



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  
September 11-15, 2017  
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

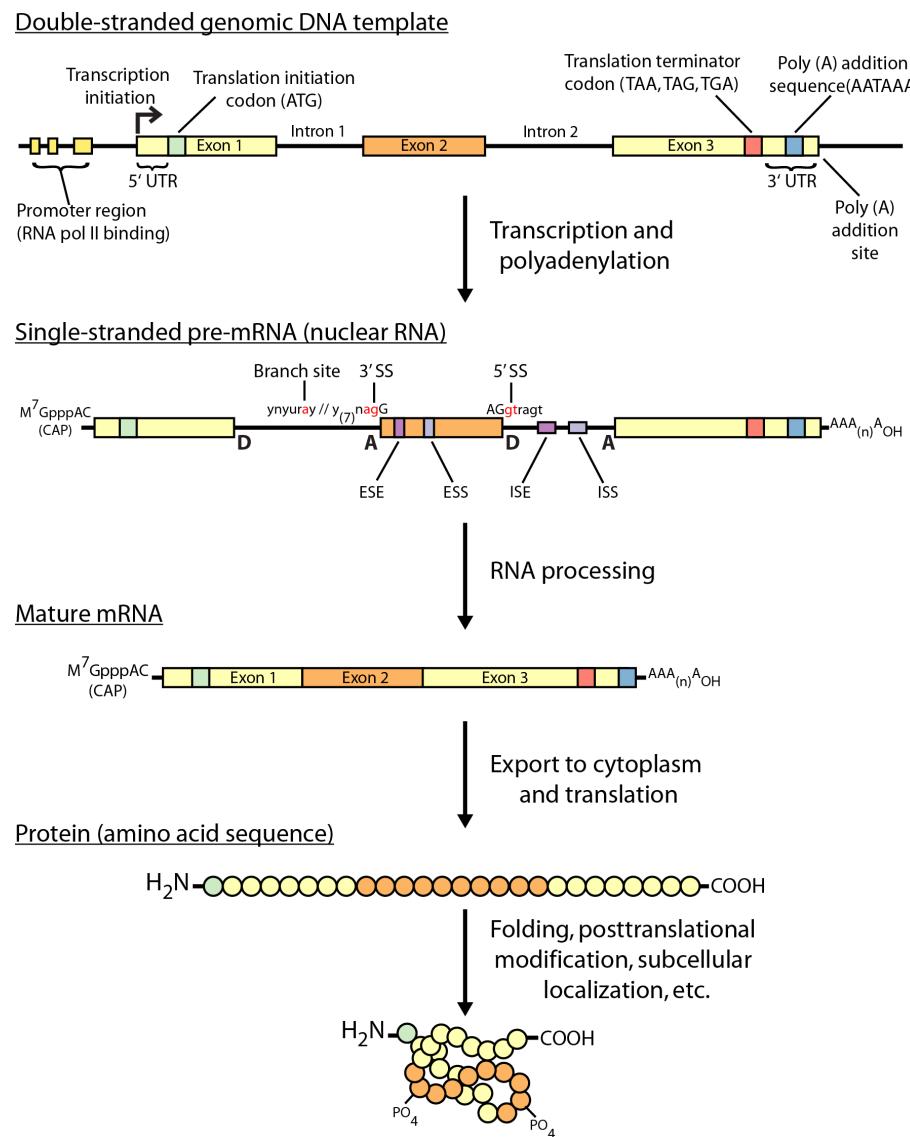
# Introduction

# 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.

# Review of central dogma

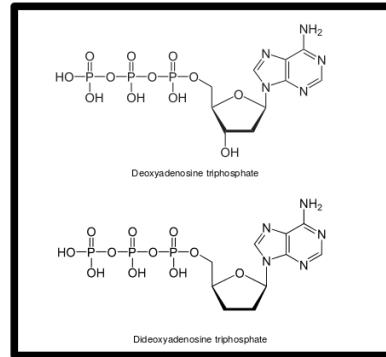
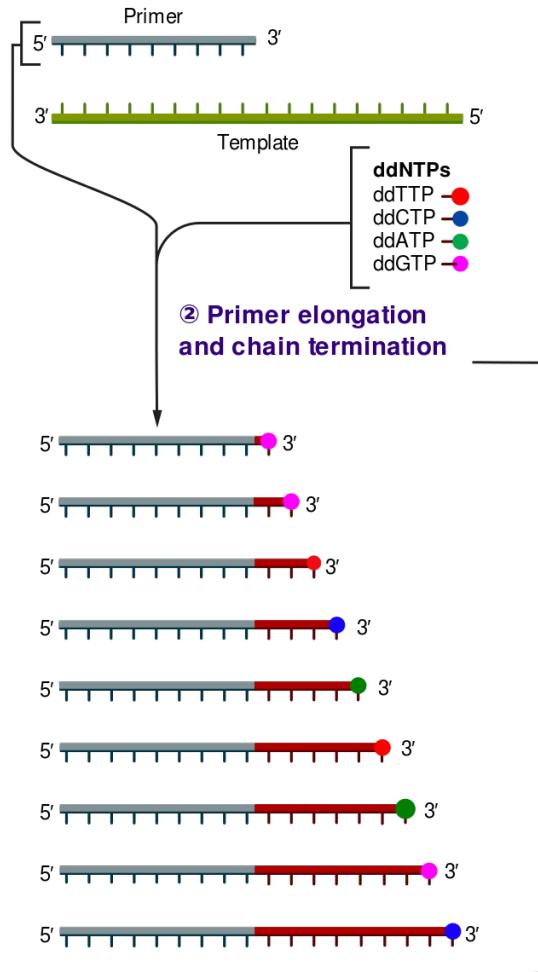


# ‘Omics technologies

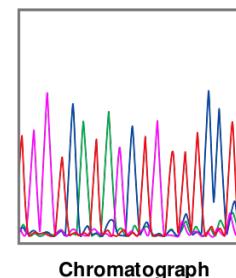
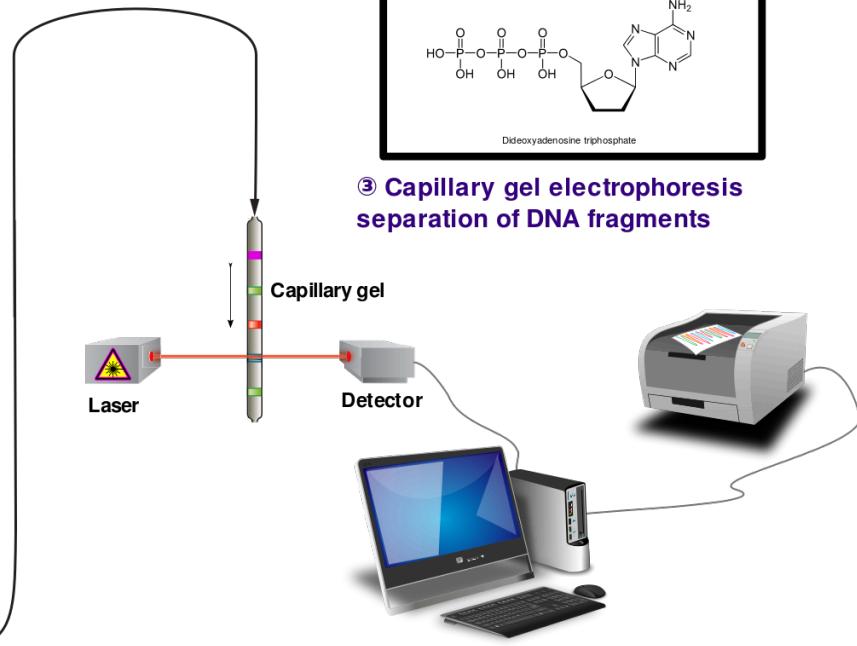
# Non-parallel (Sanger) sequencing

## ① Reaction mixture

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



## ③ Capillary gel electrophoresis separation of DNA fragments



## ④ Laser detection of flurochromes and computational sequence analysis

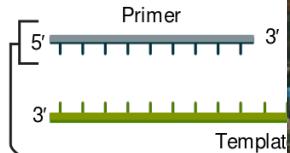
# Non-paral

15 February 2001

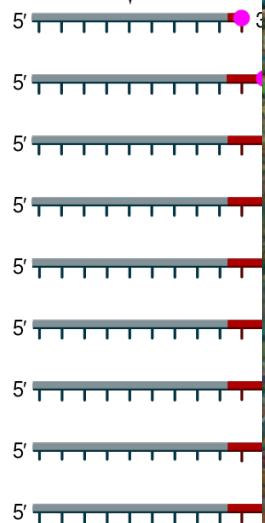
# nature

## the human genome

- ① Reaction mixture
- Primer and DNA template
- ddNTPs with flourochromes



- ② Primer and chain



### Nuclear fission

Five-dimensional energy landscapes

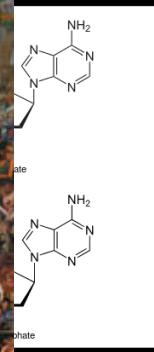
### Seafloor spreading

The view from under the Arctic ice

### Career prospects

Sequence creates new opportunities

**naturejobs**  
genomics special

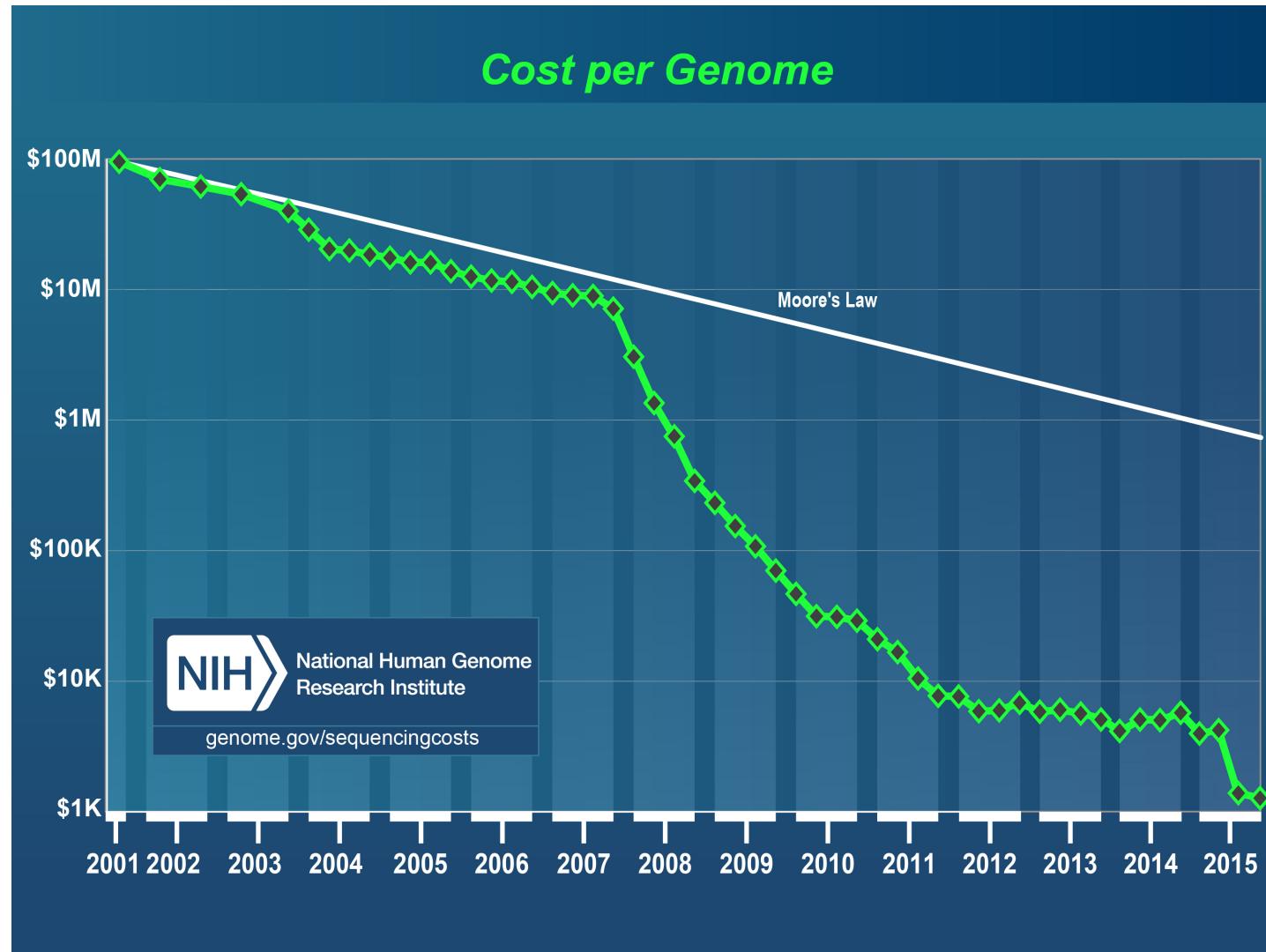


- Gel electrophoresis
- Fluorophores
- Segments

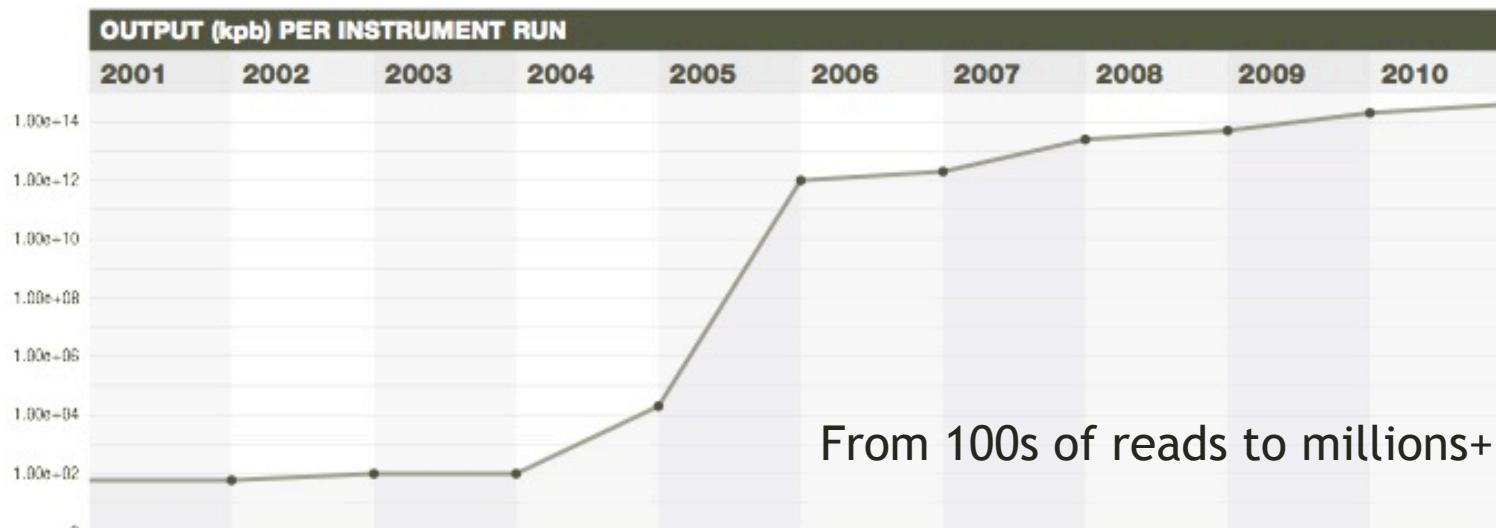


- Selection of fluorochromes
- Automation of sequence analysis

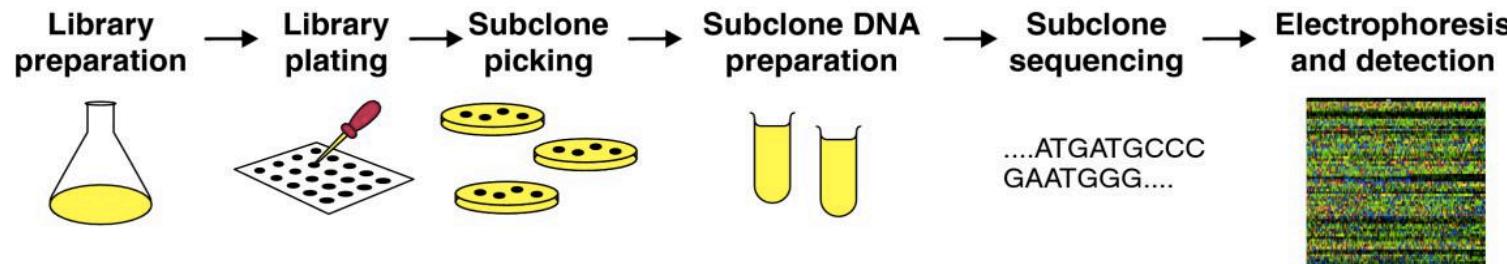
# Costs of sequencing a human genome have plummeted from \$100M to \$1000



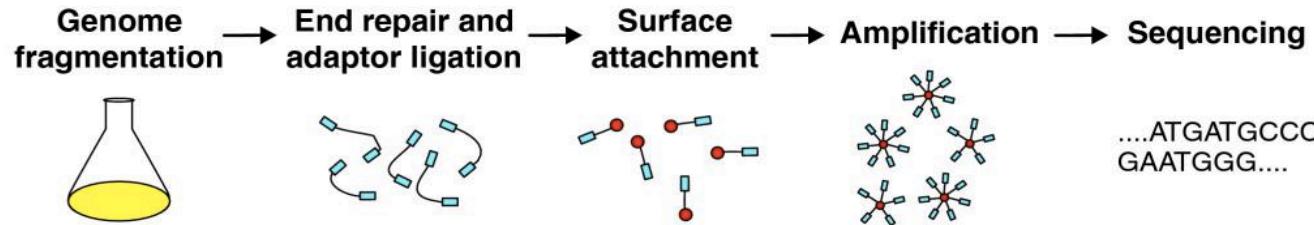
# Massively parallel sequencing (NGS) has transformed biomedical inquiry



(a)



(b)

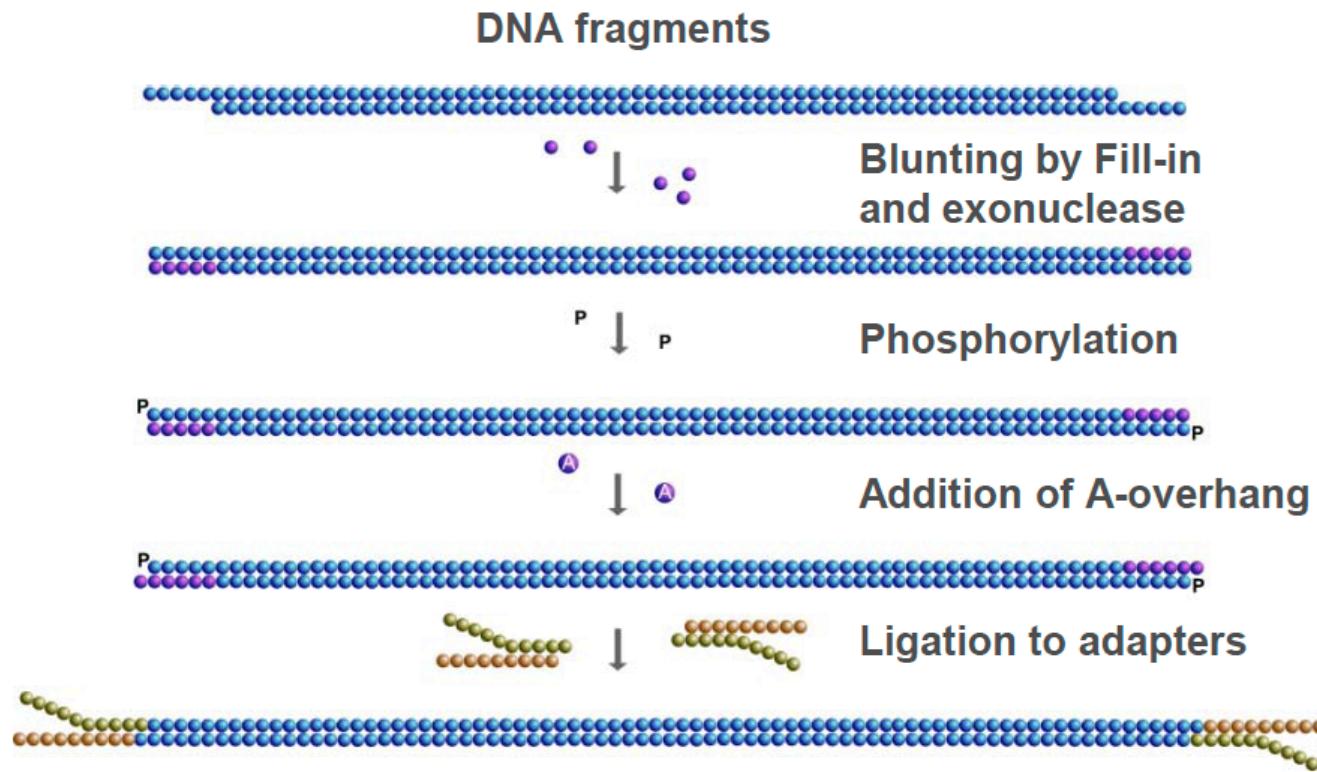


E.R. Mardis, Nature (2011) 470: 198-203, Ann. Rev. Analyt. Chem. (2013)

# 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

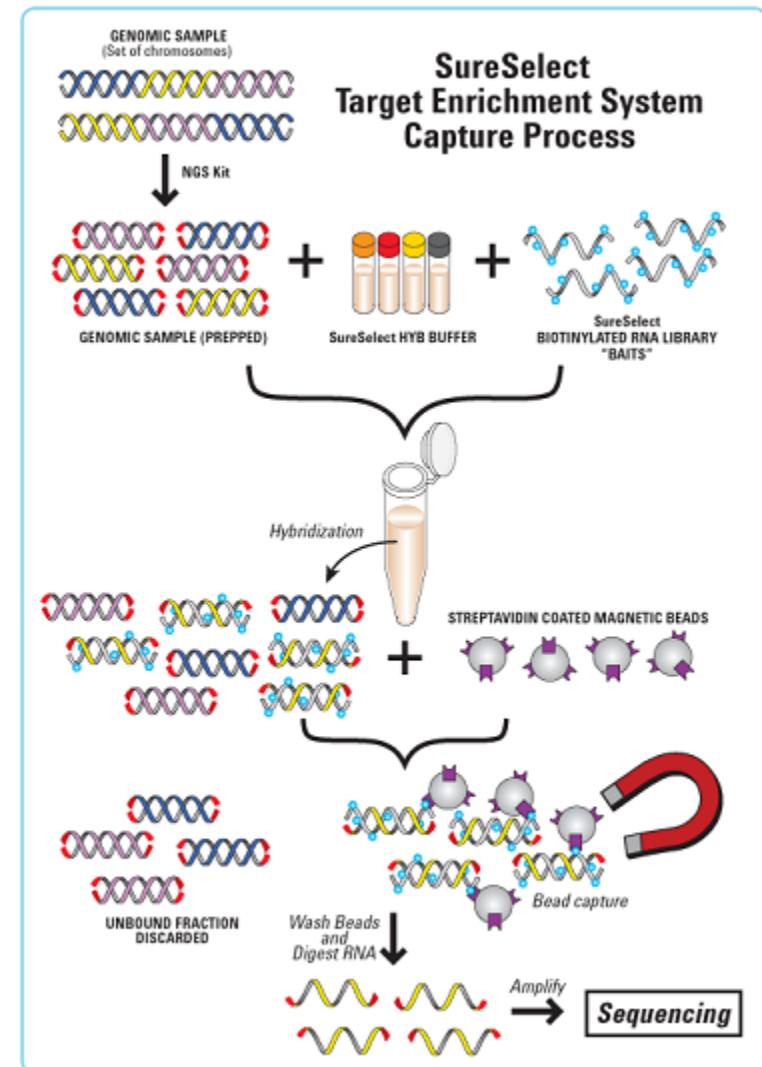
# Library Construction for MPS



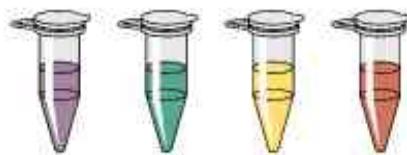
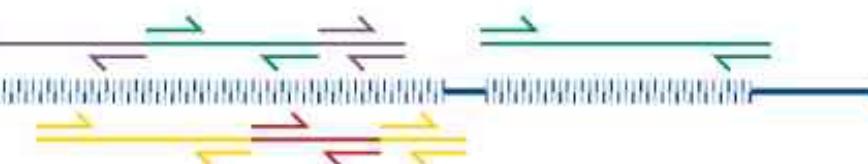
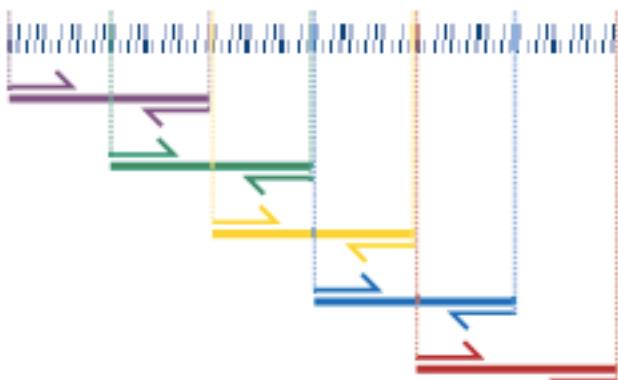
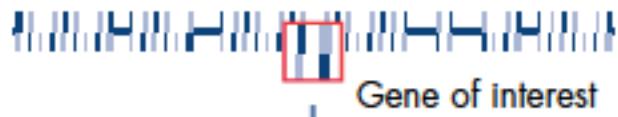
- Shear high molecular weight DNA with sonication
- Enzymatic treatments to blunt ends
- Ligate synthetic DNA adapters (each with a DNA barcode), PCR amplify
- Quantitate library
- Proceed to WGS, or perform exome or specific gene hybrid capture

# 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.



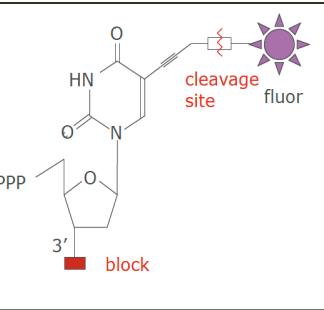
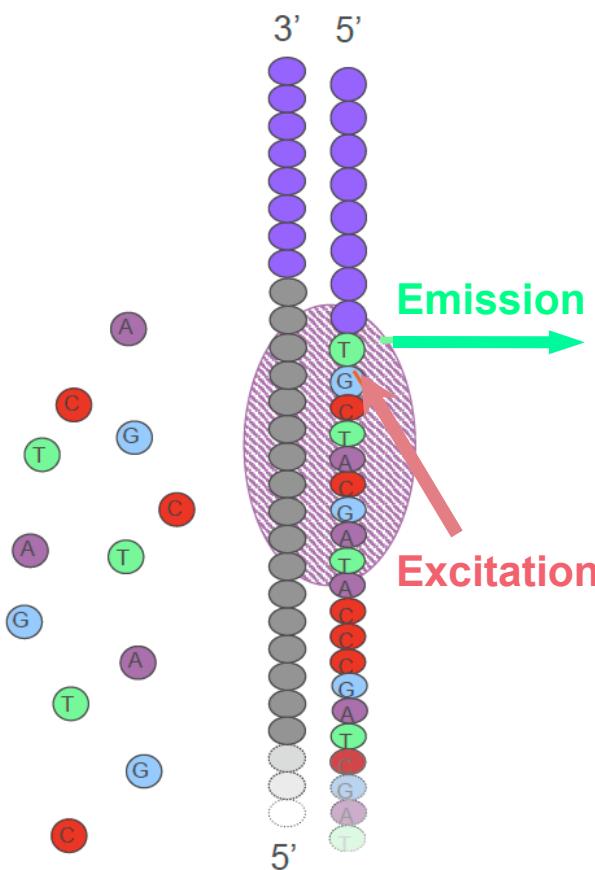
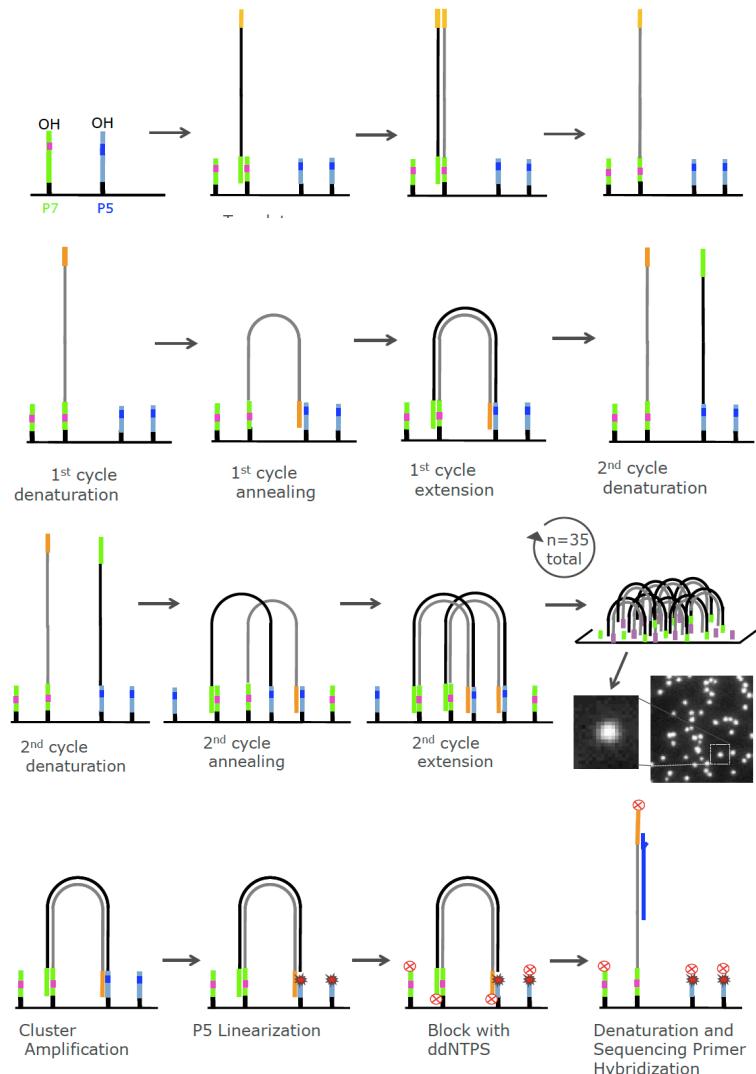
# Multiplex PCR Amplification of Targets



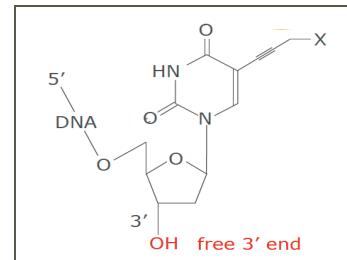
Multiplex PCR primer sets

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

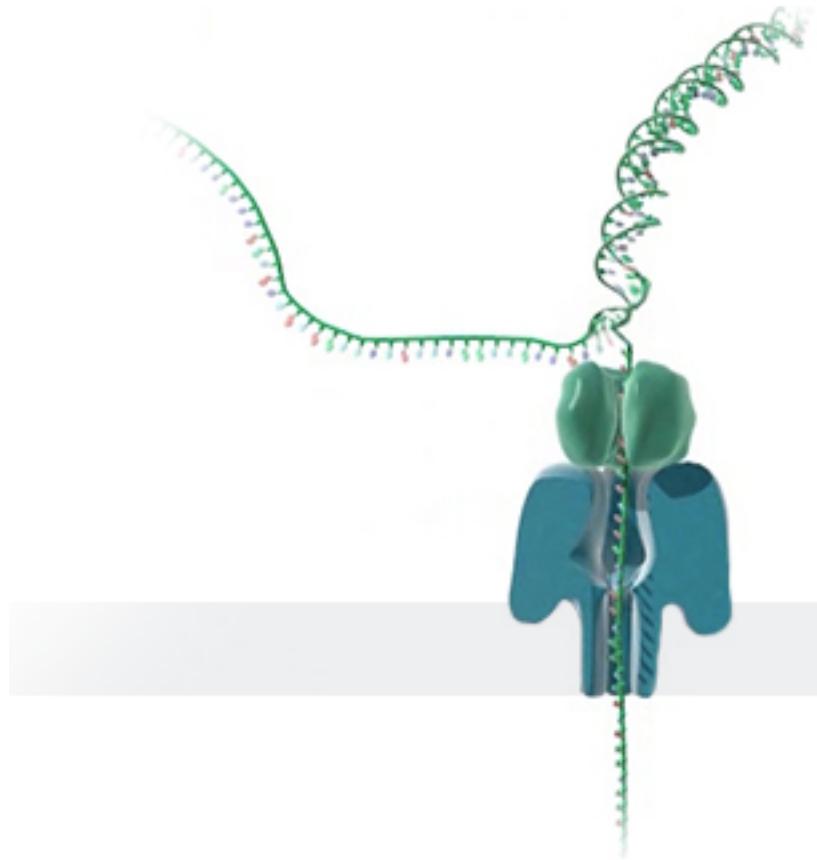


**Incorporate  
Detect  
De-block  
Cleave fluor**



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



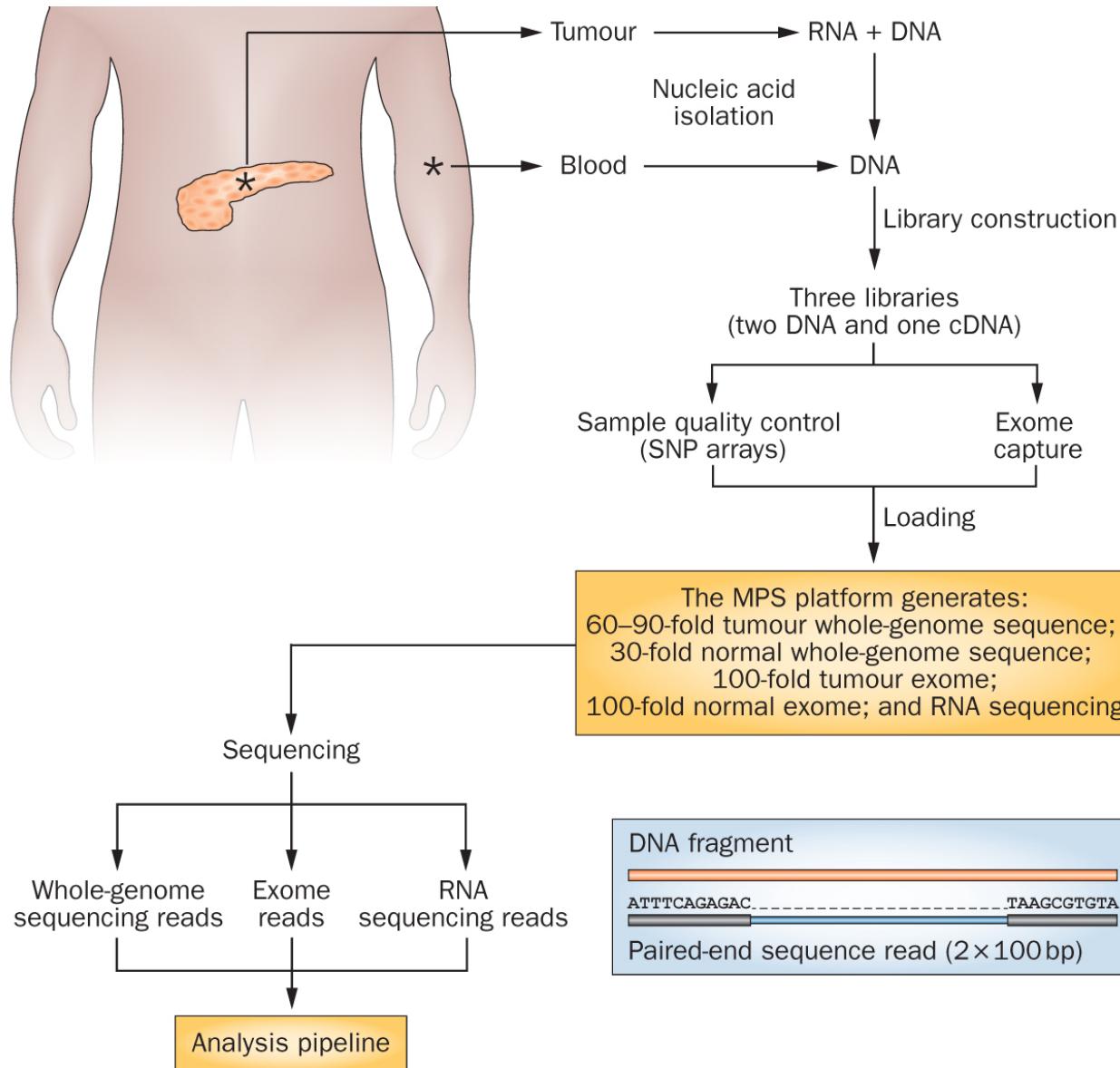
smidgION



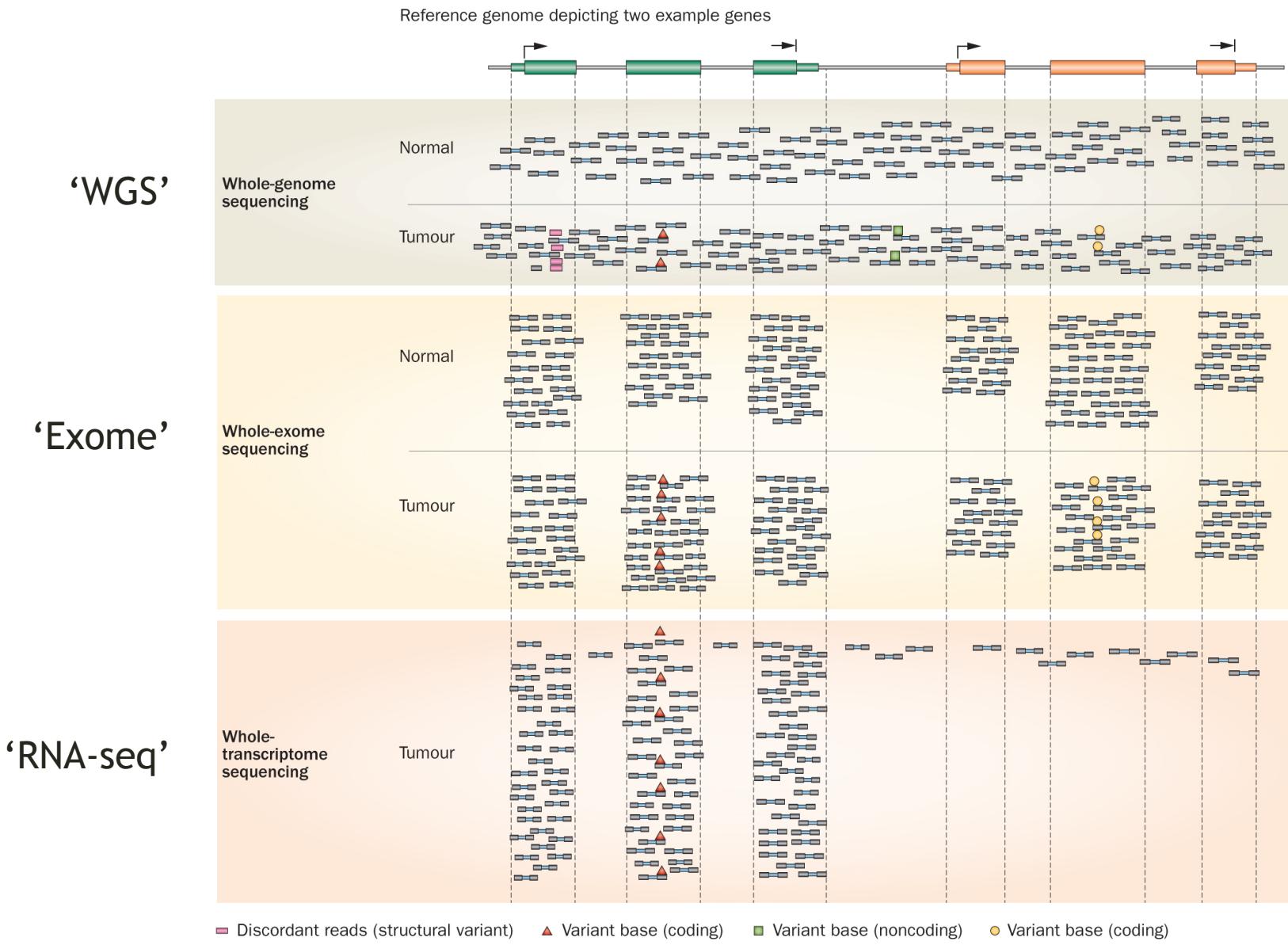
- Early access
- Price: TBD
- 48 x 3000 pores
- Array of minions

- In development
- Smart-phone compatible

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



# 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](#)
- [GTF](#)
- [BED](#)
- [VCF](#)
- [VEP](#)
- [MAF](#)
- [SAM/BAM/CRAM](#)
- Many, many custom data formats output by specialized tools ... often in TSV format

# 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: TU:FT		
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: TU:FT		
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: TU:FT		
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: TU:FT		
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: TU:FT		
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: TU:FT		
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: TU:FT		
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: TU:FT		
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: TU:FT		
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: TU:FT		
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: TU:FT		

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:a718aca6135fdca8357d5bfe9
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:2888721359 VN:1.85(exported) PP:2888721359 VN:1.85(exported) PN:MarkDuplicates Clnt_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-jwalker-1543
e10-2-5.gsc.wustl.edu-jwalker-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-liuJS/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_alignments/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=MarkDuplicates MAP_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]+.* OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 VERBOSITY=INFO QUIET=false COMPRESSION_LEVEL=5 CREATE_INDEX=false CREATE_MDS_FILE=false
mgriffit@linus270 ~$
```

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

# 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 <sup>16</sup> -1]	bitwise FLAG
3	RNAME	String	\*  [!-()+-<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 <sup>29</sup> -1]	1-based leftmost mapping POStion
5	MAPQ	Int	[0,2 <sup>8</sup> -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 <sup>29</sup> -1]	Position of the mate/next segment
9	TLEN	Int	[-2 <sup>29</sup> +1,2 <sup>29</sup> -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. CCTGTTCTCCACAAAGTGTACTTTGGATTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTGGCTGGGCATGT
11 QUAL e.g. CCCFFFFFHBBBBBHHHJJIIFIJJJJJJJJHIIJJJJJJJJJJGGHHIJHIIJJJJJJJJGHGGIJJJJJJIJEHHHHFFFFCDDDDDDB@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

# 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

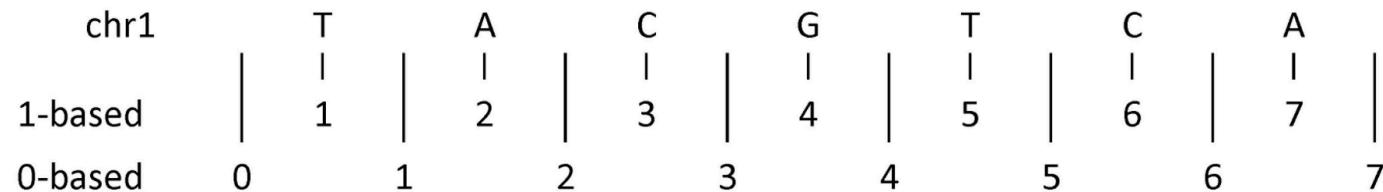
---

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

# Common problems

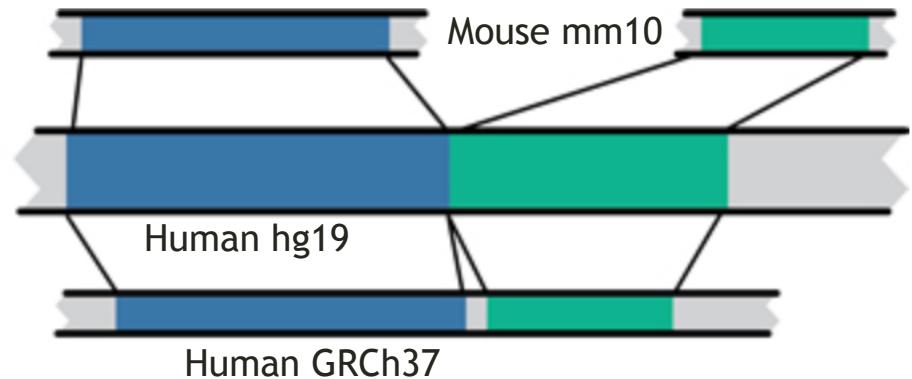
# Common problems

- Genomic coordinate systems
  - 1-based vs. 0-based



- Differences in carriage returns

- Genome builds
  - And annotation builds
  - “Liftover” tools



# Genome browsers

# Genome browsers - Ensembl

**e!Ensembl** BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search: All species  for   
e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

### Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotates genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

### Favourite genomes

 **Human**  
GRCh38.p10

 **Mouse**  
GRCm38.p5

 **Zebrafish**  
GRCz10

[Edit favourites](#)

### All genomes

-- Select a species --

[View full list of all Ensembl species](#)

### Find a Data Display



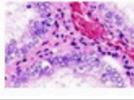
Not sure how to find the data visualisation you need? With our new [Find a Data Display](#) page, you can choose a gene, region or variant and then browse a selection of relevant visualisations

[Try it now!](#)

### Variant Effect Predictor



### Gene expression in different tissues



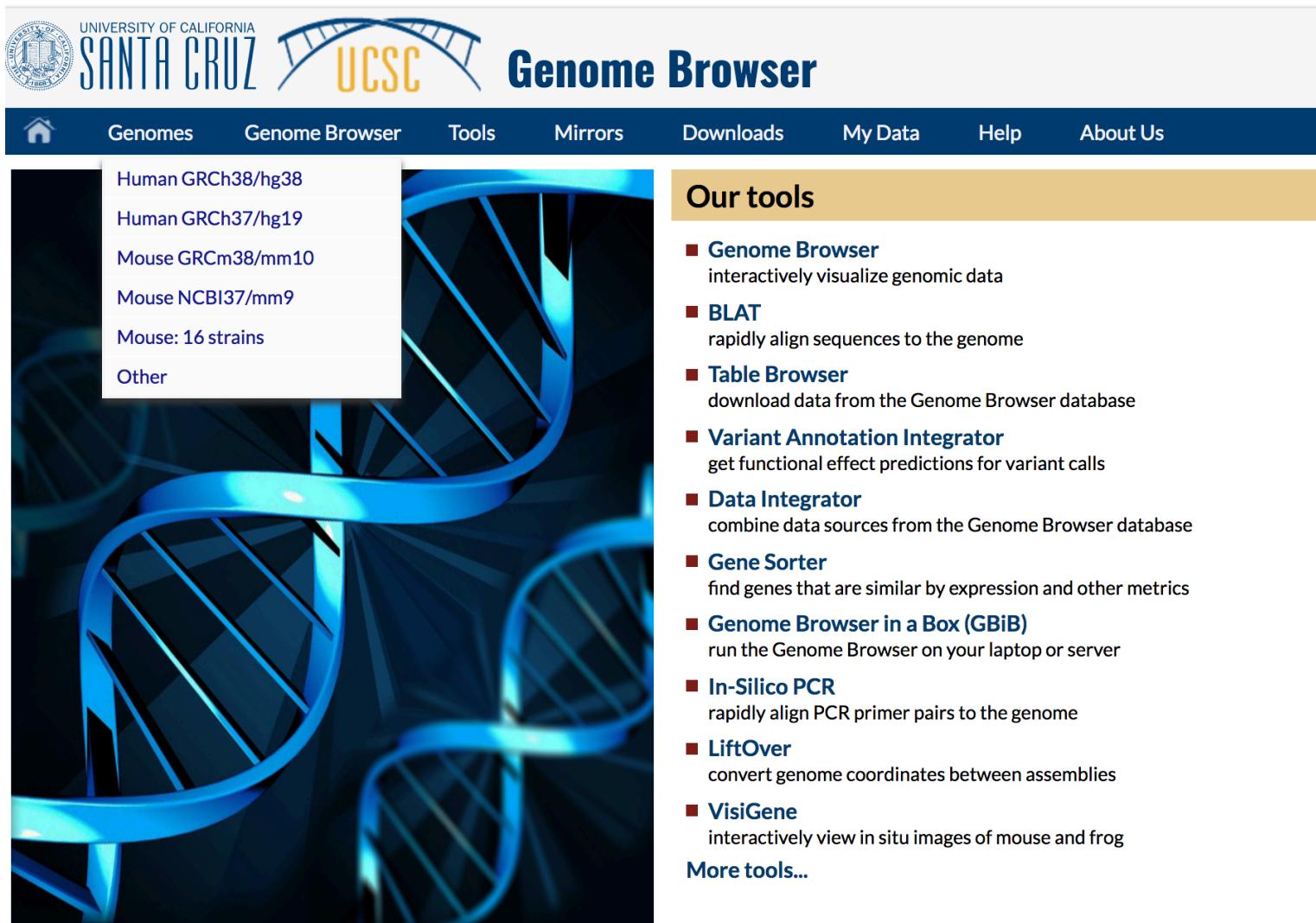
### Retrieve gene sequence

```
GCCTGACTTCGGGTGGC  
GGGCTTGTGGCGCGAGC  
GCGCCTCTGTCGGCCTC  
AGGGGACAGATTTGTGA  
CACCTCTGGAGCGGGTTG  
CCCAGTCAGCGTGGCG
```

### Compare genes across species



# Genome browsers - UCSC

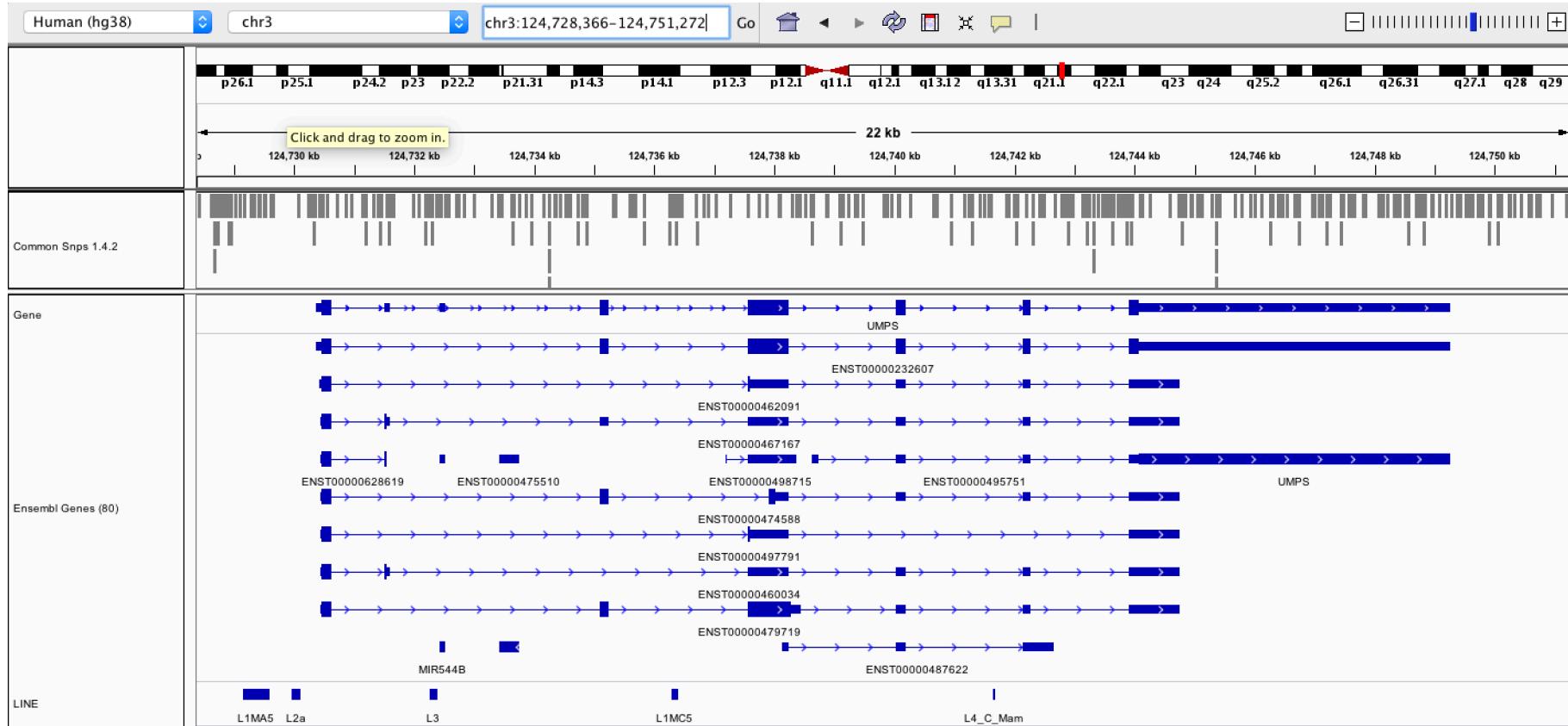


The image shows the homepage of the UCSC Genome Browser. At the top left is the UCSC logo with the text "UNIVERSITY OF CALIFORNIA SANTA CRUZ". To the right is the "UCSC" logo with a blue arch graphic. The title "Genome Browser" is centered above a navigation bar. The navigation bar includes links for "Home", "Genomes", "Genome Browser", "Tools", "Mirrors", "Downloads", "My Data", "Help", and "About Us". A dropdown menu under "Genomes" lists "Human GRCh38/hg38", "Human GRCh37/hg19", "Mouse GRCm38/mm10", "Mouse NCBI37/mm9", "Mouse: 16 strains", and "Other". The main content area features a large blue DNA helix graphic. On the right, a yellow sidebar titled "Our tools" lists ten tools with descriptions:

- Genome Browser**: interactively visualize genomic data
- BLAT**: rapidly align sequences to the genome
- Table Browser**: download data from the Genome Browser database
- Variant Annotation Integrator**: get functional effect predictions for variant calls
- Data Integrator**: combine data sources from the Genome Browser database
- Gene Sorter**: find genes that are similar by expression and other metrics
- Genome Browser in a Box (GBiB)**: run the Genome Browser on your laptop or server
- In-Silico PCR**: rapidly align PCR primer pairs to the genome
- LiftOver**: convert genome coordinates between assemblies
- VisiGene**: interactively view *in situ* images of mouse and frog

[More tools...](#)

# 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](#)
- Resources for choosing colors
  - <http://colorbrewer2.org/>
  - <http://mkweb.bcgsc.ca/color/>
  - [Understanding and using Color Palettes](#)
- 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.
- Choose 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

---

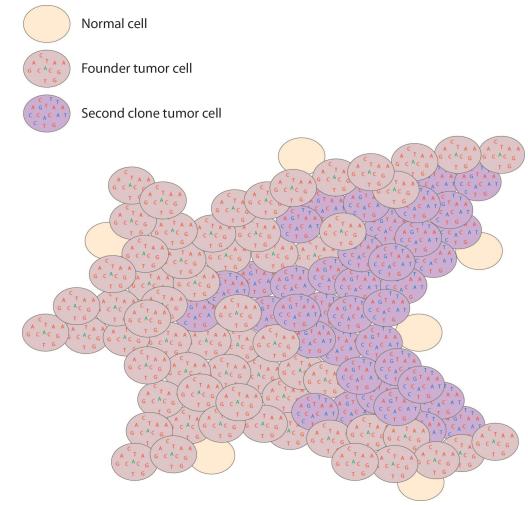
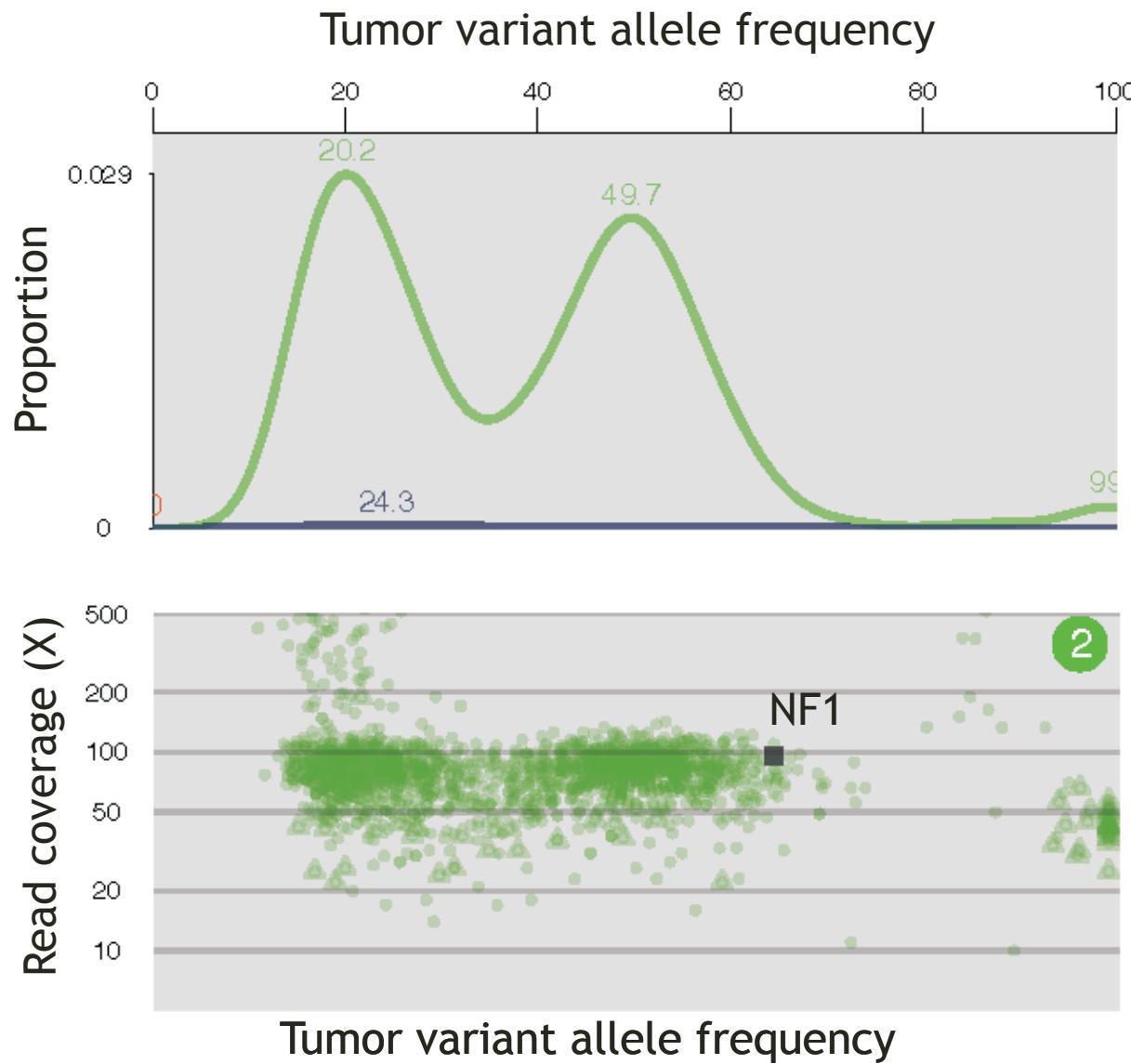
## Ten Simple Rules for Better Figures (Rougier et al. 2014):

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

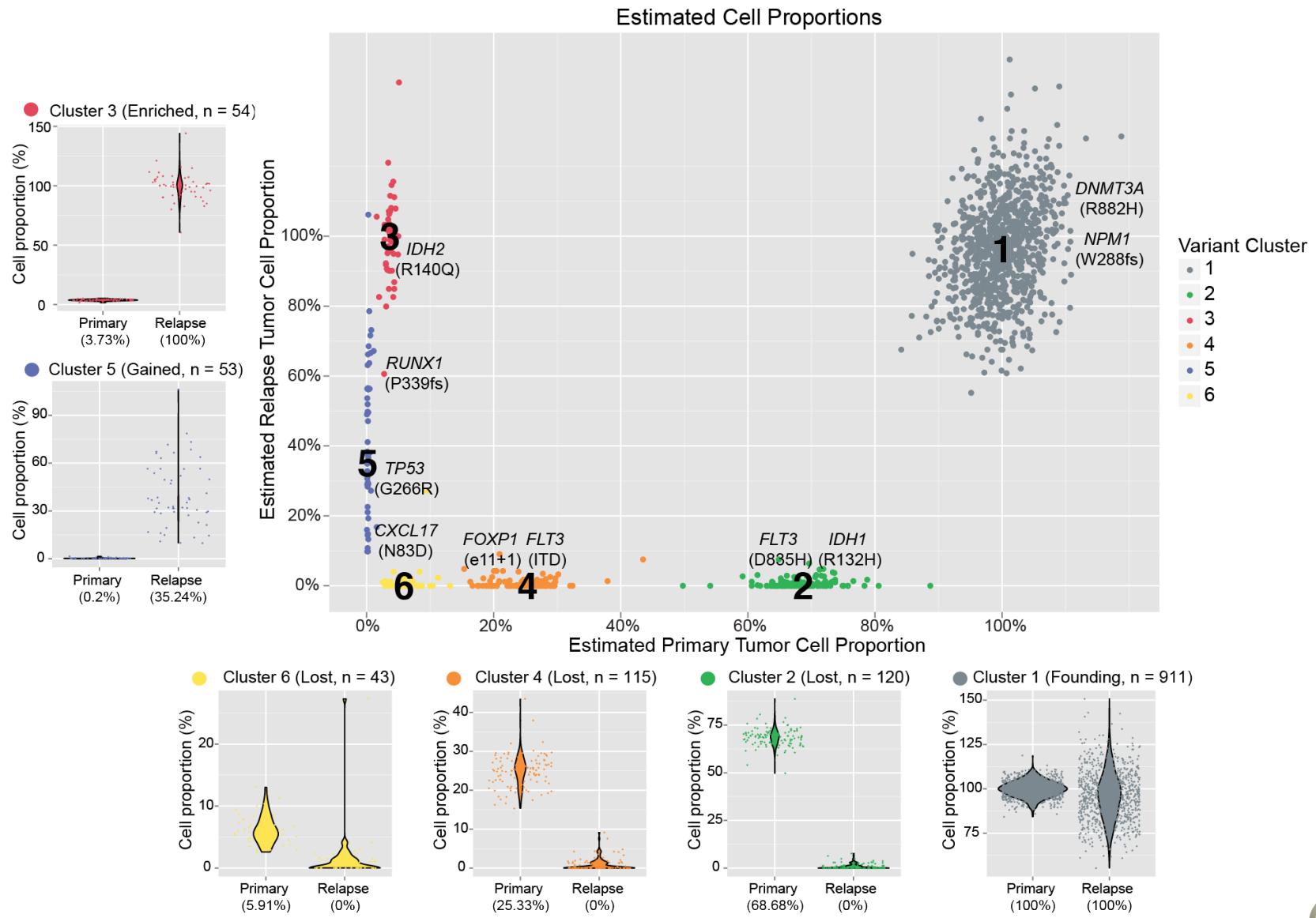
1. Know Your Audience
2. Identify Your Message
3. Adapt the Figure to the Support Medium
4. Captions Are Not Optional
5. Do Not Trust the Defaults
6. Use Color Effectively
7. Do Not Mislead the Reader
8. Avoid “Chartjunk”
9. Message Trumps Beauty
10. Get the Right Tool

## Example visualizations using R

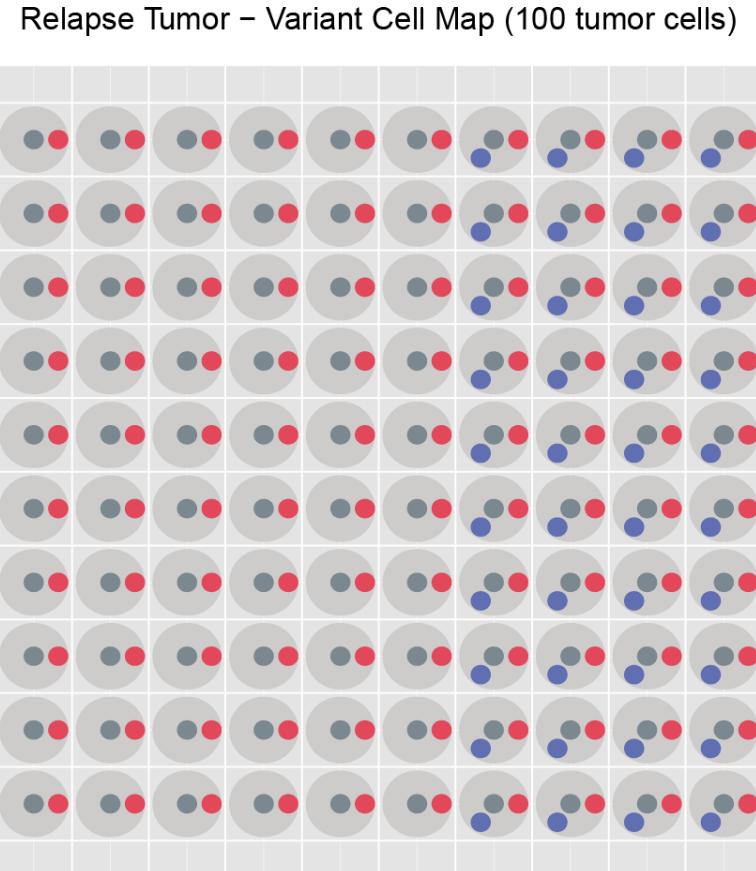
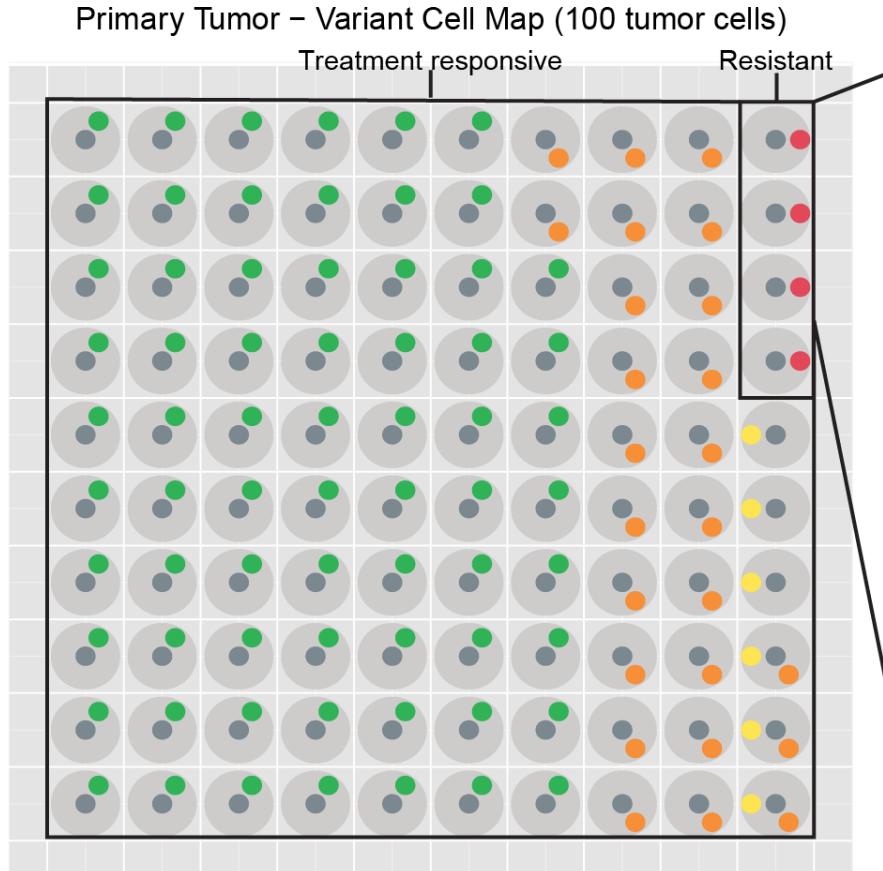
# Histogram and scatterplot to define cell populations



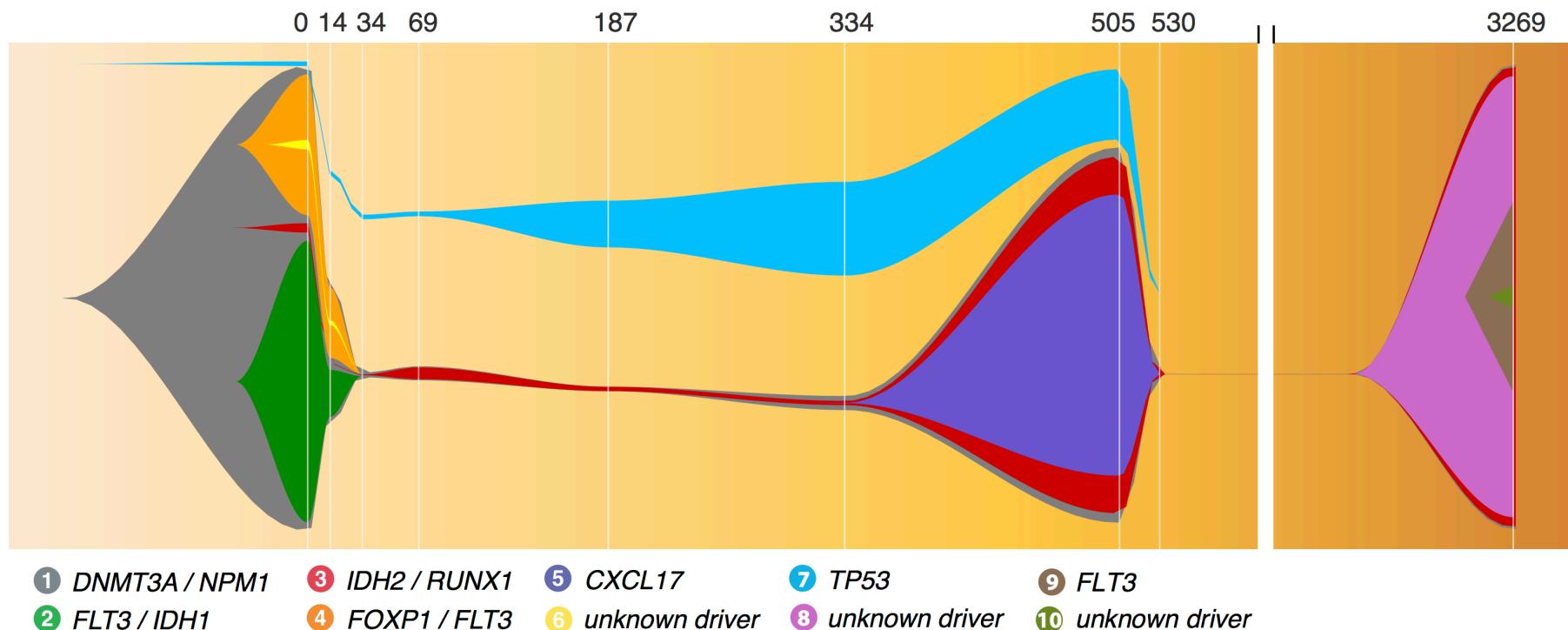
# Clustering to define two distinct cell populations



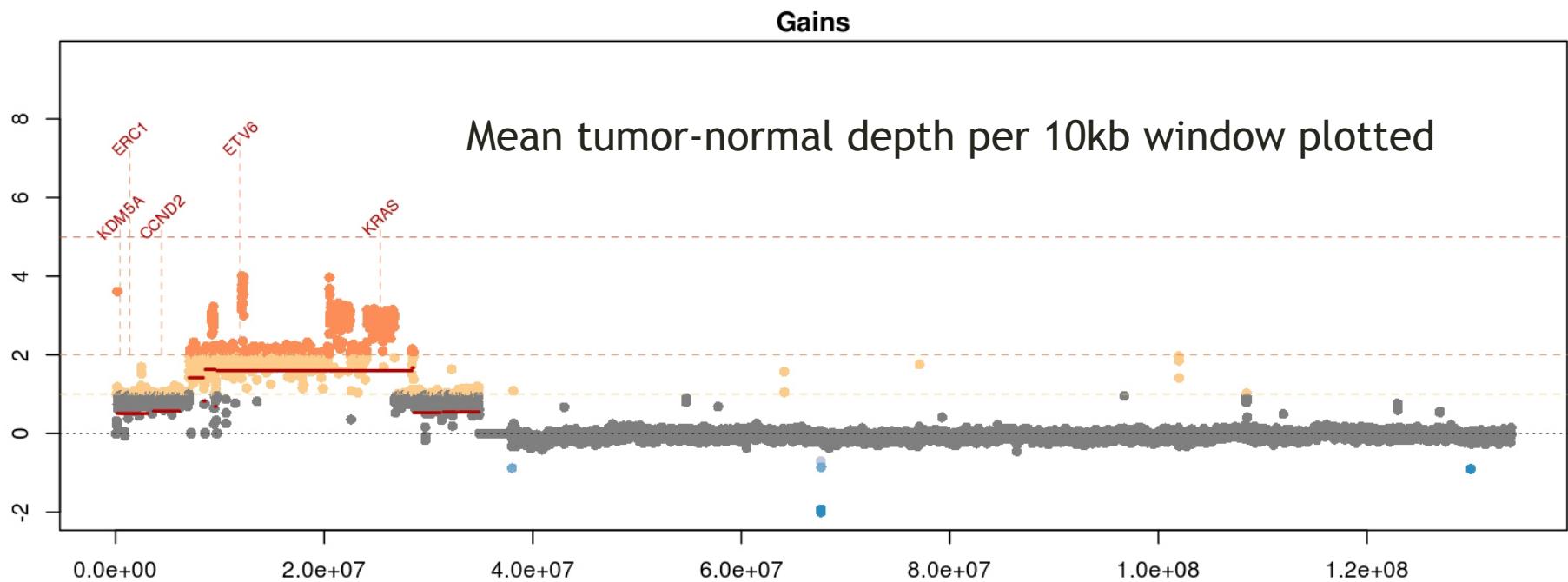
# Using a “cell map” to represent the variant clusters in 100 hypothetical cells of a tissue



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

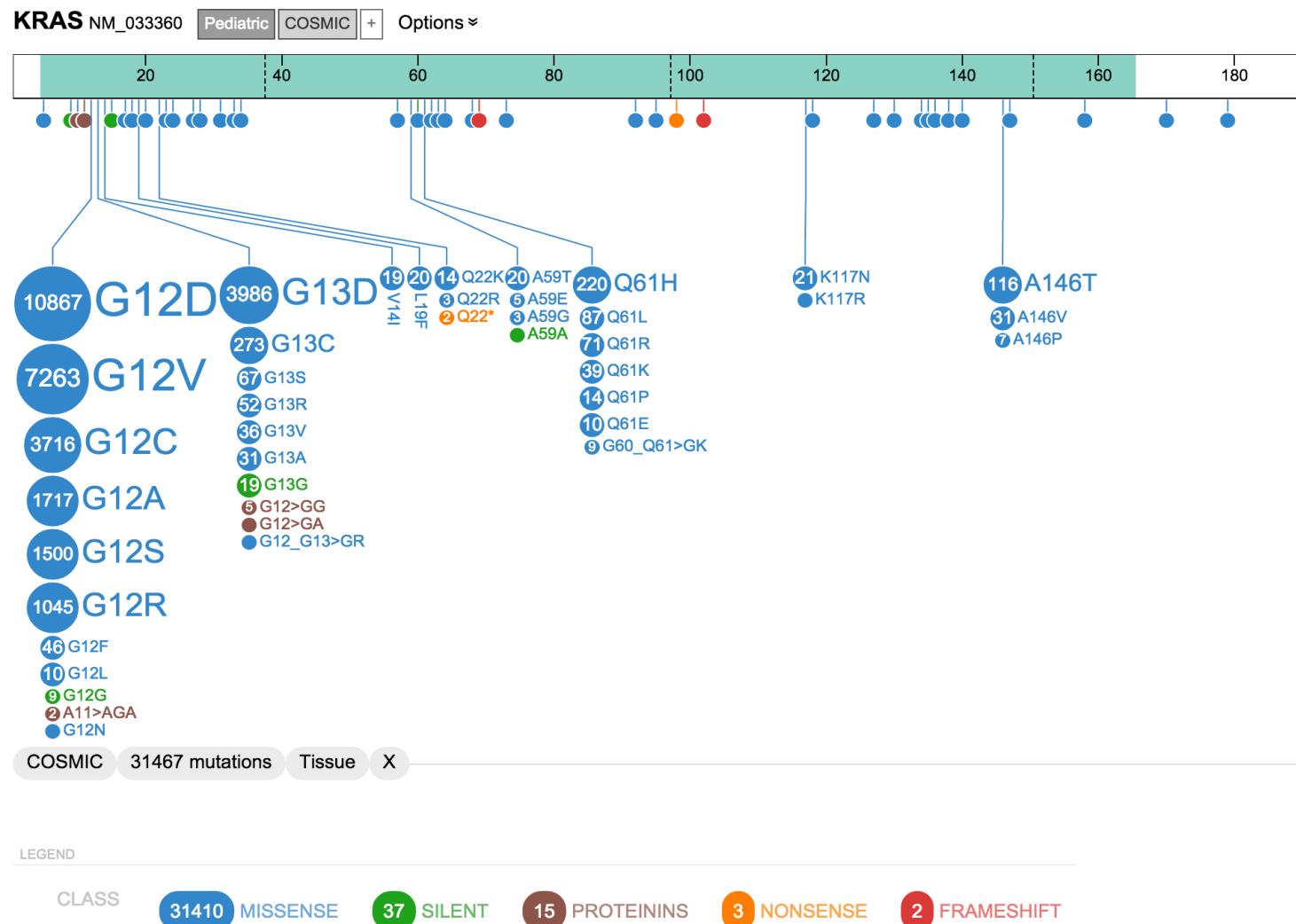


# Visualizing copy number variation



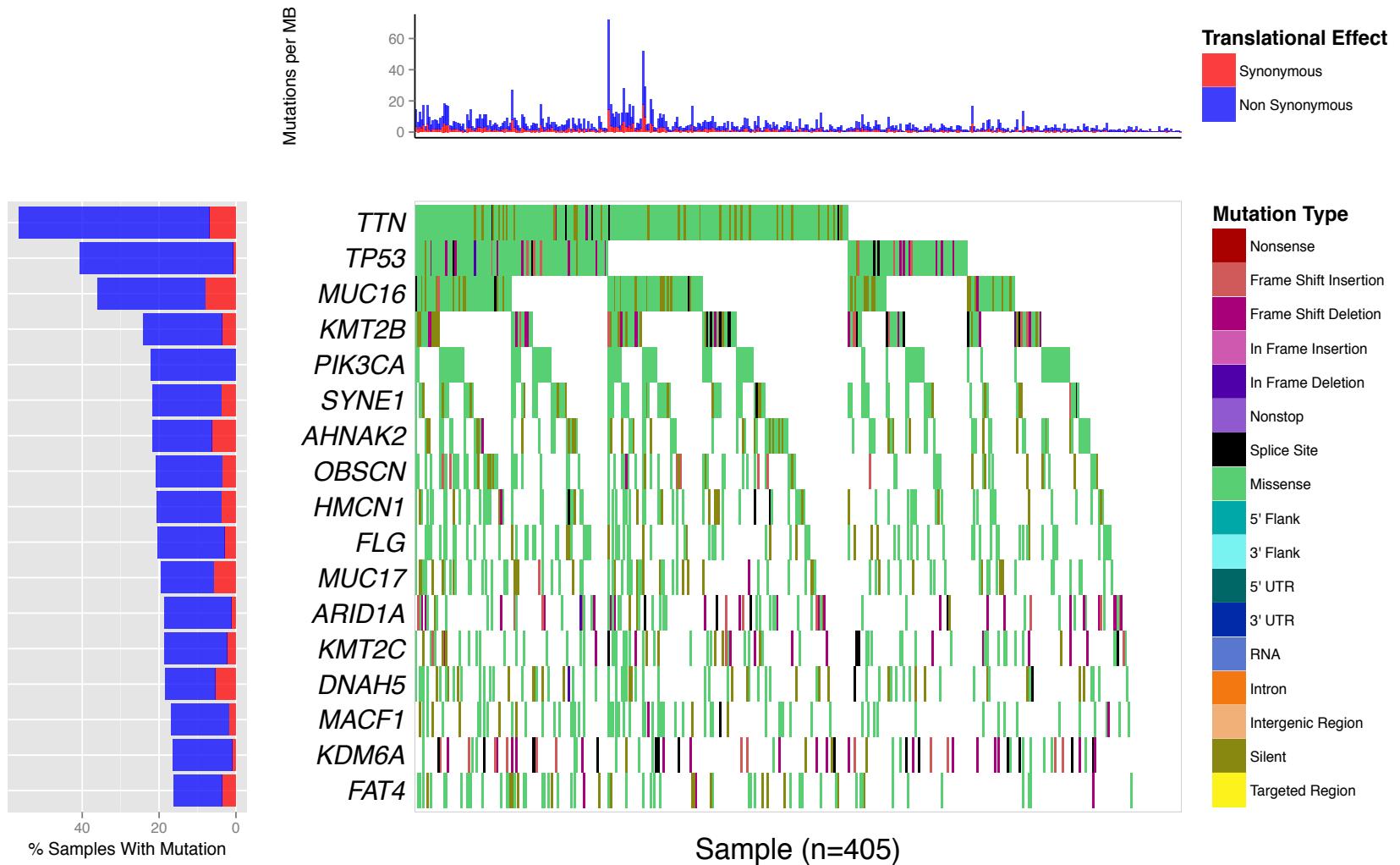
KRAS amplification in a metastatic breast cancer

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



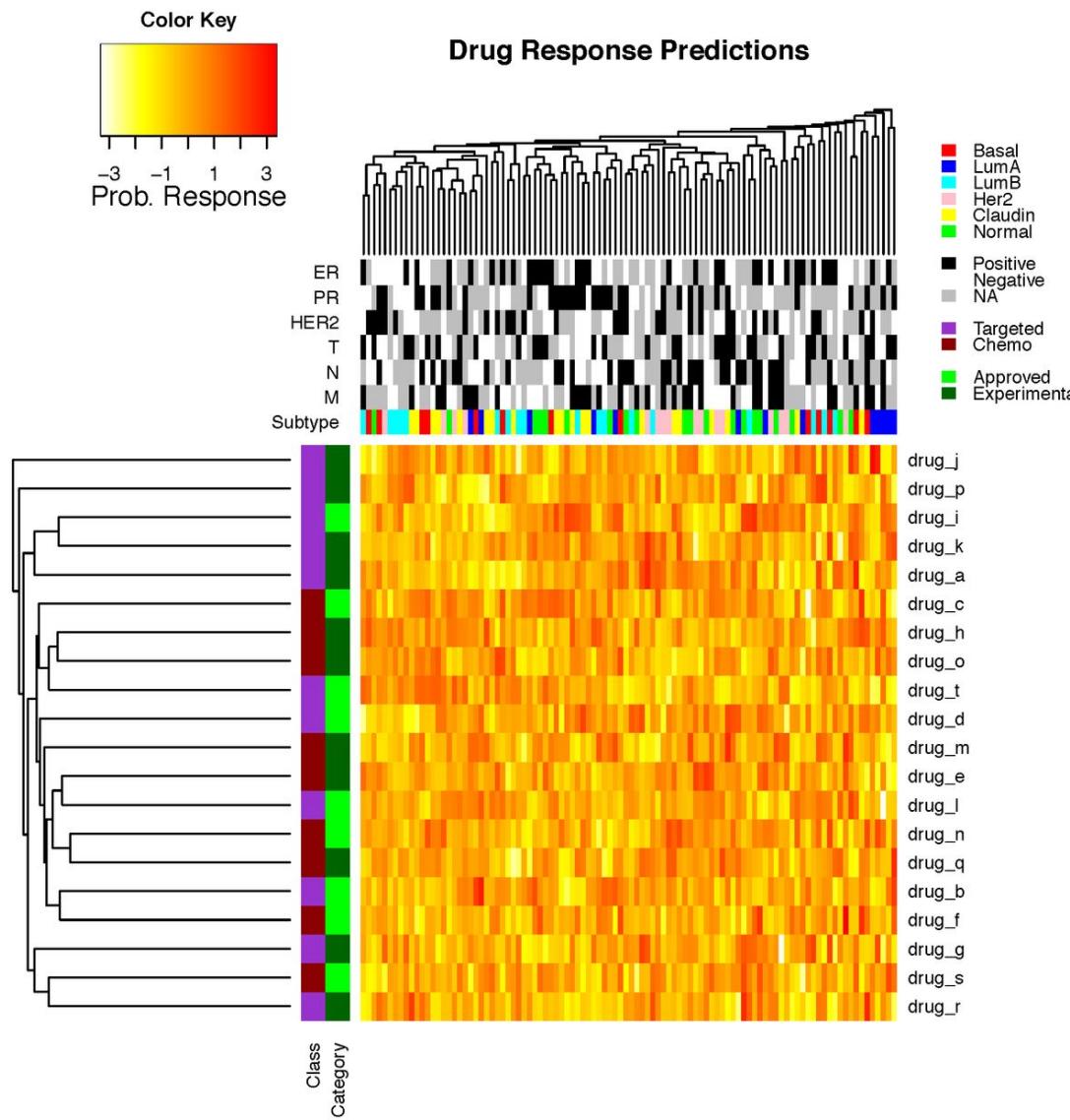
\* Actually this one used “ProteinPaint” a web based tool we will try

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

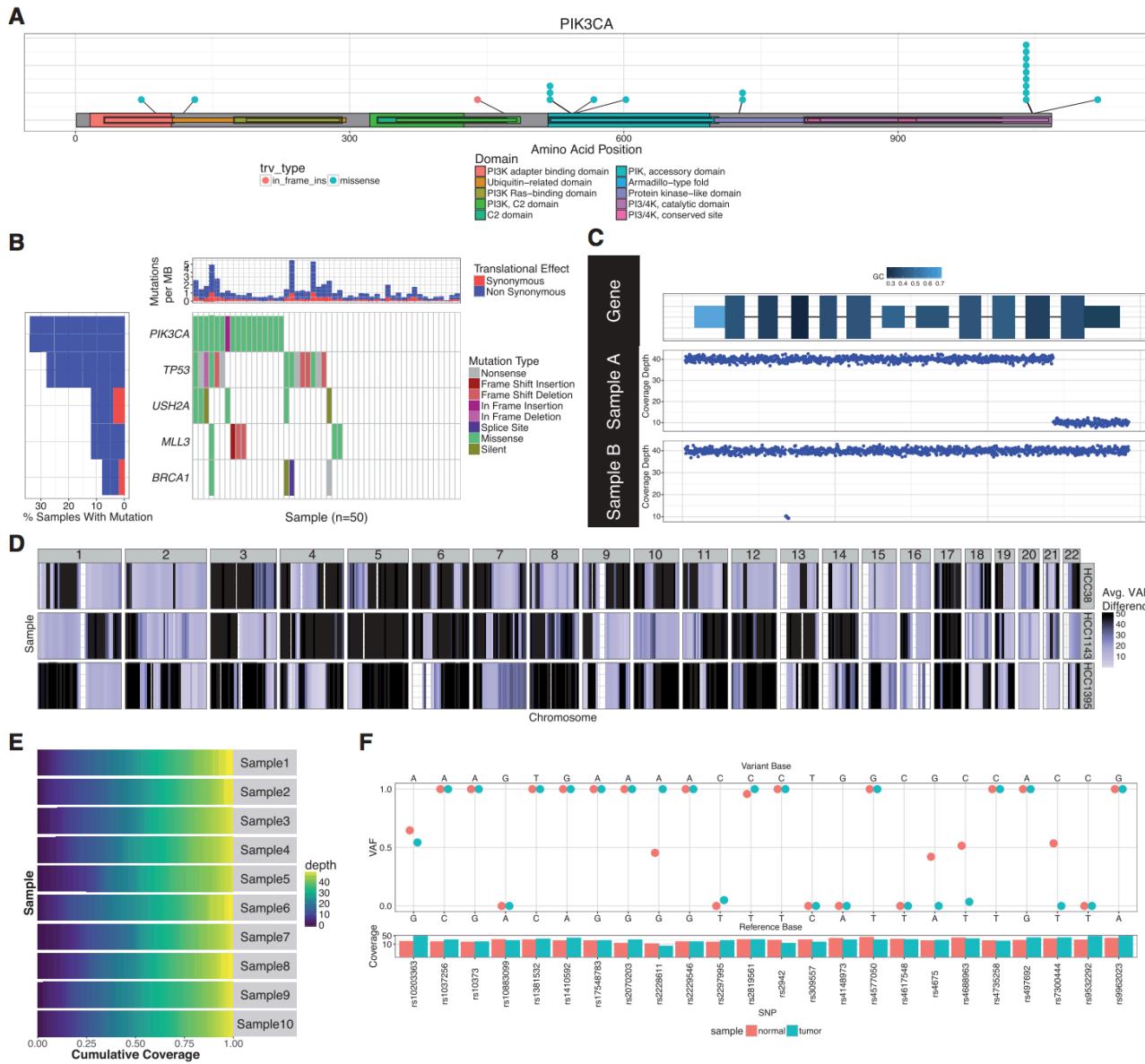


<https://github.com/griffithlab/GenVisR>

# 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](http://www.genviz.org)

# Tutorial for module 1

---

[www.genviz.org/course/#module-1](http://www.genviz.org/course/#module-1)