



The Elizabeth H.
and James S. McDonnell III

**McDONNELL
GENOME INSTITUTE**
at Washington University



Washington
University in St. Louis

SCHOOL OF MEDICINE

PMBIO Module 02

Inputs. References, Annotations, and Raw Data

Malachi Griffith, Obi Griffith, Zachary Skidmore, Huiming Xia
Introduction to bioinformatics for DNA and RNA sequence analysis (IBDR01)

29 October - 2 November, 2018
Glasgow



Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.



The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:



Attribution — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



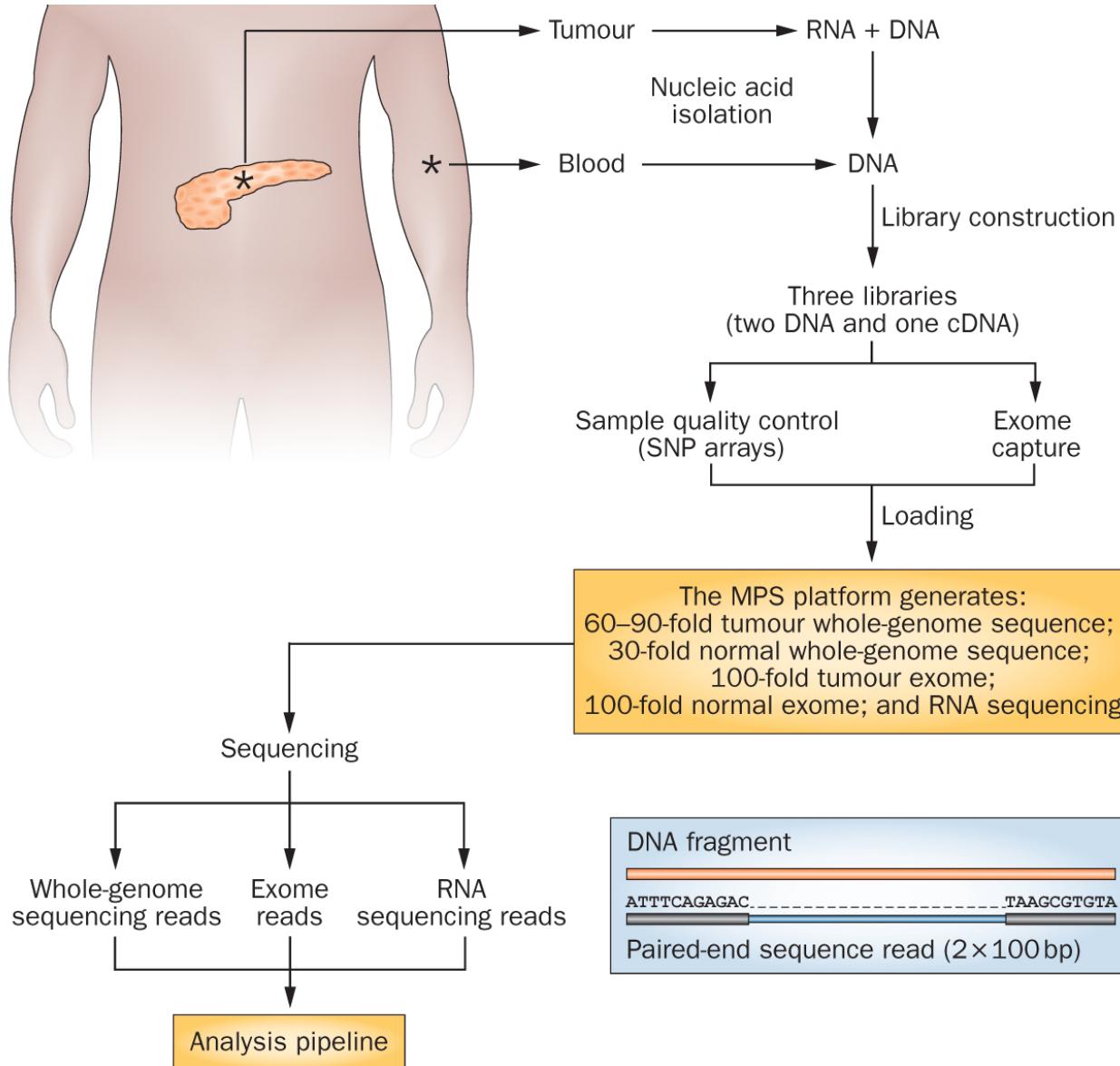
ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

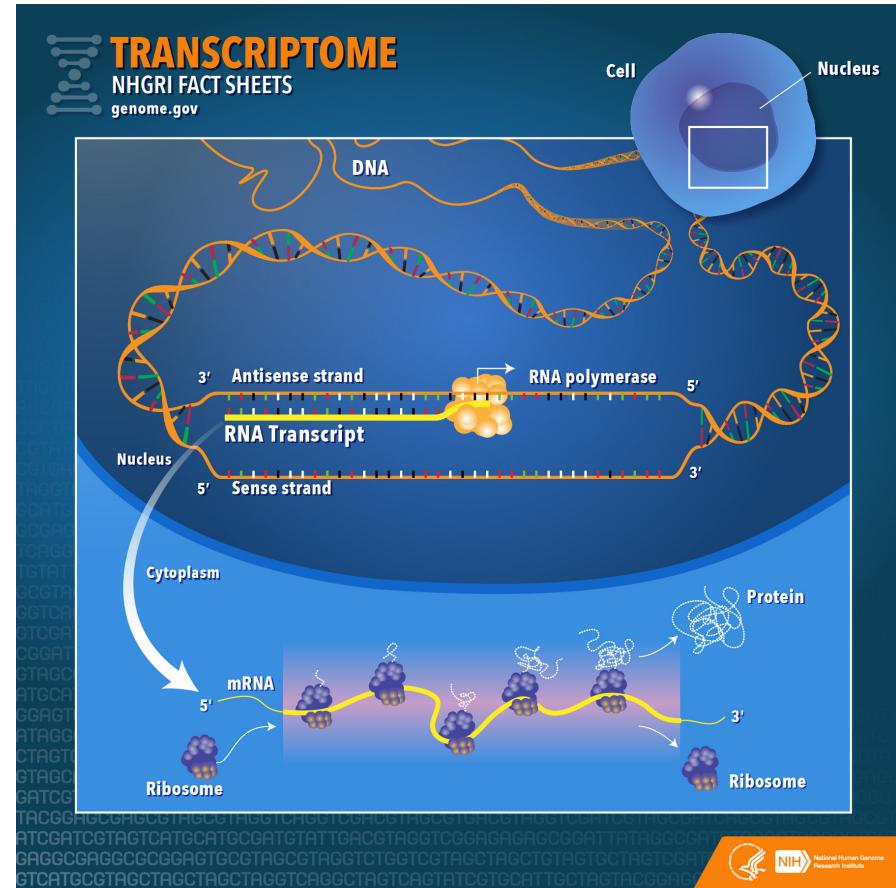
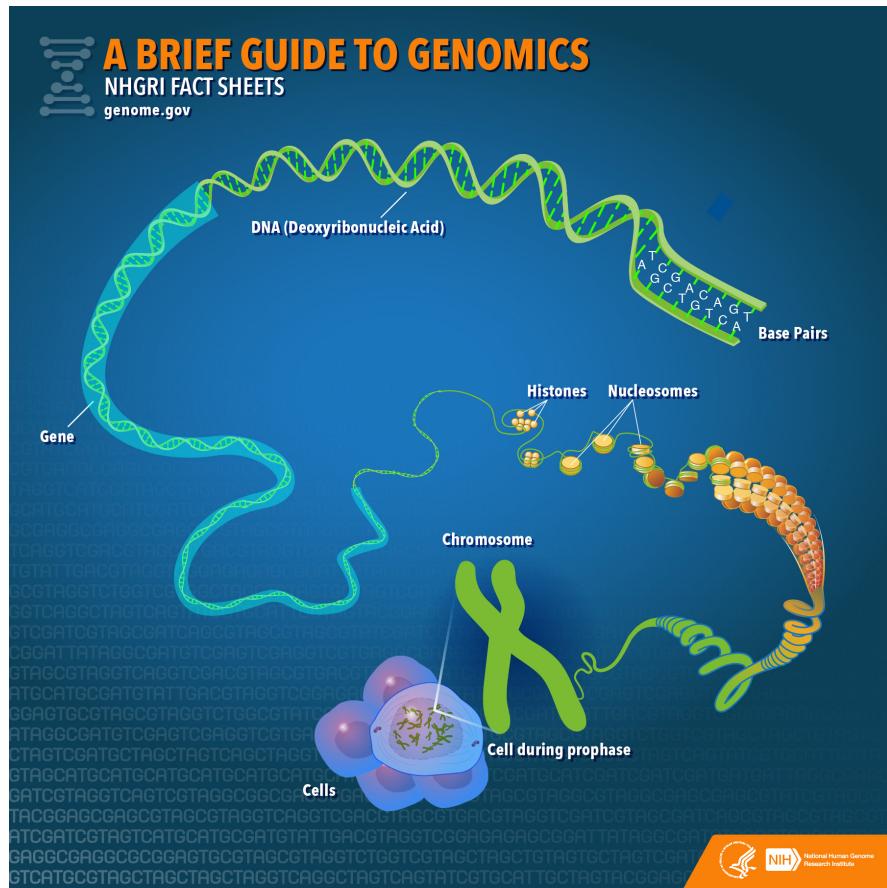
Learning objectives of module 02: Inputs

- **Key concepts:** Central dogma (chromosomes, genes, transcripts, proteins), reference genome assemblies, reference genome versions, FASTA file format, gene/transcript annotation pipelines (Refseq, Ensembl, UCSC, Gencode), GTF file format, sequence data generation, NGS reads, FASTQ file format, raw data QC.
- Obtain reference genome and annotation files and understand the standard formats used to represent them
- Index large files for more efficient access/analysis
- Download and explore raw data files
- Review experimental details for a proof-of-principle personalized genomics exercise
- Perform a raw data quality assessment and discuss any data quality issues that are observed. What are their implications for interpretation of the results?

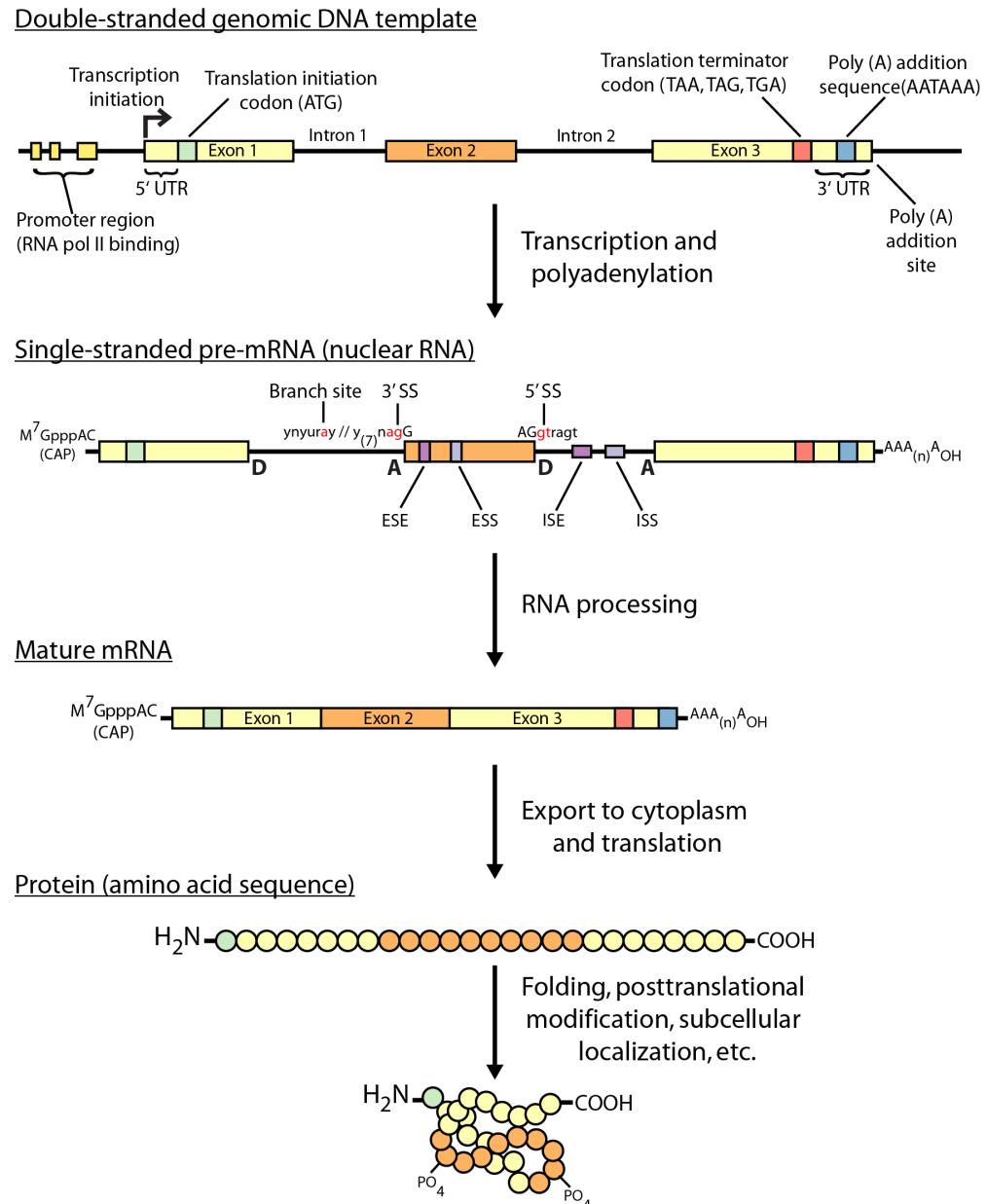
Cancer genomics data has exploded with rapid advances in sequencing technologies



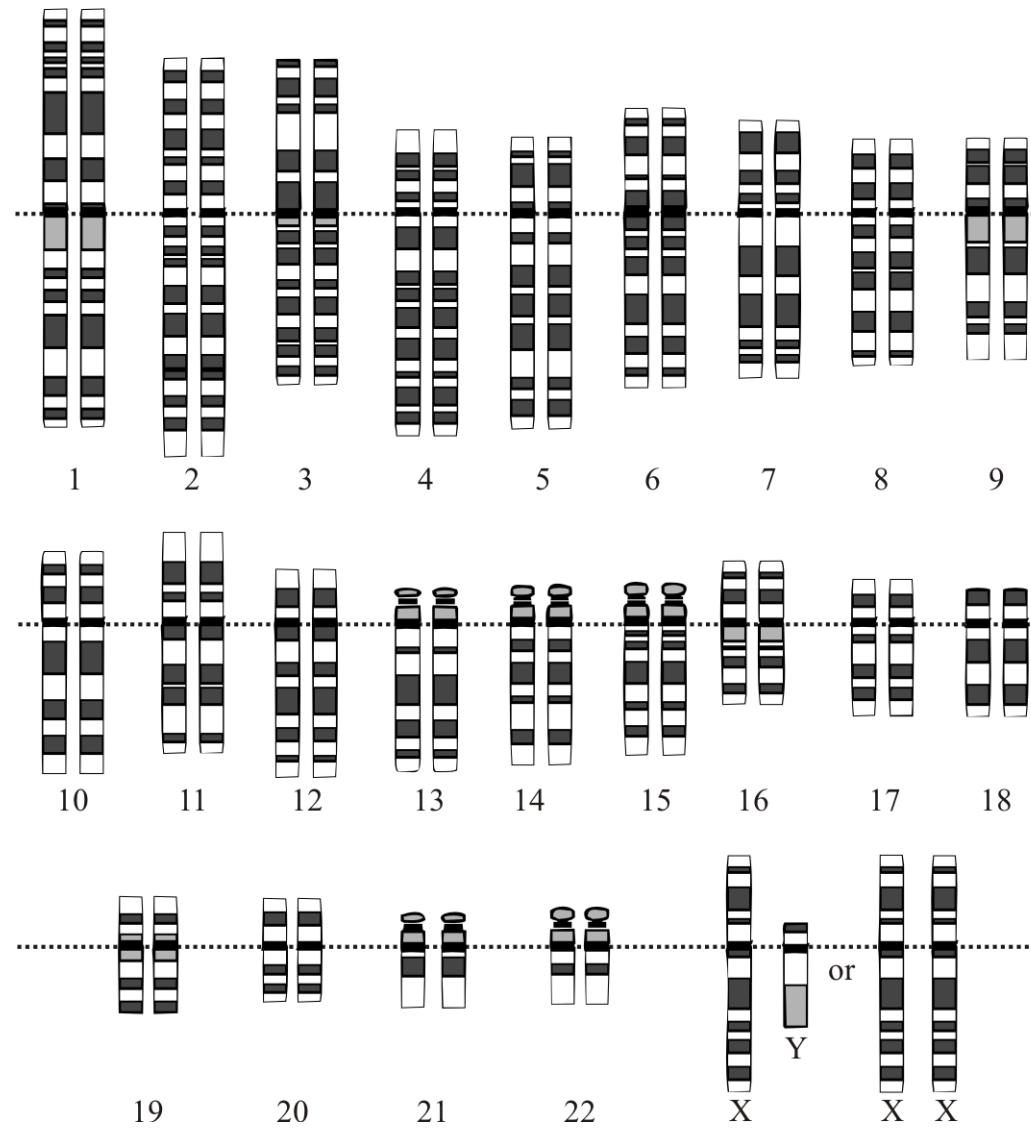
Genomes and transcriptomes



The Central Dogma



Human chromosomes (karyotype)



Our reference genome

- All reference files were obtained from the 1000 genomes project
 - The GRCh38 build of the human genome is used
 - This is the latest version of the human reference
- For the tutorial, two chromosomes are used (chr. 6 and chr. 17)
 - The reason for this is to reduce run time for the tutorial
 - Performing this analysis on the complete genome reference would require only minor modification of the commands
 - Would also require more storage, compute resources, and time

Reference genome assemblies/versions/builds

- NOTE!: Probably ~50% of all bioinformatics results problems where something unexpected is happening involve genome coordinate build incompatibility
 - e.g. reads were aligned to build 37 but you are using transcript annotations based on build 38 coordinates
- Learn to use ‘lift-over’ tools
 - <https://www.biostars.org/p/65558/>

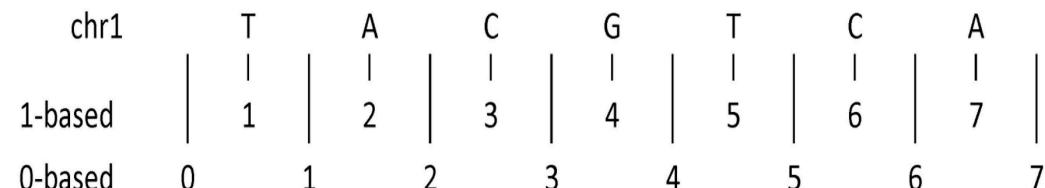
1- versus 0-based coordinates

- NOTE!: The other ~50% relate to 1 versus 0 based coordinates or conceptually similar data parsing issues
- Learn the basics of these two coordinate systems that are both used ubiquitously in genomics
 - <https://www.biostars.org/p/84686/>

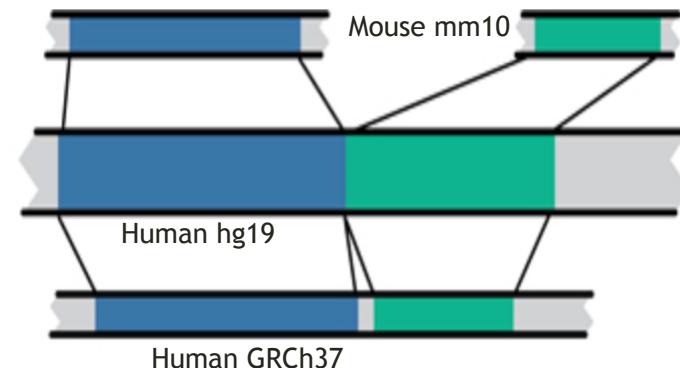
Common sources of confusion

- Genomic coordinate systems

- 1-based vs. 0-based

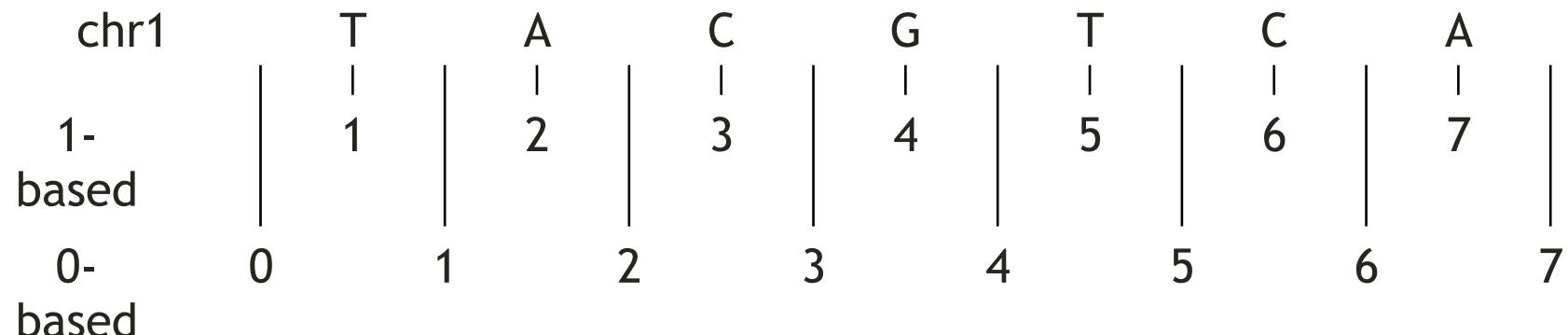


- Genome builds
 - And annotation builds
 - “Liftover” tools
- Left-shifted vs right-shifted



REF	CAC	6	CAC	C	Not left aligned but parsimonious
ALT	C				Not right trimmed
REF	GCACA	3	GCACA	GCA	Not left trimmed
ALT	GCA				
REF	GGCA	2	GGCA	GG	
ALT	GG				
REF	GCA	3	GCA	G	Normalized (left aligned & parsimonious)
ALT	G				

0-based vs 1-based method to indicate a single nucleotide or variant



	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

- 1-based coordinate system
 - Single nucleotides, variant positions, or ranges are specified directly by their corresponding nucleotide numbers
 - GFF, SAM, VCF, Ensembl browser, ...
- 0-based coordinate system
 - Single nucleotides, variant positions, or ranges are specified by the coordinates that flank them
 - BED, BAM, UCSC browser, ...

Common genomic data file formats

- FASTA - sequences (DNA, RNA, protein)
 - FASTQ - raw sequence data (with qualities)
 - SAM/BAM/CRAM - aligned sequence data
 - GTF - gene/transcript annotations
 - BED - other genome features
 - VCF - variant calls (individual, multi-individual)
 - MAF - aggregated variant information (project, population)
-
- Many, many custom data formats output by specialized tools ... often in TSV format

FASTA file format example

```
>18 ENST00000603290 chr6+ 319926-320364,335114-335163,345854-345912
ATAGCAGGGATCGAAAATGTCTAGTAATCGCTTTCTTGAAGCAAGCTGTATCTAGGT
TTCTCGTGTGGCAGTCGGGTGTGTAAGAGCAGGCCCTCCACTGGCTGGCCTGCTACCAC
AGACTCACTTGCAAATGGAGCCCTAGAACTCCTGCATGGCTGGACTCAGAGGGCTTGAGA
TCTTCCAGTACTTAACTGAGTCTAGGGAGGACATGGTTTTAAACTTCTAAATCAAAGA
ACTCTTTGGCCTGGCTTCCAGAGGTACCCCTGTCCTCAGCTGGGACCTCTCTCTAG
GGCTCGTGCTCACTGACCAGCCTCTCCTACAACACTGAAGACTTTGCAGGACACTCTTCC
TAGATGTTGCCTTCATGTTGAGTCCCAGGTGTGAGCATTCTCCTGACGGCTGTACTT
GAGCAGTCACCTGGCACAGGGAGTTAAATACCTGTGCATCCCAGCAGCGGATTACCATC
TCAAAACCTGACAAGACATTCAAAGAAAGTATTAAATTCACTCACGAGTGCCGGCTCCG
CGGTGAGA
>19 ENST00000605391 chr6+ 319952-320364,345854-345928,348103-348161
AATCGCTCTTCTTGAAGCAAGCTGTATCTAGGTTCTCGTGTGGCAGTCGGGTGTGTAAG
GAGCGGCCCTCCACTGGCTGGCCTGCTACCAACAGACTCACTGCAAATGGAGCCCTAG
AACTCCTGCATGGCTGGACTCAGAGGGCTTGAGATCTTCCAGTACTTAAC TGAGTCTAGG
GAGGACATGGTTTTAAACTCTAAATCAAAGAACTCTTGGCCTGGCTTCAGGAGG
TACCCCTGTCTTCAGCTGGGACCTCTCTAGGGCTCGTGCTCACTGACCAGCCTCTC
CTTACAAC TGAAAGACTTTGCAGGACACTCTCCTAGATGTTGCCTCATGTTGAGTCCC
GGTGTCTGAGCATTCTCCTGACGGCTGTACTTGAGCAGTCACCTGGCACAGGACAAGA
CATTCAAAGAAAGTATTAAATTCACTCACGAGTGCCGGCTCCGCGGTGAGAGCTGCCTT
GTACACTGCCTGGCCGGGTCTCCAGGAGCGTGACACTGGTGATCGCATACTCATGACC
GTCACTG
```



Raw data

- For purposes of the tutorial, the test data has been pre-filtered
 - Identified reads that appear to match transcripts on the selected chromosomes
- The test data corresponds to a single hypothetical human breast cancer patient (cell lines)
 - HCC1395 and HCC1395/BL
 - Genomic DNA and RNA were isolated
 - The RNA samples also included one of two ERCC RNA “spike-in” mixes (Mix1 or Mix2)
- The input data is provided in ‘fastq’ format:
 - http://en.wikipedia.org/wiki/FASTQ_format

FASTQ file format example

Fastq files represent raw sequence - base calls and qualities

@HWI-ST718_146963544:6:1213:8996:10047/1

CTTTTTTATTTTGCTGACTGGGTTGATTCAAAA

+

CCCFFFFFHHHHGJHIIJHIHIIIFHIJJJJGIBBFGE

@HWI-ST718_146963544:5:2303:11793:37095/1

ATGAATTATAGGGCTGTATTTAATTTGCATTTAA

+

@@??BDDFFF<FHEGFFGGIEBGHIIIIIBEHIIGIH<FHE

Phred scores and ASCII glyphs

Phred Q	Probability (P) of Wrong Base	Base Call Accuracy	Sanger “Q + 33” Shift	Sanger “Q + 33” Shift ASCII glyph
0	1	0	33	!
1	0.794	0.206	34	“
2	0.631	0.369	35	#
10	0.1	0.90	43	+
20	0.01	0.99	53	5
30	0.001	0.999	63	?

Encoding History:

- Sanger Format (shown above): Q of 0 to 93 using ASCII 33 to 126
 - Sanger data, SAM format, Illumina 1.8+
- Solexa/Illumina 1.0: Q of -5 to 62 using ASCII 59 to 126
- Illumina 1.3 to 1.8: Q of 0 to 62 using ASCII 64 to 126
- Illumina 1.5 to 1.8: Phred scores 0 to 2 have a slightly different meaning

Known transcript annotations

- All annotation files are obtained from Ensembl
 - <http://useast.ensembl.org/info/data/ftp/index.html>
 - There are many other ways to obtain gene annotation files. For example:
 - UCSC Genome Browser, Ensembl API, BioMart, Entrez, Galaxy, etc. could also be used
- You will download GTF files describing human transcripts (exon coordinates, gene ids, gene symbols, etc.)
- Descriptions of the GTF file format can be found here:
 - <http://genome.ucsc.edu/FAQ/FAQformat.html#format4>

GTF file format example

chr6	havana	gene	167607	170631	.	-	.	gene_id "ENSG00000218577"; gene_ver	
chr6	havana	transcript	167607	170631	.	-	.	.	gene_id "ENSG00000218577";
chr6	havana	exon	170352	170631	.	-	.	gene_id "ENSG00000218577"; gene_ver	
chr6	havana	exon	169545	169753	.	-	.	gene_id "ENSG00000218577"; gene_ver	
chr6	havana	exon	167607	167826	.	-	.	gene_id "ENSG00000218577"; gene_ver	
chr6	havana	gene	181466	205484	.	-	.	gene_id "ENSG00000263667"; gene_ver	
chr6	havana	transcript	181466	205484	.	-	.	.	gene_id "ENSG00000263667";
chr6	havana	exon	204887	205484	.	-	.	gene_id "ENSG00000263667"; gene_ver	
chr6	havana	exon	198165	198509	.	-	.	gene_id "ENSG00000263667"; gene_ver	
chr6	havana	exon	189476	189649	.	-	.	gene_id "ENSG00000263667"; gene_ver	
chr6	havana	exon	187736	187906	.	-	.	gene_id "ENSG00000263667"; gene_ver	
chr6	havana	exon	186356	186561	.	-	.	gene_id "ENSG00000263667"; gene_ver	
chr6	havana	exon	184593	184809	.	-	.	gene_id "ENSG00000263667"; gene_ver	
chr6	havana	exon	182366	182474	.	-	.	gene_id "ENSG00000263667"; gene_ver	
chr6	havana	exon	181466	181695	.	-	.	gene_id "ENSG00000263667"; gene_ver	

Indexing reference genomes

- Before sequences can be mapped to the genome, the reference genome files must be ‘indexed’ in a way that is compatible with the aligner being used
 - Each alignment algorithm generally requires a custom index to be built for that purpose
 - Do not use an index created for another aligner
- In general we will encounter the concept of “indexing” many times throughout the course
 - Reference genomes (FASTA files), reference transcriptomes (GTF files), alignments (BAM files), variants (VCF files), etc.

Pre-Alignment QC

