An Introduction to Genome Versions and File Formats

Henry Farmery

PhD Student Tavaré Group
Supervised by Dr Andy Lynch

Overview

- The Human Genome
- The Genome Reference Consortium
- What’s new in GRCh38?
- What next for the human genome reference?
- File Formats Overview

A Very Brief History of the Human Genome

Caenorhabditis elegans

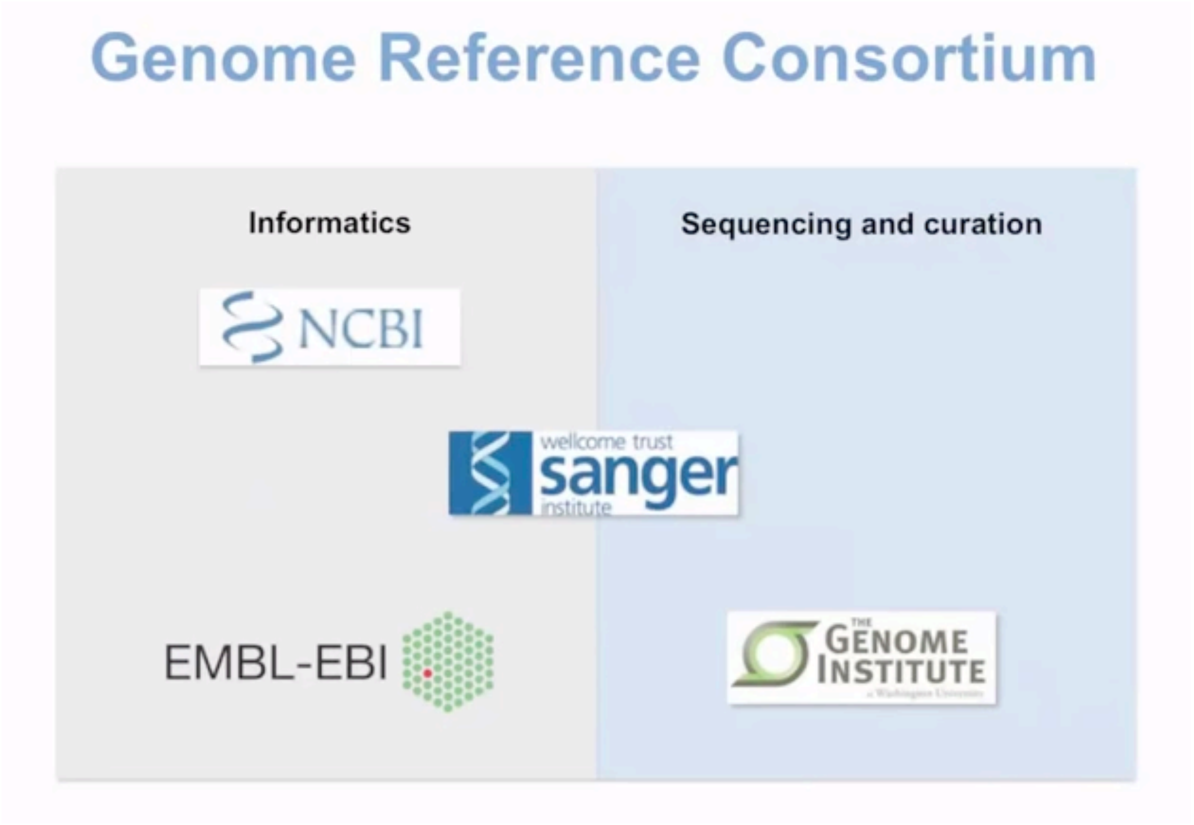
Alfred Sturtevant

- Reference genomes trace their origins to gene maps, first seen around 1911
- In 1998 C. elegans became the first multi cellular organism with a fully assembled reference genome
- In February 2001 the Human Genome Project published its first results
- As of February 2014 there were 38 major updates to the human genome

The Human Genome

- **Breaking news**: the human genome isn't really complete!
- Instead *most* of the human genome is assembled
- New iterations are released periodically by Genome Reference Consortium
- The most recent is GRCh38

Who are the GRC?



Genome Reference Consortium

Informatics

Sequencing and curation

- Make changes to the genomes in two phases
  - Minor release: coordinate unchanged
  - Major release: coordinates changed

What’s new in GRCh38?

- Alternate Sequences:
  - Some parts of the genome can’t be represented by a single sequence
  - GRCh38 includes 261 alternative loci
- Centromere Modeling:
  - Centromeres are highly repetitive regions
  - GRCh38 attempts to model regions in order to attract centrosome reads and reduce noise

What’s new in GRCh38?

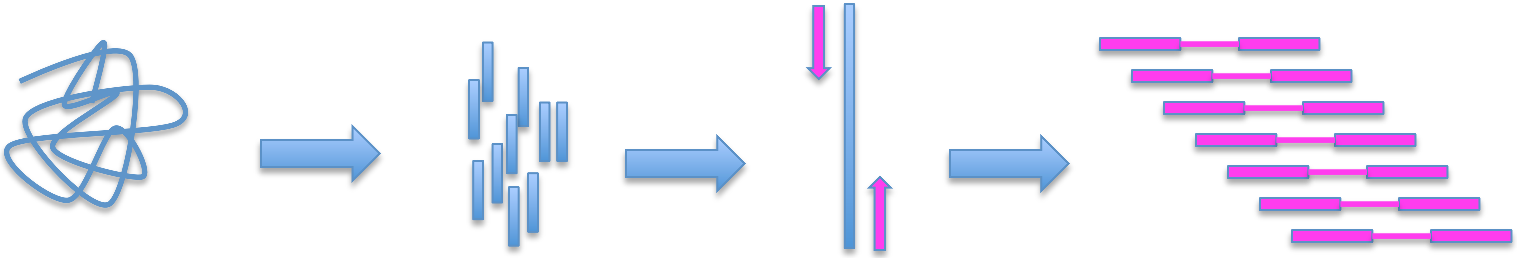
- Mitochondrial genome:
  - Updated to reflect most recent work done by MITOMAP
- Sequence Error Correction:
  - Data from 1000 genome project used to correct errors in previous release
  - 6183 SNVs, 489 Insertions, 910 deletions

What next for the Human Genome reference?

- No more “major releases”
  - As aligners become more competent with alternate loci the need for full scale revisions lessons
- A greater range of human assemblies
  - CHM1
  - NA12878
- More patches
  - First patch for GRCh38 released Sept 2014

How is the reference used by bioinformaticians?

When we conduct a whole genome sequencing experiment we align “reads” back to the reference.



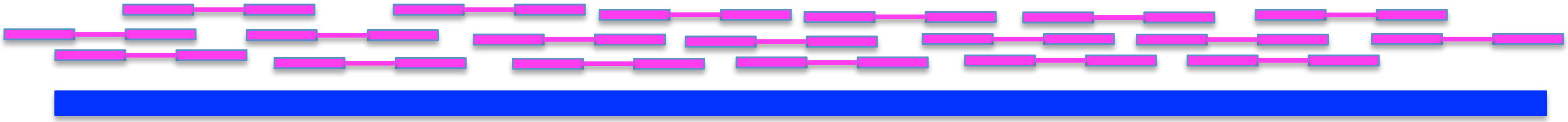
DNA

PCR Amplified Fragmented DNA

Sequencing the fragments

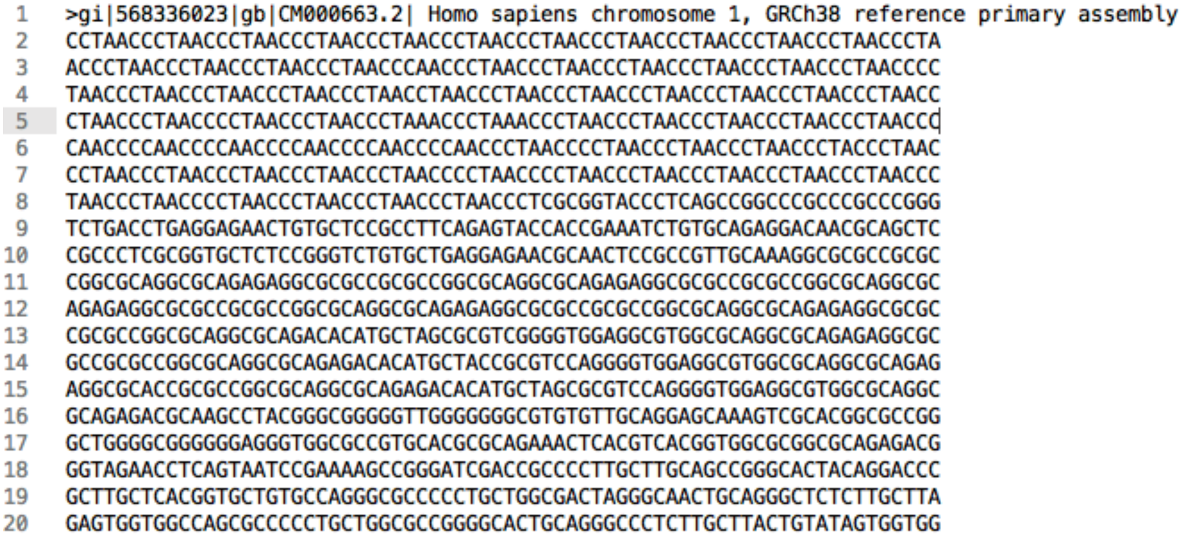
Sequencing Reads

Reads aligned to the reference:



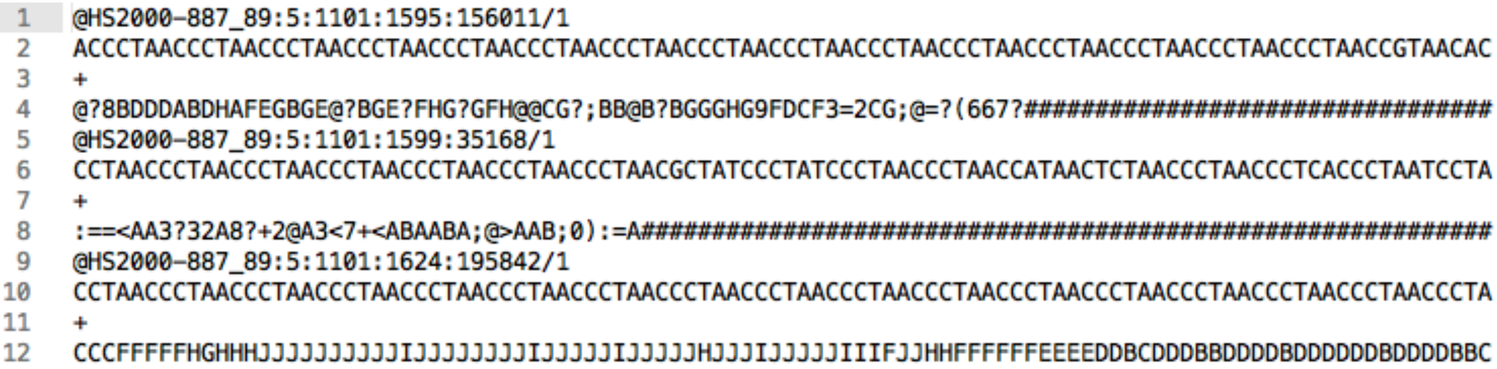
FASTA and FASTQ

- Used for: DNA Sequences
- FASTA looks like this:



```
>gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCC
TAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
CAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACGCAGCTC
CGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGC
CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGC
CGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGC
GCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAG
AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGC
GCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGG
GCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGGCGCAGAGACG
GGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCACTACAGGACCC
GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCTTGCTTA
GAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAGTGGTGG
```

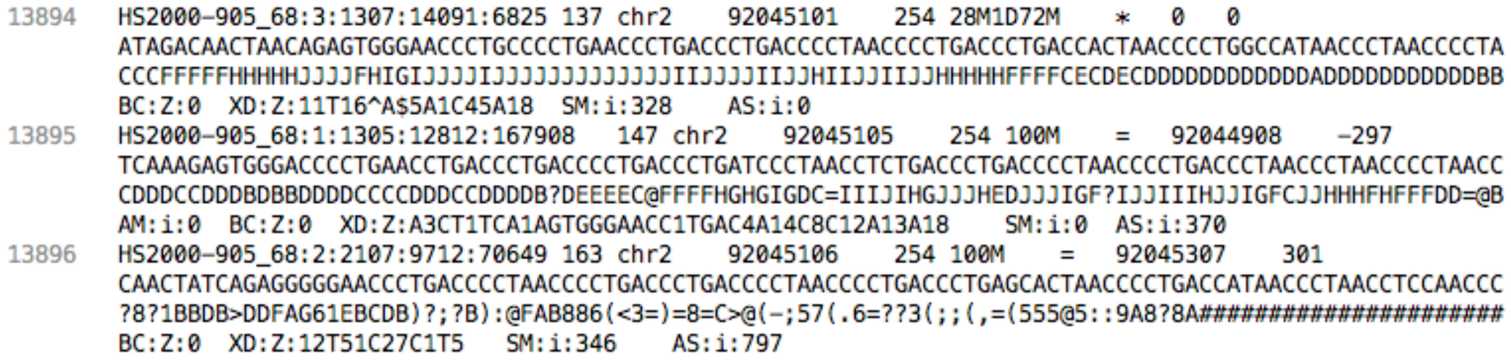
- FASTQ Looks like this:



```
@HS2000-887_89:5:1101:1595:156011/1
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCGTAACAC
+
@?8BDDDABDHAFEGBGE@?BGE?FHG?GFH@@CG?;BB@B?BGGGHG9FDCF3=2CG;@=?(667?#################################
@HS2000-887_89:5:1101:1599:35168/1
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACGCTATCCCTATCCCTAACCCTAACCATAACTCTAACCCTAACCCTCACCCTAATCCTA
+
:==<AA3?32A8?+2@A3<7+<ABAABA;@>AAB;0):=A############################################################
@HS2000-887_89:5:1101:1624:195842/1
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
+
CCCFFFFFHGHHHJJJJJJJJJJIJJJJJJJJIJJJJJIJJJJJHJJJIJJJJJIIIFJJHHFFFFFFEEEEDDBCDDDBBDDDDBDDDDDDBDDDDBBC
```

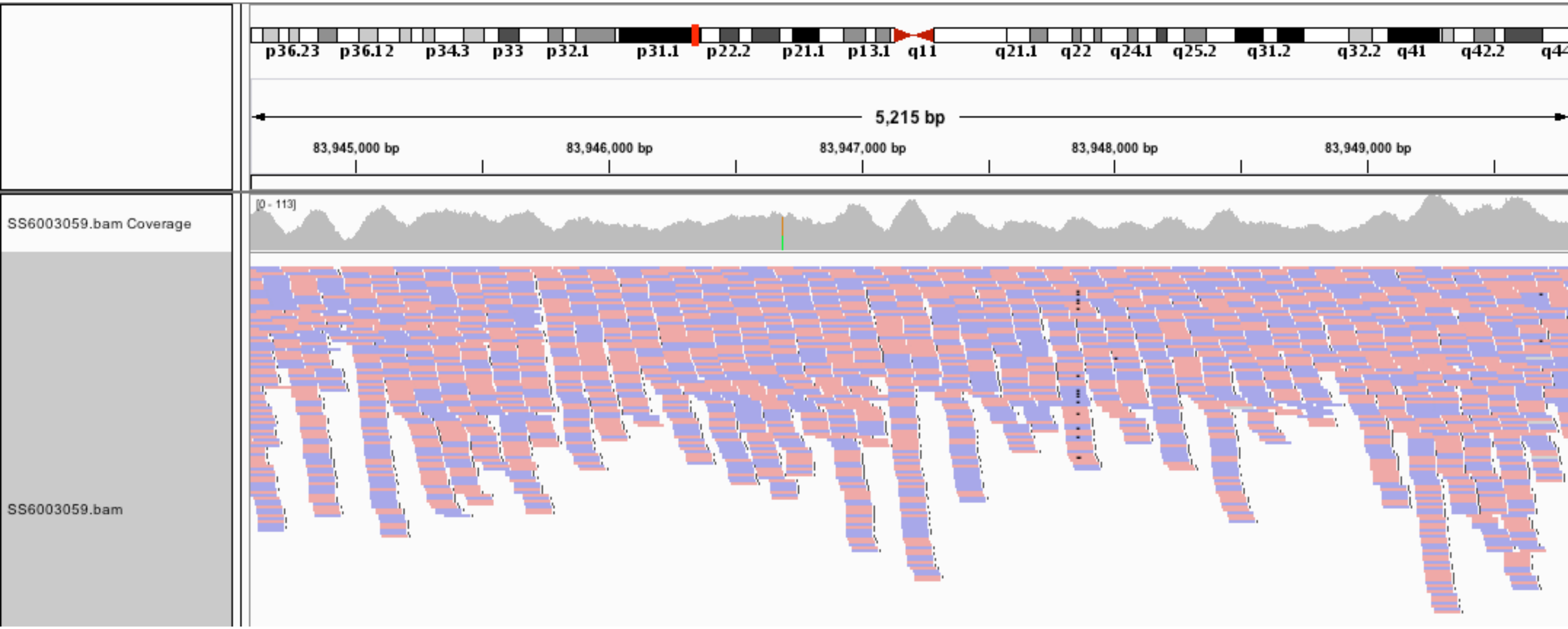
BAM and SAM

- Used for: Aligned Sequences
- They look like this:



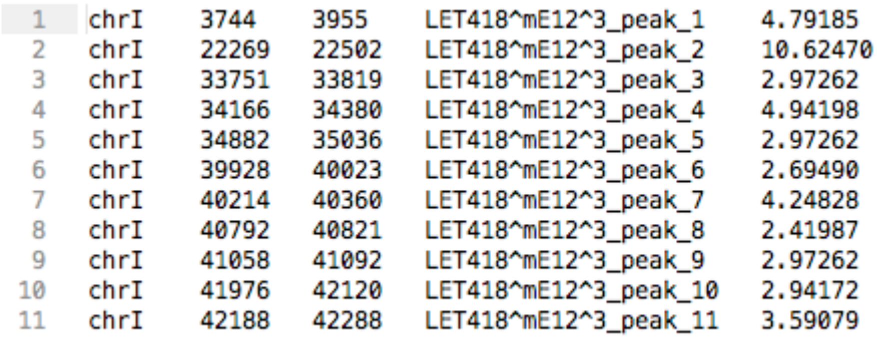
```
HS2000-905_68:3:1307:14091:6825 137 chr2     92045101     254 28M1D72M     *    0    0
ATAGACAACTAACAGAGTGGGAACCCTGCCCCTGAACCCTGACCCTGACCCCTAACCCCTGACCCTGACCACTAACCCCTGGCCATAACCCTAACCCCTA
CCCFFFFFHHHHHJJJJFHIGIJJJJIJJJJJJJJJJJJJJIIJJJJIIJJHIIJJIIJJHHHHHFFFFCECDECDDDDDDDDDDDDADDDDDDDDDDDBB
BC:Z:0  XD:Z:11T16^A$5A1C45A18  SM:i:328     AS:i:0
HS2000-905_68:1:1305:12812:167908    147 chr2     92045105     254 100M     =    92044908     -297
TCAAAGAGTGGGACCCCTGAACCTGACCCTGACCCCTGACCCTGATCCCTAACCTCTGACCCTGACCCCTAACCCCTGACCCTAACCCTAACCCCTAACC
CDDDCCDDDBDBBDDDDCCCCDDDCCDDDDB?DEEEEC@FFFFHGHGIGDC=IIIJIHGJJJHEDJJJIGF?IJJIIIHJJIGFCJJHHHFHFFFDD=@B
AM:i:0  BC:Z:0  XD:Z:A3CT1TCA1AGTGGGAACC1TGAC4A14C8C12A13A18     SM:i:0  AS:i:370
HS2000-905_68:2:2107:9712:70649 163 chr2     92045106     254 100M     =    92045307     301
CAACTATCAGAGGGGGAACCCTGACCCCTAACCCCTGACCCTGACCCCTAACCCCTGACCCTGAGCACTAACCCCTGACCATAACCCTAACCTCCAACCC
?8?1BBDB>DDFAG61EBCDB)?;?B):@FAB886(<3=)=8=C>@(-;57(.6=??3(;;(,=(555@5::9A8?8A######################
BC:Z:0  XD:Z:12T51C27C1T5    SM:i:346     AS:i:797
```

- In IGV they look like this:



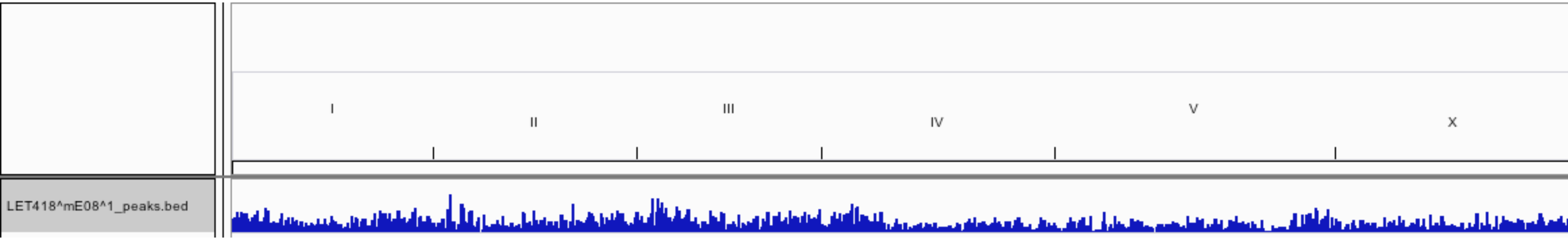
BED Files + Big BED

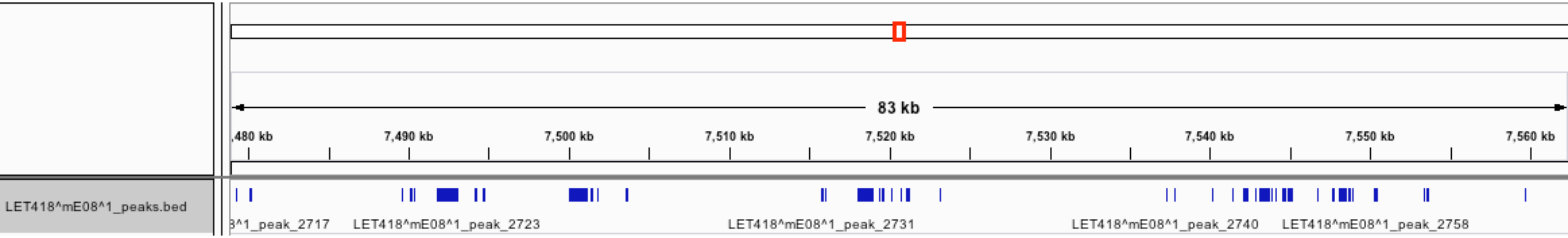
- Used for: Specifying Regions
- They look like this:



```
chrI    3744    3955    LET418^mE12^3_peak_1     4.79185
chrI    22269   22502   LET418^mE12^3_peak_2     10.62470
chrI    33751   33819   LET418^mE12^3_peak_3     2.97262
chrI    34166   34380   LET418^mE12^3_peak_4     4.94198
chrI    34882   35036   LET418^mE12^3_peak_5     2.97262
chrI    39928   40023   LET418^mE12^3_peak_6     2.69490
chrI    40214   40360   LET418^mE12^3_peak_7     4.24828
chrI    40792   40821   LET418^mE12^3_peak_8     2.41987
chrI    41058   41092   LET418^mE12^3_peak_9     2.97262
chrI    41976   42120   LET418^mE12^3_peak_10    2.94172
chrI    42188   42288   LET418^mE12^3_peak_11    3.59079
```

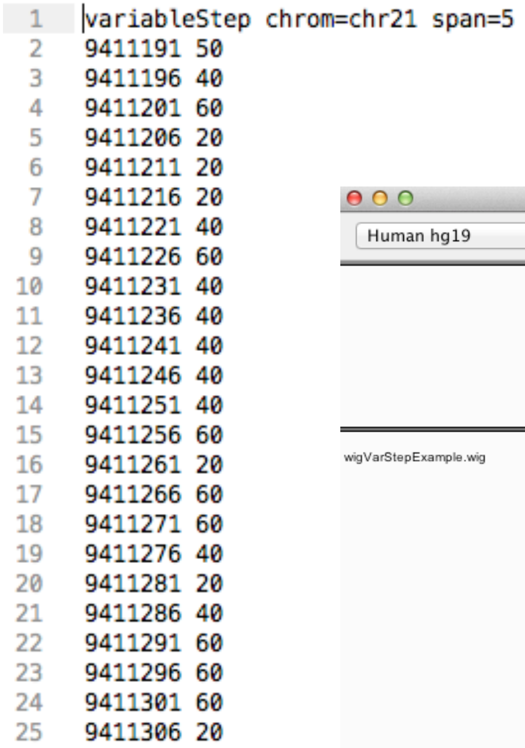
- In IGV they look like this:



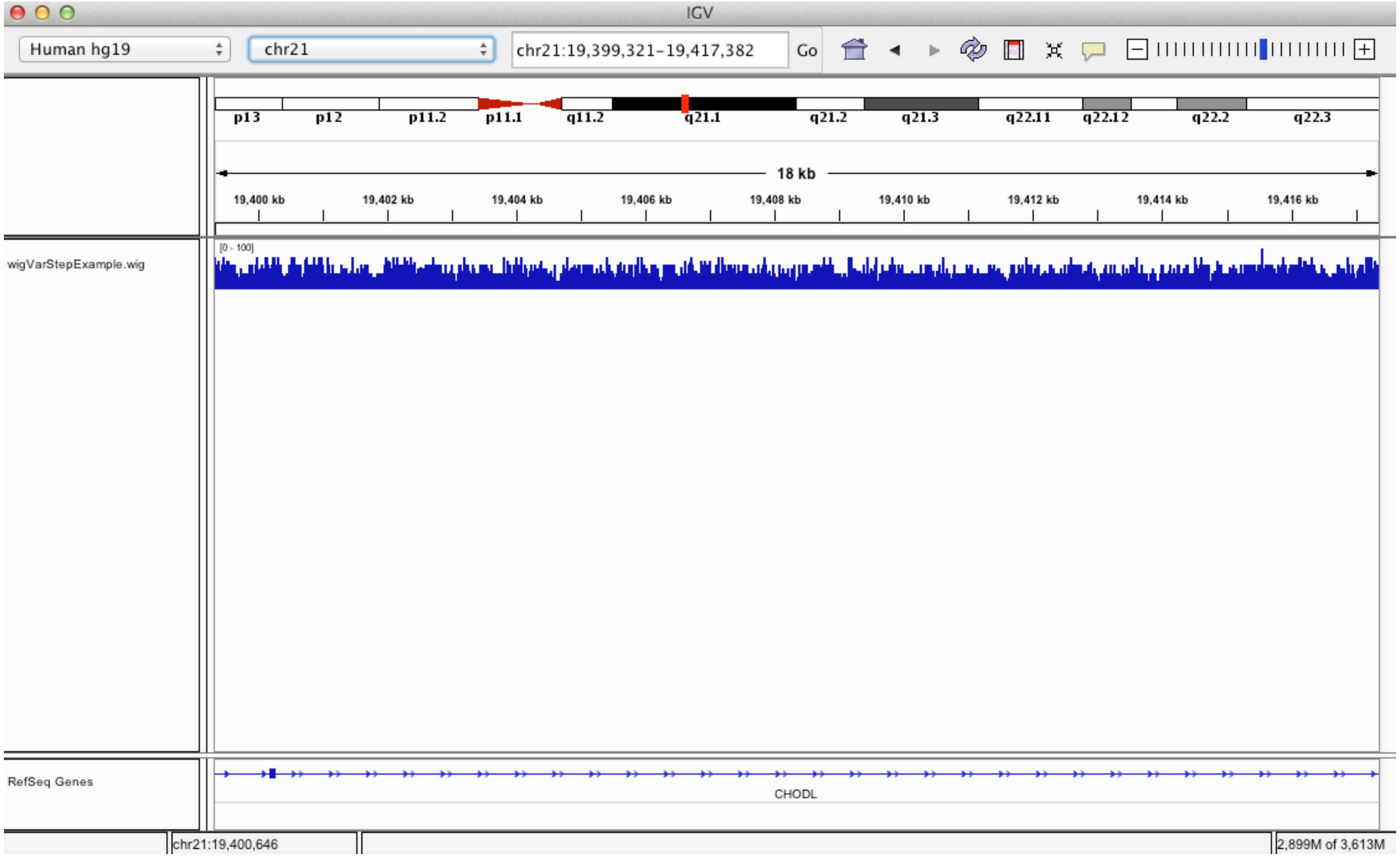


Wig and BigWig

- Displaying dense genomic data in density format
- Wig files look like this:

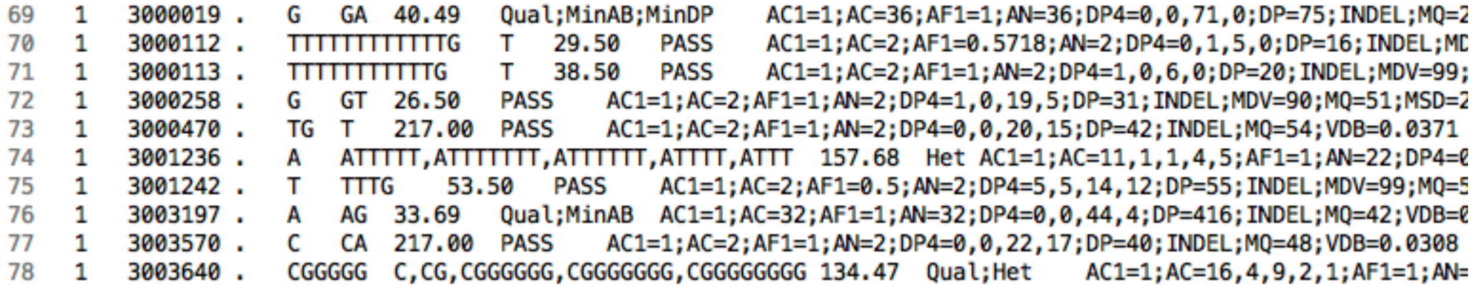


```
variableStep chrom=chr21 span=5
9411191 50
9411196 40
9411201 60
9411206 20
9411211 20
9411216 20
9411221 40
9411226 60
9411231 40
9411236 40
9411241 40
9411246 40
9411251 40
9411256 60
9411261 20
9411266 60
9411271 60
9411276 40
9411281 20
9411286 40
9411291 60
9411296 60
9411301 60
9411306 20
```

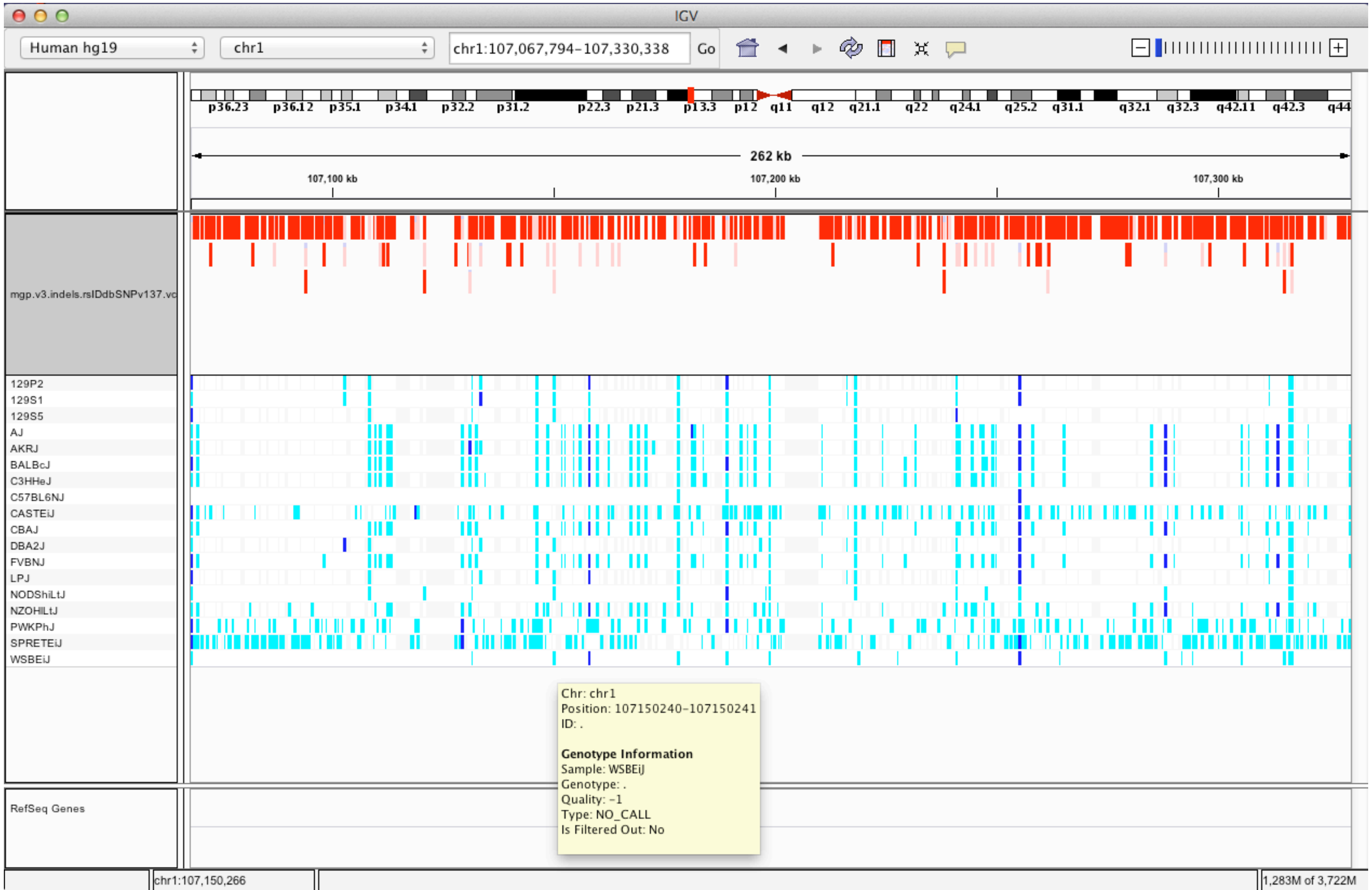


VCF

- Used for specifying variants (SNPs,SNVs and Indels)
- VCF files look like this:



```
1   3000019 .   G   GA  40.49   Qual;MinAB;MinDP    AC1=1;AC=36;AF1=1;AN=36;DP4=0,0,71,0;DP=75;INDEL;MQ=2
1   3000112 .   TTTTTTTTTTTTG   T   29.50   PASS    AC1=1;AC=2;AF1=0.5718;AN=2;DP4=0,1,5,0;DP=16;INDEL;MD
1   3000113 .   TTTTTTTTTTTG    T   38.50   PASS    AC1=1;AC=2;AF1=1;AN=2;DP4=1,0,6,0;DP=20;INDEL;MDV=99;
1   3000258 .   G   GT  26.50   PASS    AC1=1;AC=2;AF1=1;AN=2;DP4=1,0,19,5;DP=31;INDEL;MDV=90;MQ=51;MSD=2
1   3000470 .   TG  T   217.00  PASS    AC1=1;AC=2;AF1=1;AN=2;DP4=0,0,20,15;DP=42;INDEL;MQ=54;VDB=0.0371
1   3001236 .   A   ATTTTT,ATTTTTTT,ATTTTTTT,ATTTT,ATTT  157.68  Het AC1=1;AC=11,1,1,4,5;AF1=1;AN=22;DP4=0
1   3001242 .   T   TTTG    53.50   PASS    AC1=1;AC=2;AF1=0.5;AN=2;DP4=5,5,14,12;DP=55;INDEL;MDV=99;MQ=5
1   3003197 .   A   AG  33.69   Qual;MinAB  AC1=1;AC=32;AF1=1;AN=32;DP4=0,0,44,4;DP=416;INDEL;MQ=42;VDB=0
1   3003570 .   C   CA  217.00  PASS    AC1=1;AC=2;AF1=1;AN=2;DP4=0,0,22,17;DP=40;INDEL;MQ=48;VDB=0.0308
1   3003640 .   CGGGGG  C,CG,CGGGGGG,CGGGGGGG,CGGGGGGGG 134.47  Qual;Het    AC1=1;AC=16,4,9,2,1;AF1=1;AN=
```



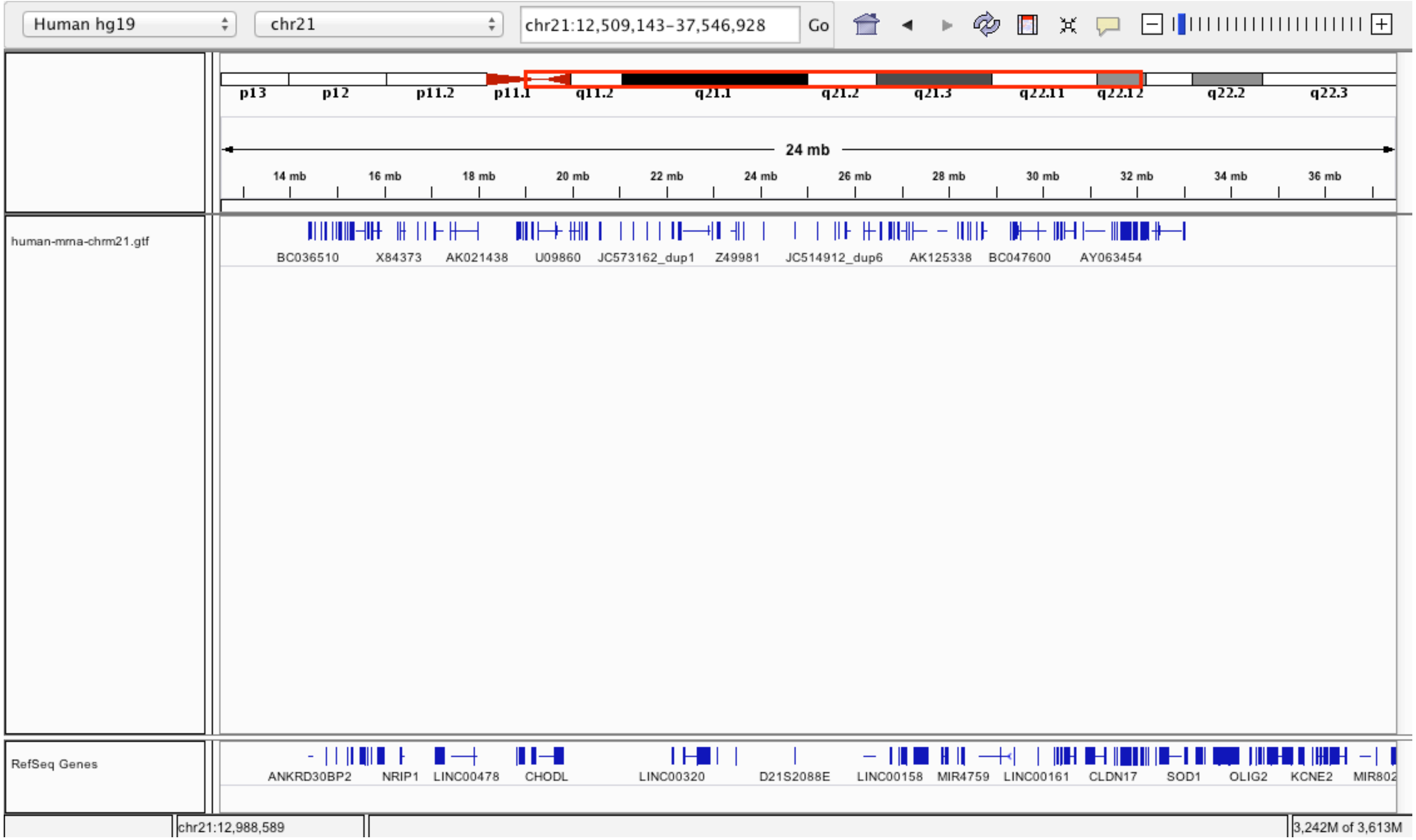
Chr: chr1
Position: 107150240-107150241
ID: .

Genotype Information
Sample: WSBEiJ
Genotype: .
Quality: -1
Type: NO_CALL
Is Filtered Out: No

GFF3 (GTF)

- Used to display genomic features
- GFF files look like this:

```
##gff-version 3
ctg123  Genbank  exon  1300  1500  1  +  0  ID=exon00001
ctg123  Genbank  exon  1050  1500  1  +  0  ID=exon00002
ctg123  Genbank  exon  3000  3902  23 +  0  ID=exon00003
ctg123  Genbank  exon  5000  5500  1  +  0  ID=exon00004
ctg123  Genbank  exon  7000  9000  1  +  0  ID=exon00005
```



Thanks for listening…

…now for something more interesting!