The Elizabeth H.
and James S. McDonnell III
McDonnell Genome Institute
at Washington University

GenViz Module 1: Introduction to genomic data visualization and interpretation

Malachi Griffith, Obi Griffith, Zachary Skidmore

Genomic Data Visualization and Interpretation
April 8-12, 2019
Freie Universität Berlin

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 the course

- **Module 1: Introduction to genomic data visualization and interpretation**
- Module 2: Using R for genomic data visualization and interpretation
- Module 3: Introduction to GenVisR
- Module 4: Expression profiling, visualization, and interpretation
- Module 5: Variant annotation and interpretation
- Module 6: Q & A, discussion, integrated assignments, and working with your own data

- Tutorials
  - Provide working examples of data visualization and interpretation
  - Self contained, self explanatory, portable

Learning objectives of module 1

- Introduction to genomic data visualization and interpretation

- Introduction
- ‘Omics technologies
- Common genomic file formats
- Common problems
- Genome browsers
- Fundamentals of data visualization
- Best practices in visualization
- Examples of visualization in R

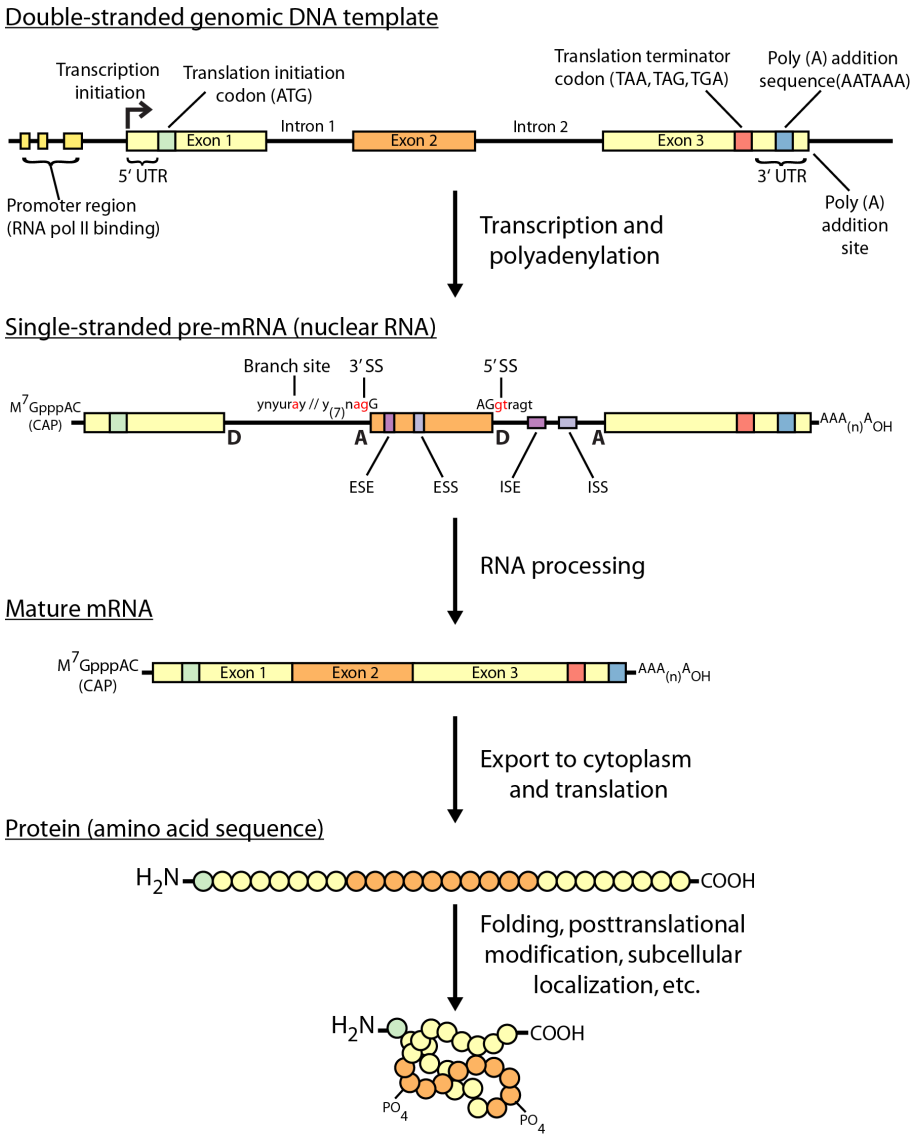
Introduction

Why do we create visualizations of genomic data?

- Data exploration and interpretation of results
  - QC analysis
  - Understanding whether/how an experiment worked
  - Discovery

- Communication
  - Slides for presentations
    - e.g. Keynote, Powerpoint, etc.
  - Figures for publications
    - e.g. PDFs, PNGs, etc.
      - Illustrator, Gimp, Inkscape, etc.
      - R and R Studio

Review of central dogma

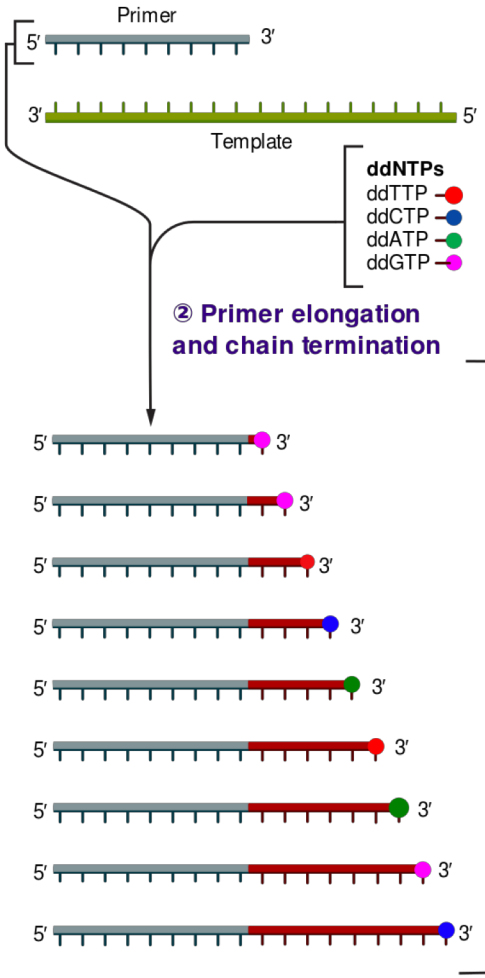


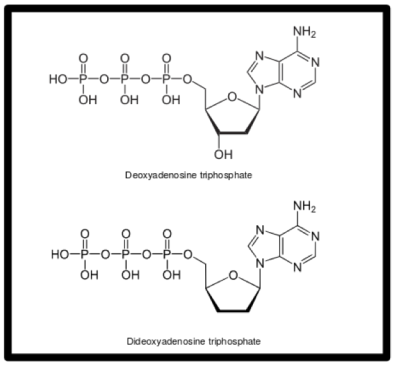
‘Omics technologies

Non-parallel (Sanger) sequencing

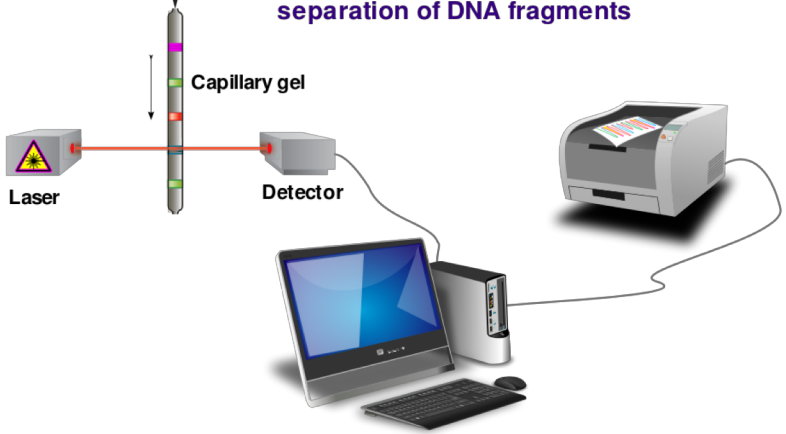
① Reaction mixture

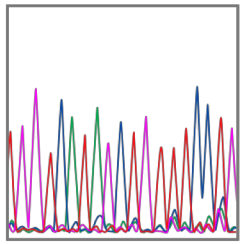
- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)





③ Capillary gel electrophoresis separation of DNA fragments



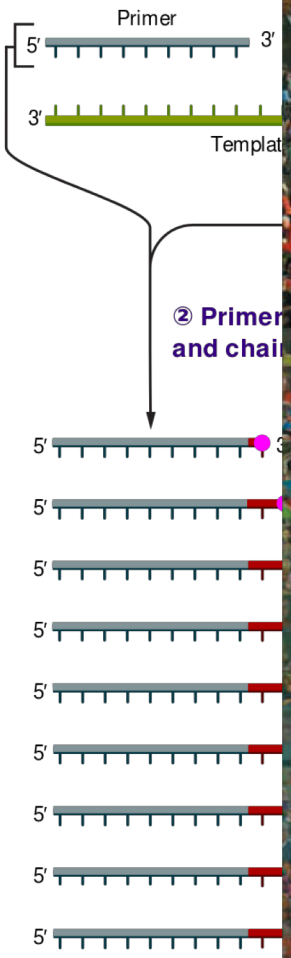


Chromatograph

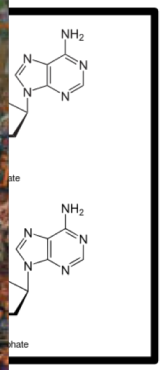
④ Laser detection of flourochromes and computational sequence analysis

Non-paral

① Reaction mixture
- Primer and DNA temp
- ddNTPs with flouroch



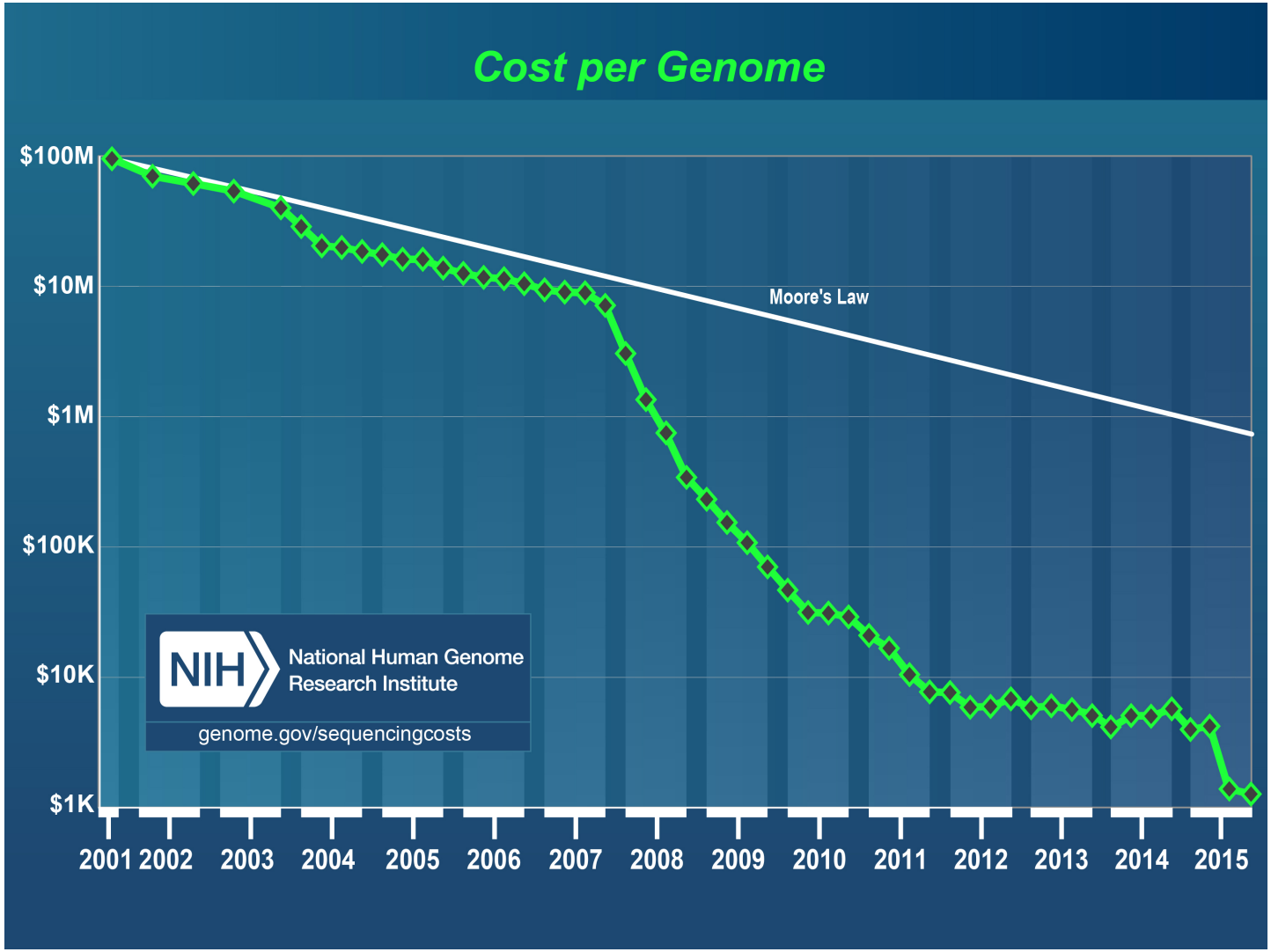
② Primer
and chai



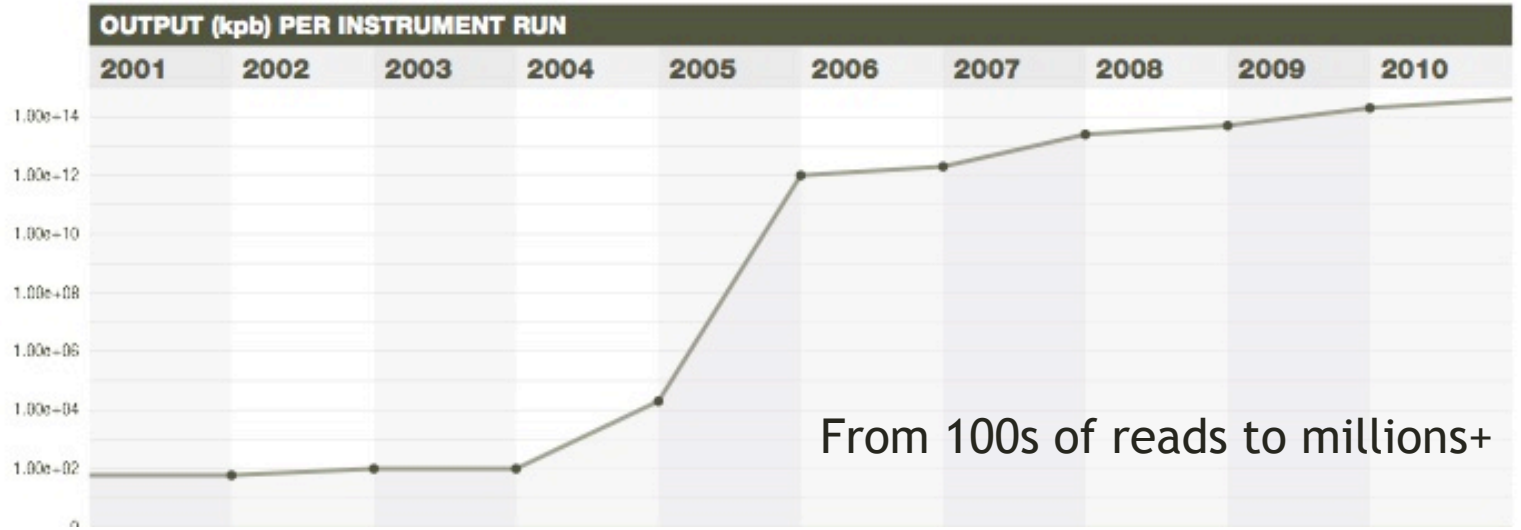
rophoresis
gments

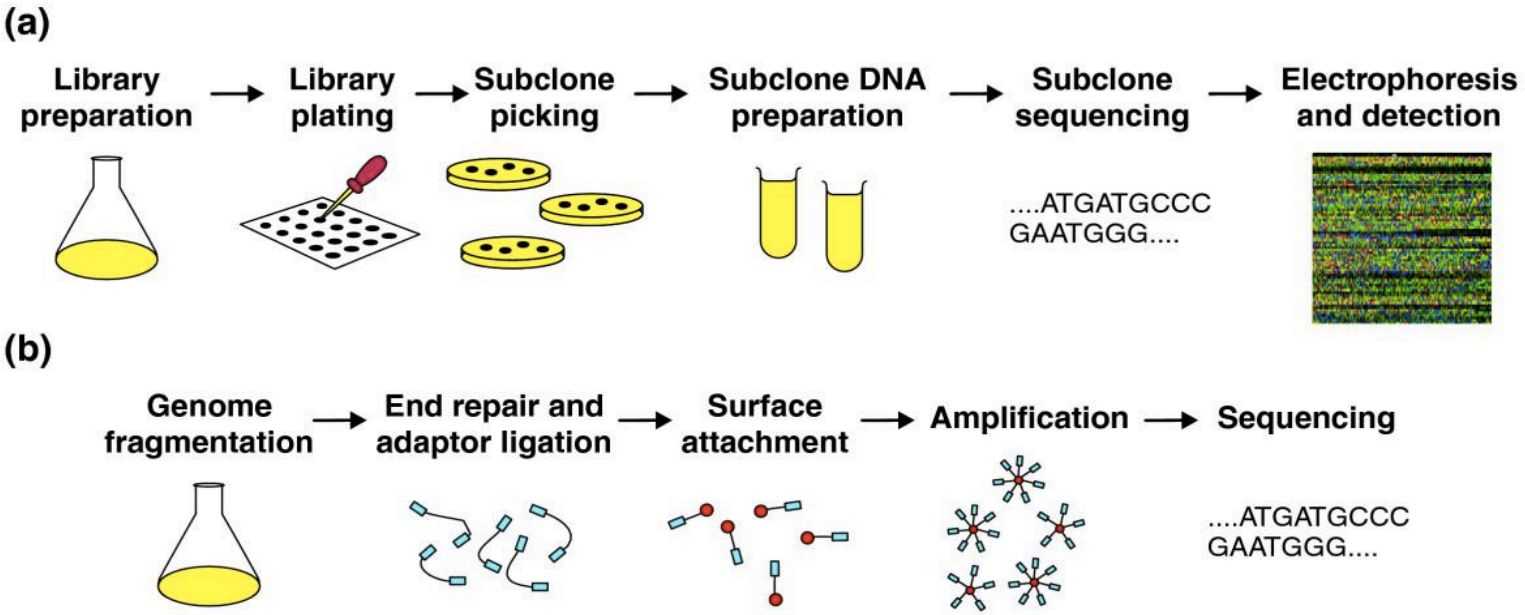
ion of flourochromes
nal sequence analysis

Costs of sequencing a human genome have plummeted from $100M to $1000



Massively parallel sequencing (NGS) has transformed biomedical inquiry



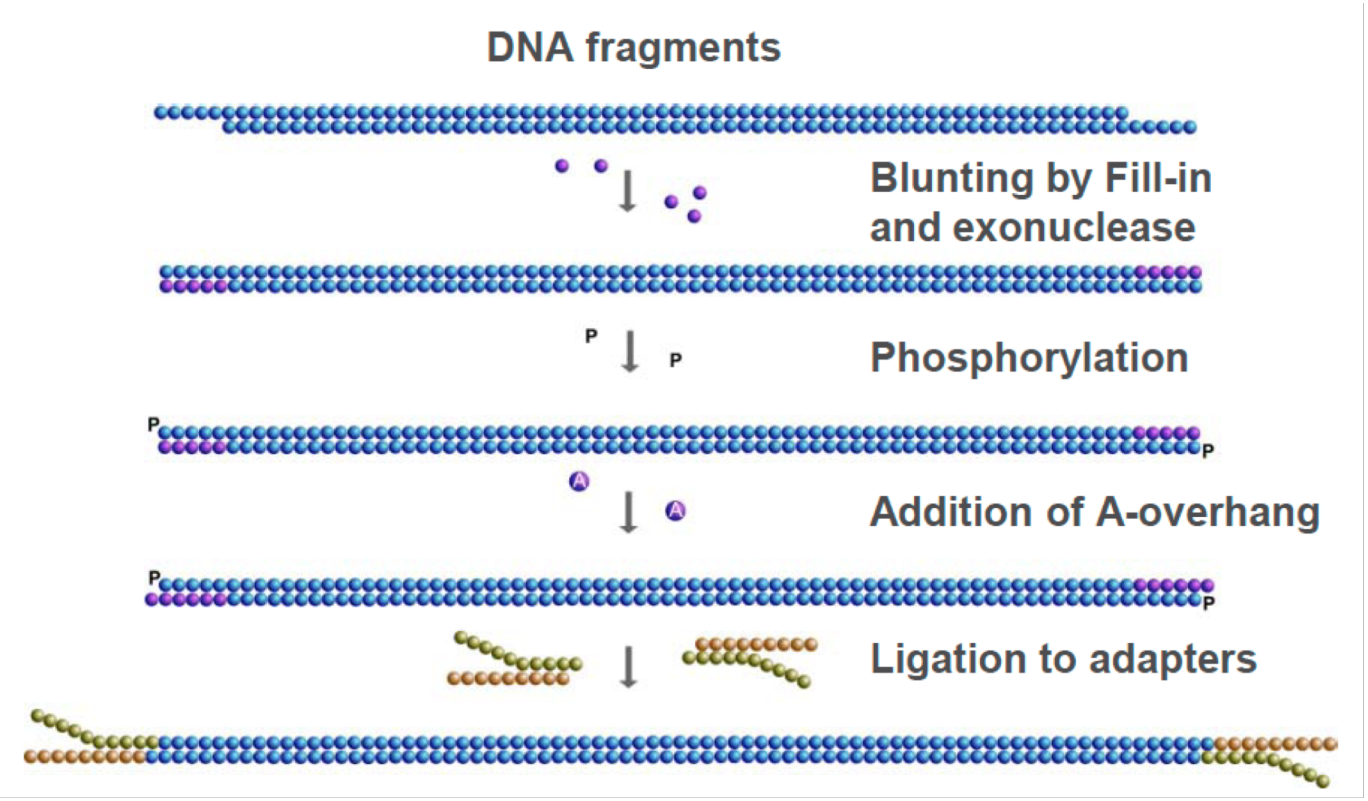


E.R. Mardis, Nature (2011) 470: 198-203, Ann. Rev. Analyt. Chem. (2013)

Massively Parallel DNA sequencing instruments

- All MPS platforms require a library obtained either by amplification or ligation with custom linkers (adapters)
- Each library fragment is amplified on a solid surface (either bead or flat surface) with covalently attached adapters that hybridize the library adapters
- Direct step-by-step detection of the nucleotide base incorporated by each amplified library fragment set
- Hundreds of thousands to hundreds of millions of reactions detected per instrument run = “massively parallel sequencing”
- A “digital” read type that enables direct quantitative comparisons
- Shorter read lengths than capillary sequencers

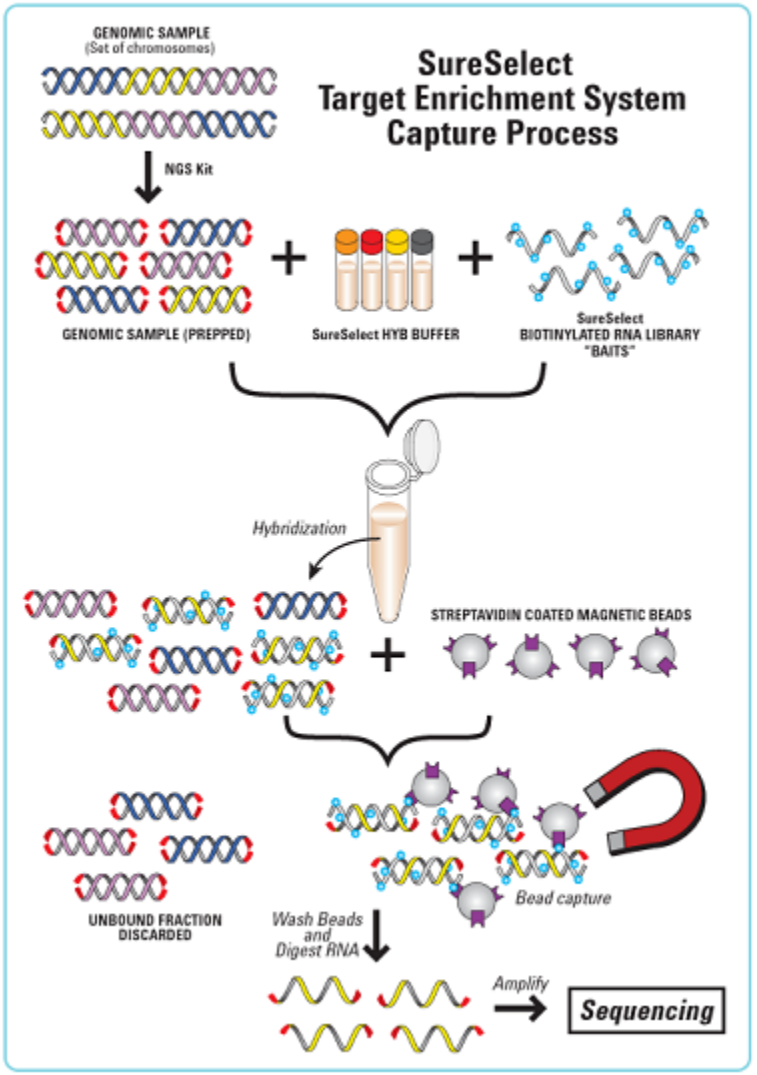
Library Construction for MPS



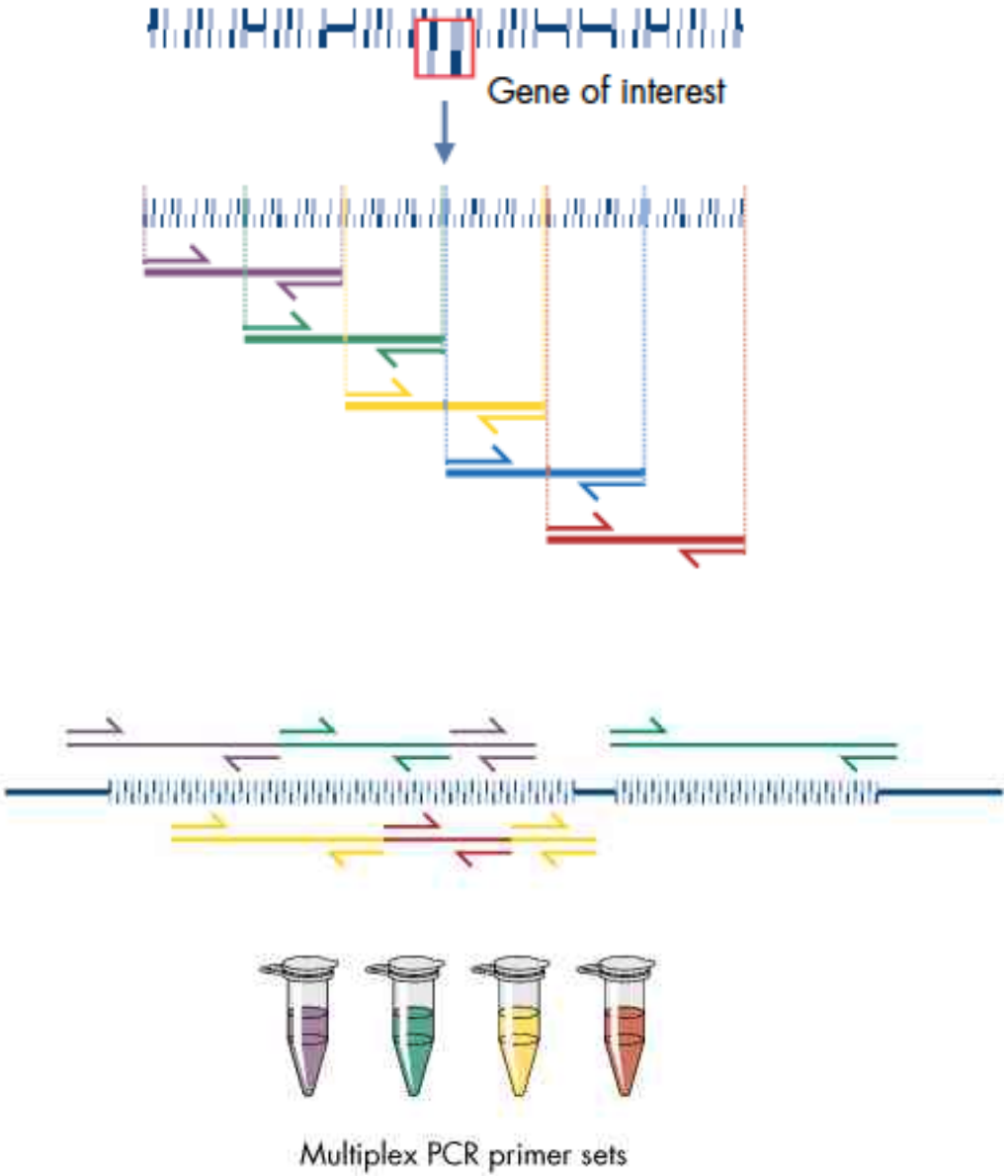
- Shear high molecular weight DNA with sonication
- Enzymatic treatments to blunt ends
- Ligate synthetic DNA adapters (each with a DNA barcode), PCR amplify
- Quantitate library
- Proceed to WGS, or perform exome or specific gene hybrid capture

Hybrid Capture

- **Hybrid capture** - fragments from a whole genome library are selected by combining with probes that correspond to most (not all) human exons or gene targets.
- The probe DNAs are biotinylated, making selection from solution with streptavidin magnetic beads an effective means of purification.
- An "**exome**" by definition, is the exons of all genes annotated in the reference genome.
- **Custom capture reagents** can be synthesized to target specific loci that may be of clinical interest.

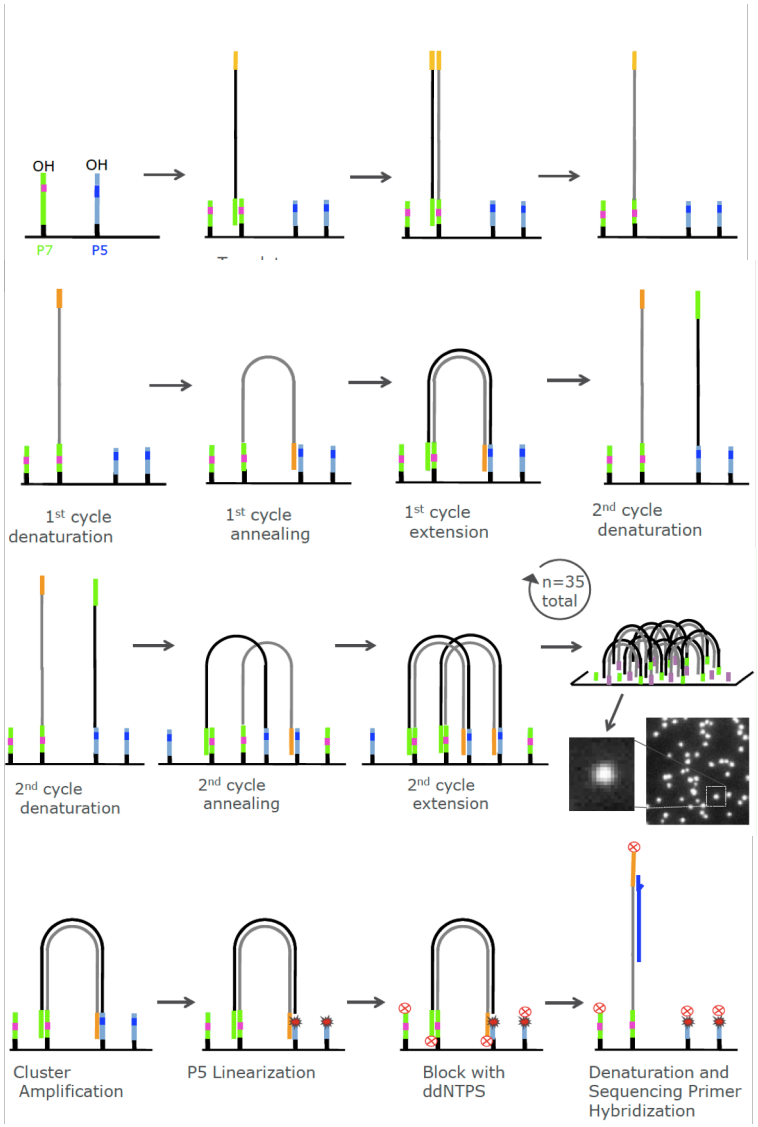


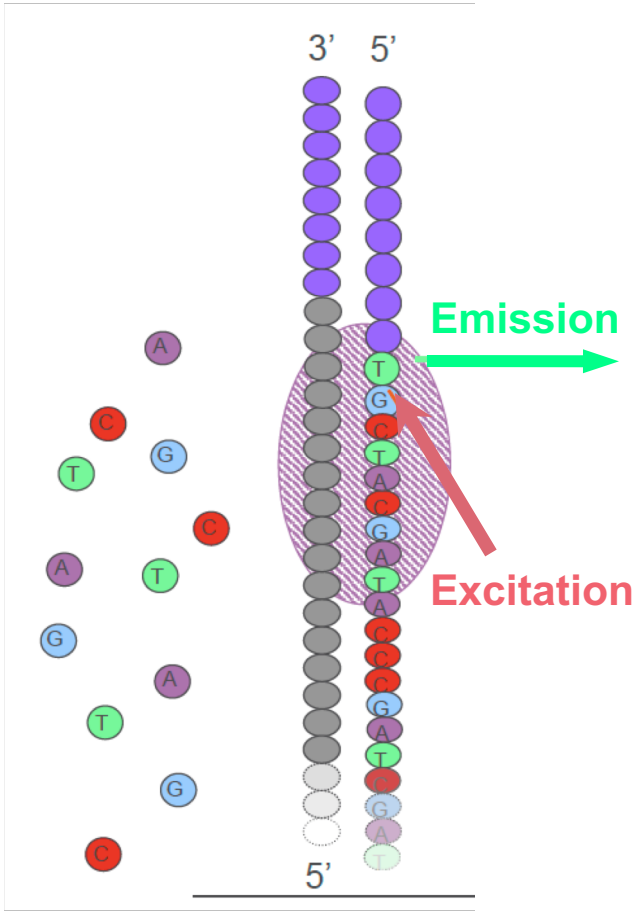
Multiplex PCR Amplification of Targets

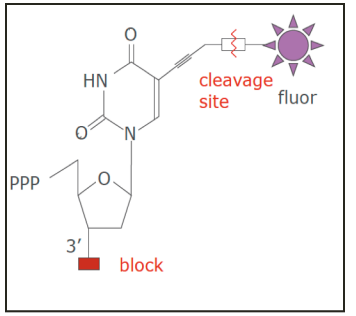


1. Design amplification primer pairs for exons of genes of interest; tile primers to overlap fragments in larger exons
2. Group primer pairs according to G+C content, Tm and reaction condition specifics
3. Amplify genomic DNA to generate multiple products from each primer set; pool products from each set
4. Create library by ligation or tail platform adaptors on the primer ends
5. Sequence

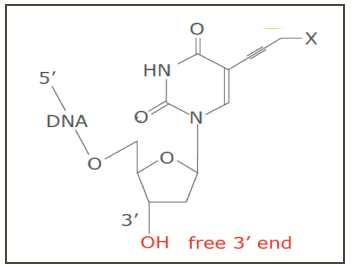
Illumina: Massively Parallel Sequencing by Synthesis





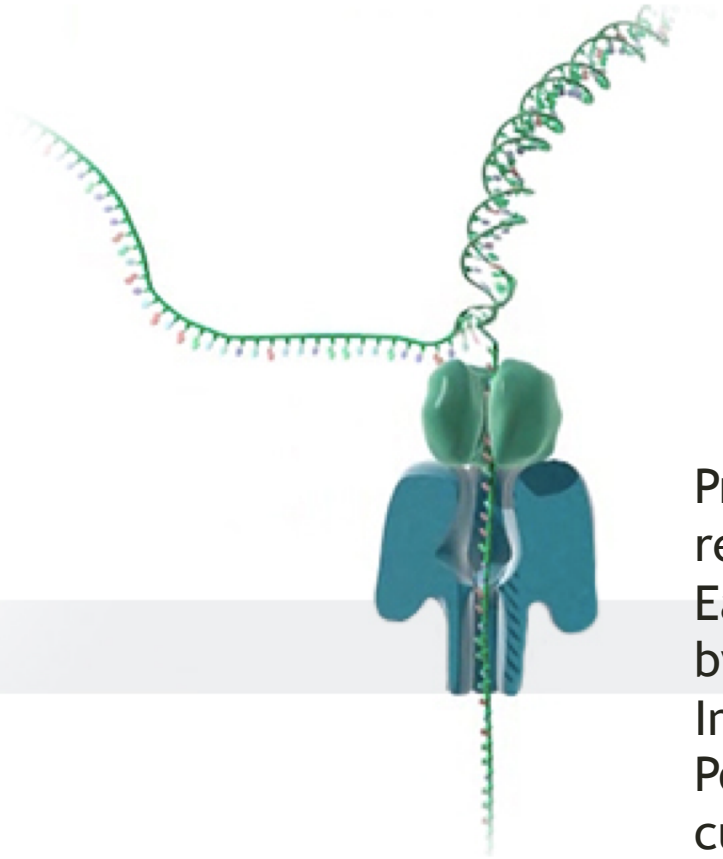


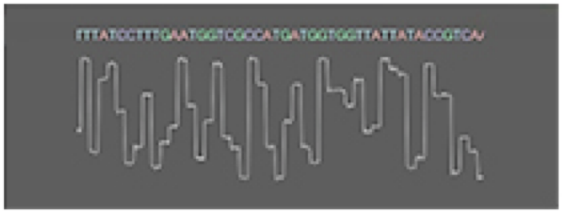
Incorporate
Detect
De-block
Cleave fluor



Additional amplification on flowcell to form clusters

Platforms: Oxford Nanopore Sequencing





Protein nanopore set in an electrically resistant polymer membrane
Each nanopore is controlled and measured by a matched ASIC (Application Specific Integrated Circuit)
Post-run data analysis compares pore current changes to a model of all possible multimers

- Variable read lengths
- Electrical current-based detection of multiplex nucleotides in pore
- Error rate is around 10-20% with newest pore/software
- Usage to-date is for bacterial and viral genome sequencing, RNA sequencing

Nanopore Sequencing Devices

- Commercially available
- $1000 starter pack
- 512 pores
- Simple 10-minute sample prep
- 10-40k reads
- 20—100MB per run

- Size of a “usb stick”
- Long reads (median 1kb, max ~100kb)
- Relatively high error rate (38% in 2015)
  - Improving rapidly with 2D reads and other optimization

Future nanopore developments

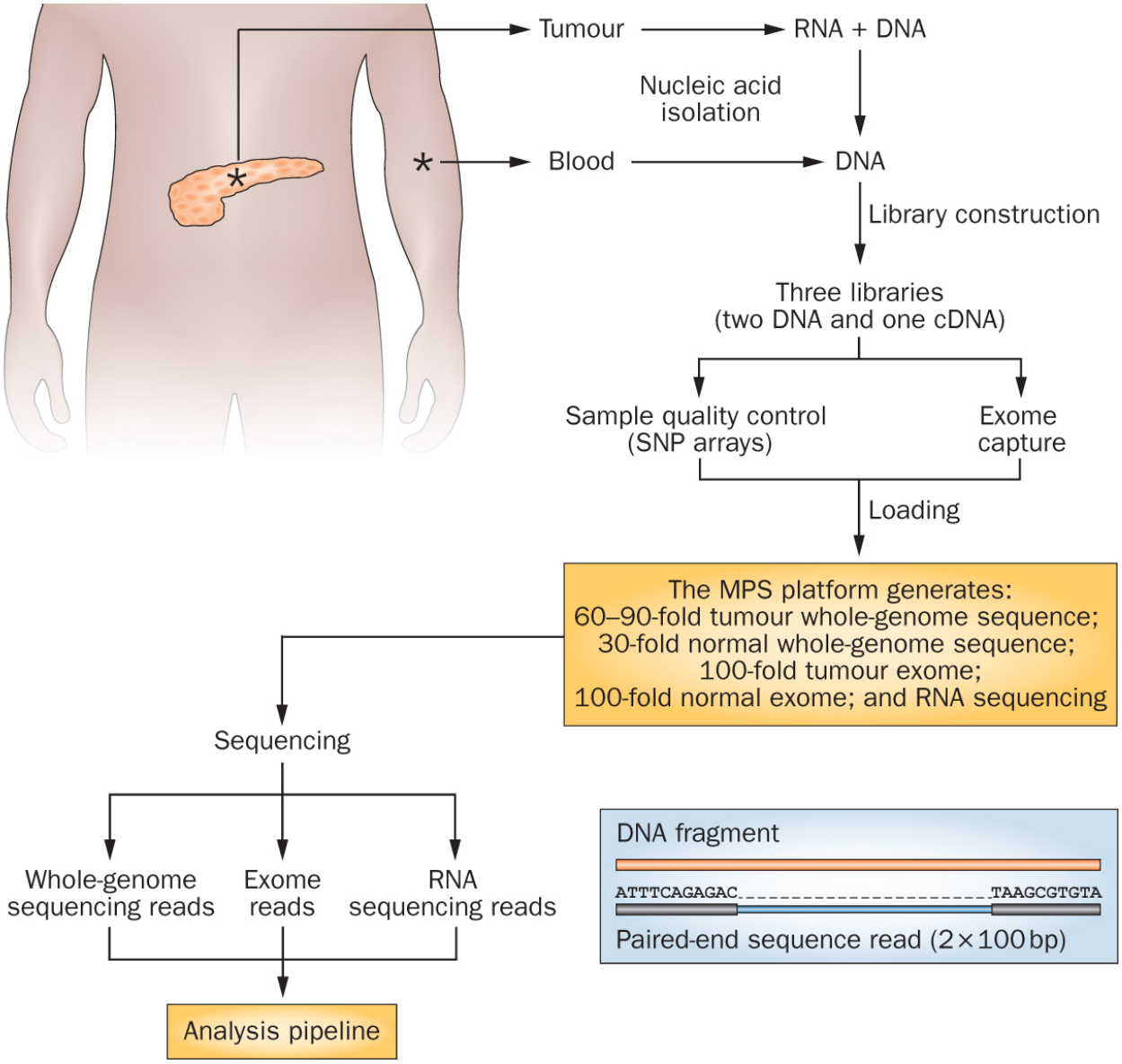
promethION

- Early access
- Price: TBD
- 48 x 3000 pores
- Array of minions

smidgION

- In development
- Smart-phone compatible

Genomics research has exploded with the rapid advances in DNA sequencing technologies



Short read alignments – The fundamental unit of most omics assays

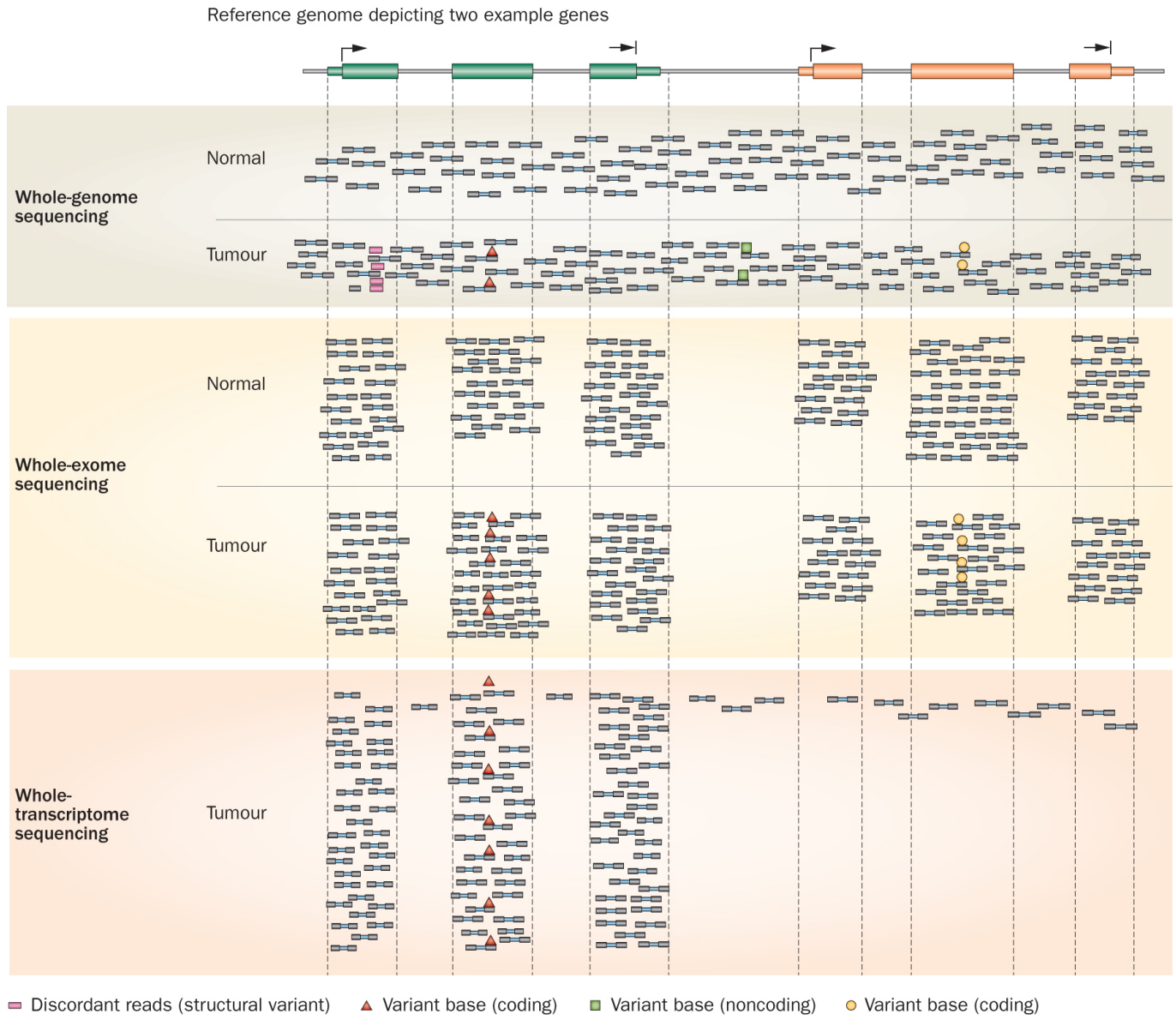
- Alignment is about fitting individual pieces (reads) into the correct part of the puzzle
- The human genome project gave us the picture on the box cover (the reference genome)
- Imperfections in how the pieces fit can indicate changes to a copy of the picture

Reference: AGCCTGAGACCGTAAAAAAGTCAAG

A read sequence: GAGACCGTAAAAACGTC

A variant!

Whole genome, exome and transcriptome sequencing allows us to detect and confirm many different variant types



Common genomic file formats

Common genomic data file formats

- FASTA
- FASTQ
- GTF
- BED
- VCF
- VEP
- MAF
- SAM/BAM/CRAM

- Many, many custom data formats output by specialized tools … often in TSV format

Fasta – format for representing nucleic acid or amino acid sequences

```
>AY274119.3 Severe acute respiratory syndrome-related coronavirus
isolate Tor2, complete genome

ATATTAGGTTTTTACCTACCCAGGAAAAGCCAACCAACCTCGATCTCTTGTAGATCTGTTCTCTAAACGA
ACTTTAAAATCTGTGTAGCTGTCGCTCGGCTGCATGCCTAGTGCACCTACGCAGTATAAACAATAATAAA
TTTTACTGTCGTTGACAAGAAACGAGTAACTCGTCCCTCTTCTGCAGACTGCTTACGGTTTCGTCCGTGT
TGCAGTCGATCATCAGCATACCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTTC
TTGGTGTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTCCTTCAGGTTAGAGACGTGCTAGTGCG
TGGCTTCGGGGACTCTGTGGAAGAGGCCCTATCGGAGGCACGTGAACACCTCAAAAATGGCACTTGTGGT
…

>FJ882960.1 SARS coronavirus ExoN1 isolate P3pp34, complete genome
CGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTAGCTGTCGCTCGGCTGCATGCCTA
GTGCACCTACGCAGTATAAACAATAATAAATTTTACTGTCGTTGACAAGAAACGAGTAACTCGTCCCTCT
TCTGCAGACTGCTTACGGTTTCGTCCGTGTTGCAGTCGATCATCAGCATACCTAGGTTTCGTCCGGGTGT
…
```

First line starts with “>” header or ”Comment”; used as a summary/description, often starting with unique accession/identifier

Subsequent lines contain sequence

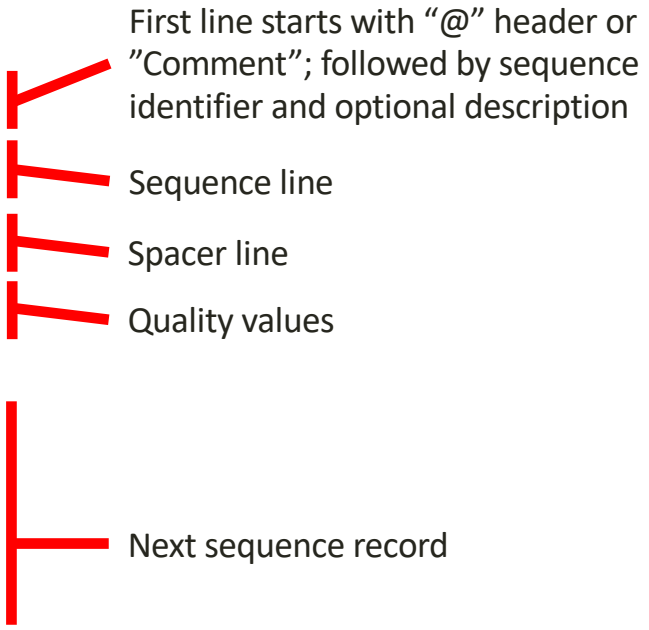
- Interleaved: sequence broken into multiple lines of characters
- Sequential: entire sequence on a single line

Multiple sequence FASTA obtained by simply concatenating multiple FASTA records together

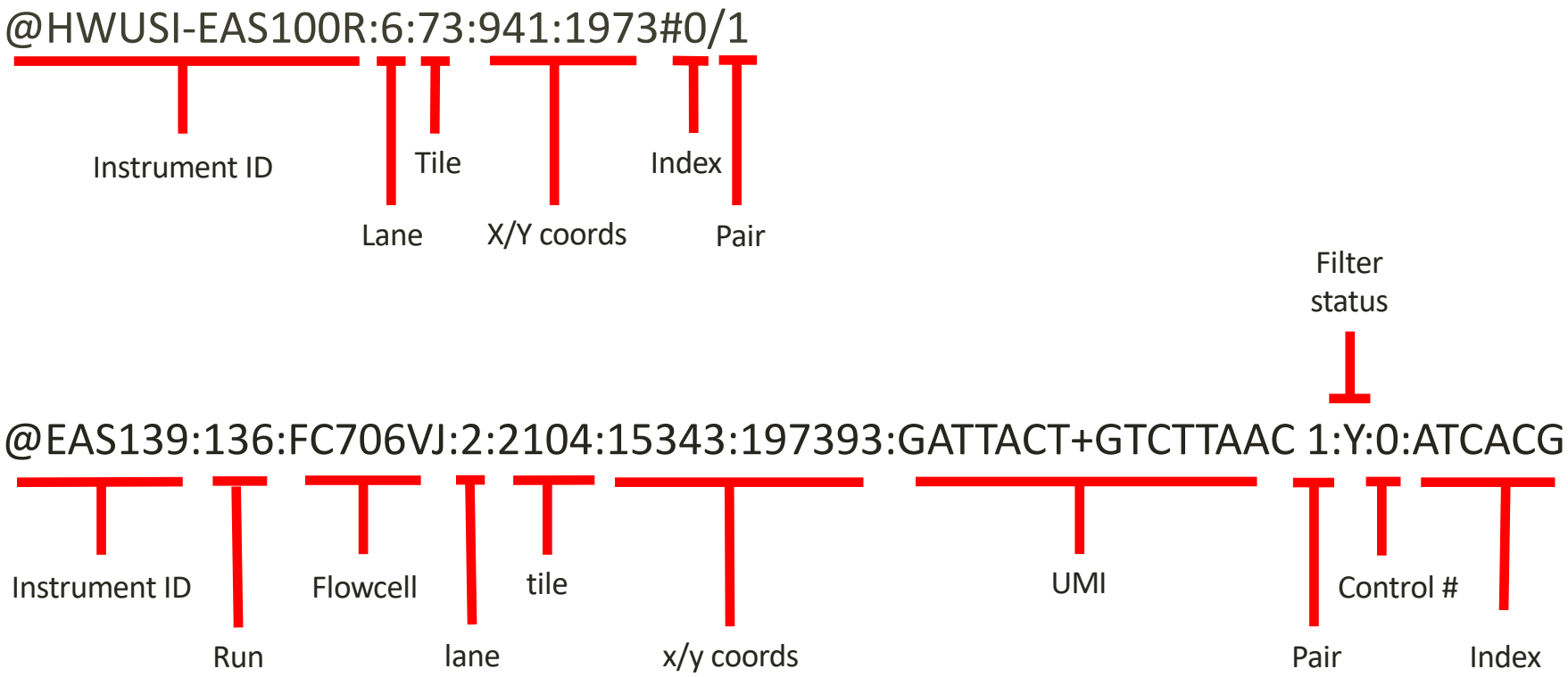
Fastq – format for representing raw sequence – base calls and quality values

```
@HWUSI-EAS100R:6:73:941:1973#0/1
CTTTTTTATTTTTGTCTGACTGGGTTGATTCAAAA
+
CCCFFFFFHHHHGJHIIJHIHIIIFHIJJJJIJJGIBBFGE

@HWUSI-EAS100R:6:2303:11793:37095#0/1
ATGAATTATAGGGCTGTATTTTAATTTTGCATTTTAA
+
@@??BDDFFF<FHEGFFGGIEBGHIIIIIBEHIIGIH<FHE
```



Read naming conventions



Quality values - 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	0.9	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 1.7: Phred scores 0 to 2 have a slightly different meaning
- Illumina 1.8+ -> Sanger Format

GFF/GTF - representing sequence features

- GFF – General/Generic Feature Format; Gene Finding Format
  - Two versions in wide use
    - GFF2 (see also GTF)
    - GFF3
      - Added formal support for multiple levels (and direction) of hierarchy
        (e.g., gene -> transcript -> exon)
- GTF – Gene Transfer Format
  - An extension of GFF2
- GFF2, GFF3 and GTF are all tab-separate files with 9 fields
  - Differing content in 9th column

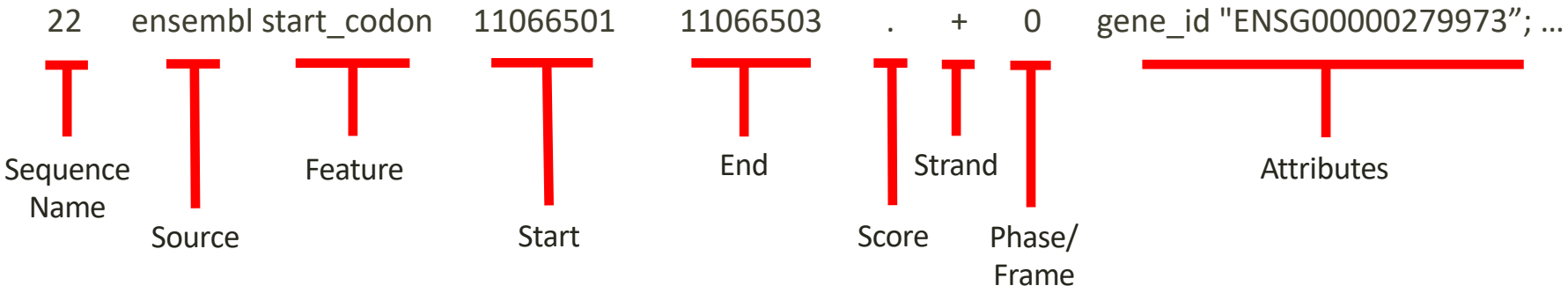
GFF/GTF – general structure

General GFF structure

Position index	Position name	Description
1	sequence	The name of the sequence where the feature is located.
2	source	Keyword identifying the source of the feature, like a program (e.g. Augustus or RepeatMasker) or an organization (like TAIR).
3	feature	The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow their parents in a single block (so all exons of a transcript are put after their parent "transcript" feature line and before any other parent transcript line). In GFF3, all features and their relationships should be compatible with the standards released by the Sequence Ontology Project.
4	start	Genomic start of the feature, with a **1-base offset**. This is in contrast with other 0-offset half-open sequence formats, like BED files.
5	end	Genomic end of the feature, with a **1-base offset**. This is the same end coordinate as it is in 0-offset half-open sequence formats, like BED files.[*citation needed*]
6	score	Numeric value that generally indicates the confidence of the source on the annotated feature. A value of "." (a dot) is used to define a null value.
7	strand	Single character that indicates the Sense (molecular biology) strand of the feature; it can assume the values of "+" (positive, or 5'->3'), "-", (negative, or 3'->5'), "." (undetermined).
8	phase	phase of CDS features; it can be either one of 0, 1, 2 (for CDS features) or "." (for everything else). See the section below for a detailed explanation.
9	Attributes.	All the other information pertaining to this feature. The format, structure and content of this field is the one which varies the most between the three competing file formats.

https://en.wikipedia.org/wiki/General_feature_format

Ensembl GTF example record



Example of attributes string:

gene_id "ENSG00000279973”; gene_version "1"; transcript_id "ENST00000624155"; transcript_version "1"; exon_number "1"; gene_name "BAGE5"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_name "BAGE5-201"; transcript_source "ensembl"; transcript_biotype "protein_coding"; tag "basic"; transcript_support_level "1";

Note: there will be many GTF records/rows per transcript per gene (UTRs, start_codon, exons, etc)

VCF file

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	H_TU-GTB15-3685	H_TU-GTB15-M1501867
1	1026106	.	G	T	.	PASS	NT=ref;QSS=18;QSS_NT=18;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1216591	.	G	A	.	PASS	NT=ref;QSS=120;QSS_NT=108;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1249123	.	G	T	.	PASS	NT=ref;QSS=16;QSS_NT=16;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1262394	.	G	T	.	PASS	NT=ref;QSS=34;QSS_NT=34;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1326886	.	C	T	.	PASS	NT=ref;QSS=199;QSS_NT=157;SGT=CC->CT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1391597	.	T	C	.	PASS	NT=ref;QSS=32;QSS_NT=32;SGT=TT->CT;TQSS=2;TQSS_NT=2	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1904481	.	G	T	.	PASS	NT=ref;QSS=24;QSS_NT=24;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1912142	.	G	T	.	PASS	NT=ref;QSS=33;QSS_NT=33;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	1919717	.	G	A	.	PASS	NT=ref;QSS=17;QSS_NT=17;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	2319028	.	C	T	.	PASS	NT=ref;QSS=76;QSS_NT=76;SGT=CC->CT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	2333646	.	G	T	.	PASS	NT=ref;QSS=26;QSS_NT=26;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3328555	.	G	T	.	PASS	NT=ref;QSS=20;QSS_NT=20;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3350384	.	G	A	.	PASS	NT=ref;QSS=33;QSS_NT=33;SGT=GG->AG;TQSS=2;TQSS_NT=2	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3388456	.	C	T	.	PASS	NT=ref;QSS=55;QSS_NT=55;SGT=CC->CT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3662615	.	G	T	.	PASS	NT=ref;QSS=18;QSS_NT=18;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	3774072	.	G	T	.	PASS	NT=ref;QSS=21;QSS_NT=21;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	6021727	.	G	A	.	PASS	NT=ref;QSS=16;QSS_NT=16;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	6271112	.	G	T	.	PASS	NT=ref;QSS=52;QSS_NT=52;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	6278217	.	G	T	.	PASS	NT=ref;QSS=30;QSS_NT=30;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	6609812	.	G	A	.	PASS	NT=ref;QSS=74;QSS_NT=74;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	9338624	.	G	A	.	PASS	NT=ref;QSS=15;QSS_NT=15;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU:TU:FT		
1	10678477	.	G	T	.	PASS	NT=ref;QSS=26;QSS_NT=26;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	10720178	.	G	T	.	PASS	NT=ref;QSS=33;QSS_NT=33;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	11140620	.	A	C	.	PASS	NT=ref;QSS=20;QSS_NT=20;SGT=AA->AC;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	11194363	.	G	T	.	PASS	NT=ref;QSS=19;QSS_NT=19;SGT=GG->GT;TQSS=2;TQSS_NT=2	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	11294450	.	C	T	.	PASS	NT=ref;QSS=35;QSS_NT=35;SGT=CC->CT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	11561899	.	G	A	.	PASS	NT=ref;QSS=32;QSS_NT=32;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	11595041	.	G	A	.	PASS	NT=ref;QSS=137;QSS_NT=105;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	11735264	.	G	T	.	PASS	NT=ref;QSS=170;QSS_NT=122;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	11852226	.	G	T	.	PASS	NT=ref;QSS=39;QSS_NT=39;SGT=GG->GT;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	11855448	.	G	A	.	PASS	NT=ref;QSS=32;QSS_NT=32;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		
1	12198424	.	G	A	.	PASS	NT=ref;QSS=24;QSS_NT=24;SGT=GG->AG;TQSS=1;TQSS_NT=1	GT:AD:BQ:SS:DP:FDP:SDP:SUBDP:AU:CU:GU		

Details of the VCF file format: hts-specs, VCF-v4.2.pdf

Example of SAM/BAM file format

Example SAM/BAM header section (abbreviated)

```
mgriffit@linus270 ~> samtools view -H /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam  | grep -P "SN\:22|HD|RG|PG"
@HD     VN:1.4  SO:coordinate
@SQ     SN:22   LN:51304566     UR:ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa.gz AS:GRCh37-lite  M5:a718acaa6135fdca8357d5bfe9
4211dd  SP:Homo sapiens
@RG     ID:2888721359   PL:illumina     PU:D1BA4ACXX.3  LB:H_KA-452198-0817007-cDNA-3-lib1      PI:365  DS:paired end   DT:2012-10-03T19:00:00-0500     SM:H_KA-452198-0817007  CN:WUGSC
@PG     ID:2888721359   VN:2.0.8        CL:tophat --library-type fr-secondstrand --bowtie-version=2.1.0
@PG     ID:MarkDuplicates       PN:MarkDuplicates       PP:2888721359   VN:1.85(exported)       CL:net.sf.picard.sam.MarkDuplicates INPUT=[/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blad
e10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300.bam] OUTPUT=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jw
alker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300-post_dup.bam METRICS_FILE=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-1543
4-136080019/staging-1iuJS/H_KA-452198-0817007-cDNA-3-lib1-2888360300.metrics REMOVE_DUPLICATES=false ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=9500 TMP_DIR=[/gscmnt/gc13001/info/build_merged_al
ignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y] VALIDATION_STRINGENCY=SILENT MAX_RECORDS_IN_RAM=500000    PROGRAM_RECORD_ID=MarkDuplicates PROGRAM_GROUP_NAME=Mark
Duplicates MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=50000 SORTING_COLLECTION_SIZE_RATIO=0.25 READ_NAME_REGEX=[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).* OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 VERBOSITY=INFO
QUIET=false COMPRESSION_LEVEL=5 CREATE_INDEX=false CREATE_MD5_FILE=false
mgriffit@linus270 ~>
```

Example SAM/BAM alignment section (only 10 alignments shown)

```
mgriffit@linus270 ~> samtools view -f 3 -F 1804 /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam | head
HWI-ST495_129147882:3:2114:15769:38646  99      1       11306   3       100M    =       11508   302     ACTGCGGGGCCCTCTTGCTTACTGTATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCAGCACGC
CCFFFFFHHHGHJJJJJJJIJJJHGIJJIJJHIIJJJJJJHFDDDDDDDDDDDDCDDDDDDDDDDDDDD<CDDDDDDDDCCCC>@>CDDDDDD?1<B       CC:Z:15 MD:Z:5A94       PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102519765  AS:i:-5 XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:2114:15769:38646  147     1       11508   3       100M    =       11306   -302    ACTCCTAAATATGGGATTCCTGGGTTTAAAAGTATAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAATTGTACTGTTCTGTATCCCACCAG5
;5:CDCDCDECEFCD@9E=??EEIIIIHCEGGIJJJJIIJJIHF@?00IHHFFGG?*JJJIJGHGEIJJIJJJJJJIHHCIEJJJHFHHGHFFEDFCCB       CC:Z:15 MD:Z:34A65      PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102519563  AS:i:-6 XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1210:1257:16203   163     1       11810   3       100M    =       12055   345     CCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGC
CCFFFFFHFHAFGGIIIJJJEEHGIGGGIJIJJGI?@EHIGIJDGHIHIGGIJJJJJJJJIJGHHHGHFFFCDDDDDDCDCCCCCA;>@>@AA@:AA>AA       CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:102519261  AS:i:0  XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:1210:1257:16203   83      1       12055   3       100M    =       11810   -345    GAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATAC
CC>4C>DCCCACACDCC?BDCEE@ECFFFFHHHHHIJJJIIJJIIIHHEHIIGJIJIJJIGHIIIJJJJJJIIJJJJJIIJJIJIJJHGHHHDFEFFCCC       CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:102519016  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:2111:3117:78828   163     1       12634   3       100M    =       12746   212     GCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCACAGGTGAGAGGAGAG<
@@FFFFFDHHHH9FHGIIFGAFDHEGII>GHIIIIIIIIIIIIIIIFHDDFFEEECEECCCACCCCCC:AADCCBCC>CAC<CCCCCC:@CB@@BAB##       CC:Z:15 MD:Z:85G14      PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102518437  AS:i:-5 XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:2111:3117:78828   83      1       12746   3       100M    =       12634   -212    GGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCATCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTD
DCABDBDDDDDDDDDDDCDDDDBDB@BDDDB@;CCCCCDEFD@;.?<HIGGEIGEHIGJJJIIGIGIIHEGFEHFJIIIIIGJJJJHHHHHFFFFC@@       CC:Z:15 MD:Z:37G62      PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102518325  AS:i:-5 XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:1102:4242:26638   99      1       13503   3       100M    =       13779   376     CGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAC
CCFFFFFHHHHHJJJIJJJJJJJJJJJJJJGIIIIJJFHGGIJGIJJJEGIJIJJHHIHHGHFFEFDEEEECCCAACDDACDCDDDDDB?8?<B>A@CDC       CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357414  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1309:15328:74082  99      1       13534   3       100M    =       13780   346     AGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCACAGG@
CCFFFADHHHHFIJJJJJIJIJIHIJJJIHJJIJJIIIJJIJIJJJJJBFHIIJJJJJIJHH=EEFFFFCEEECEDCDCDDDDDDDDDDDDDDDBDCCD       CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357383  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1308:10126:19636  99      1       13779   3       100M    =       14027   348     CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTGC
CCFFFFFHHGHHJJJJJJJJJJIJJJJJJIJJJJJJJJJIJIJJJJJJJJJHJJJHGHHFFFFFFEEEEEEDDDDCDCDCDFADCACCACDDCCCCCCD       CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357140  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1102:4242:26638   147     1       13779   3       100M    =       13503   -376    CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTG#
##DCCDDDCCBBBABCCDDDCBDDBBDHC?=GIIJIIIIJIGIIIIJJHJJIJJIGCIIJJJJJIGHGJJIJIJJJJJIJIIIGGFGHHHHFFFFFCCC       CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates      RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357140  AS:i:0  XS:A:+  YT:Z:UU
mgriffit@linus270 ~>
```

Introduction to the SAM/BAM format

- The specification
  - http://samtools.sourceforge.net/SAM1.pdf
- The SAM format consists of two sections:
  - Header section
    - Used to describe source of data, reference sequence, method of alignment, etc.
  - Alignment section
    - Used to describe the read, quality of the read, and nature alignment of the read to a region of the genome
- BAM is a compressed version of SAM
  - Compressed using lossless BGZF format
  - Other BAM compression strategies are a subject of research. See 'CRAM' format for example
- BAM files are usually 'indexed'
  - A '.bai' file will be found beside the '.bam' file
  - Indexing aims to achieve fast retrieval of alignments overlapping a specified region without going through the whole alignments. BAM must be sorted by the reference ID and then the leftmost coordinate before indexing

SAM/BAM header section

- Used to describe source of data, reference sequence, method of alignment, etc.
- Each section begins with character '@' followed by a two-letter record type code. These are followed by two-letter tags and values
  - @HD The header line
    - VN: format version
    - SO: Sorting order of alignments
  - @SQ Reference sequence dictionary
    - SN: reference sequence name
    - LN: reference sequence length
    - SP: species
  - @RG Read group
    - ID: read group identifier
    - CN: name of sequencing center
    - SM: sample name
  - @PG Program
    - PN: program name
    - VN: program version

SAM/BAM alignment section

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
★ 2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
★ 6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Example values

```
1       QNAME  e.g.  HWI-ST495_129147882:1:2302:10269:12362 (QNAME)
2       FLAG   e.g.  99
3       RNAME  e.g.  1
4       POS    e.g.  11623
5       MAPQ   e.g.  3
6       CIGAR  e.g.  100M
7       RNEXT  e.g.  =
8       PNEXT  e.g.  11740
9       TLEN   e.g.  217
10      SEQ    e.g.  CCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGT
11      QUAL   e.g.  CCCFFFFFHHHHHJJIJFIJJJJJJJJJJJJHIJJJJJJJIJJJJJGGHIJHIJJJJJJJJJGHGGIJJJJJJIJEEHHHHFFFFCDCDDDDDDDB@ACDD
```

SAM/BAM flags explained

- http://broadinstitute.github.io/picard/explain-flags.html
- 12 bitwise flags describing the alignment
- These flags are stored as a binary string of length 11 instead of 11 columns of data
- Value of '1' indicates the flag is set. e.g. 00100000000
- All combinations can be represented as a number from 1 to 2048 (i.e. 2^{11}-1). This number is used in the BAM/SAM file. You can specify 'required' or 'filter' flags in samtools view using the '-f' and '-F' options respectively

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Note that to maximize confusion, each bit is described in the SAM specification using its hexadecimal representation (i.e., '0x10' = 16 and '0x40' = 64).

CIGAR strings explained

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- The CIGAR string is a sequence of base lengths and associated 'operations' that are used to indicate which bases align to the reference (either a match or mismatch), are deleted, are inserted, represent introns, etc.
- e.g. 81M859N19M
  - A 100 bp read consists of: 81 bases of alignment to reference, 859 bases skipped (an intron), 19 bases of alignment

CRAM files

- CRAM is an ultra-compressed version of a BAM file
  - Usually between 30-60% smaller than the corresponding BAM
- Stores “diffs” from the reference genome
  - requires the matching reference genome to restore original data!
- Base quality binning may be used as well
- Some tools still require conversion back to bam

Quality Score Bins	Example of Empirically Mapped Quality Scores*
N (no call)	N (no call)
2–9	6
10–19	15
20–24	22
25–29	27
30–34	33
35–39	37
≥ 40	40

By replacing the quality scores between 19 and 25 with a new score of 22, data storage space is conserved.
*The mapped quality score of each bin (except “N”) is subject to change depending on individual Q-tables.

Introduction to the BED format

- When working with BAM files, it is very common to want to examine a focused subset of the reference genome
  - e.g. the exons of a gene
- These subsets are commonly specified in 'BED' files
  - https://genome.ucsc.edu/FAQ/FAQformat.html#format1
- Many BAM manipulation tools accept regions of interest in BED format
- Basic BED format (tab separated):
  - Chromosome name, start position, end position
  - Coordinates in BED format are 0 based

Manipulation of SAM/BAM and BED files

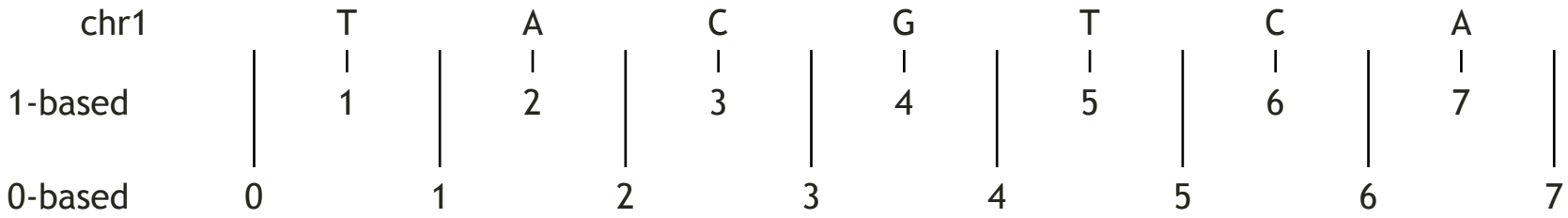
- Several tools are used ubiquitously in sequence analysis to manipulate these files
- SAM/BAM files
  - samtools
  - bamtools
  - picard
- BED files
  - bedtools
  - bedops

Common problems

Common sources of confusion

- Genomic coordinate systems
- Genome builds
- Variant representation

Genomic coordinates – 1 vs 0 based



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

Genome builds

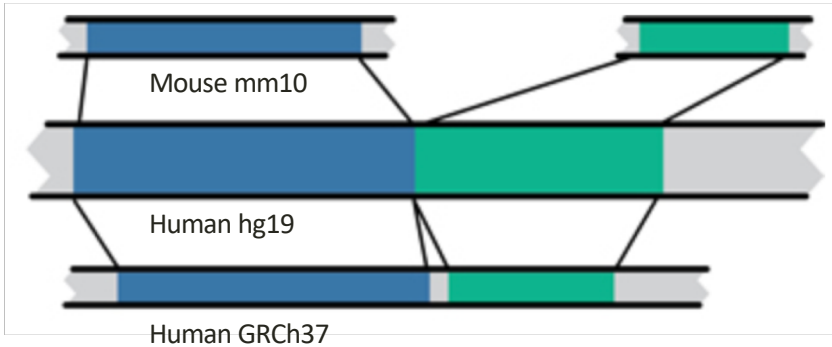
Reference Genome builds

Current human: GRCh38, hg38, b38
alternate: GRCh38v2_ccdg

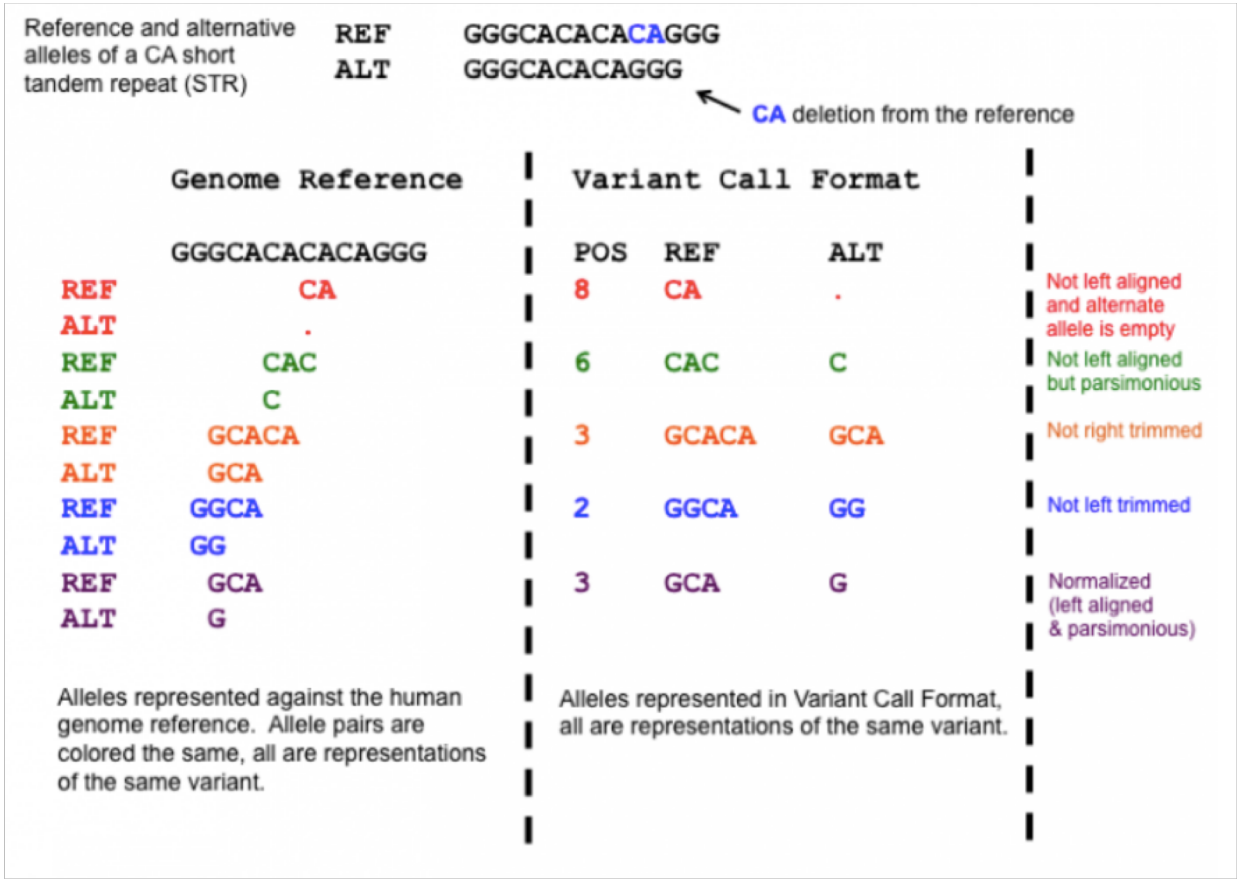
Previous human: GRCh37, hg19, b37

Current mouse: GRCm38, mm10

Lift-over



Variant shifting (alignment) and parsimony/trimming



Parsimony: representing variant in as few nucleotides as possible without reducing the length of any allele to 0

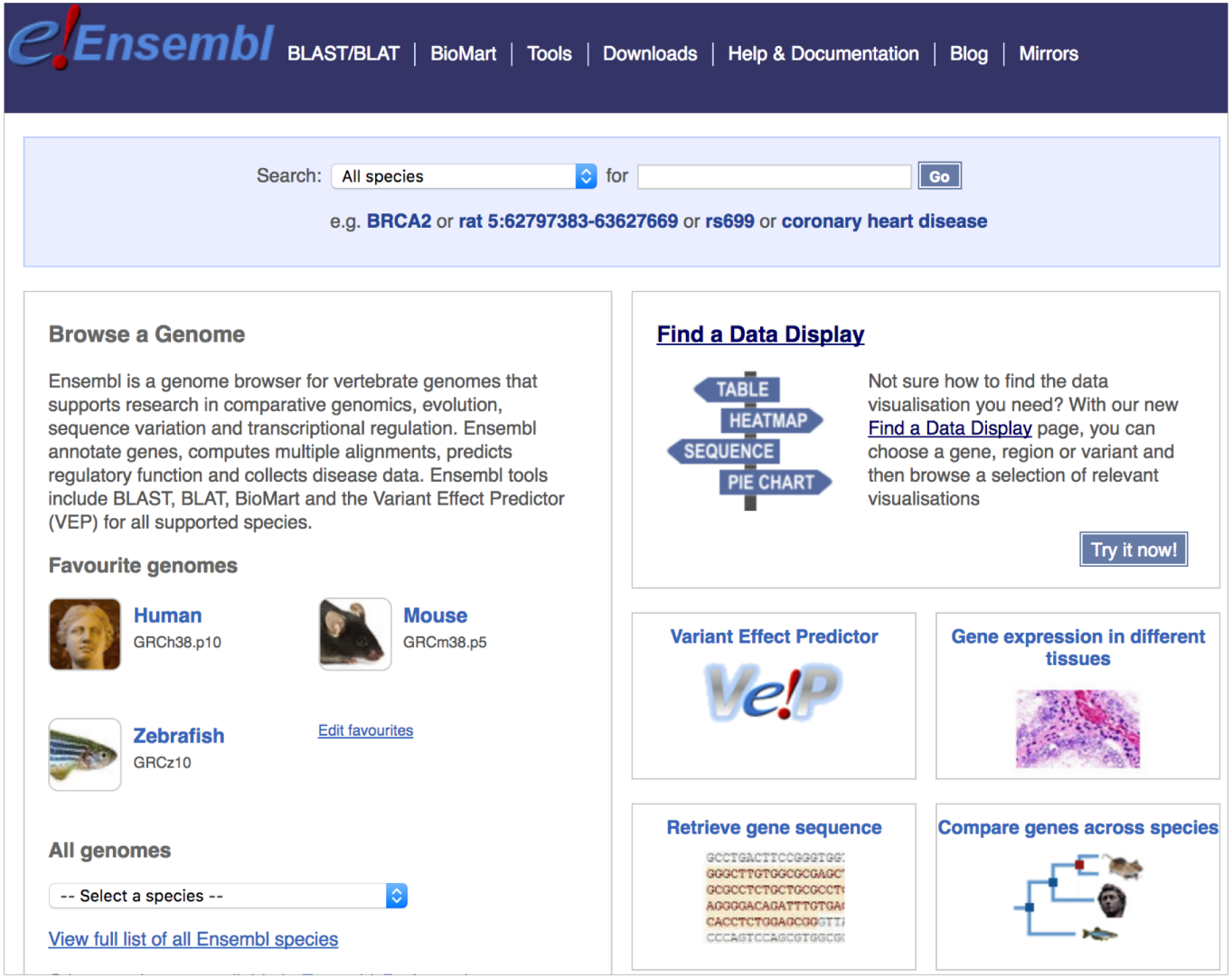
Left (right) aligning = shifting the start position of a variant as far to the left (right) as possible

How should I sort my SAM/BAM file?

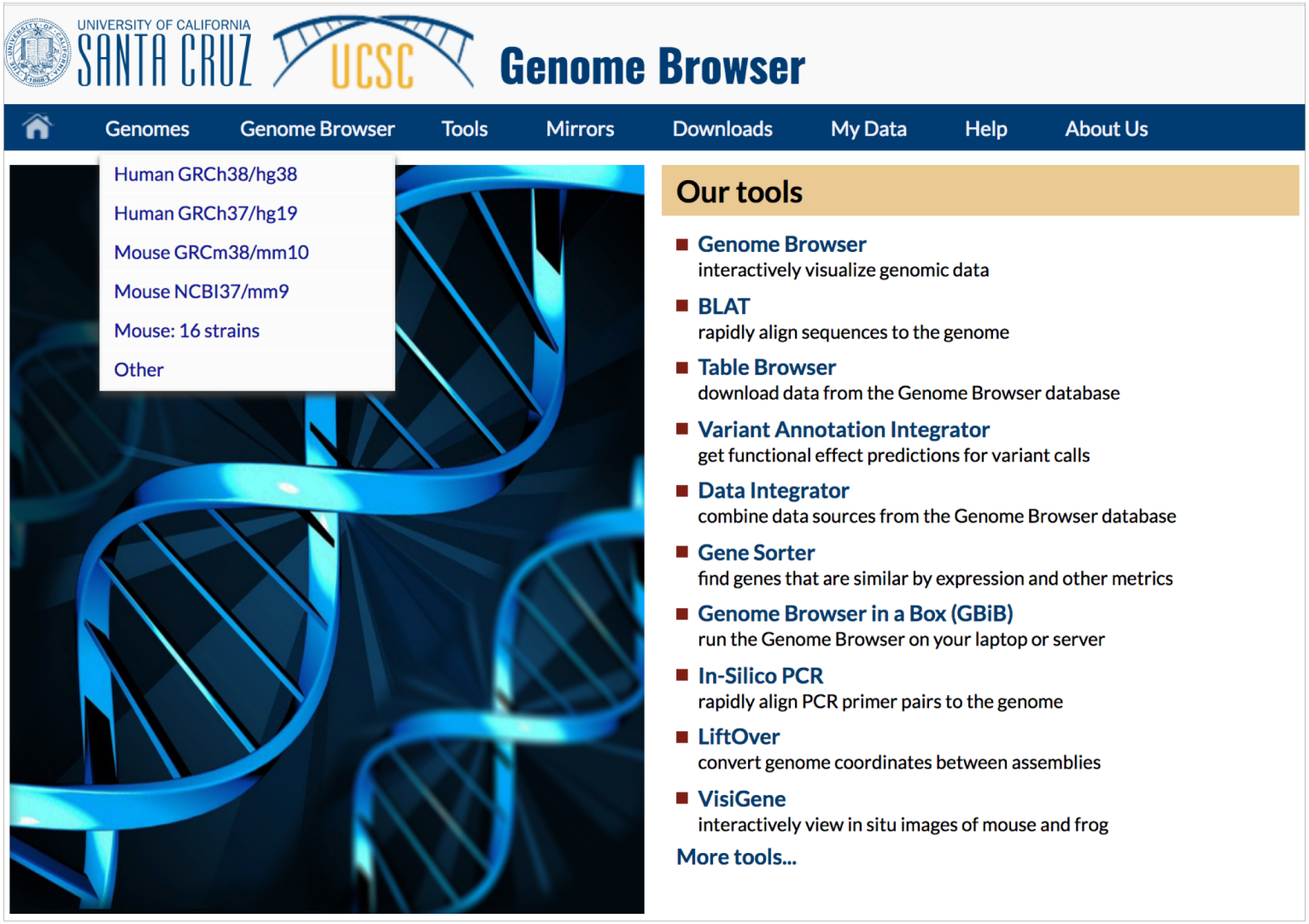
- Generally BAM files are sorted by position
  - This is for performance reasons
    - When sorted and indexed, arbitrary positions in a massive BAM file can be accessed rapidly
- Certain tools require a BAM sorted by read name
  - Usually this is when we need to easily identify both reads of a pair
    - The insert size between two reads may be large
    - In fusion detection we are interested in read pairs that map to different chromosomes

Genome browsers

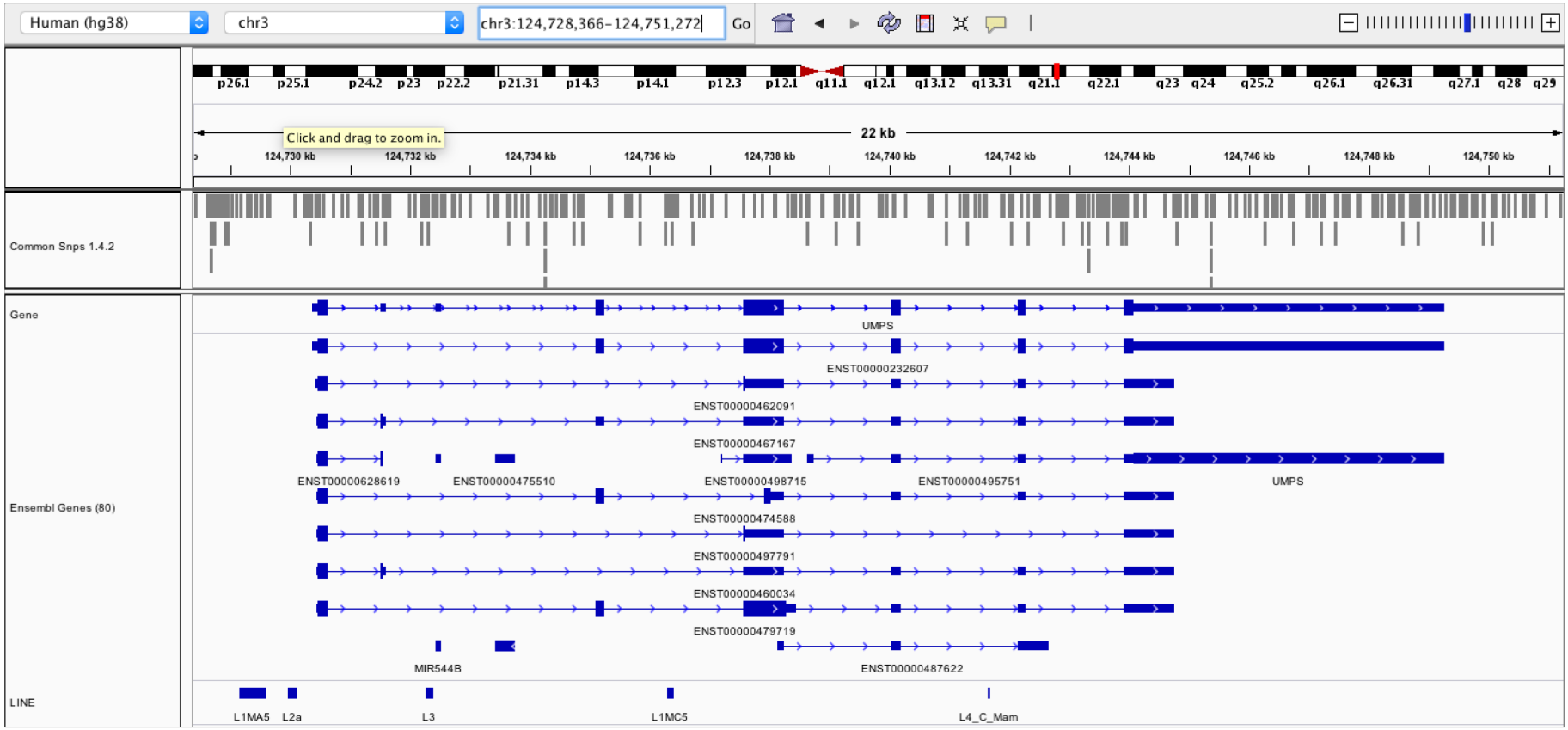
Genome browsers - Ensembl



Genome browsers - UCSC



Genome browsers - IGV

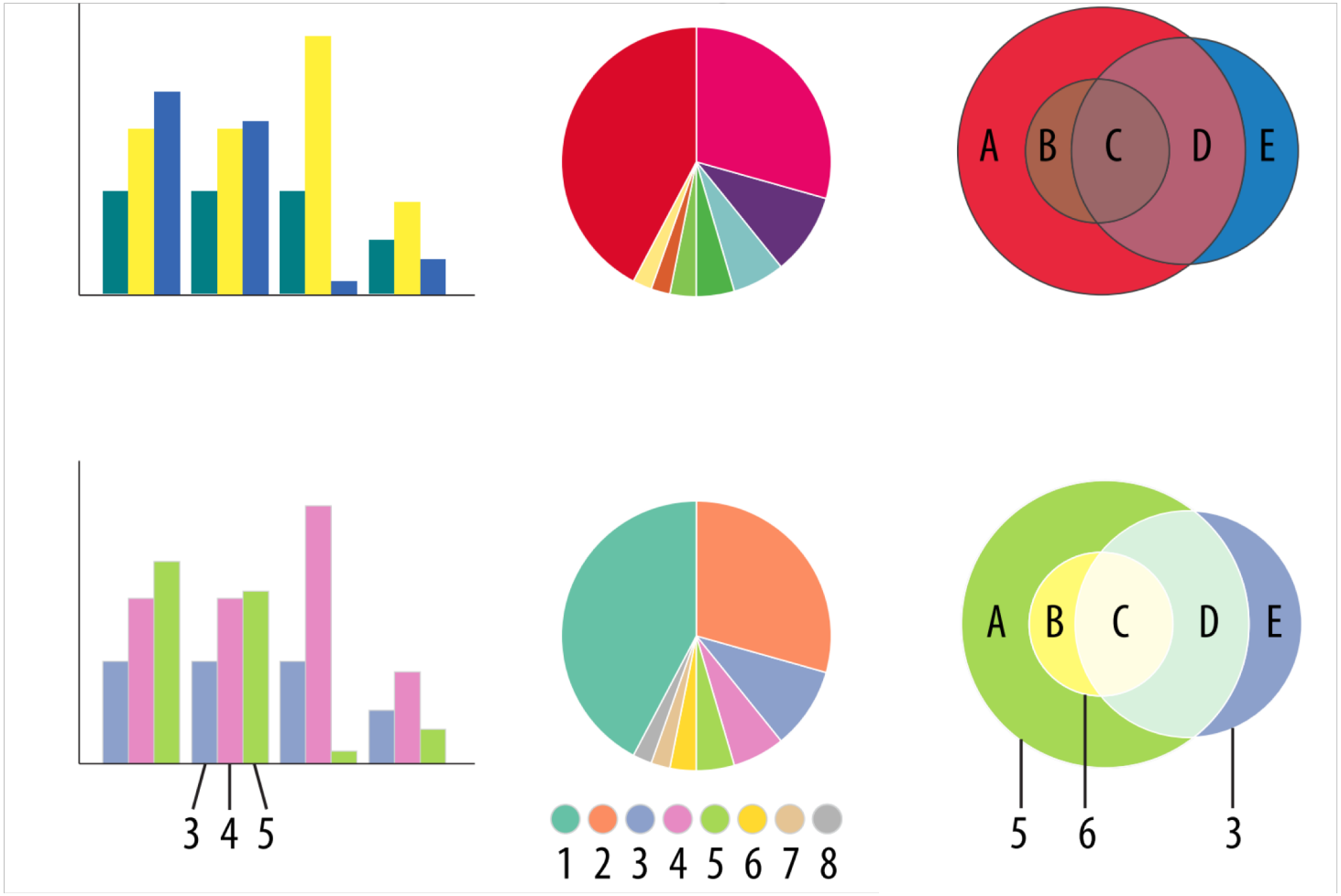


Fundamentals of data visualization

Fundamentals of data visualization

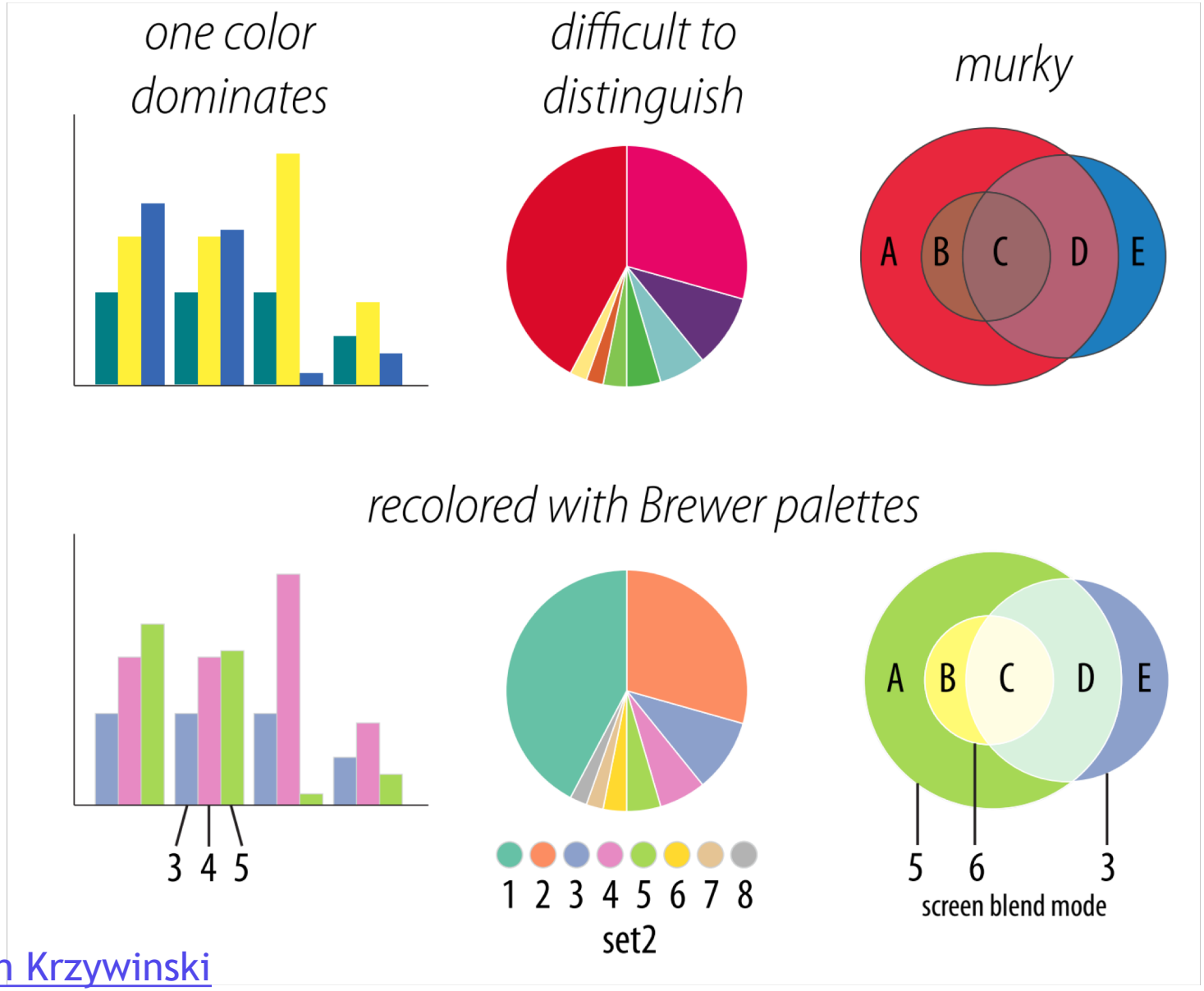
- Where to learn more about the art and science of visualization:
  - Collection of 40 Nature Methods articles on data visualization: "Points of View" Articles
  - Visual design principles lecture

Which series is more effective (top or bottom)?



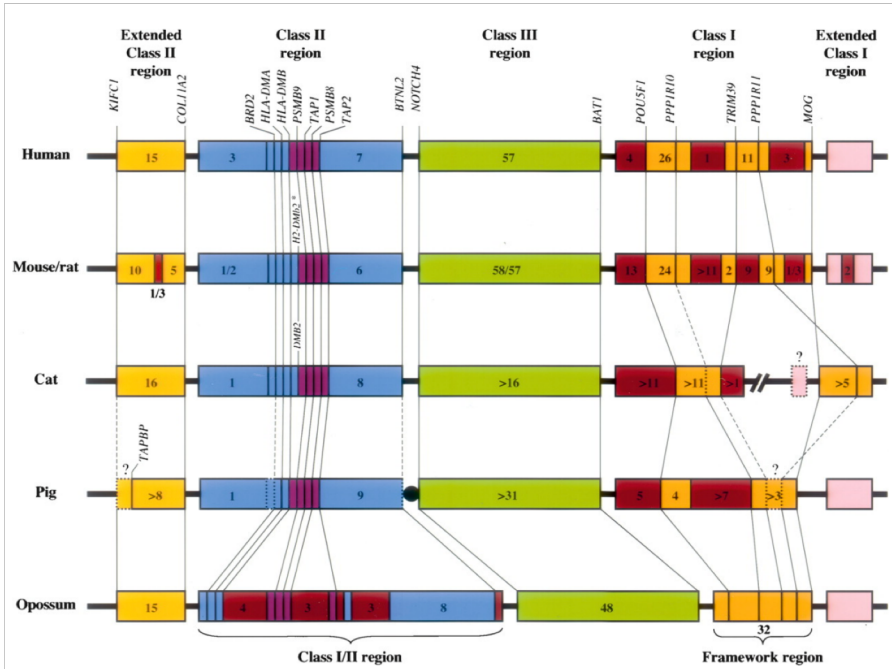
Martin Krzywinski

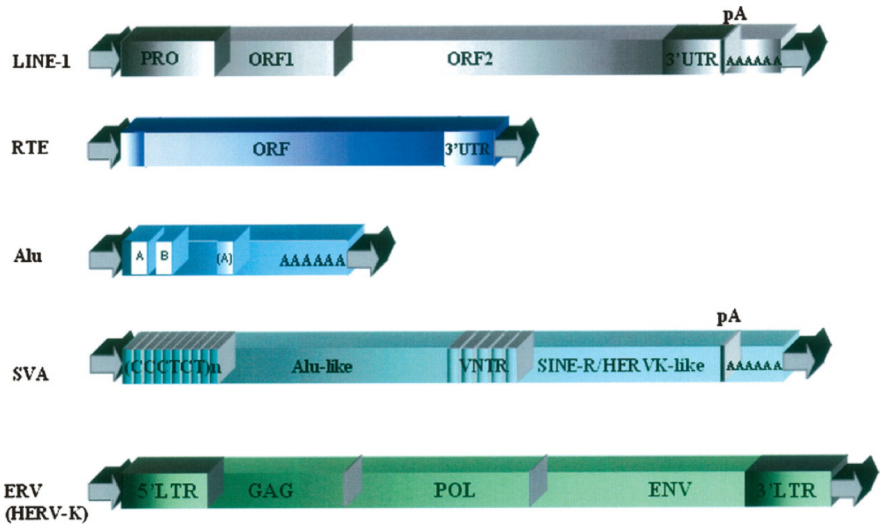
Which series is more effective (top or bottom)?



Martin Krzywinski

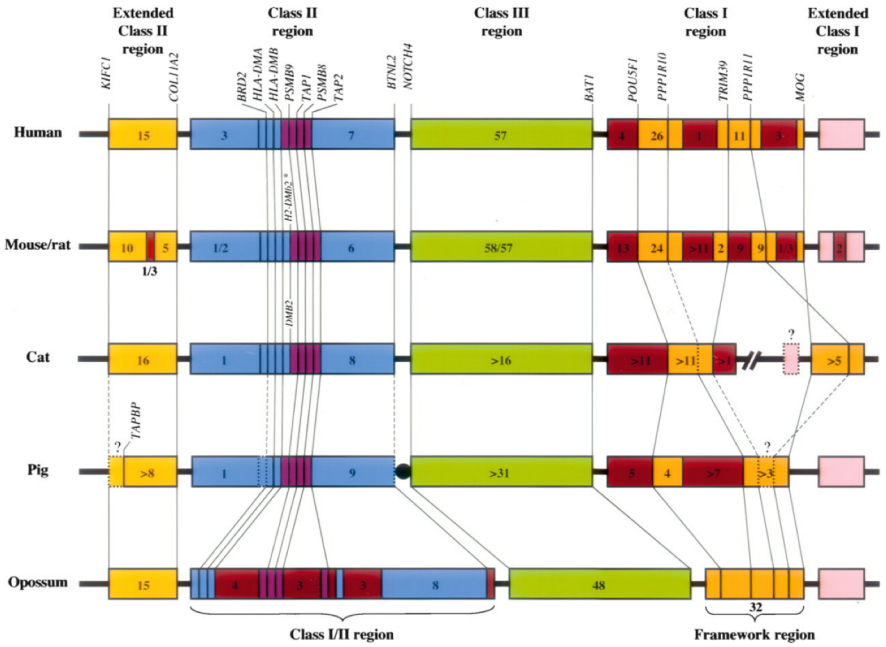
Which is more effective (left or right?)

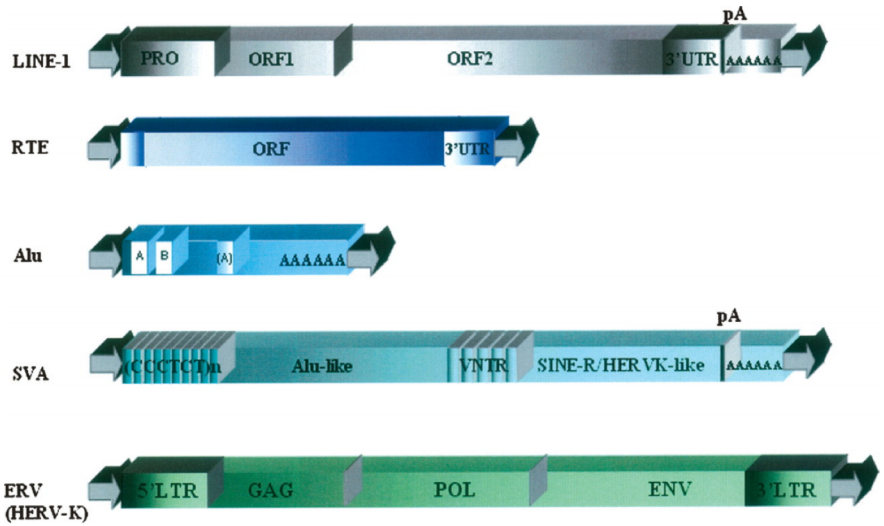




Martin Krzywinski

Which is more effective (left or right?)





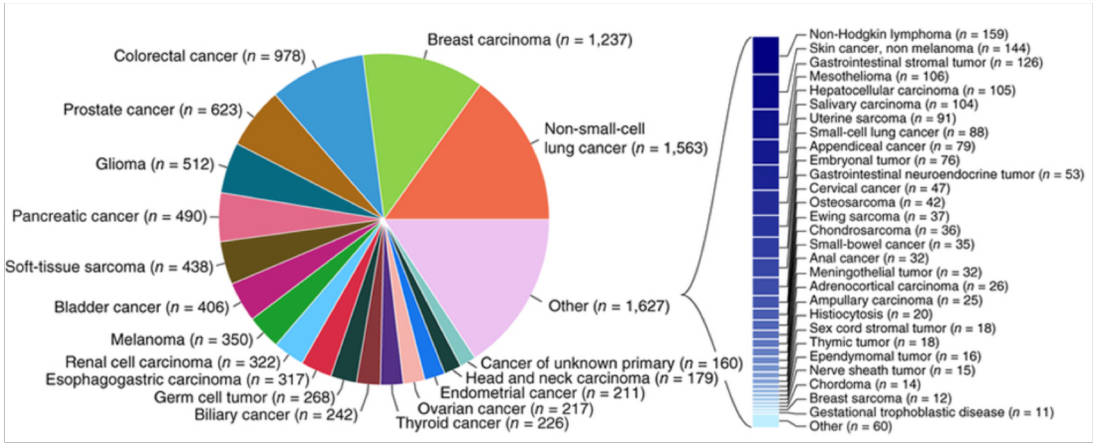
Excellent organization and consistency. Vertical lines cue continuity. Good use of color.

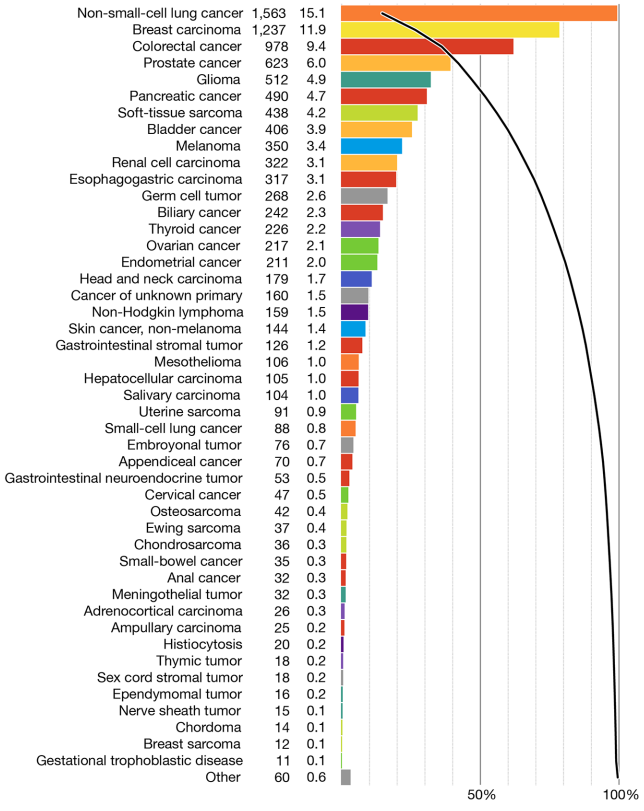
Samollow, P.B., The opossum genome: insights and opportunities from an alternative mammal. Genome Res, 2008. 18(8): p. 1199-215.

Chartjunk plentiful. Screaming ornamental and redundant elements. Text inconsistent and illegible.

Gentles, A.J., et al., Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. Genome Res, 2007. 17(7): p. 992-1004.

Which is more effective (left or right?)

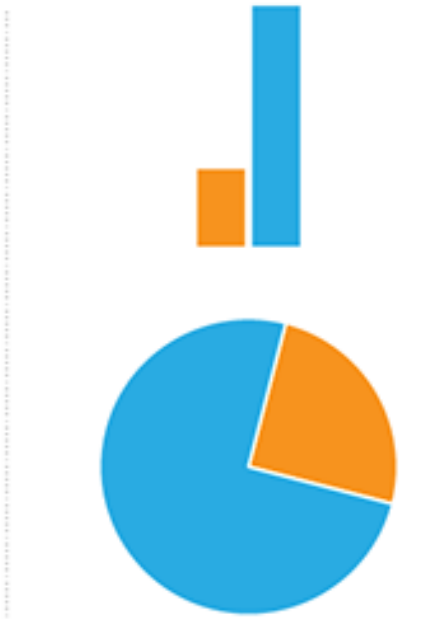




Martin Krzywinski

When to use a pie chart

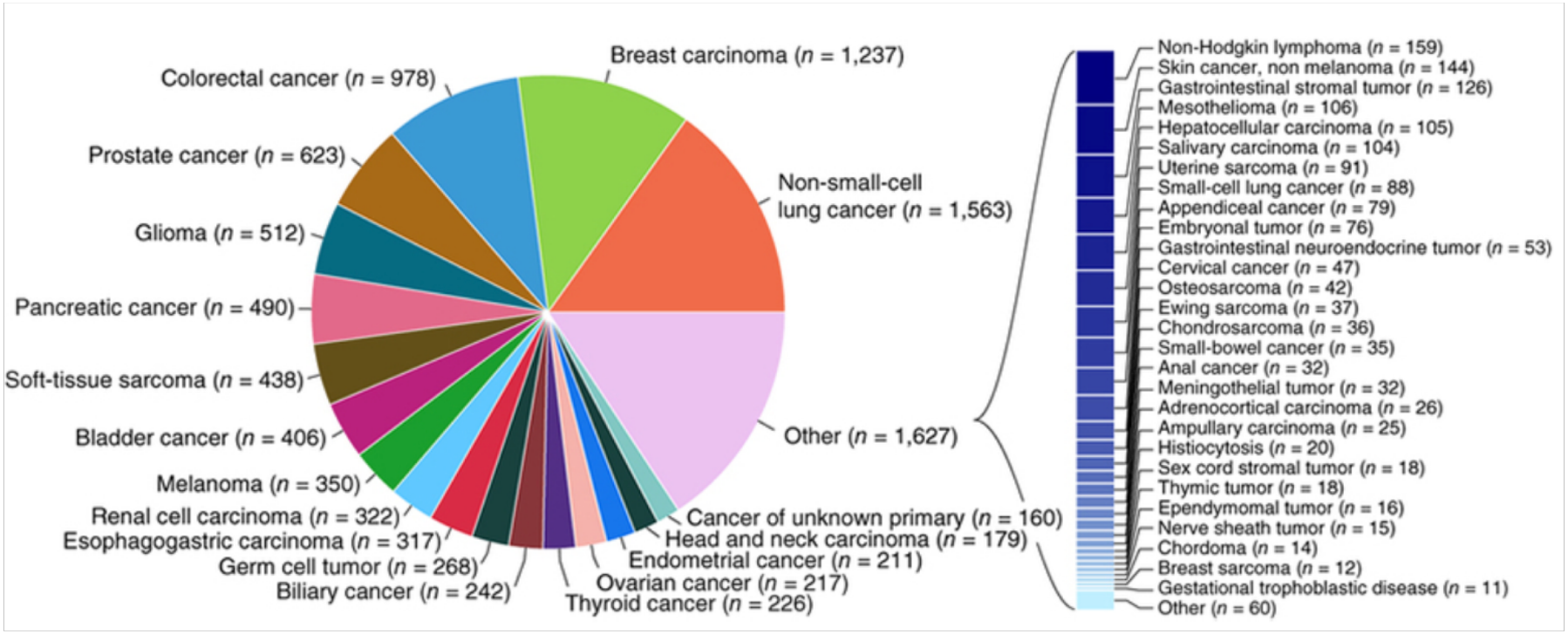
Pie charts are good at precisely showing 1:3 proportions



...but not if the slice is rotated

Martin Krzywinski

When to not use a pie chart

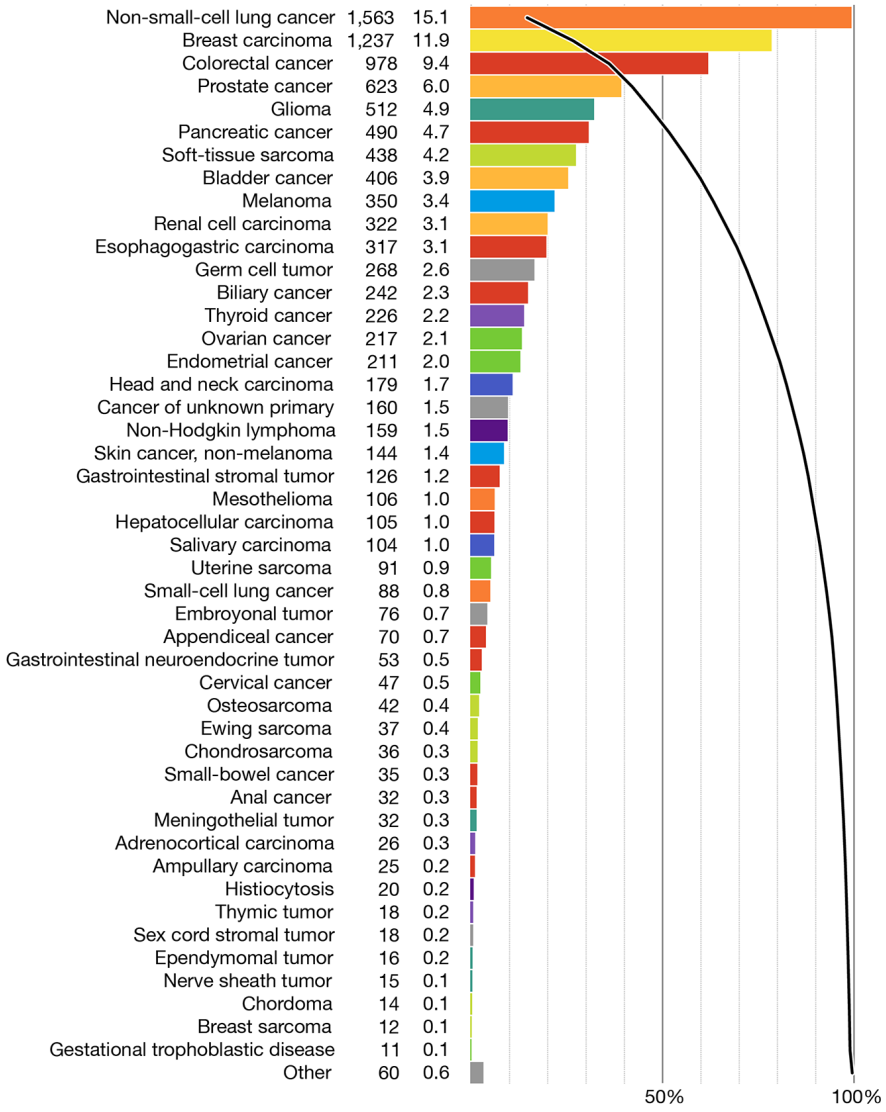


Ahmet Zehir et al. (2017) Nature Medicine doi:10.1038/nm.4333

- Hard to judge proportions
- Poor use of color
- Hard to read labels
- Over ½ of the categories had to broken out of the pie chart

Martin Krzywinski

Same data with a redesigned approach



Martin Krzywinski

Selected articles on fundamentals of data viz

- Visualizing samples with box plots
- Circos plots
- When to use (and not use) pie charts

- Resources for choosing colors
  - http://colorbrewer2.org/
  - http://mkweb.bcgsc.ca/color/
  - Understanding and using Color Palettes
  - Color palettes for color blindness
  - Names for >9000 colors
    - Including 40 beer colors

- Credit to Martin Krzywinski for his extensive work in this area and many of the above resources

Best practices in visualization

Best practices from this workshop

- *Always* label axes
- Consider readability of font size
- Avoid vertical or angled text if possible
- Avoid unnecessary use of color, point shapes, etc.
- Chose colors wisely
- If individual data points are being plotted and have started to really pile up on top of each other consider using a density function
- Always be transparent about what data manipulation is taking place (e.g. log scale, filtering of outliers, etc.)

Best practices from the experts

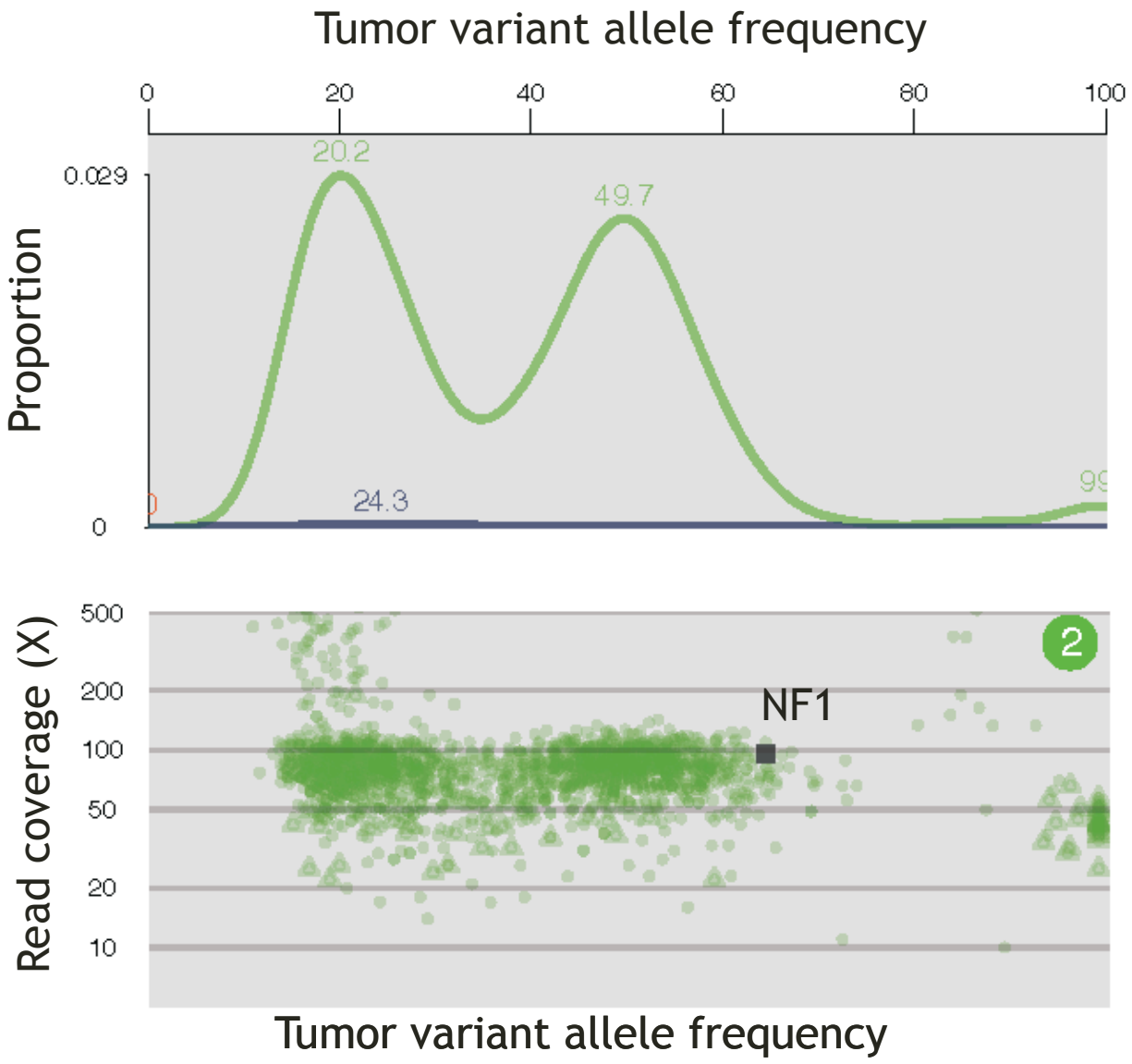
Ten Simple Rules for Better Figures (Rougier et al. 2014):

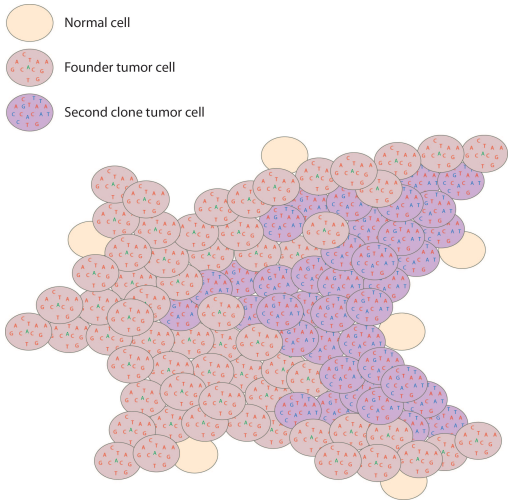
Scientific visualizations should act as a “a graphical interface between people and data”. Try to follow these rules.

1. Know Your Audience
2. Identify Your Message
3. Adapt the Figure to the Support Medium
4. Captions Are Not Optional
5. Do Not Trust the Defaults
6. Use Color Effectively
7. Do Not Mislead the Reader
8. Avoid “Chartjunk”
9. Message Trumps Beauty
10. Get the Right Tool

Example visualizations using R

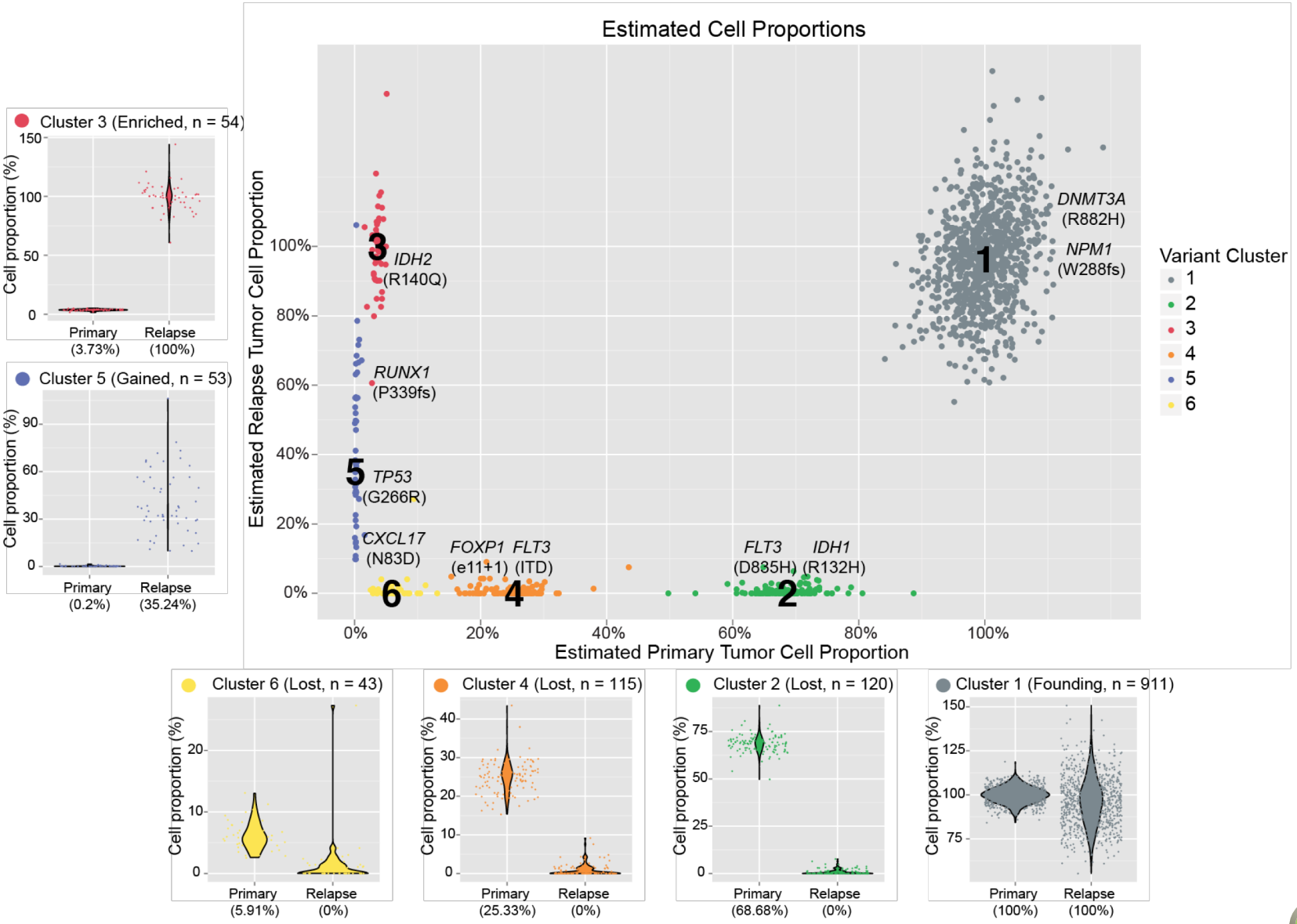
Histogram and scatterplot to define cell populations



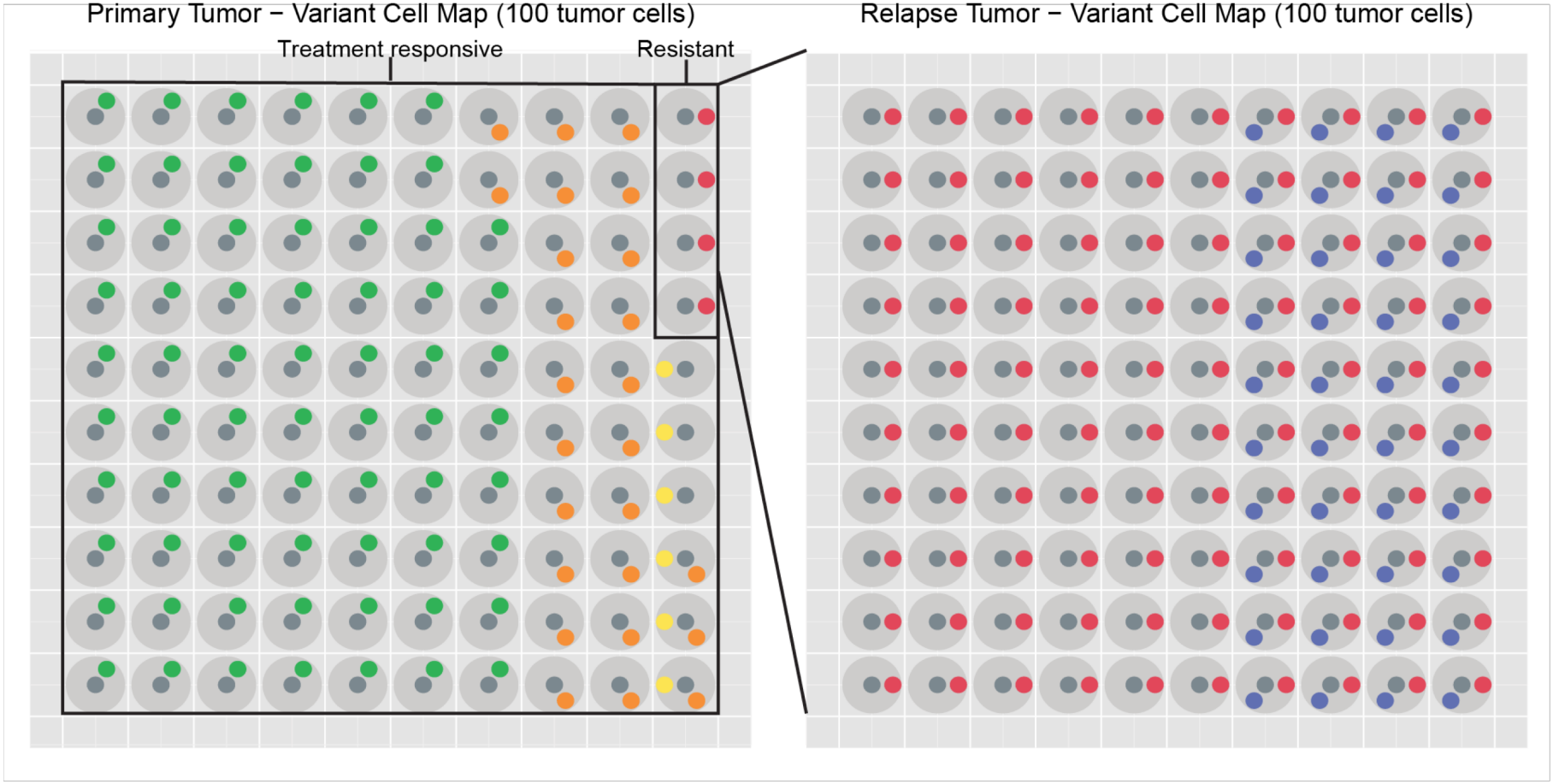


WGS data

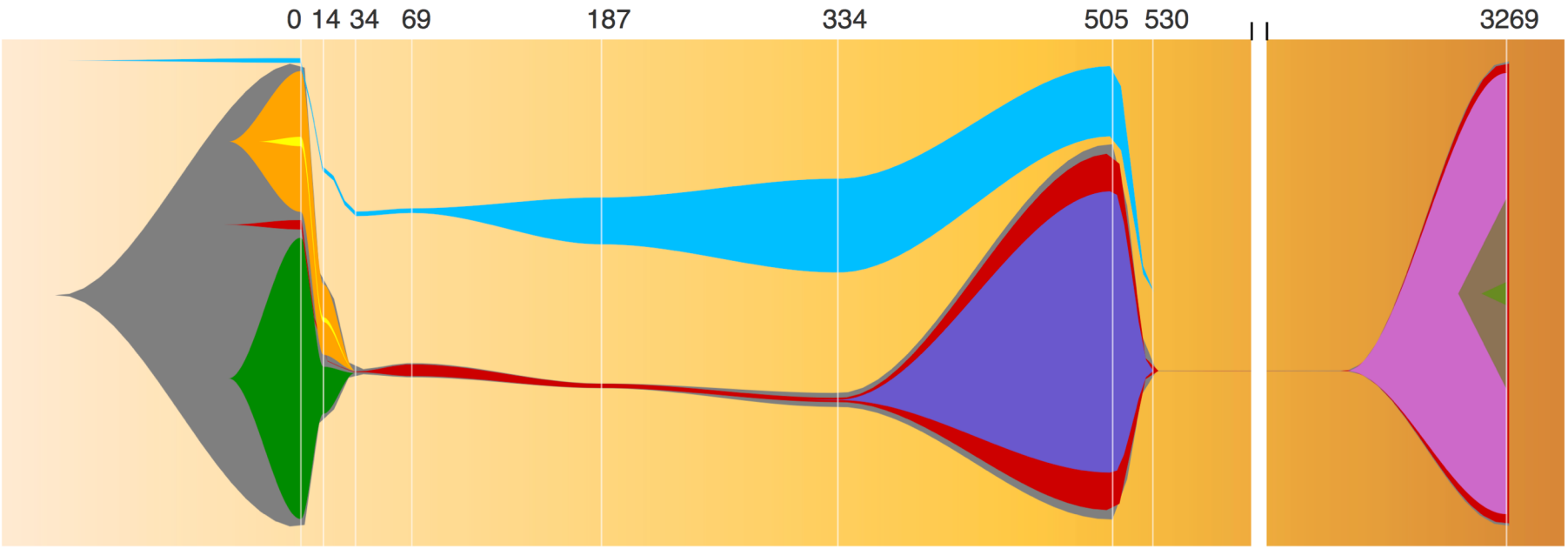
Clustering to define to distinct cell populations



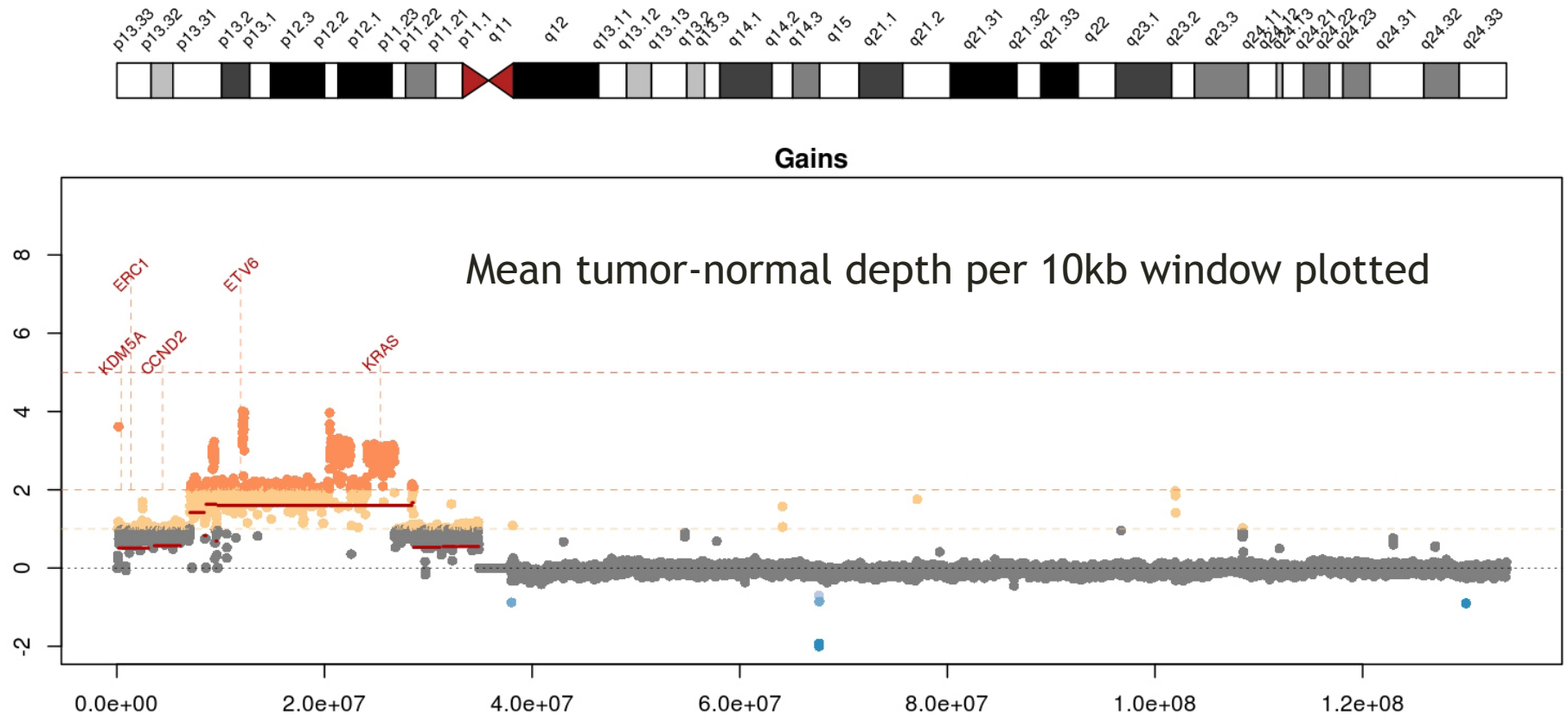
Using a “cell map” to represent the variant clusters in 100 hypothetical cells of a tissue



A ‘fish’ plot is used to represent sub-clones lost and gained over time

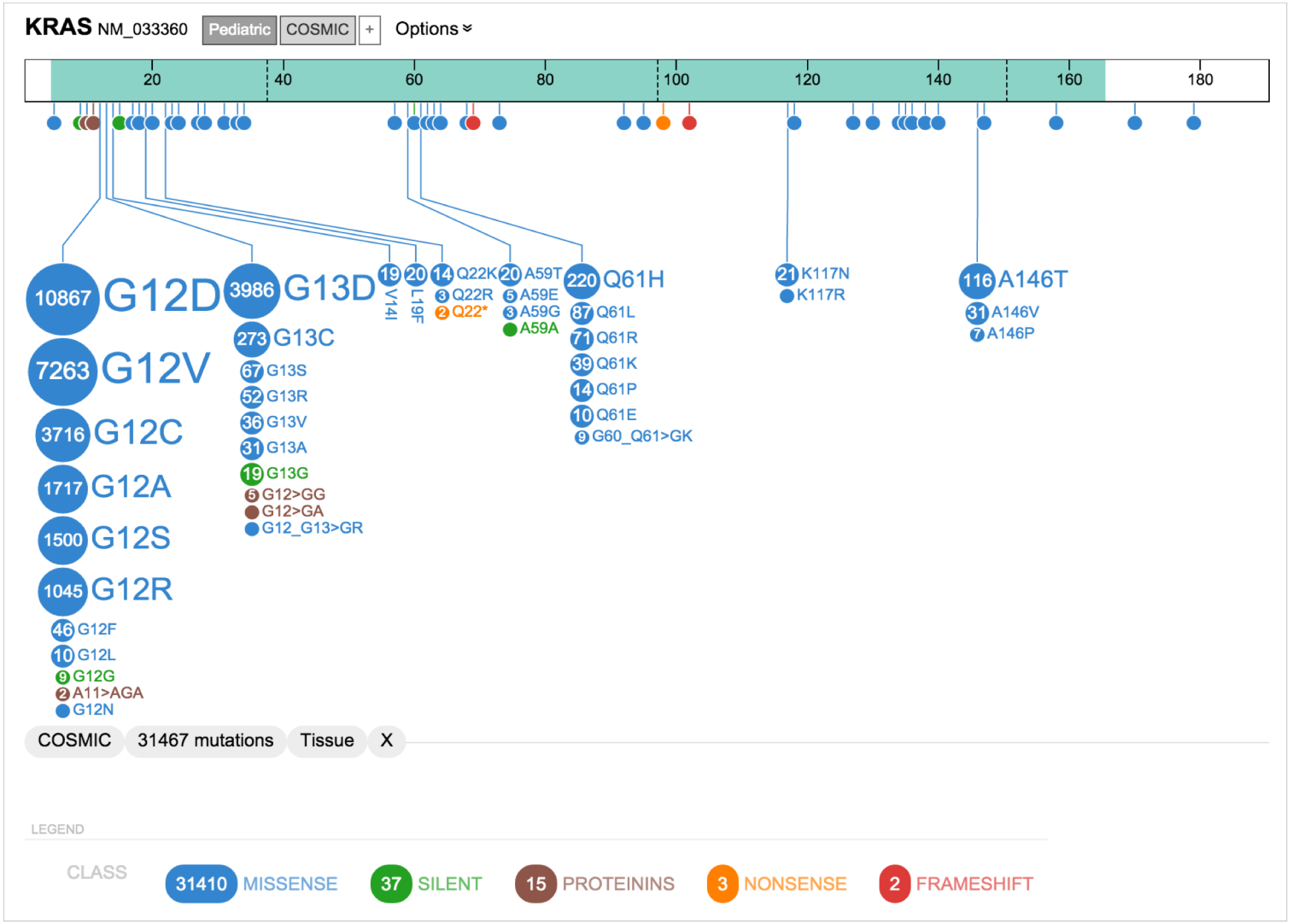


Visualizing copy number variation



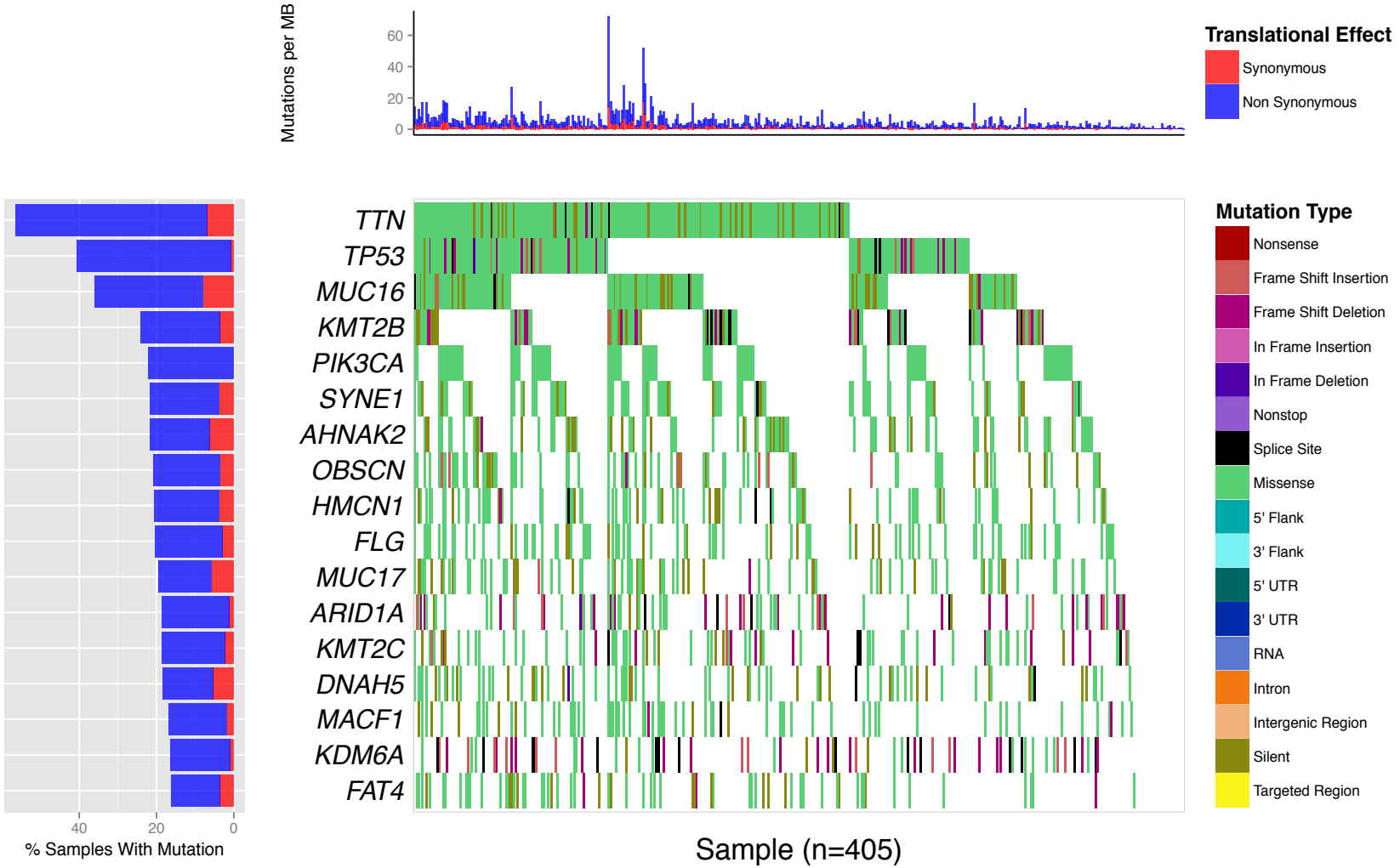
KRAS amplification in a metastatic breast cancer

A ‘lollipop’ plot is used to variant recurrence in a gene



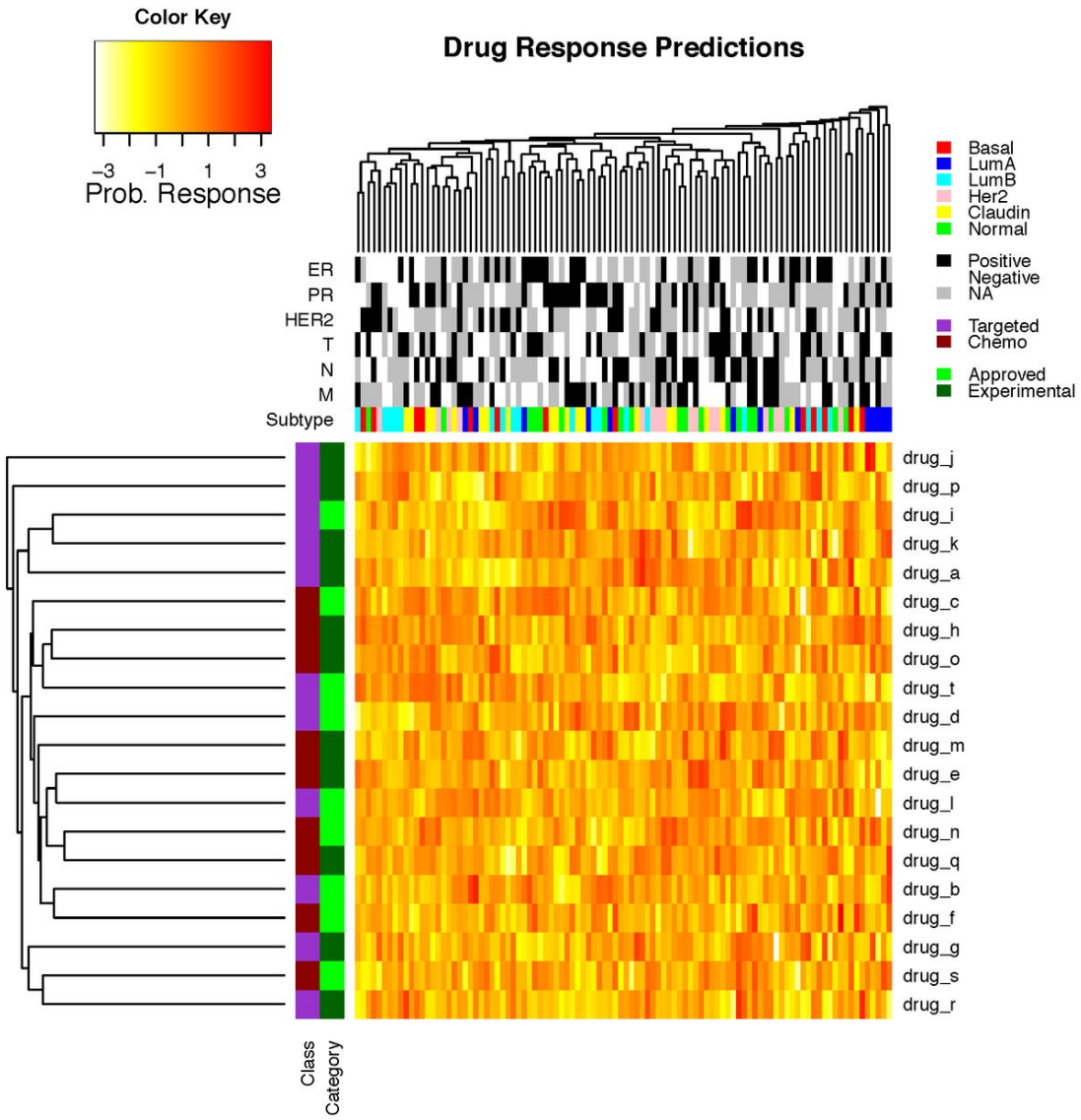
* Actually this one used “ProteinPaint” a web based tool we will try

A ‘waterfall’ plot is one way to visualize the pattern of recurrence in a cohort

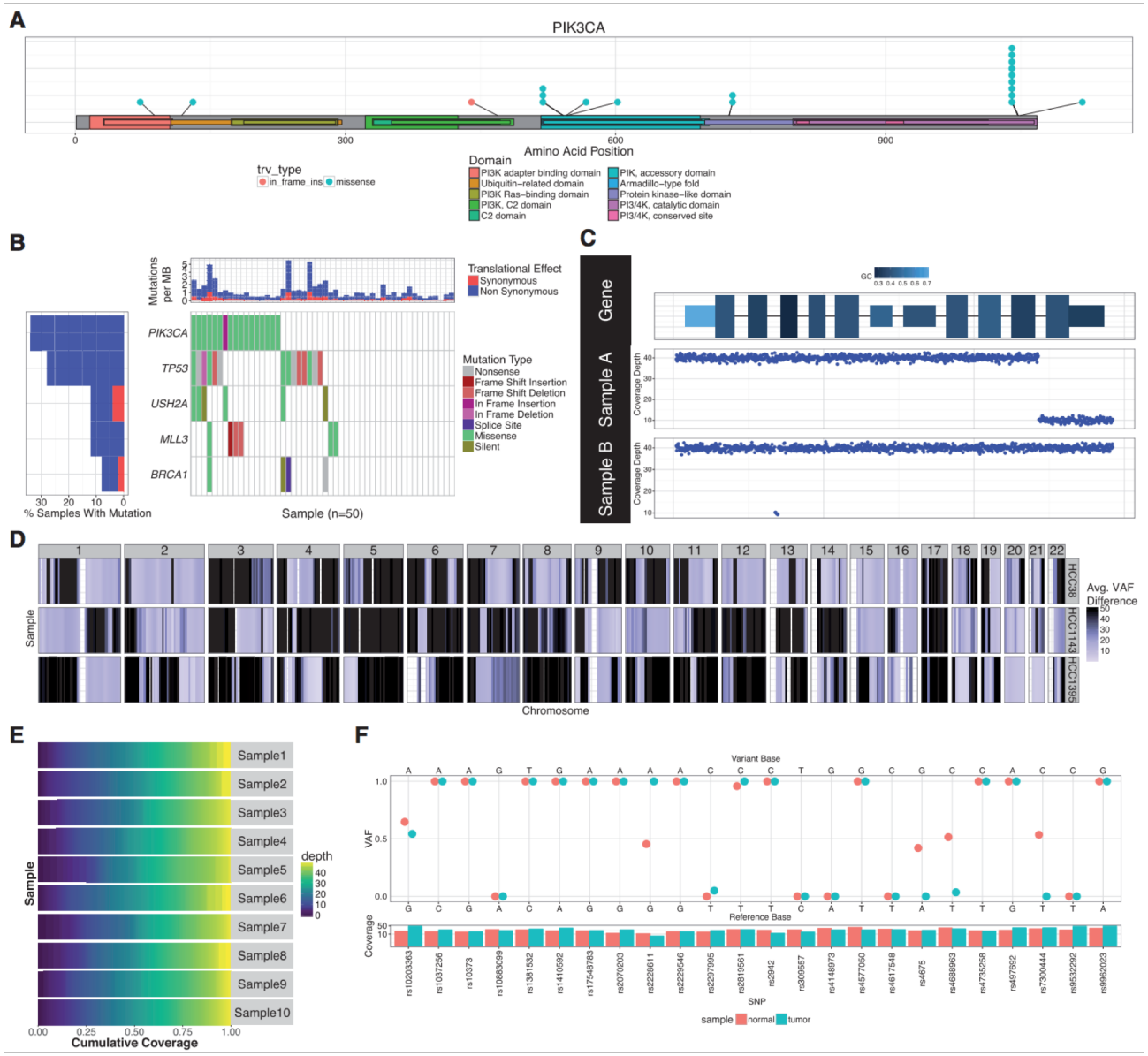


https://github.com/griffithlab/GenVisR

Heatmaps are a common way to simultaneously visualize multiple features of a dataset



GenVisR was created to help others make common genomic visualizations



Introduction to GenViz course site

www.genviz.org

Tutorial for module 1

www.genviz.org/course/#module-1