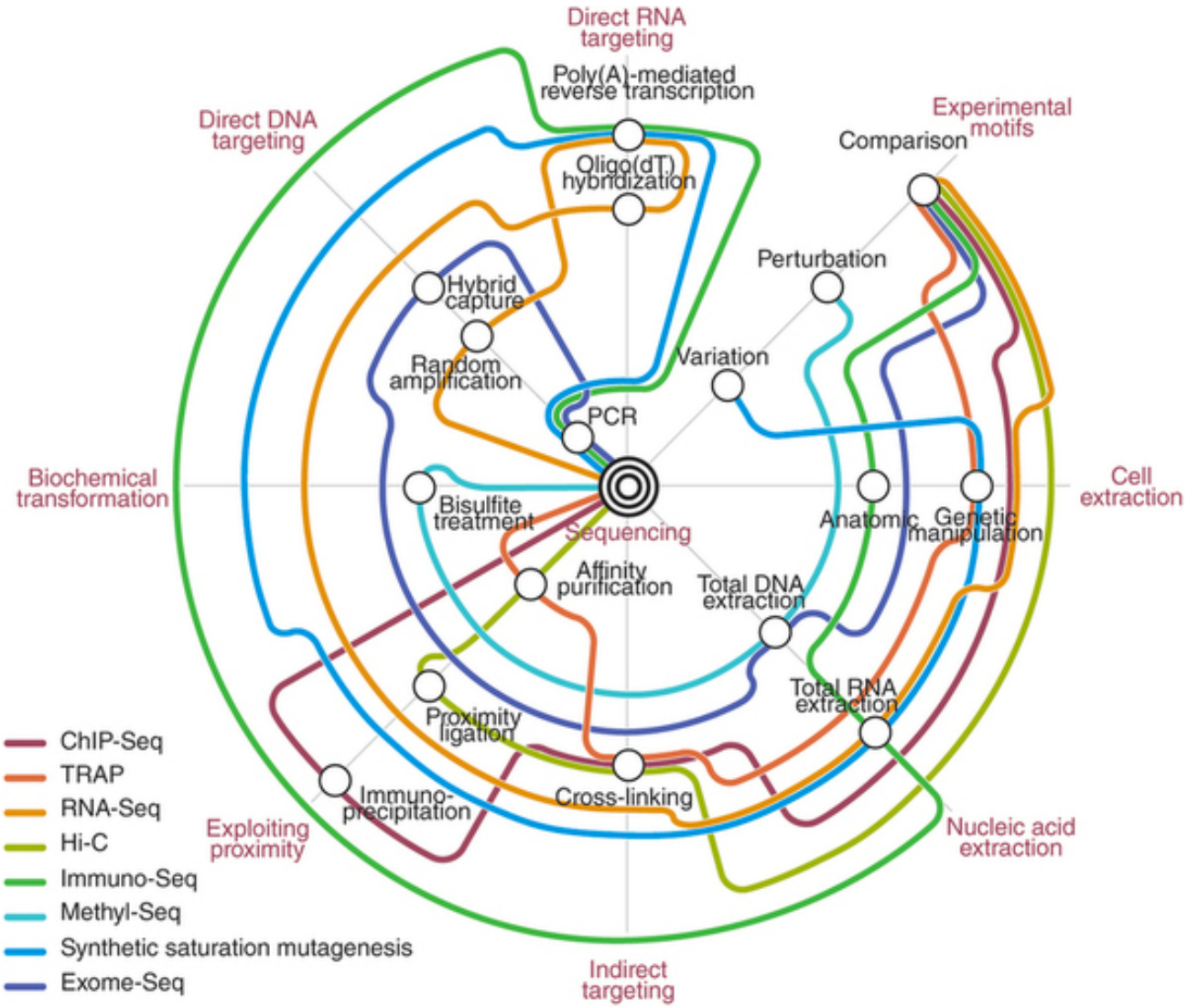
Tab-delimited with:

Column 1 – query name (e.g., read name from sequencing run)
Column 2 – binary flags
Column 3 – reference sequence name (e.g., “1” for chromosome 1)
Column 4 – position on reference
Column 5 – mapping quality score
Column 6 – Alignment string (encodes insertions/deletions)
Column 10 – query sequence
Column 11 – query base quality scores

Visualizing sequence alignments



Alignment data as a common starting point



Shendure & Lieberman Aiden, 2012

Common formats for genomic data

Machine-readable

Human-readable

Sequence alignments

cram

bam

sam

Variant/genotype calls

bcf

vcf

Sequence data

fastg

fasta

fastq

fastq.gz

2bit

Track data (location + value)

bed

wig

bigBed

bigWig

tdf

gff

gtf

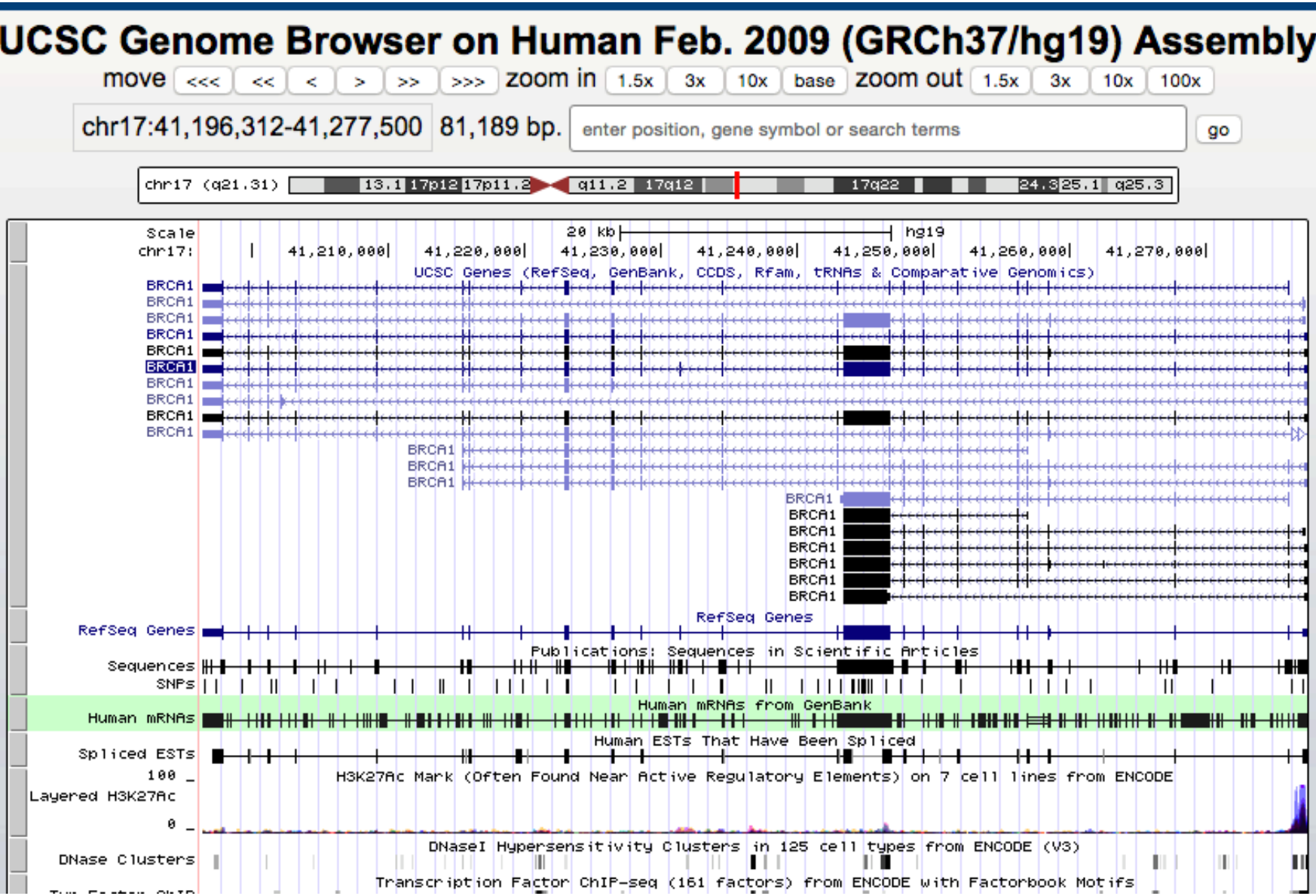
genePred

Gene/functional annotations

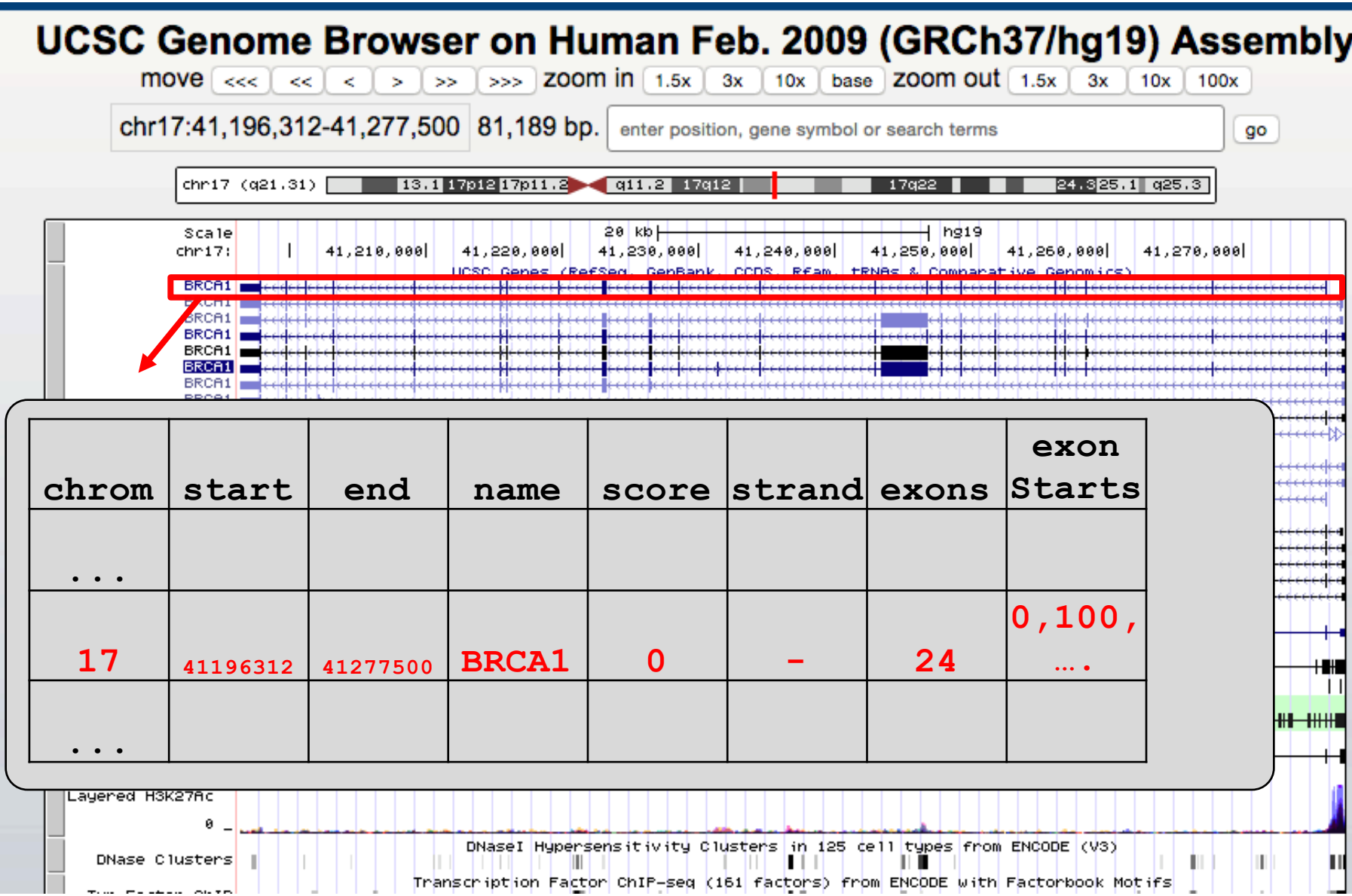
https://genome.ucsc.edu/FAQ/FAQformat.html (bed, wig, genePred)

https://samtools.github.io/hts-specs/ (SAM, BAM, VCF, BCF)

Annotation data



Annotation data



chrom	start	end	name	score	strand	exons	exon Starts
...							
17	41196312	41277500	BRCA1	0	-	24	0,100, ….
...							

Common formats for genomic data

Machine-readable

Human-readable

Sequence alignments

cram

bam

sam

Variant/genotype calls

bcf

vcf

Sequence data

fastg

fasta

fastq

fastq.gz

2bit

Track data (location + value)

bed

wig

bigBed

bigWig

tdf

Gene/functional annotations

gff

gtf

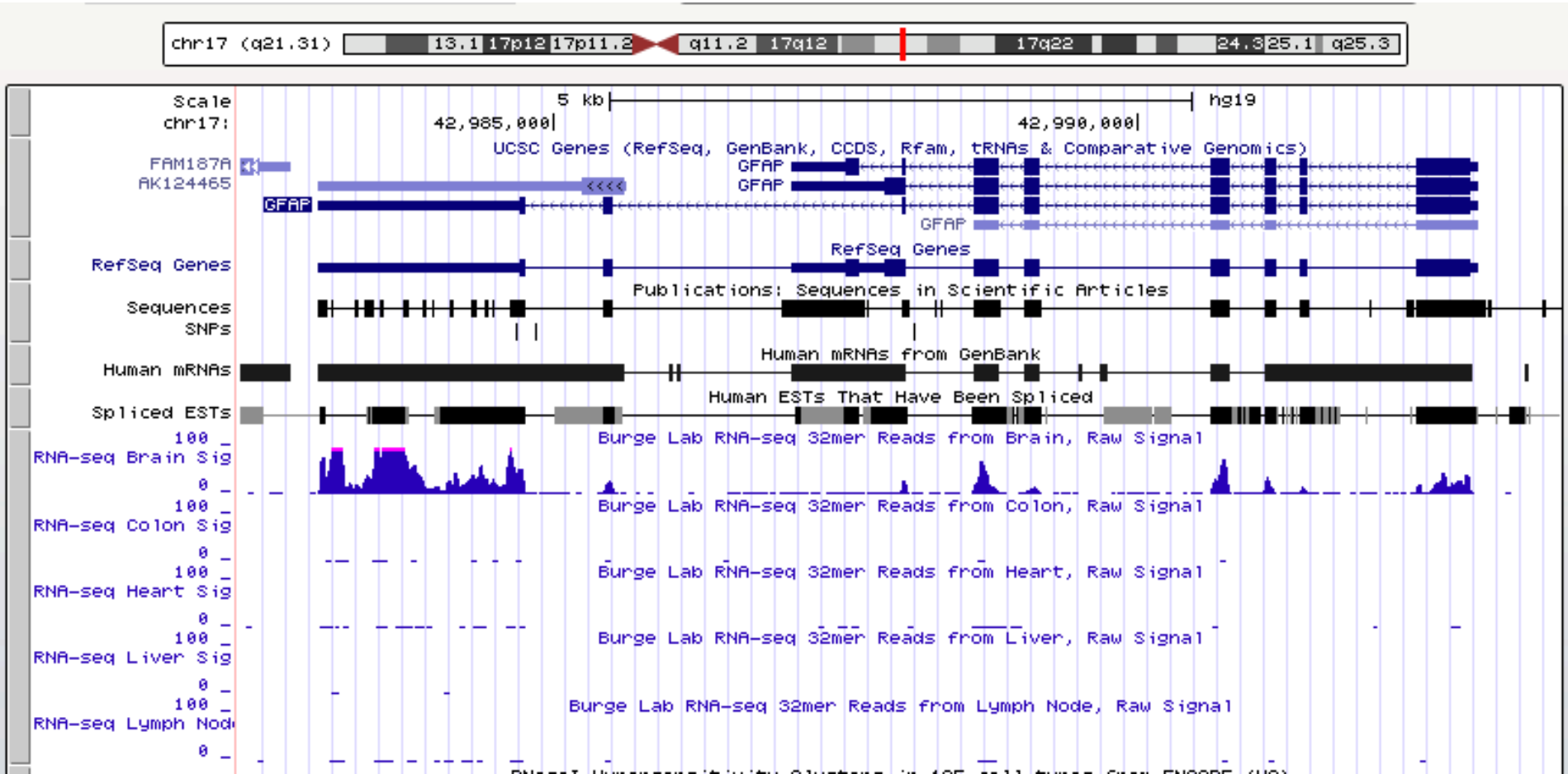
genePred

https://genome.ucsc.edu/FAQ/FAQformat.html (bed, wig, genePred)

https://samtools.github.io/hts-specs/ (SAM, BAM, VCF, BCF)

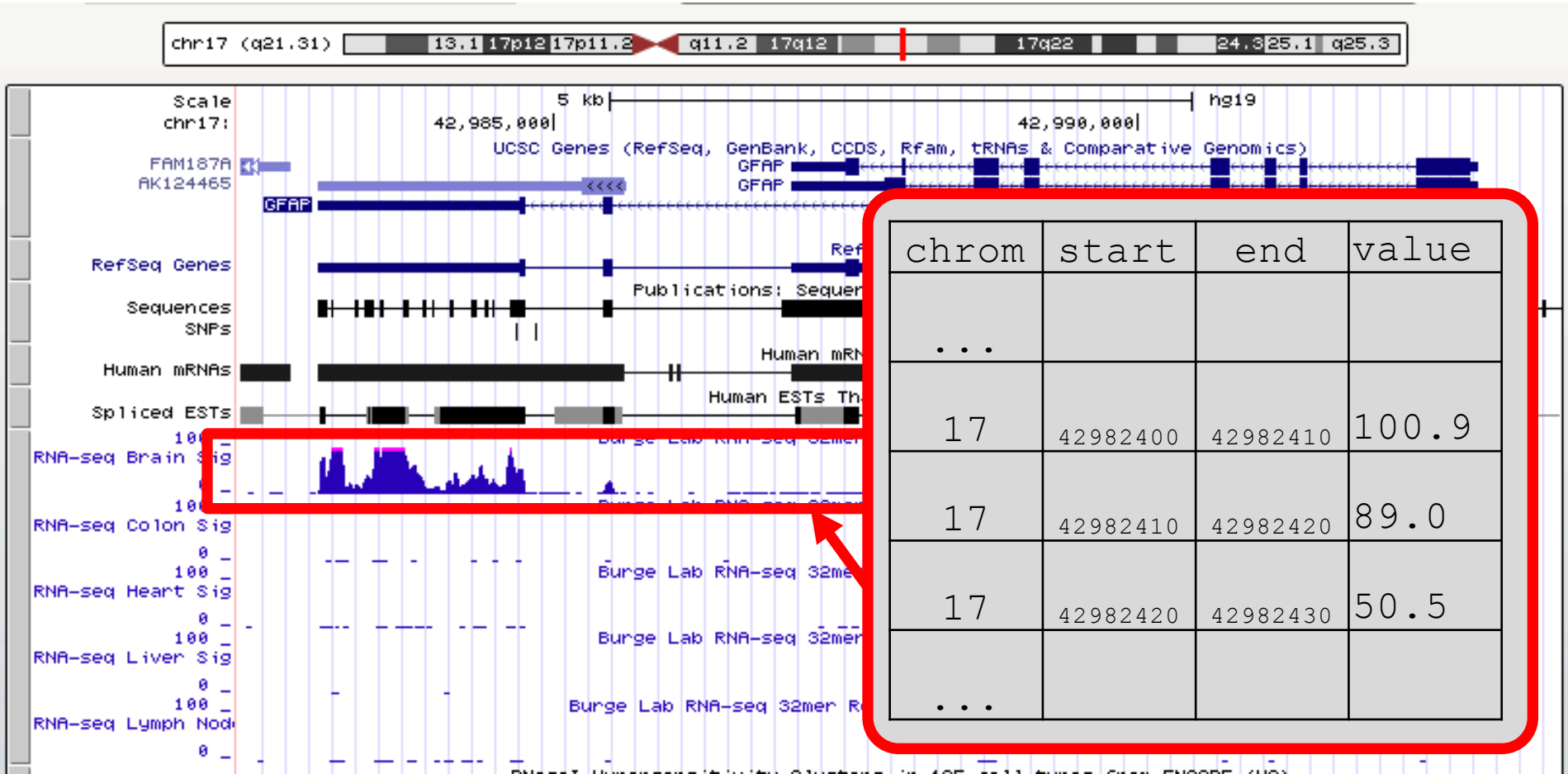
Track data

- A value for every location (single base or window) in the genome.
- e.g., for RNA-seq, read density as a measure of gene expression



Track data

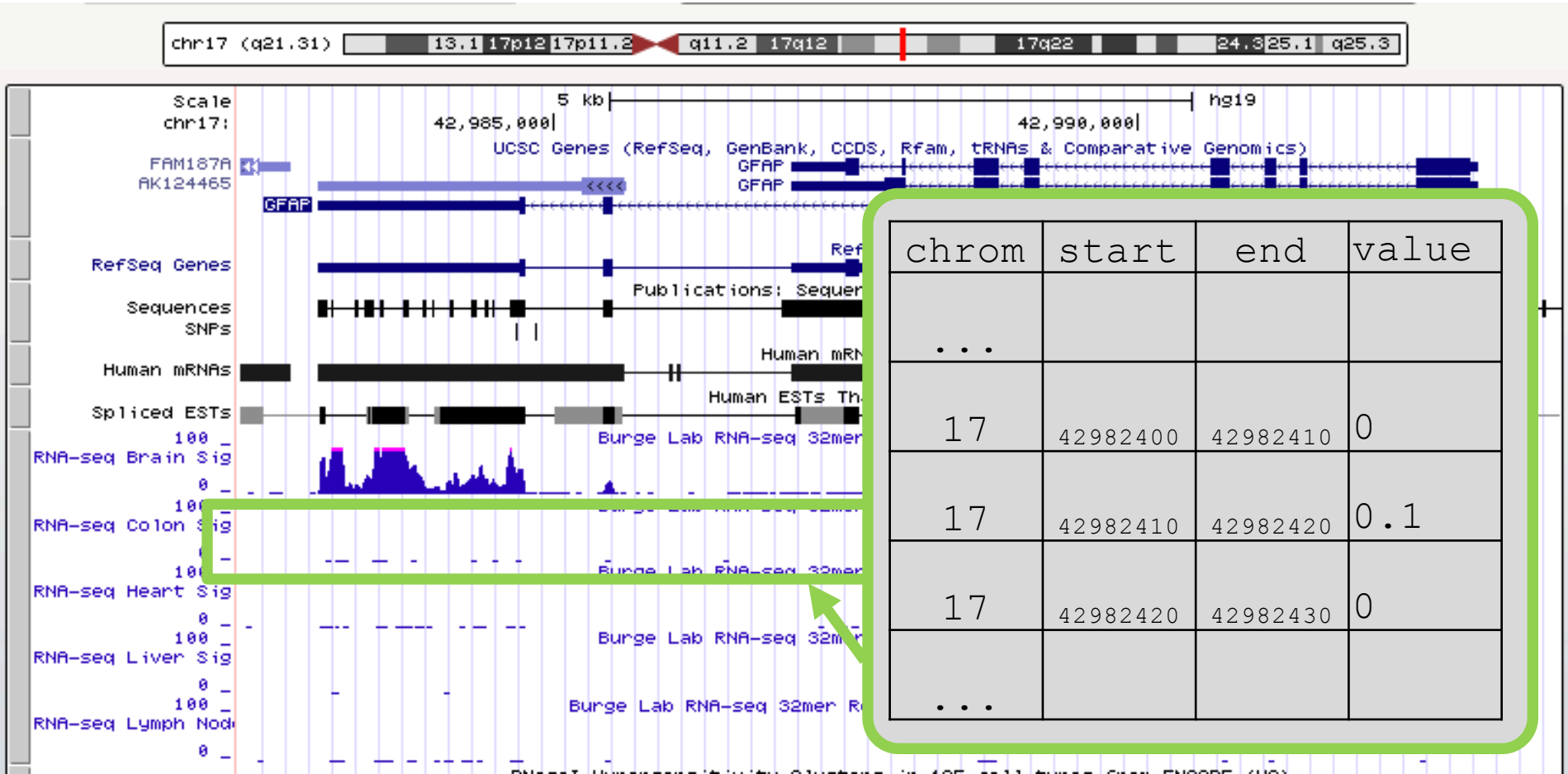
- A value for every location (single base or window) in the genome.
- e.g., for RNA-seq, read density as a measure of gene expression



chrom	start	end	value
...			
17	42982400	42982410	100.9
17	42982410	42982420	89.0
17	42982420	42982430	50.5
...			

Track data

- A value for every location (single base or window) in the genome.
- e.g., for RNA-seq, read density as a measure of gene expression



chrom	start	end	value
...			
17	42982400	42982410	0
17	42982410	42982420	0.1
17	42982420	42982430	0
...			

Track data

- Wiggle track (WIG/bigWig)
  - Stores one continuous-valued measurement at a regular step
    - E.g., depth of sequencing reads over every single base in the genome
    - Or, count of (C+G) bases in windows of 500 bp
- Bedgraph
  - Stores a measurement for each given chrom/start/end interval
- BED track (BED/bigBed)
  - Intended for storing a list of intervals, each potentially associated with a value
    - E.g., linkage interval and LOD score
    - Or, a ChIP-seq peak and associated score

Common formats for genomic data

Machine-readable

Human-readable

Sequence alignments

cram

bam

sam

Variant/genotype calls

bcf

vcf

Sequence data

fastg

fasta

fastq

fastq.gz

2bit

Track data (location + value)

bed

wig

bigBed

bigWig

tdf

gff

gtf

genePred

Gene/functional annotations

https://genome.ucsc.edu/FAQ/FAQformat.html (bed, wig, genePred)

https://samtools.github.io/hts-specs/ (SAM, BAM, VCF, BCF)

Variation data

- Genetic variants, defined by:
  - Location (chromosome, start, stop)
  - Allele
    - Reference allele plus alternate allele***(s)***
  - Per-variant annotations
    - Confidence that it is a true variant
    - Number of samples carrying this variant
    - Name in dbSNP or other databases
- Per-sample genotypes

  0/0 = homozygous REF

  0/1 = heterozygous

  1/1 = homozygous ALT

  ./. = missing

VCF – Variant call format

meta-information lines

- Info about this file
- How it was generated (processing options, etc.)
- Meaning and type (numerical vs categorical) of information fields

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

VCF – Variant call format

VCF is tab-separated (from header onward)

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="A
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbS
##INFO=<ID=H2,Number=0,Type=Flag,Description="Hap
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of sa
##FORMAT=<ID=GT,Number=1,Type=String,Description=
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description
##FORMAT=<ID=DP,Number=1,Type=Integer,Description
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description
#CHROM  POS      ID         REF   ALT     QUAL  FILTER  INFO                              FORMAT      NA00001         NA00002         NA00003
20      14370    rs6054257  G     A       29    PASS    NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
20      17330    .          T     A       3     q10     NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50  0|1:3:5:65,3    0/0:41:3
20      1110696  rs6040355  A     G,T     67    PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27  2|1:2:0:18,2    2/2:35:4
20      1230237  .          T     .       47    PASS    NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60  0|0:48:4:51,51  0/0:61:2
20      1234567  microsat1  GTC   G,GTCT  50    PASS    NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4        0/2:17:2        1/1:40:3
```

 # header line

Columns #1-8 always the same

Column #9 = FORMAT of the per-sample columns

Column #10… = one per sample

VCF – Variant call format

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Each row = 1 variant

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

VCF – Variant call format

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Each row = 1 variant

Chromosome 20 position 1110696 reference allele is A

- *short-hand for reference = 0*

We have called two alternate alleles:

G and T

- *short-hand for first ALT = 1, second ALT = 2, …*

VCF – Variant call format

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Look down to columns
to get each sample's genotype

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Sample NA00001 is heterozygous for each ALT allele

"1|2" → "G/T"

Sample NA00003 is homozygous for ALT allele #2

"2/2" → "T/T"

Visualizing VCF files

- UCSC and IGV browsers will load VCFs
- Here, IGV displaying a 100 kbp interval
- Color = allele; Ref, Het, Hmz

Manipulating VCF files

- VCFs are just tab delimited files
- Could parse with linux command line tools
- Or, write a script to read each line, extract the needed information.
- But they’re huge!

Analogy to reading an encyclopedia

To go through and find everything having to do with dinosaurs, would you:

Dinosaurs …….…….... pg. 194, 535, 789, 1004, 2039

Read end-to-end, or use the index?

Manipulating VCF files

- Compressing and indexing are keys to efficiently extracting records from VCF and similar files
- This strategy requires the input to be sorted by chromosome and coordinate
  - This guarantees that the record for chr1:10,000,00 is after the index entry for chr1:9,000,000 but before the index entry for chr1:11,000,000
- A package called tabix will construct the index and allow for fast lookup
- Can use from the command line and via parsers in Python, C++, Perl, etc.