Common data types and file formats

- Is there a specific file type for storing sequencing data?
- What about for genome sequences?
- And for genomic coordinates?
- What about alignment outputs?
- Do different alignment tools have the same output format?
- Are these formats standardized across the multitude of tools?

Common data types and file formats

- Is there a specific file type for storing sequencing data? **Yes**
- What about for genome sequences? **Yes**
- And for genomic coordinates? **Yes**
- What about alignment outputs? **Yes**
- Do different alignment tools have the same output format? **Yes**
- Are these formats standardized across the multitude of tools? **Yes**

Common data types and file formats

- You will encounter 2 major types of data in the world of sequencing data:
    - Sequence data
    - Genome feature data (genomic coordinates)
- Specific file formats represent these data types in a structured manner, and can combine multiple data types in one file.
- Some file formats are not human-readable (***binary***).
- Many are human readable, but extremely large; never use Word or Excel to open these!
- File formats are standardized

Simple sequence formats

- FASTA (simple representation of sequence data: protein & nucleotide)
- FASTQ (complex, includes data quality information: raw sequencing)

Simple sequence formats :: FASTA

```
>SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAAAGCAATGCC
AATA
```

```
>gi|340780744|ref|NC_015850.1| Acidithiobacillus caldus SM-1 chromosome, complete genome
ATGAGTAGTCATTCAGCGCCGACAGCGTTGCAAGATGGAGCCGCGCTGTGGTCCGCCCTATGCGTCCAACTGGAGCTCGTCACGAG
TCCGCAGCAGTTCAATACCTGGCTGCGGCCCCTGCGTGGCGAATTGCAGGGTCATGAGCTGCGCCTGCTCGCCCCCAATCCCTTCG
TCCGCGACTGGGTGCGTGAACGCATGGCCGAACTCGTCAAGGAACAGCTGCAGCGGATCGCTCCGGGTTTTGAGCTGGTCTTCGCT
CTGGACGAAGAGGCAGCAGCGGCGACATCGGCACCGACCGCGAGCATTGCGCCCGAGCGCAGCAGCGCACCCGGTGGTCACCGCCT
CAACCCAGCCTTCAACTTCCAGTCCTACGTCGAAGGGAAGTCCAATCAGCTCGCCCTGGCGGCAGCCCGCCAGGTTGCCCAGCATC
CAGGCAAATCCTACAACCCACTGTACATTTATGGTGGTGTGGGCCTCGGCAAGACGCACCTCATGCAGGCCGTGGGCAACGATATC
CTGCAGCGGCAACCCGAGGCCAAGGTGCTCTATATCAGCTCCGAAGGCTTCATCATGGATATGGTGCGCTCGCTGCAACACAATAC
CATCAACGACTTCAAACAGCGTTATCGCAAGCTGGACGCCCTGCTCATCGACGACATCCAGTTCTTTGCGGGCAAGGACCGCACCC
```

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
```

Line	Description
1	Always begins with '>' and then information about the read (header)
2	The actual sequence (DNA, RNA, protein)

Simple sequence formats :: FASTQ

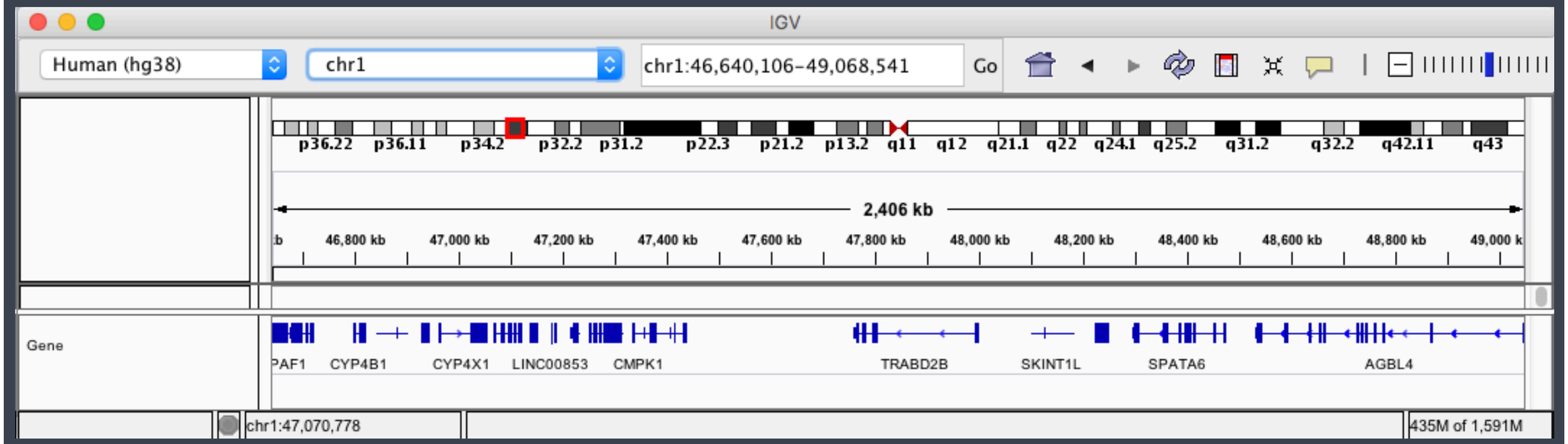
(FASTA with quality information)

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAAAGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""7F@71,'';C?,B;?6B;:EA1EA1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@/=<?7=9<2A8==
```

Line	Description
1	Always begins with '@' and then information about the read (header)
2	The actual DNA sequence
3	Always begins with a '+' and can have the header info from line 1
4	Has a string of characters which represent the quality score

Genomic feature formats (genomic coordinates)

- What are genomic coordinates?



Genomic feature formats (genomic coordinates)

- Tab-delimited (Text file separated by tabs)
- Contain specific information about genome (or assembly) coordinates
- May or may not include sequence data

Example 1: GTF - genomic coordinates of different “features” (mRNA, UTRs, miRNA)

chr1	havana	transcript	112674487	112700739	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	exon	112674487	112674729	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; exon_number "1"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; exon_id "ENSE00001450533"; exon_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	five_prime_utr	112674487	112674729	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	five_prime_utr	112674848	112674912	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	exon	112674848	112675049	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; exon_number "2"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; exon_id "ENSE00003676444"; exon_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";

Genomic feature formats (genomic coordinates)

- Tab-delimited (Text file separated by tabs)
- Contain specific information about genome (or assembly) coordinates
- May or may not include sequence data
- The chromosome names MUST match the reference sequence name
  - Tied to a specific version (assembly/release) of a reference genome
    - hg19/GRCh37
    - hg38/GRCh38
  - Not all reference genomes are the represented the same!
  - E.g. human chromosome 1
    - **UCSC – 'chr1'** versus **Ensembl/NCBI – '1'**
  - Best practice: get the GTF from the same source as the reference genome

Genomic feature formats (genomic coordinates)

- Tab-delimited (Text file separated by tabs)
- Contain specific information about genome (or assembly) coordinates
- May or may not include sequence data

Example 2: SAM/BAM - Read alignment coordinates + sequence

- SAM = Sequence Alignment/Map
- Plain text
- Files can be very large: Many 100's of GB or more
- BAM = BGZF compressed SAM
- Files are typically very large: ~ 1/5 of SAM, but still very large

Commonly used file formats

- FASTA
- FASTQ – Fasta with quality
- GTF – Gene transfer format (genome interval ++)
- SAM – Sequence Alignment/Map format
- BAM – Binary Sequence Alignment/Map format
- *Bed – Basic genome interval (ChIP-seq peaks output)*
- *VCF - Variant Call Format (variant calling output)*
- *Wiggle (wig, bigwig) – Used for visualization of information on genome browsers*

http://genome.ucsc.edu/FAQ/FAQformat.html