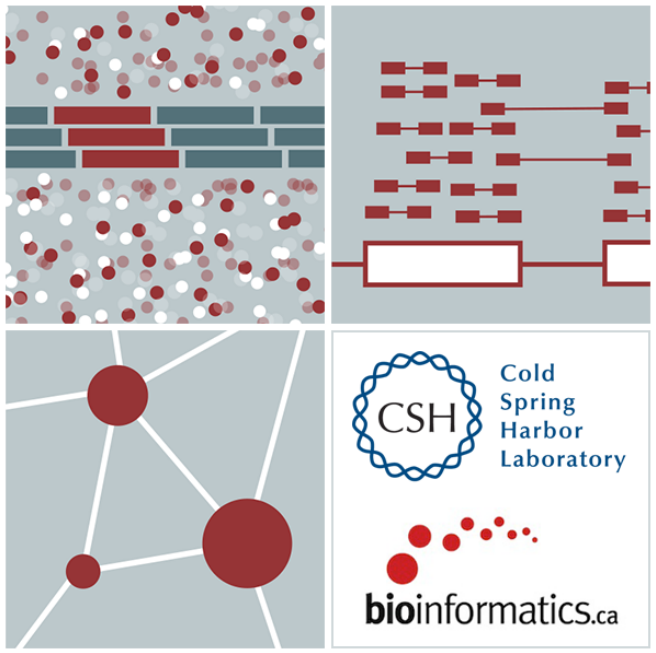
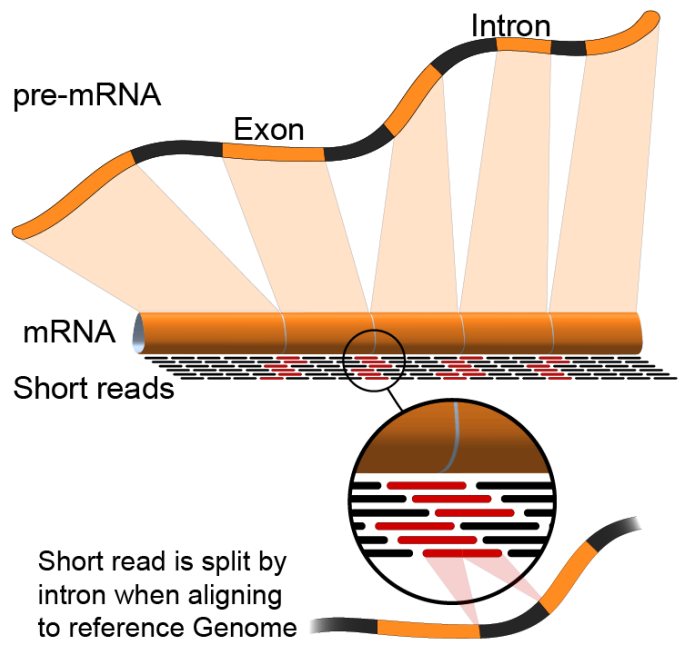
Introduction to sequencing data processing and analysis

Kelsy Cotto, Malachi Griffith, Obi Griffith,
Evelyn Schmidt, Kartik Singhal, Zach Skidmore

CRI Bioinformatics Workshop. May 17-22, 2025





Washington University in St. Louis
School of Medicine

Introductions to Bioinformatics instructors (WashU team)

Malachi Griffith
Associate Professor of Medicine
Associate Professor of Genetics
Assistant Director, MGI

Obi Griffith
Associate Professor of Medicine
Associate Professor of Genetics
Assistant Director, MGI

GRIFFITHLAB

griffithlab.org
rnabio.org genviz.org pmbio.org

Kelsy Cotto
Computational Biologist

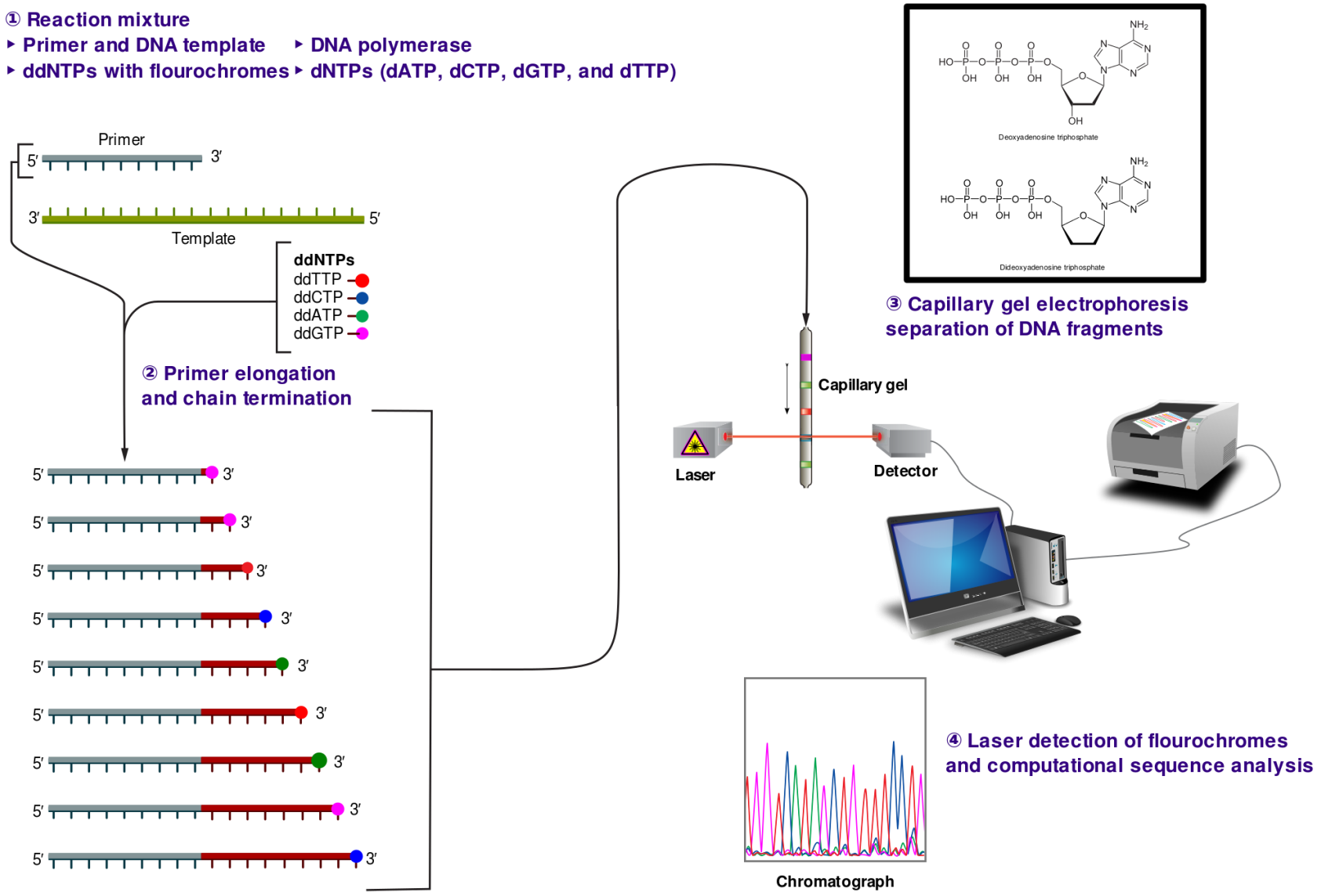
Evelyn Schmidt
Graduate Student

Zach Skidmore
Senior Quantitative Scientist

Kartik Singhal
Graduate Student

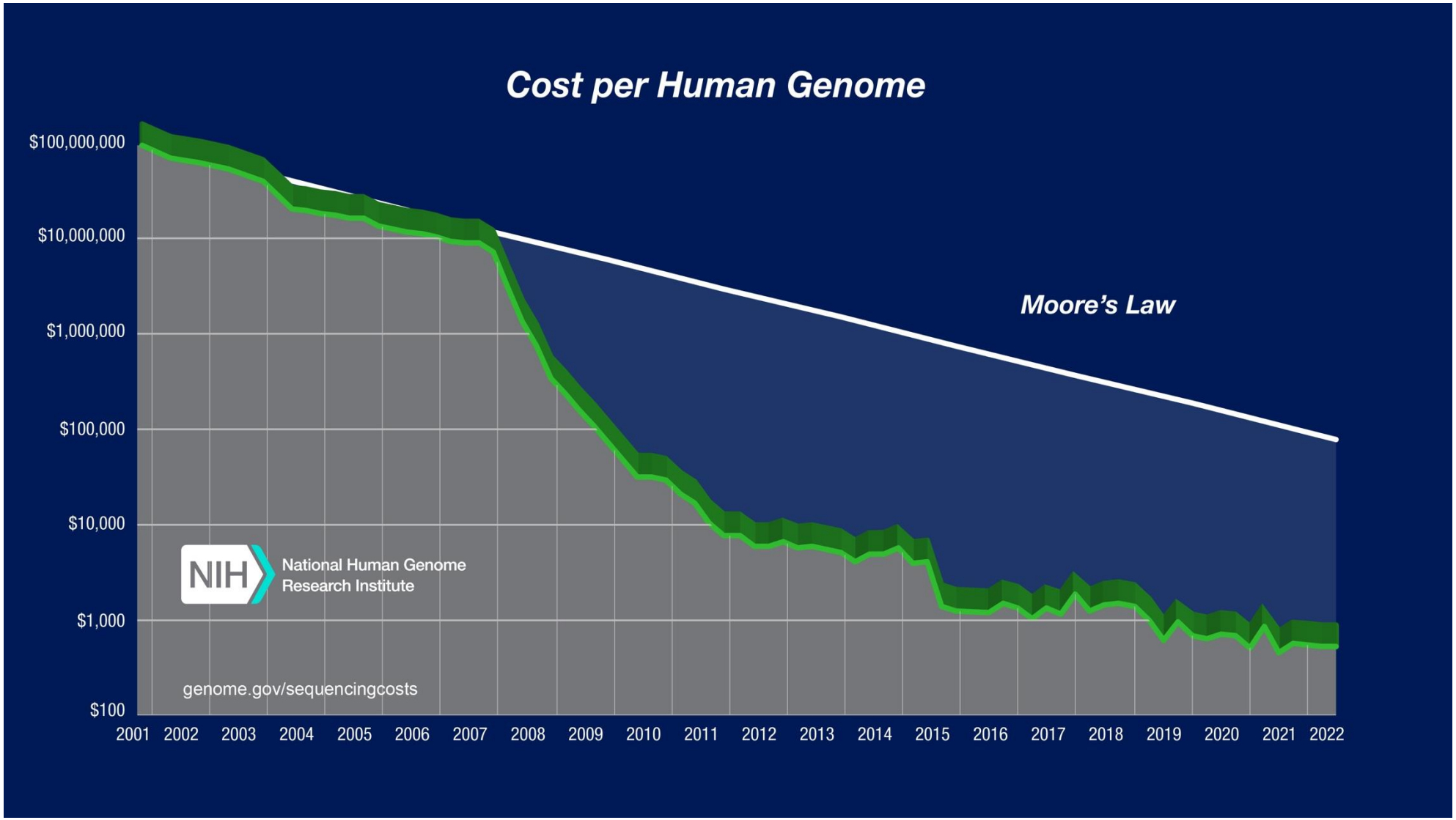
Washington University in St. Louis
School of Medicine

DNA/RNA analysis began with (Sanger) sequencing



DNA/RNA analysis began with (Sanger) sequencing

Costs of sequencing have plummeted from ~$100M to <$1000 per human genome

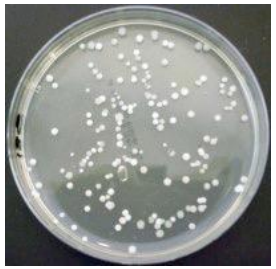


What changed? Level of parallelism

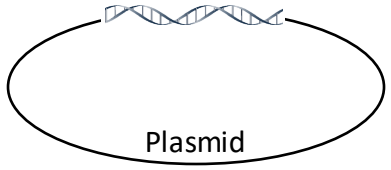
First Generation:

Fragment DNA

Clone DNA

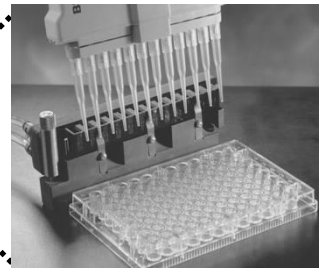


bacterial colonies



isolate 96 clones
add to sequencer

ABI 3700



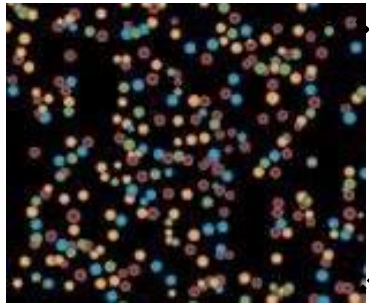
96 well plate

Second/next Generation:

Fragment DNA

Add adaptors

Add sample directly
to sequencer



3 billion “clusters” in a 4 x 1 inch flow-cell

Illumina HiSeq X

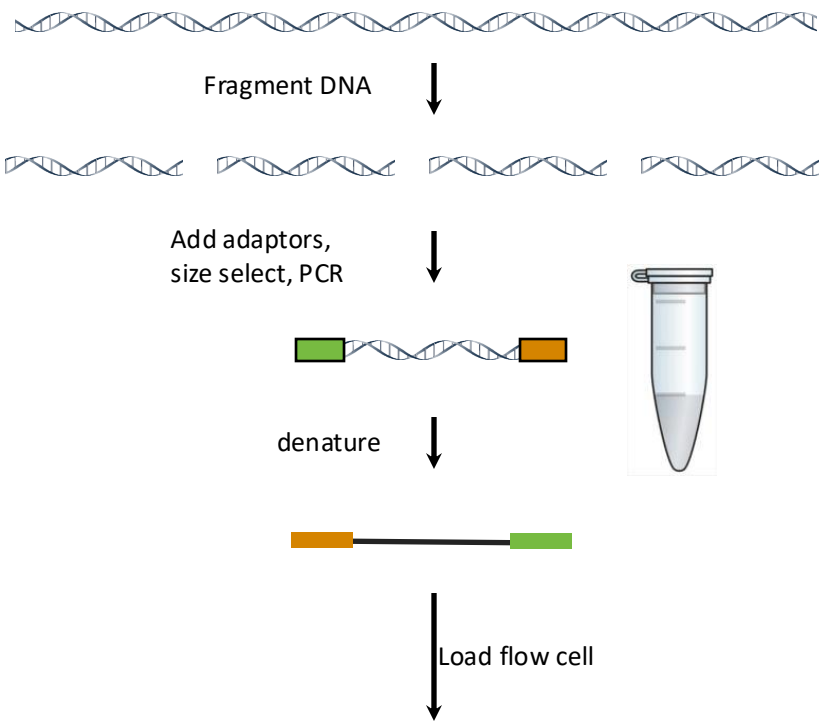
MPS (NGS) Platforms: Illumina is currently dominant




	NextSeq Series	HiSeq 4000 System	HiSeq X Series‡	NovaSeq 6000 System
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)	●	●	●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●		●
Exome Sequencing	●	●		●
Targeted Gene Sequencing (amplicon, gene panel)	●	●		●
Whole-Transcriptome Sequencing	●	●		●
Gene Expression Profiling with mRNA-Seq	●	●		●
miRNA & Small RNA Analysis	●	●		●
DNA-Protein Interaction Analysis	●	●		●
Methylation Sequencing	●	●		●
Shotgun Metagenomics	●	●		●
Run Time	12–30 hours	< 1–3.5 days	< 3 days	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	1500 Gb	1800 Gb	6000 Gb
Maximum Reads Per Run	400 million	5 billion	6 billion	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 x 250**

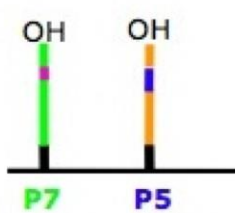
- Higher accuracy, range of capacity and throughput
- Longer read lengths on some platforms

Illumina sequencing trick #1: immobilizing single molecules on a solid surface

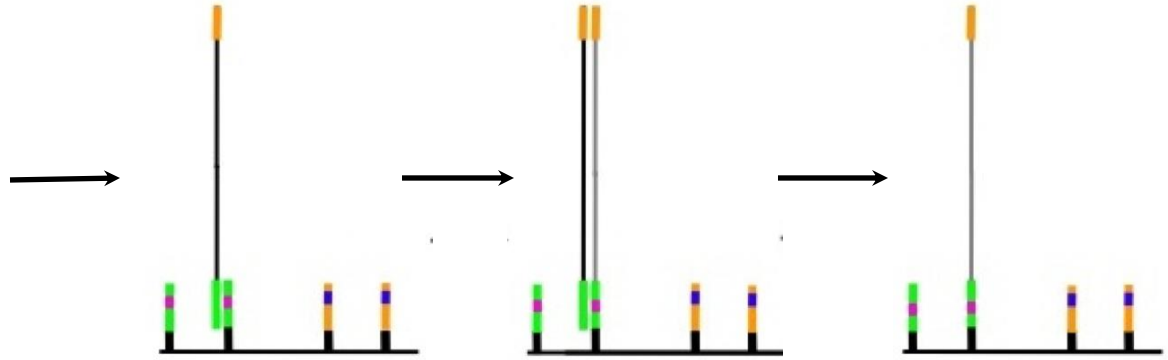


1) Prepare Sequencing Library

2) Seed Flow Cell with Single Molecules



Flow cell surface w/
attached primers

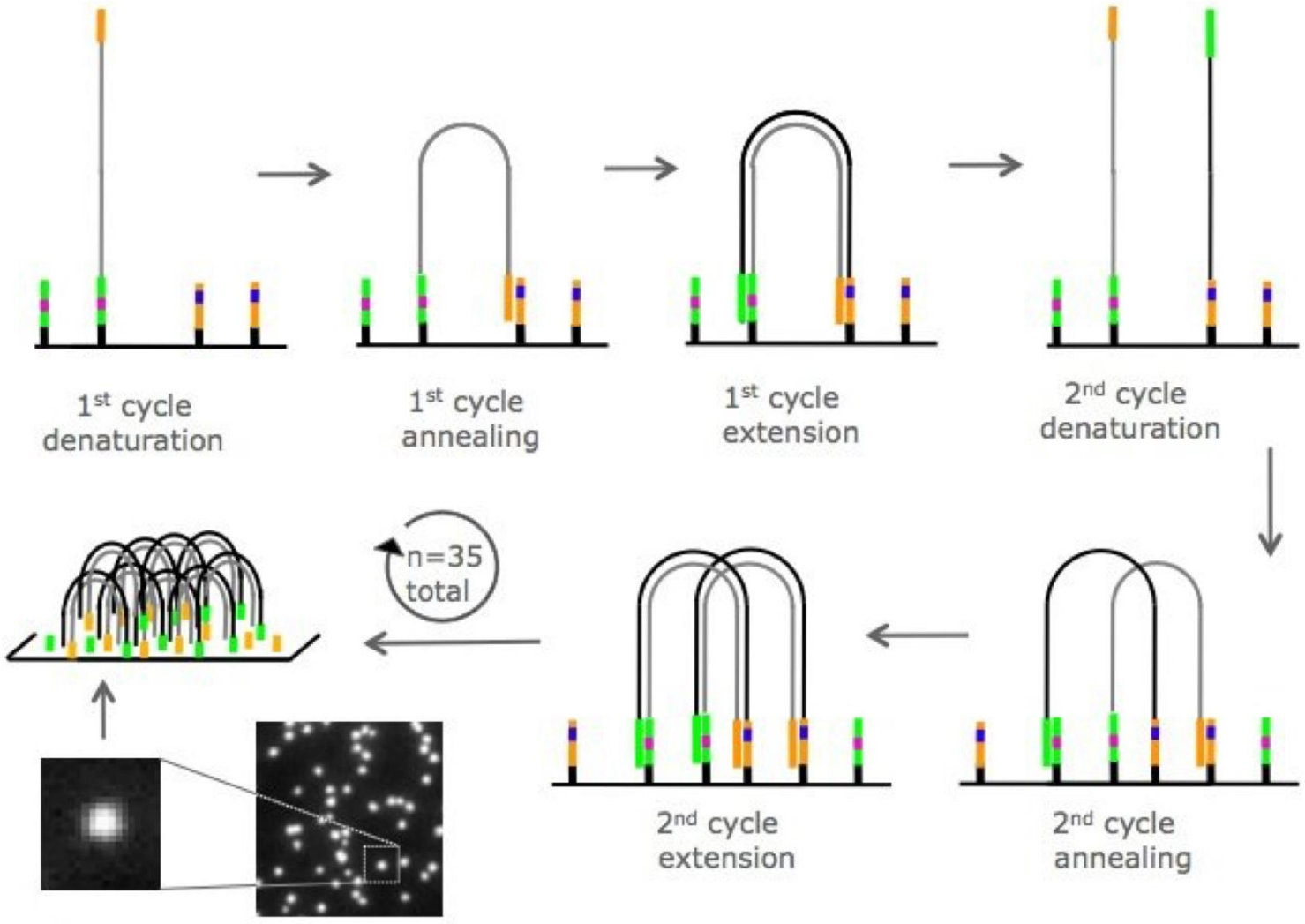


Template
Hybridization

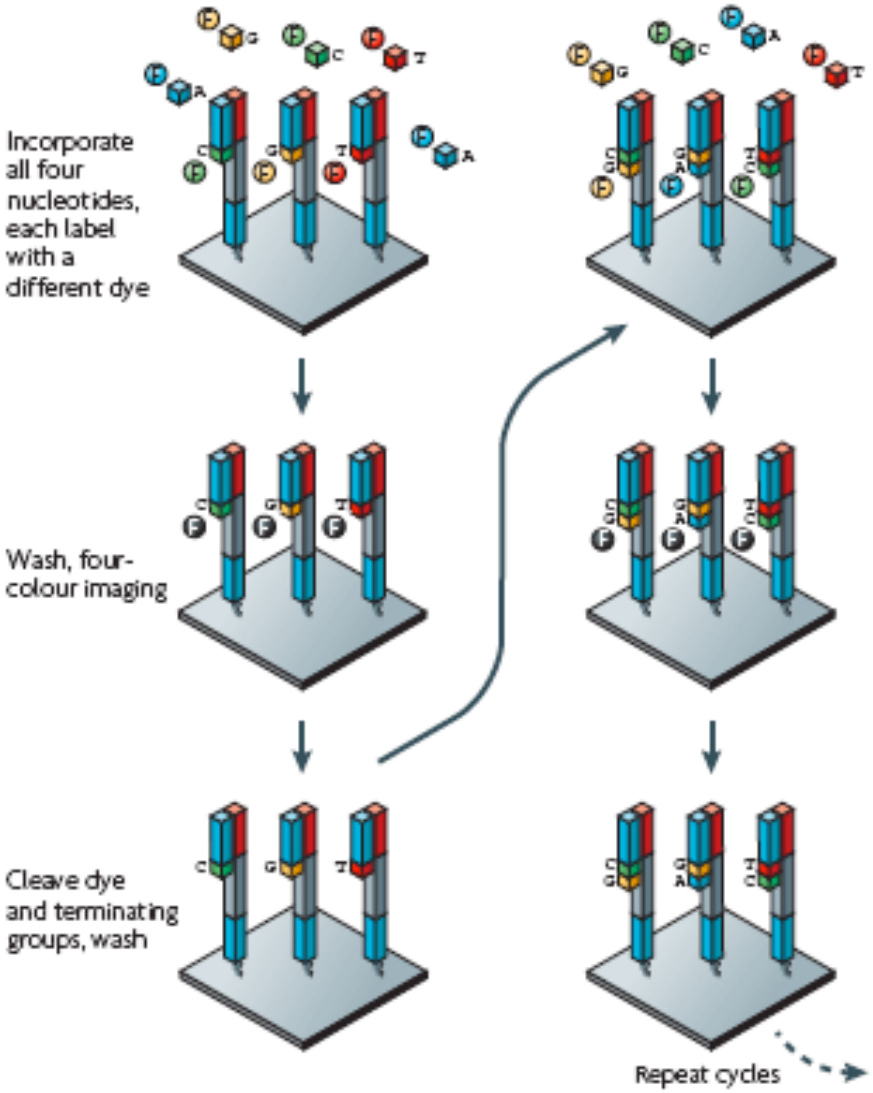
Initial Extension

Denaturation

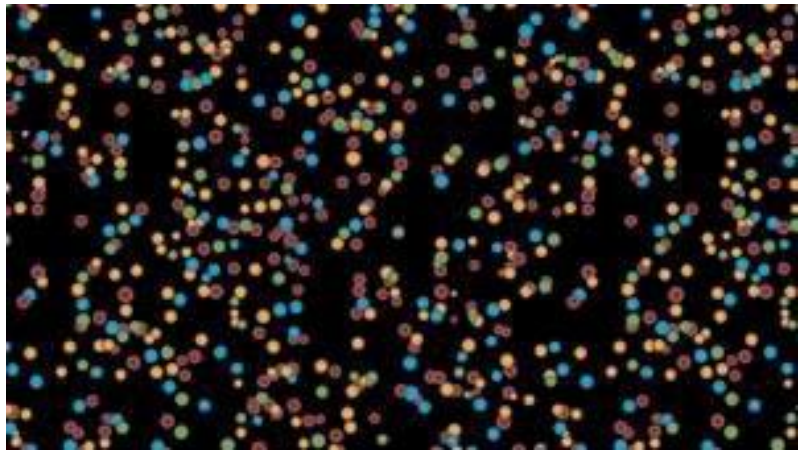
Illumina sequencing trick #2: cluster amplification by solid-phase “bridge” PCR



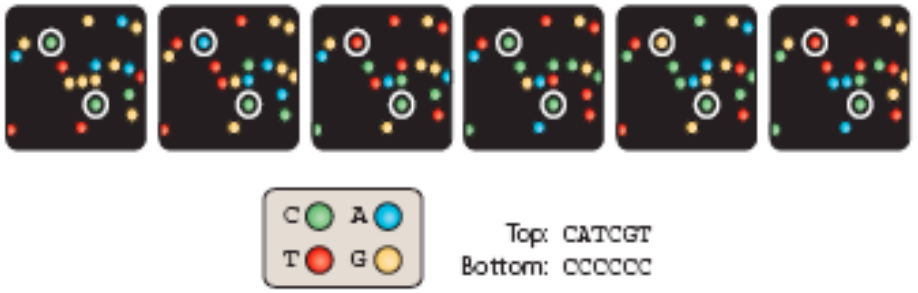
Illumina sequencing trick #3: base-by-base DNA synthesis with 4 dyes & reversible terminators



4 images merged



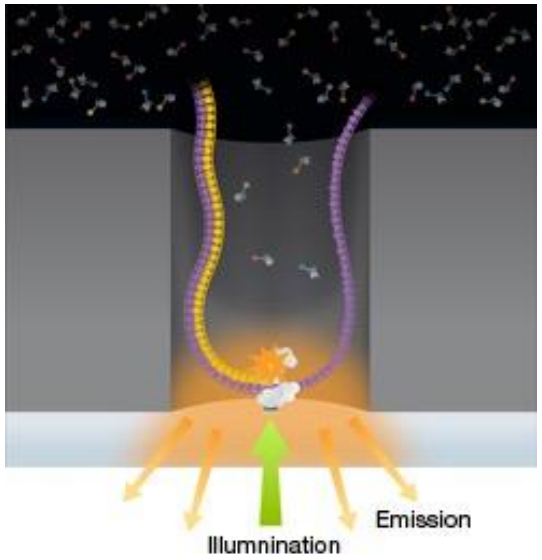
6 cycles w/ base-calling



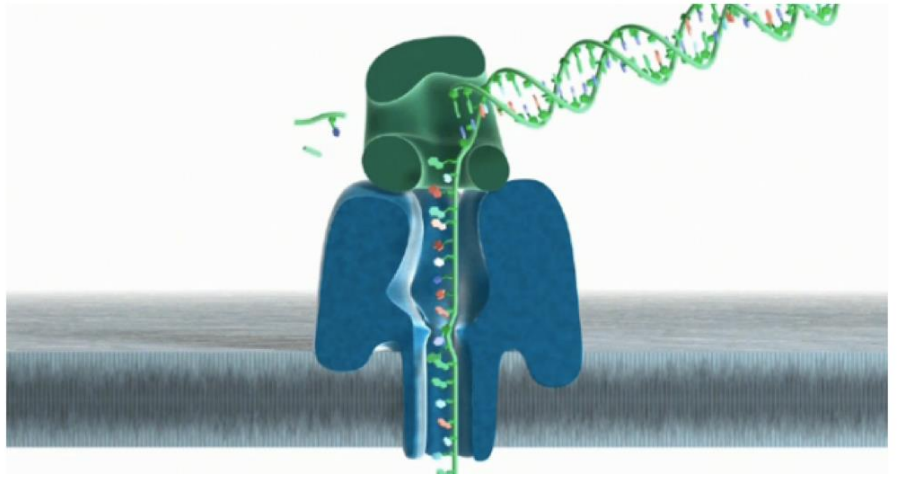
Next-next (3rd) generation sequencing platforms

Defining Characteristics: Long reads (10-100 kb) from single molecules.

Pacific Biosciences: watching a polymerase synthesize DNA in real time



Oxford Nanopore: Translocating DNA through a nanopore with electrode-based detection



The promise: Long reads will allow us to accurately sequence and assemble whole human genomes, from scratch, without using the reference genome.

Status: Currently limited by low throughput, high error rate (10-15%) and high cost (>$50,000 per human genome). 3rd generation technologies have proven useful, but generally for niche applications so far.

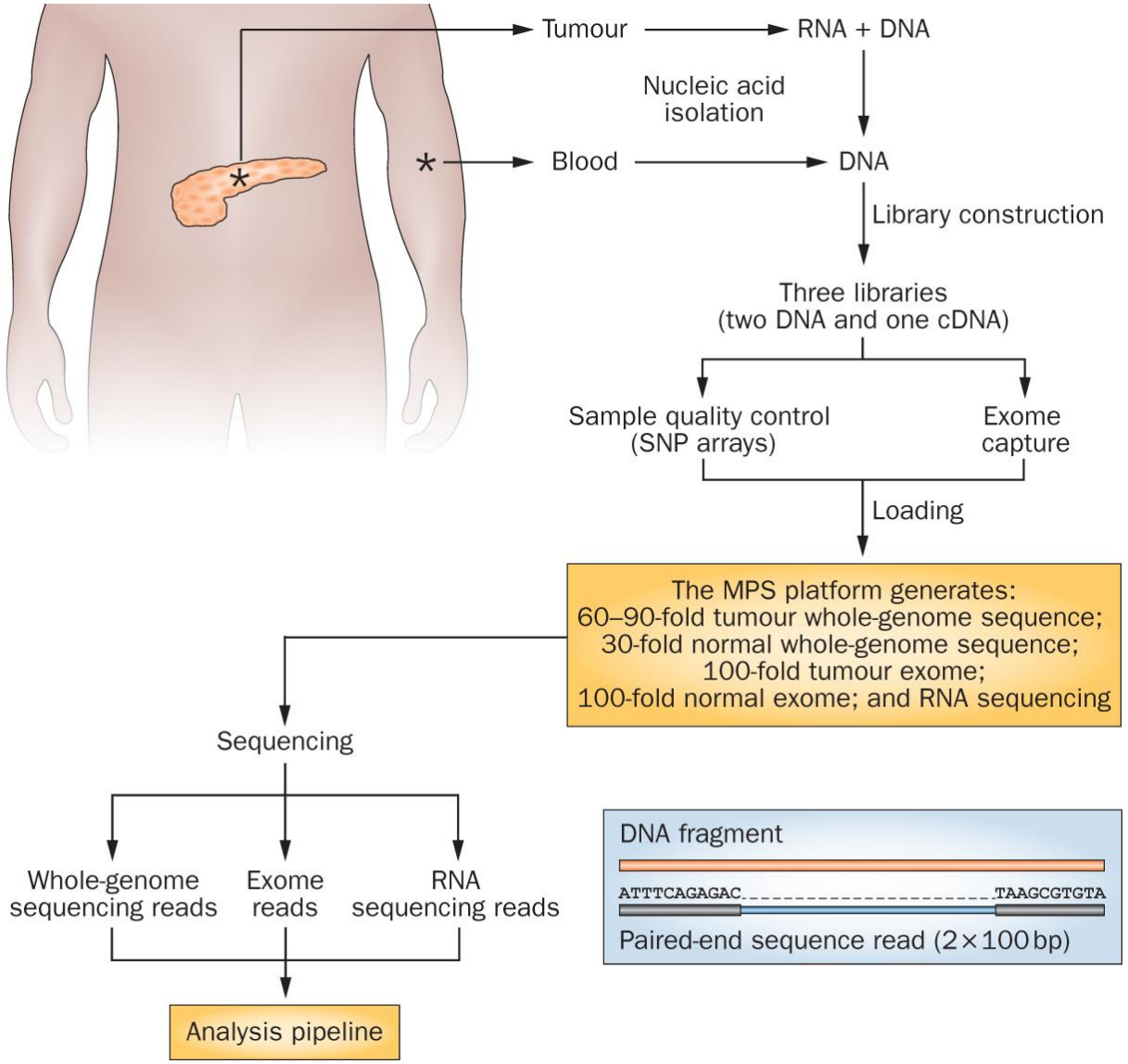
Costs of sequencing have plummeted from ~$100M to <$1000 per human genome

Human Exome DNA-seq full service, cost per sample[b]	< 12 Samples	≥ 12 Samples	≥ 1000 Samples*	≥ 2500 Samples*	~Targeted Coverage
IDT Exome, germline	$316	$165	$153	$140	50X
IDT Exome, germline	$328	$177	$165	$152	75X
IDT Exome	$340	$189	$177	$164	100X
IDT Exome, somatic	$364	$226	$214	$201	150X
IDT Exome, somatic	$388	$250	$238	$225	200X

Human Whole Genome DNA-seq full service, cost per sample[c]	Cost Per Genome	≥ 500 Samples*	≥ 1000 Samples*	≥ 1500 Samples*	~Targeted Coverage
WGS - PCRfree	$443	$432	$394	$393	25X
WGS - PCRfree	$491	$487	$449	$430	30X
WGS - PCRfree	$875	$818	$780	$722	60X

PacBio Sequencing (Revio), Full Service Whole Genome Sequencing	Cost Per Genome	~Human Coverage
Revio Full Service WGS PacBio Sequencing with HMW DNA Isolation and Enhanced Size Selection (1 SMRT Cell Coverage)	$2,990	30X
Revio Full Service WGS PacBio Sequencing without HMW DNA Isolation and Enhanced Size Selection (1 SMRT Cell Coverage)	$2,780	30X
Revio Full Service WGS PacBio Sequencing with HMW DNA Isolation and Enhanced Size Selection (2 SMRT Cell Coverage)	$5,200	60X
Revio Full Service WGS PacBio Sequencing without HMW DNA Isolation and Enhanced Size Selection (2 SMRT Cell Coverage)	$4,920	60X

Cancer genomics research has exploded with the rapid advances in DNA sequencing technologies

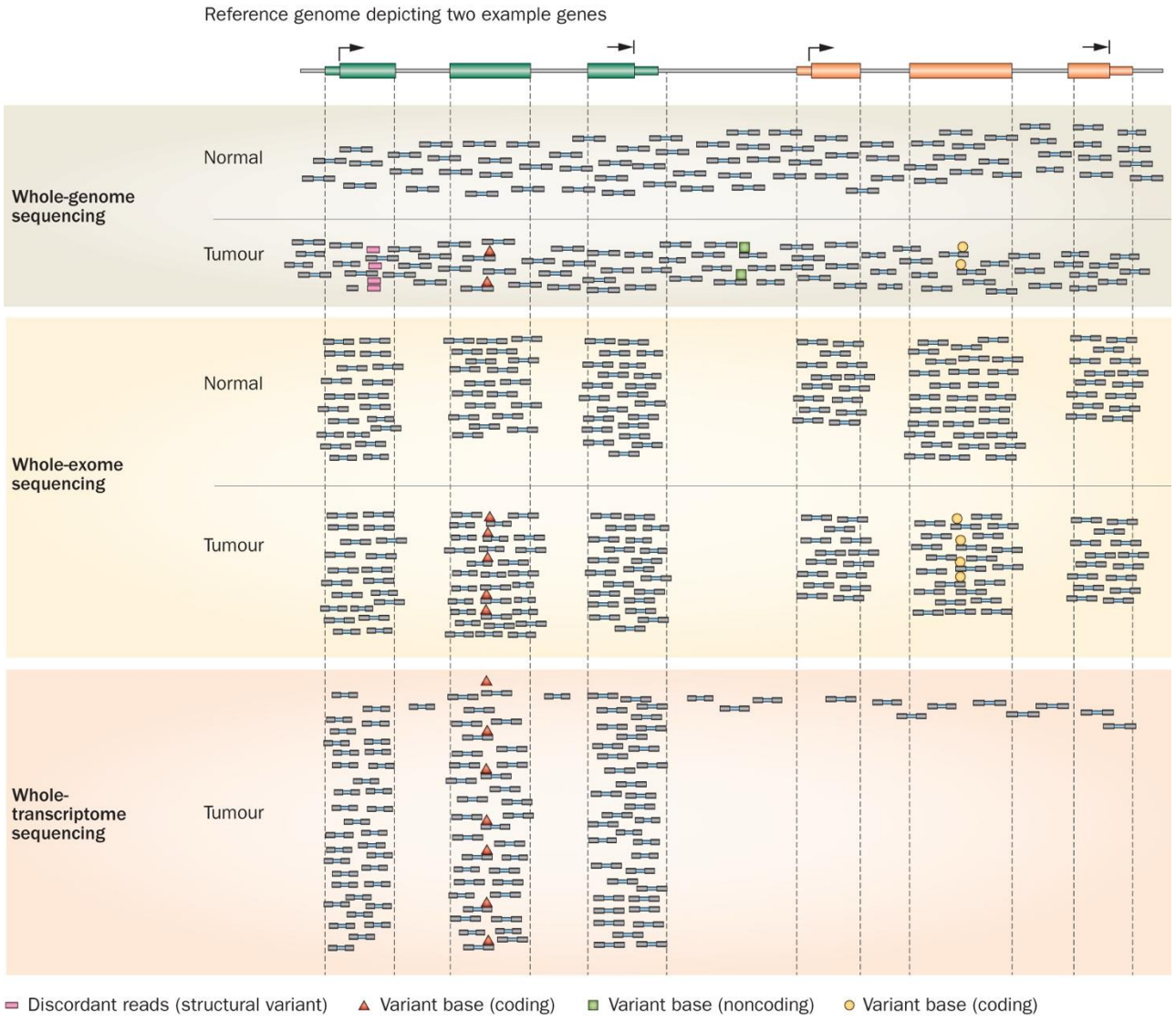


Whole genome, exome and transcriptome allow us to detect and confirm many molecular alteration types

‘WGS’

‘Exome’

‘RNA-seq’



Many other omic approaches exist

Epigenome	Epigenomics	Epigenetic modifications	Molecular genetics	Epigenomics is the study of the complete set of epigenetic modifications on the genetic material of a cell, collectively known as the epigenome
Exposome (2005)	Exposomics	An individual's environmental exposures, including in the prenatal environment	Molecular genetics	A proposed term and field of study of the disease-causing effects of environmental factors (the "nurture" component of "nature vs. nurture").[5]
Exposome (2009)		Composite occupational exposures and occupational health problems	Occupational safety and health	The proposers of this term were aware of the previous term as used above but proposed to apply the term to a new field.[6][7]
Exome	Exomics	Exons in a genome	Molecular Genetics	
Foodome	Foodomics	Food and Nutrition issues related to bioactivity, quality, safety and traceability of foods through the application and integration of advanced omics technologies to improve consumer's well-being, health, and confidence.	Nutrition	The term was first defined in 2009[8]
Genome	Genomics (Classical genetics)	Genes (DNA sequences/Chromosomes)	Genetics	"Genome" refers to the set of all genes in an organism. However, "genome" was coined decades before it was discovered that most DNA is "non-coding" and not part of a gene; thus, "genome" originally referred to the entire collection of DNA within an organism. Today, both definitions are used, depending on the context.[9]

More than 50 “omics” topics in biology defined:

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

Examples of other bulk sequence approaches that are utilized in Cancer Omics:

- Low pass WGS
- CHiP-Seq
- ATAC-seq
- Hi-C
- RiboSeq
- …

How does it work? Short read alignments are the currency of DNA/RNA analysis

- 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 damage or variation in picture

Reference: AGCCTGAGACCGTAAAAAAGTCAAG

A read sequence: GAGACCGTAAAAACGTC

A variant

Alignment is central to most genomics research

FASTQ format

Align to genome

Quality control

Find peaks

Quantify transcripts

Find SVs

.
.
.

etc.

Analysis, Interpretation, Discovery

Phase 0

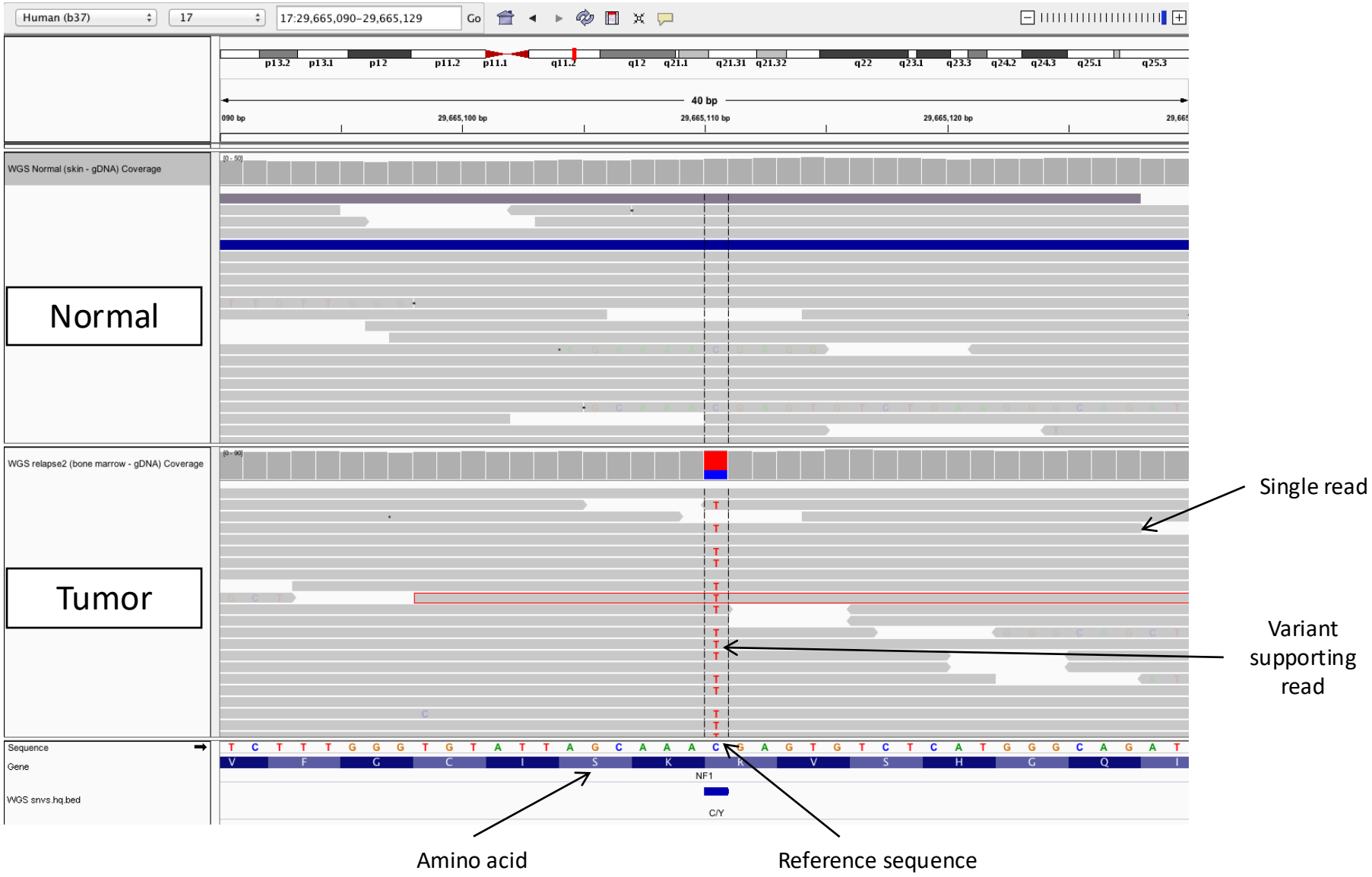
Phase1

Phase 2-100

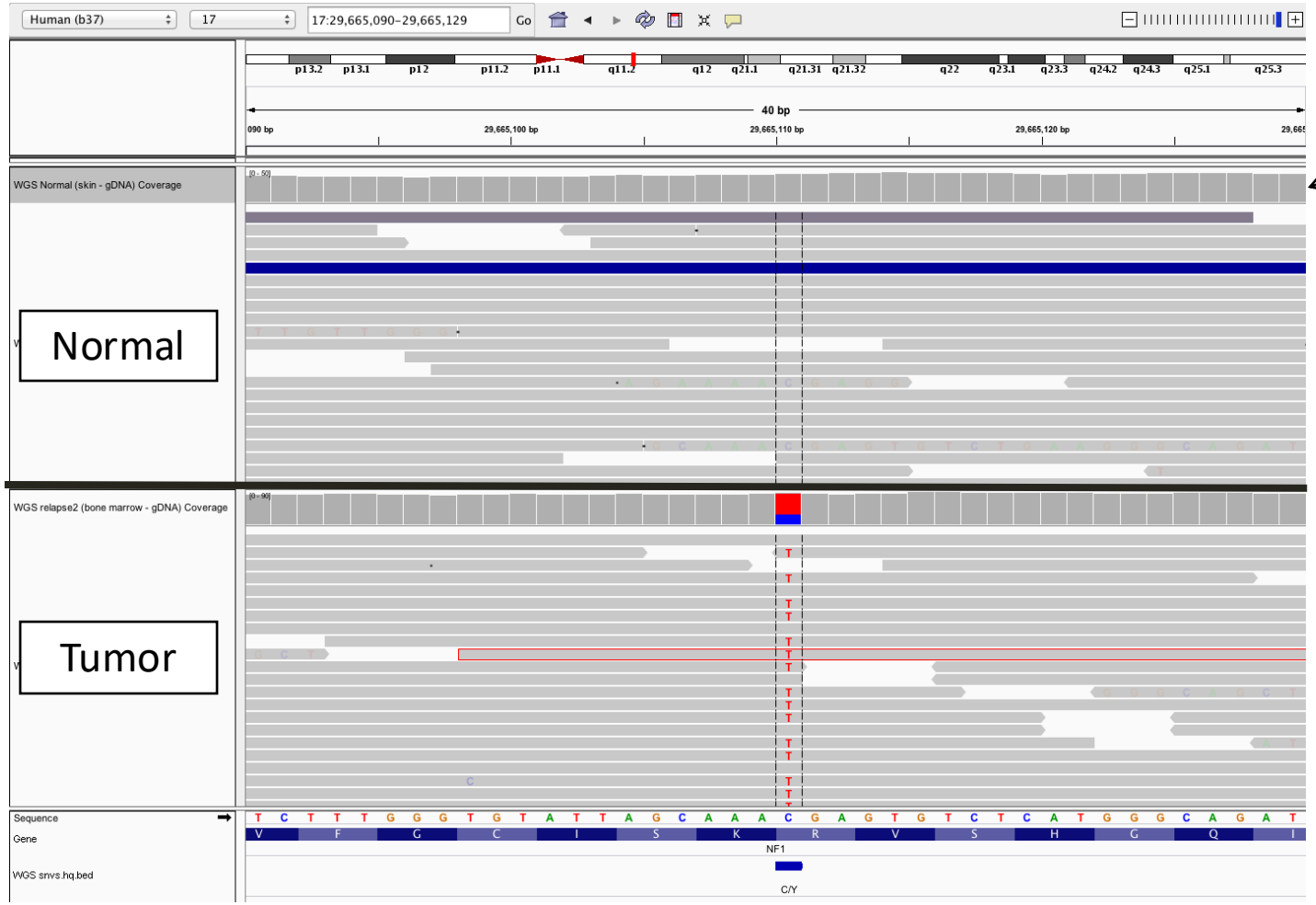
Choice of reference genome files impact analysis

- Many nuances to consider:
  - Naming convention of chromosomes (e.g. “1” vs “chr1”)
  - Build/assembly versions (GRCh37 vs GRCh38)
  - Alternative contigs and “patches”
    - Alternative assemblies
    - HLA genes
  - Decoy sequences
  - Experiment specific additions
    - Cancer associated viruses
    - Artificial constructs, genetic engineering
  - New genome assemblies (T2T, graph genomes)

Single nucleotide variants (SNVs) and insertions/deletions (indels) appear as short alignment discrepancies from reference genome



Variant allele fraction (VAF) and coverage



VAF = Variant reads / Total reads

Coverage track

VAF = 0/20 = 0%

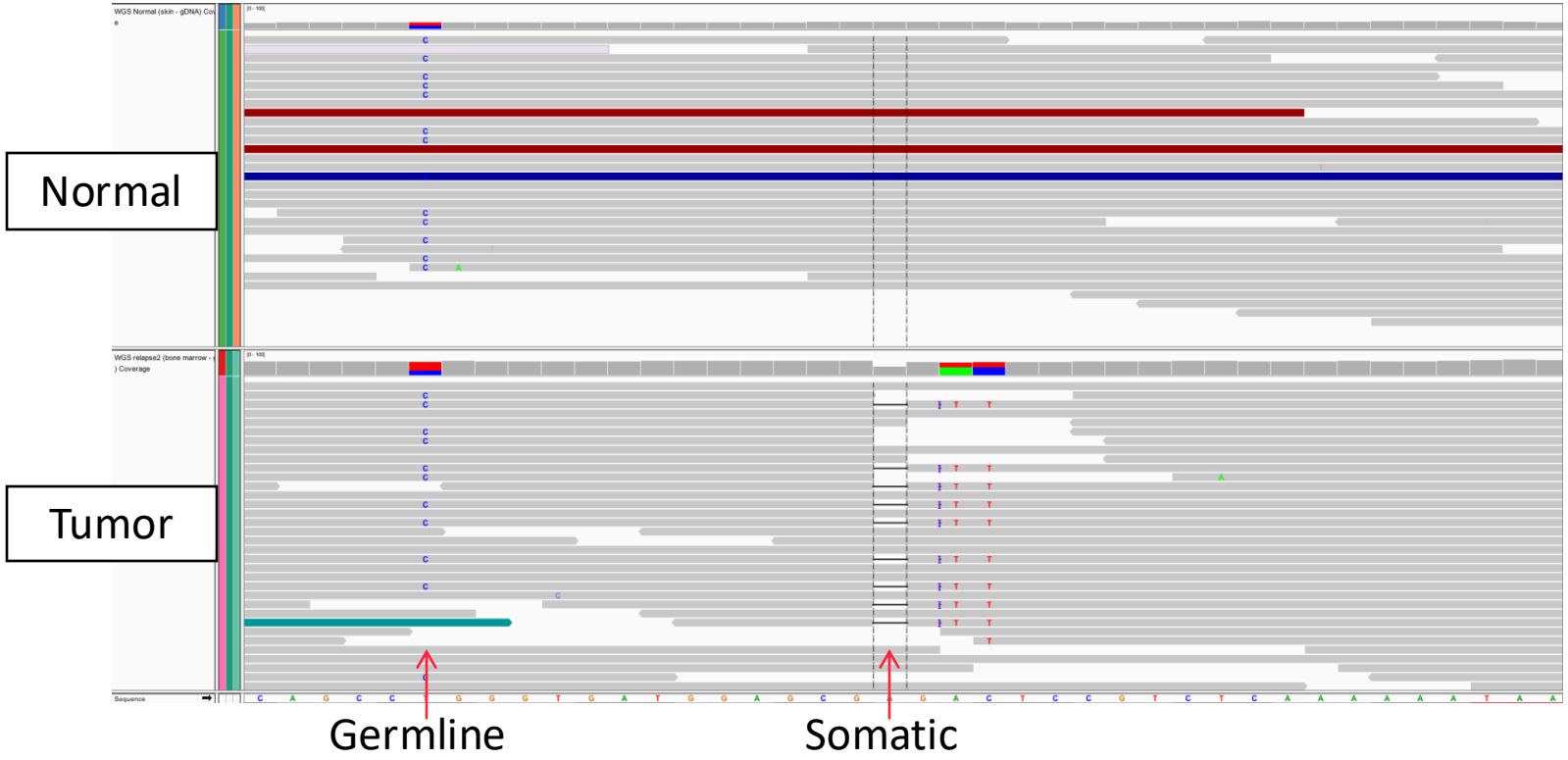
VAF = 14/20 = 70%

Total coverage=20X

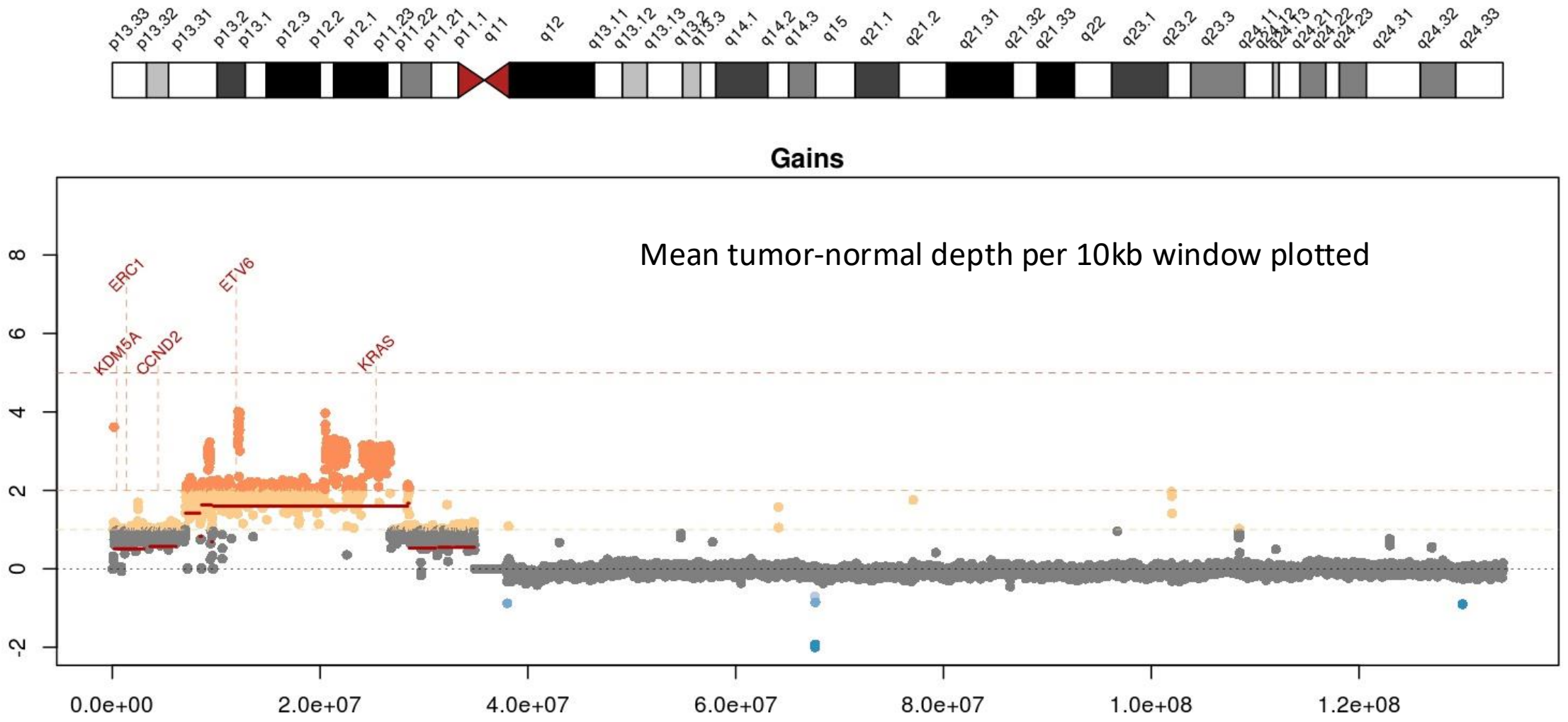
Recall that we have two copies of our genome (2n). Mutations are typically in one copy. A heterozygous variant (one copy) is expected to have VAF = 50%. Often not true due to sample purity, tumor heterogeneity, sampling error, alignment issues, copy number variation, etc.

Both somatic and germline variants are important

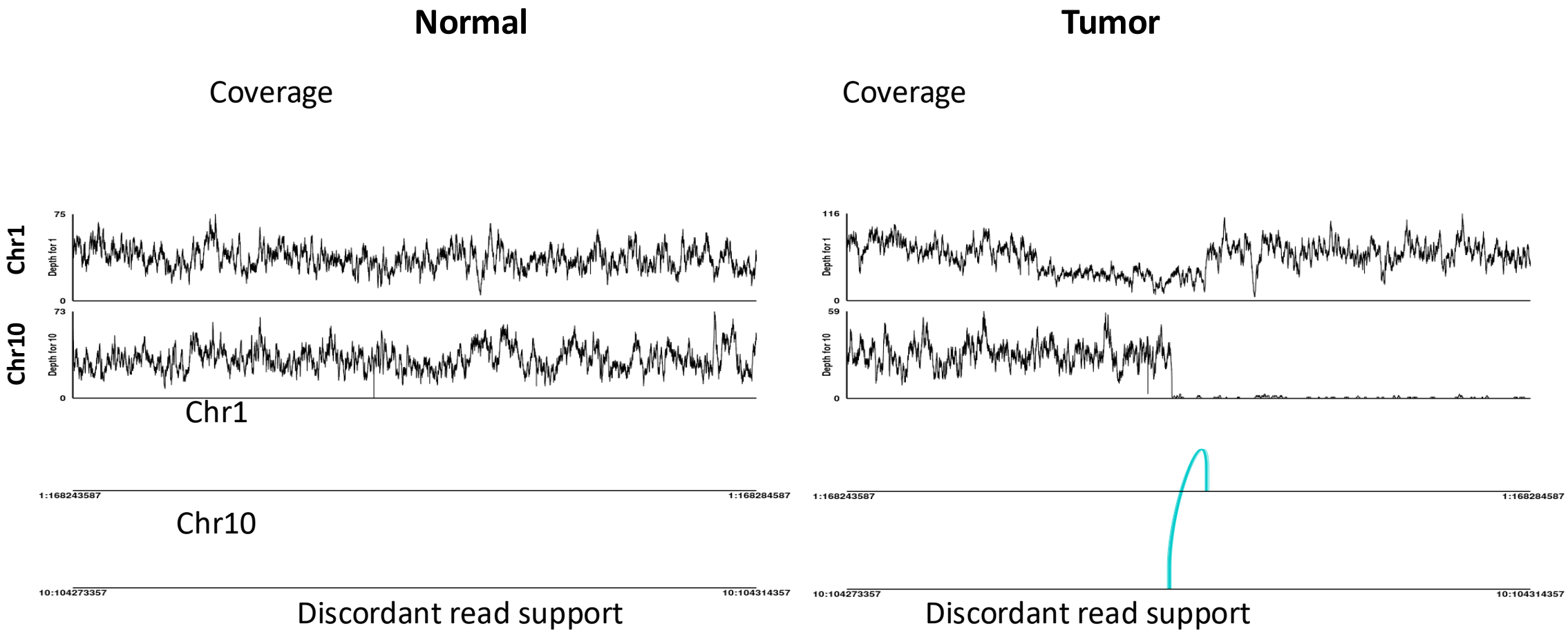
- Germline variants
  - Present in egg or sperm
    - All cells of affected offspring
  - Heritable
  - Cause of familial cancers
- Somatic variants
  - Occur in non-germline tissues
    - Only tumor cells
  - Non-heritable
  - Cause of sporadic cancers



Copy number variants (CNVs) appear as deviations from in alignment “depth” or “coverage”

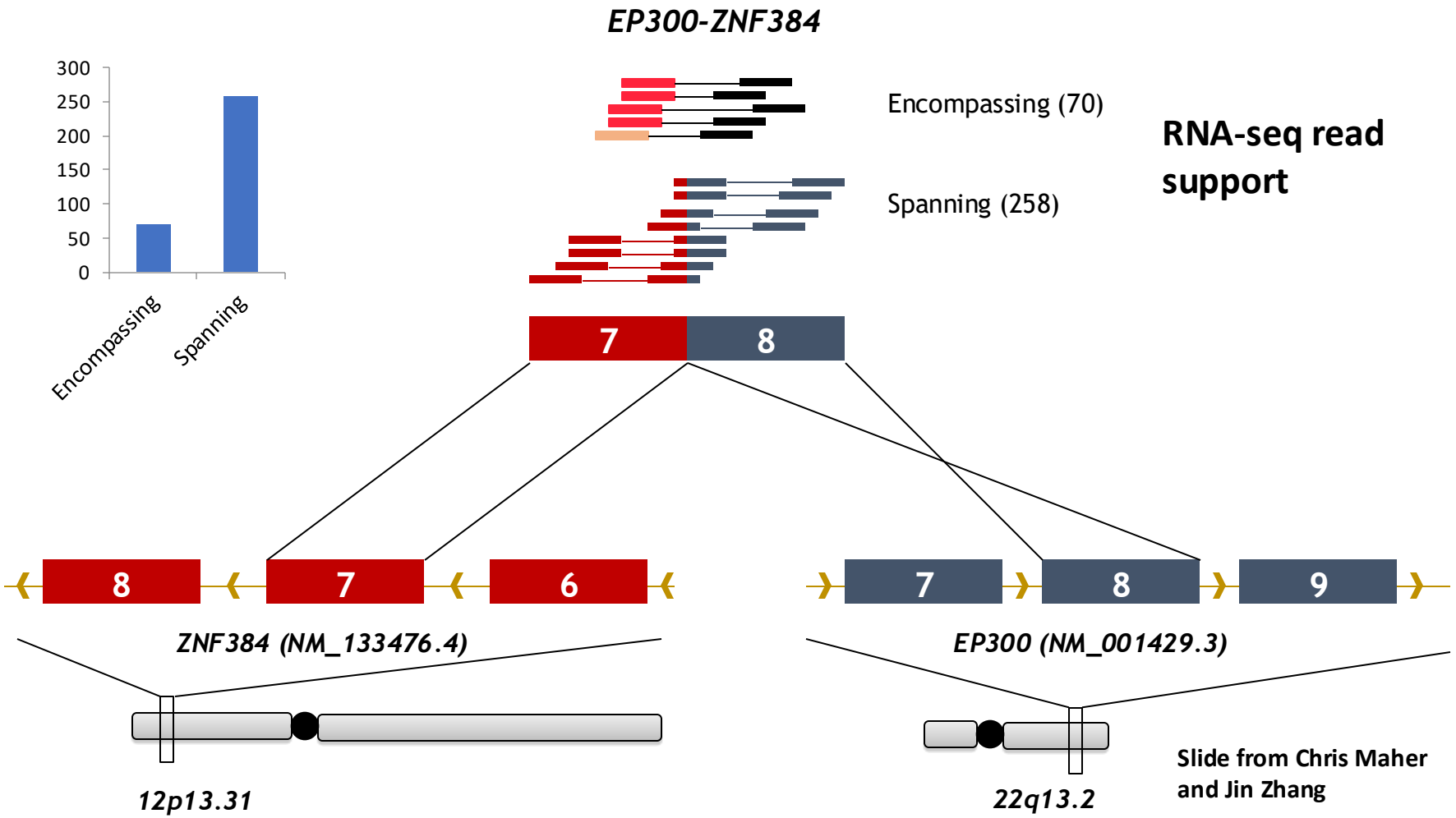


Structural variants (SVs) can be identified using a combination of coverage and discordant read alignments



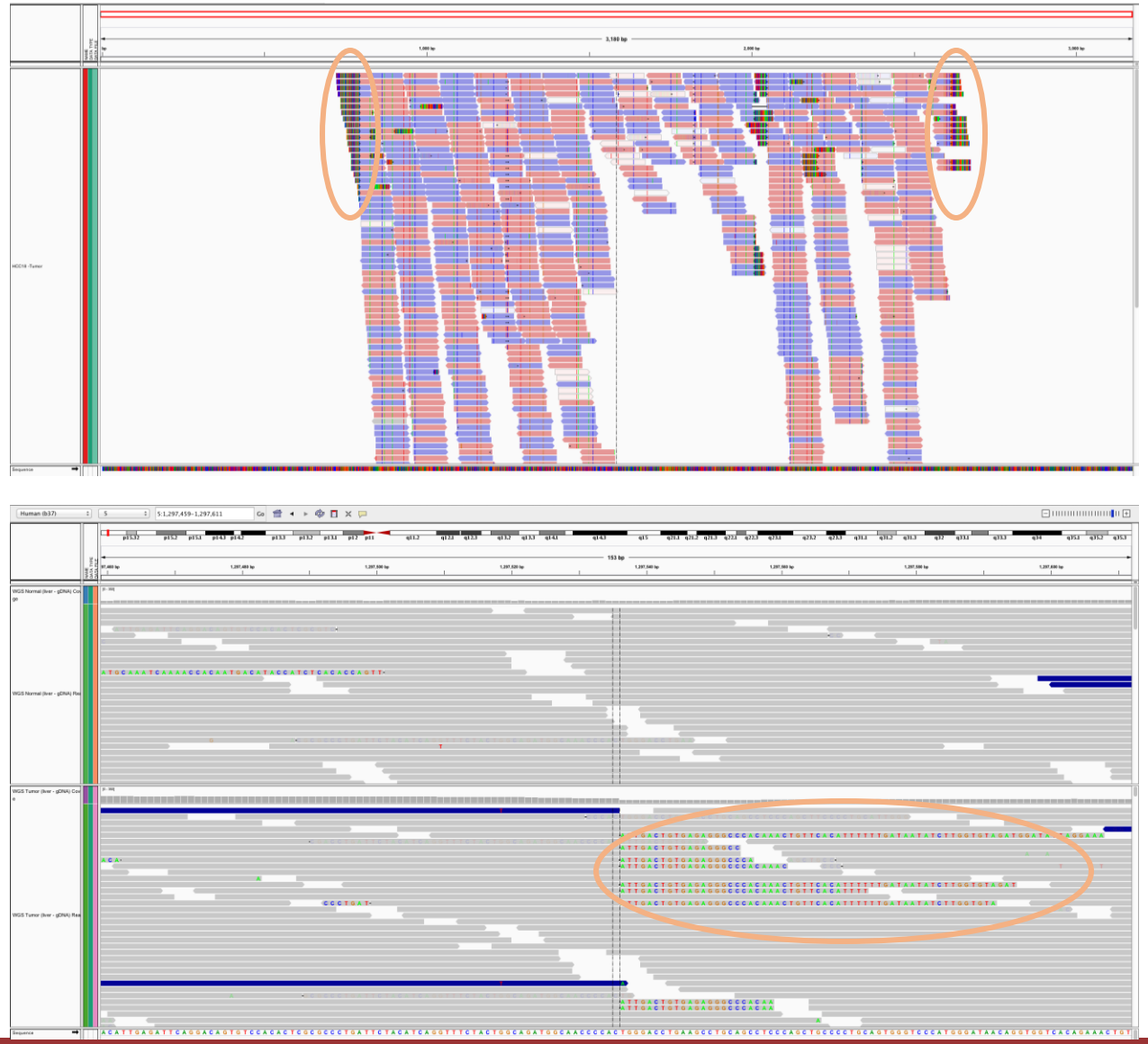
A Chr1-Chr10 (TBX19-SUFU) unbalanced translocation identified in an adult acute lymphocytic leukemia.

Expressed gene fusions can be identified by discordant read alignments spanning known exons from RNA-seq data



Exons 1-8 of EP300 fused to exons 7-10 of ZNF384 in head-to-tail fashion.

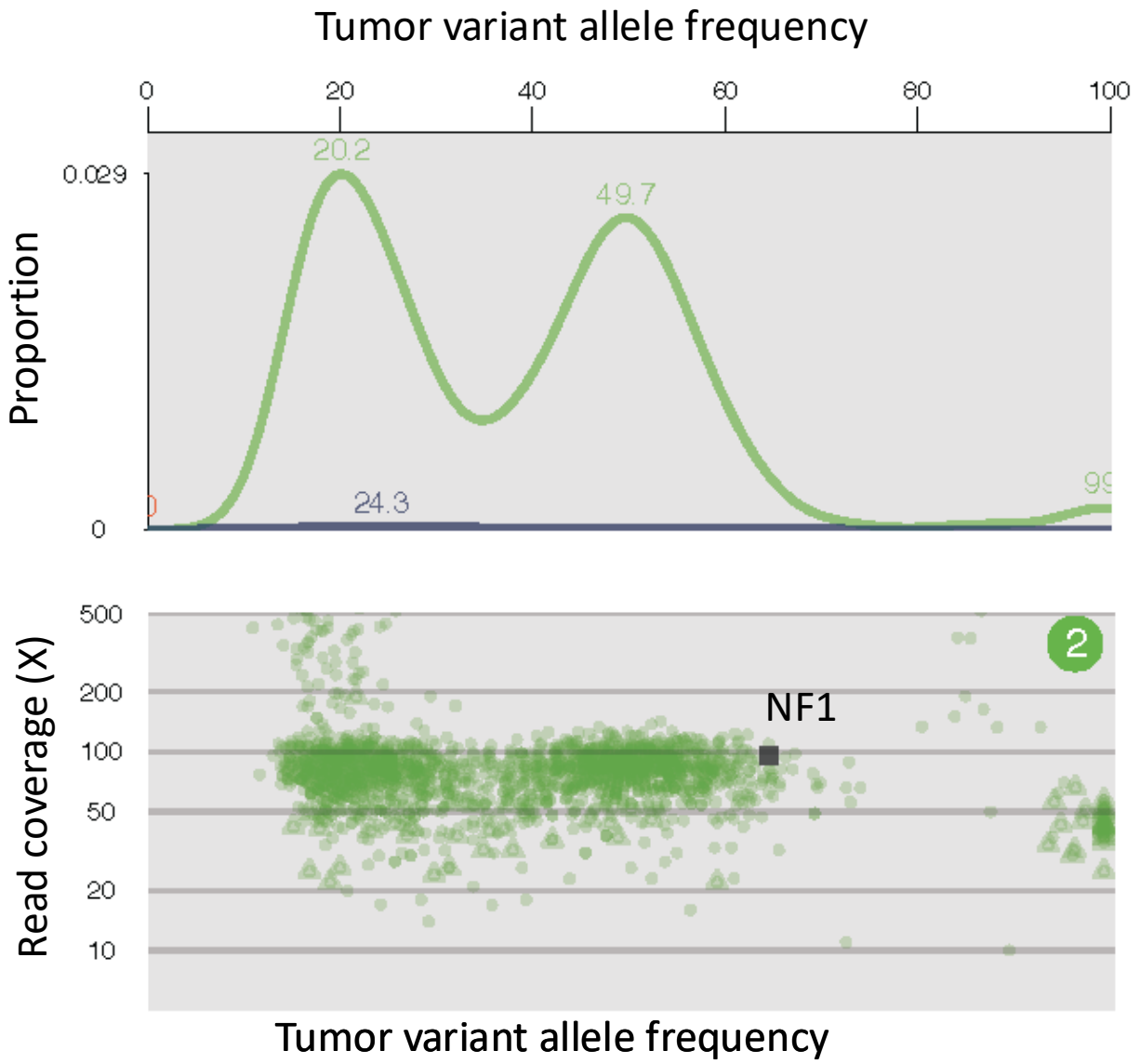
Viral expression and integrations can be identified by competitive alignment with human and viral genome databases

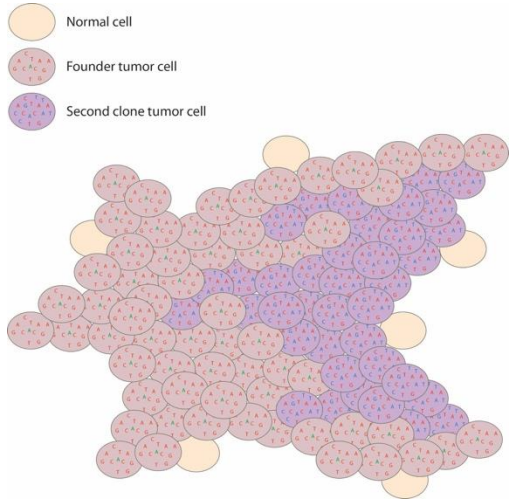


- Many reads align directly to viral genome (Hepatitis B)
- Soft-clipped reads in human reference help identify integration site

Example from HCC case thought to be virus negative

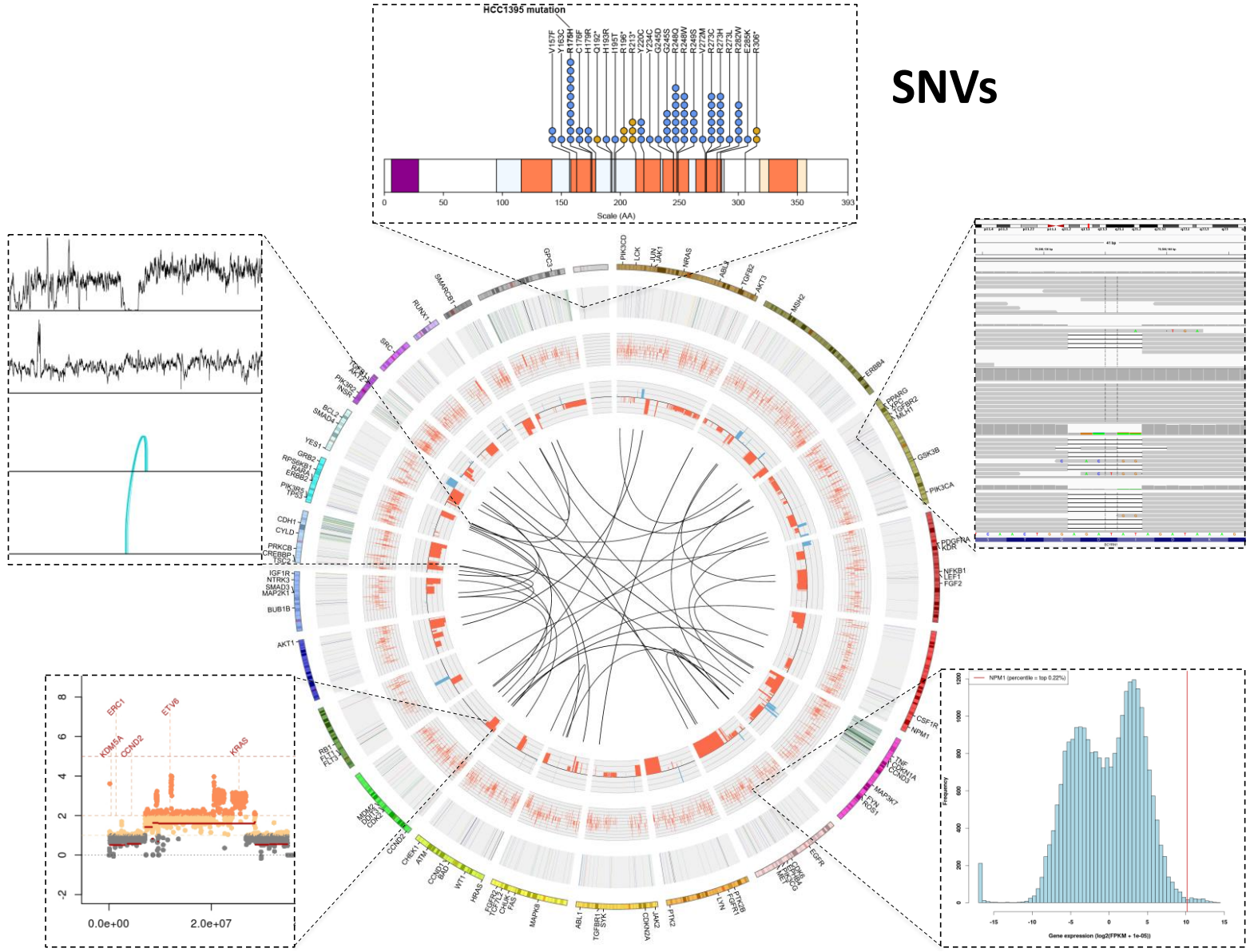
Tumor clonal architecture can be inferred from distributions of VAFs



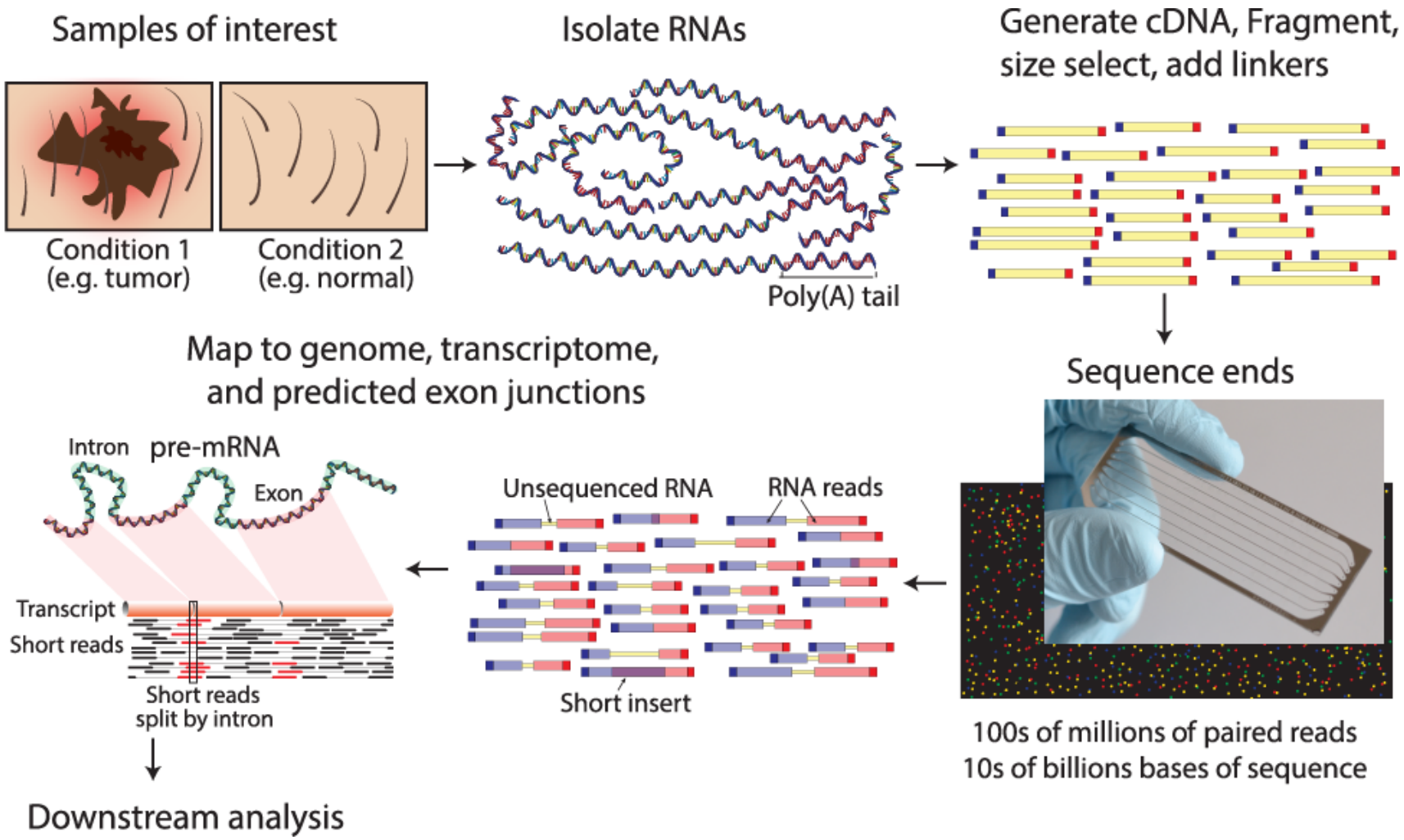


WGS data

Tumor genome analysis can reveal dozens to thousands of alterations per patient



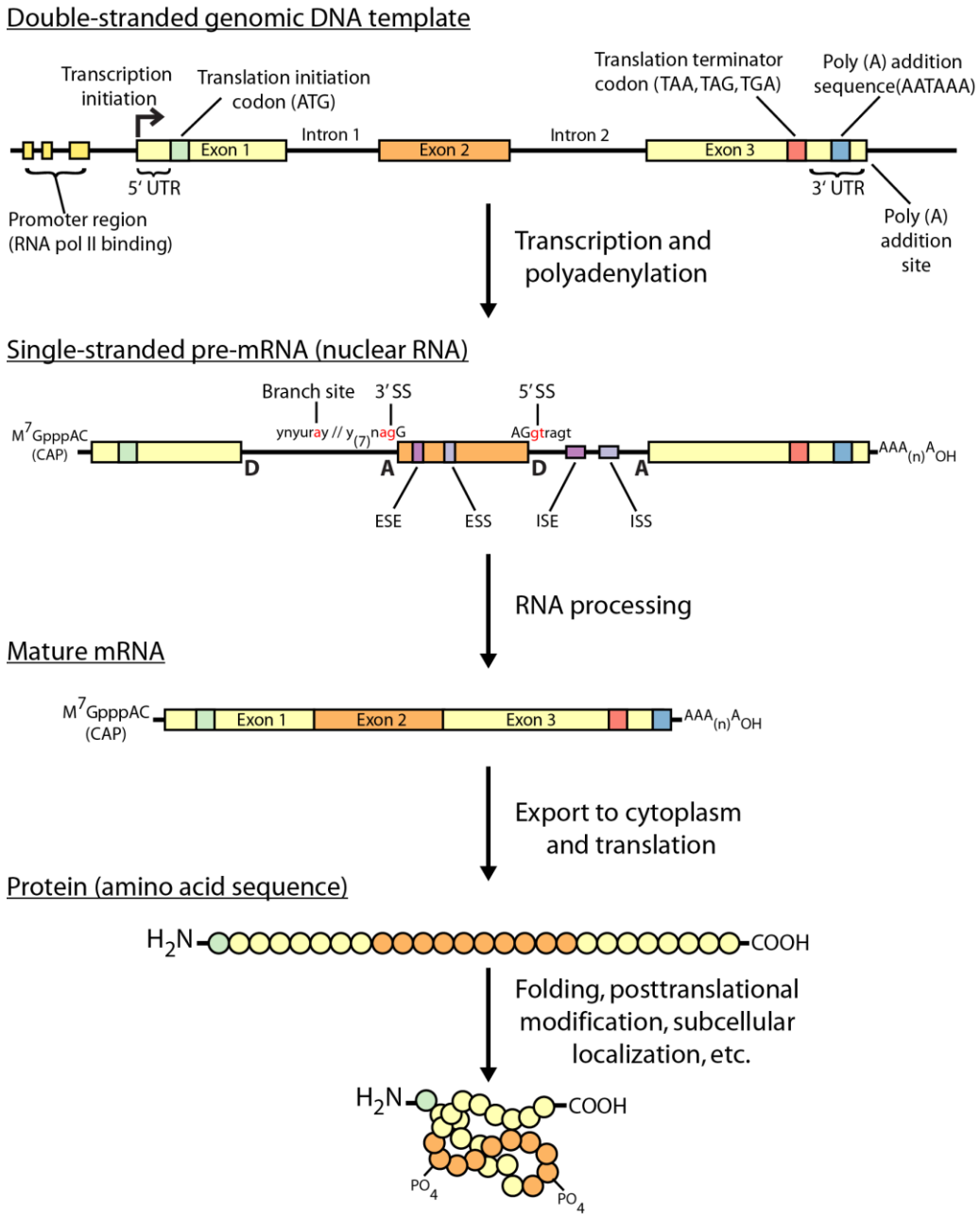
RNA sequencing



Gene expression

Thinking about the molecular biology here, what is actually being sequenced in an RNA-seq experiment?

Does it differ depending on the sequencing platform? Or for bulk vs single cell sequencing?

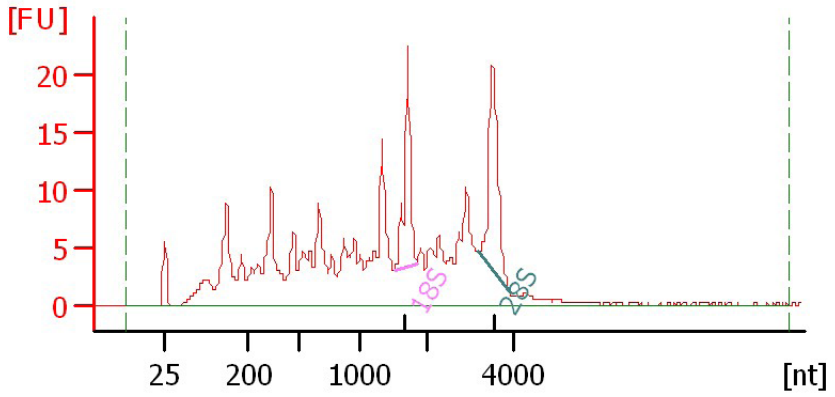


Challenges

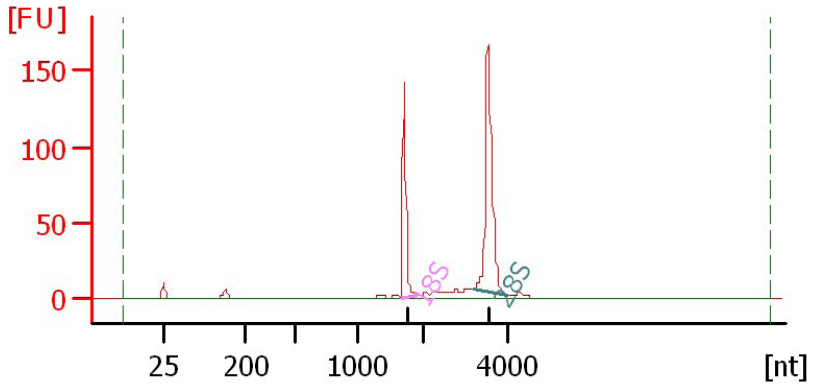
- Sample
  - Purity?, quantity?, quality?
- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
  - 10^5 – 10^7 orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction of highly expressed genes may consume the majority of reads
  - Ribosomal and mitochondrial genes
- RNAs come in a wide range of sizes
  - Small RNAs must be captured separately
  - PolyA selection of large RNAs may result in 3’ end bias
- RNA is fragile compared to DNA (easily degraded)

Agilent example / interpretation

- https://goo.gl/uC5a3C
- ‘RIN’ = RNA integrity number
  - 0 (bad) to 10 (good)



RIN = 6.0



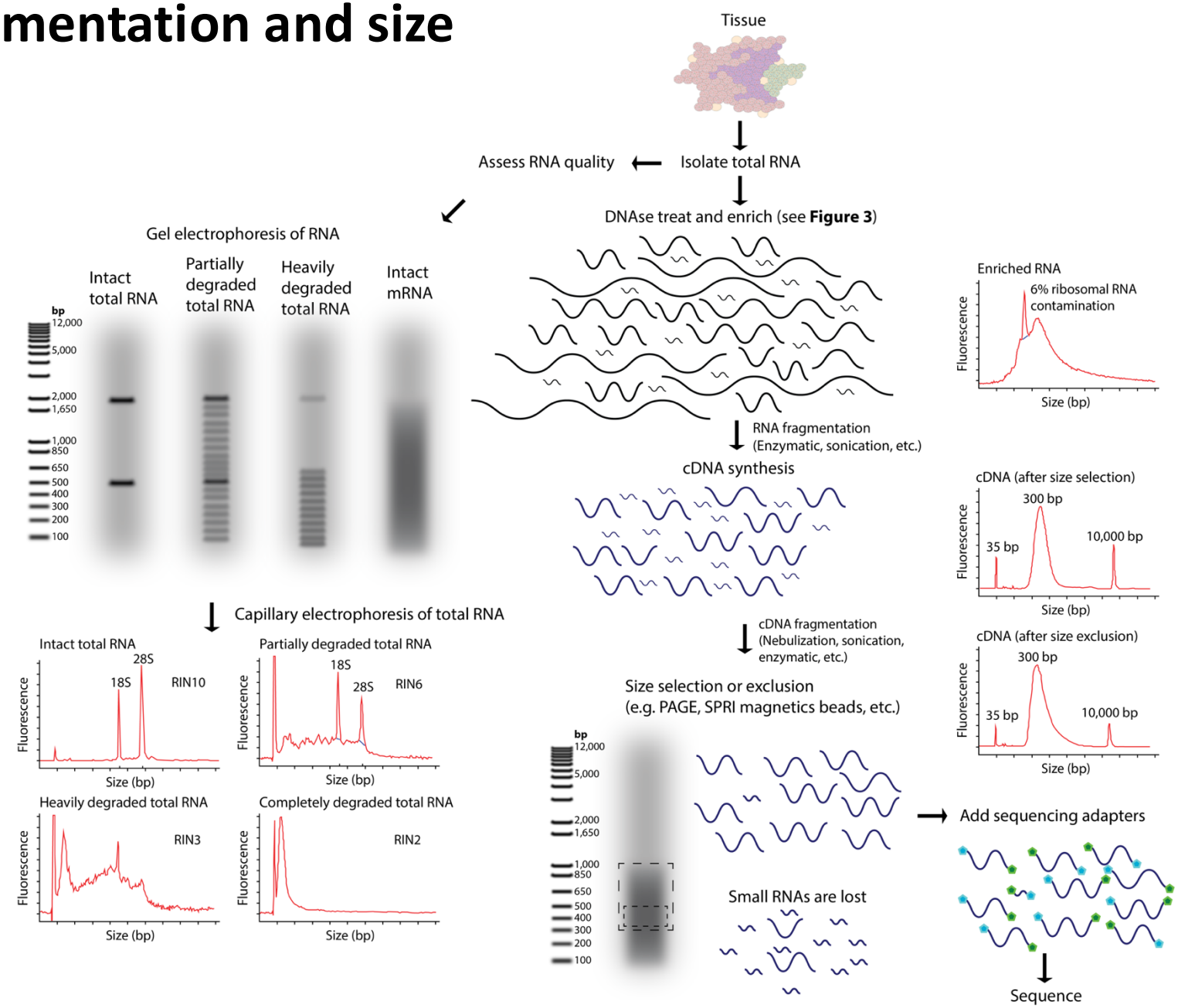
RIN = 10

There are many RNA-seq library construction strategies

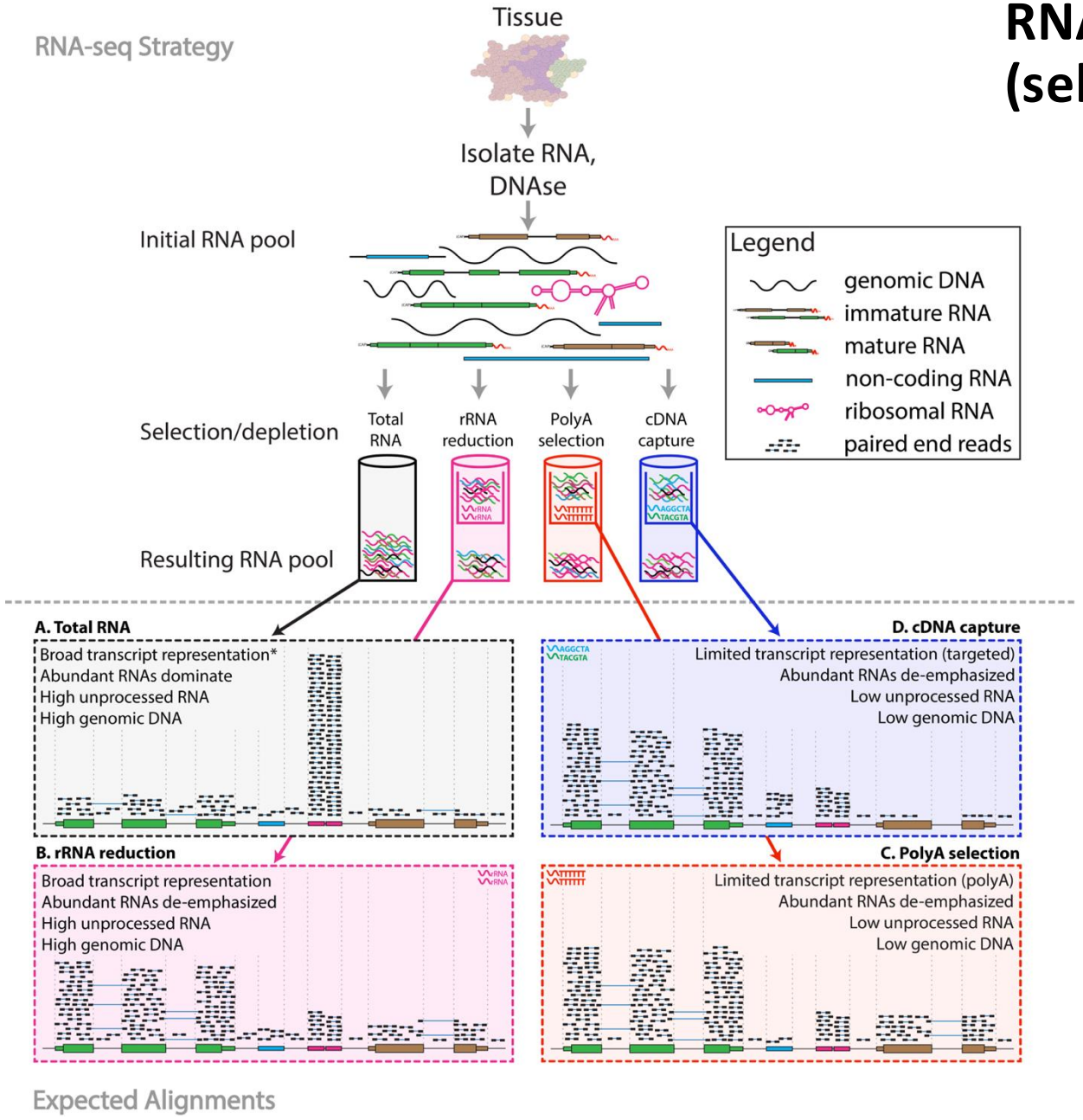
- Total RNA versus polyA+ RNA?
- Ribo-reduction?
- Size selection (before and/or after cDNA synthesis)
  - Small RNAs (microRNAs) vs. large RNAs?
  - A narrow fragment size distribution vs. a broad one?
- Linear amplification?
- Stranded vs. un-stranded libraries
- Exome captured vs. un-captured
- Library normalization?

- These details can affect analysis strategy
  - Especially comparisons between libraries

RNA-seq - Fragmentation and size selection



RNA-seq - sequence enrichment (selection/depletion)



Ordering RNA-seq data, “coverage”, and cost?

RNA-seq full service, cost per sample[a]	< 12 Samples	≥ 2500 Samples*	~Targeted Coverage
PolyA selection	$287	$215	30M reads
Ribosomal depletion, RiboErase (H/M/R)	$297	$226	30M reads
Ribosomal depletion, FastSelect (H/M/R)	$268	$205	30M reads
Ribosomal depletion, FastSelect (H/M/R+Globin)	$272	$213	30M reads
Ribosomal depletion, Watchmaker (H/M/R+Globin)	$291	NA	30M reads
Low input - Takara SMARTseq mRNA	$267	$203	30M reads
Low input - Sigma Seqplex	$273	$202	30M reads

- An example menu from a sequencing core facility (circa 2024)
- Options primarily relate to method of enrichment and input amounts
- “Coverage” is a non-intuitive concept in bulk-RNAseq.
  - 30M reads is sufficient for gene abundance estimation (increase for other applications)

Common analysis goals of RNA-Seq analysis (what can you ask of the data?)

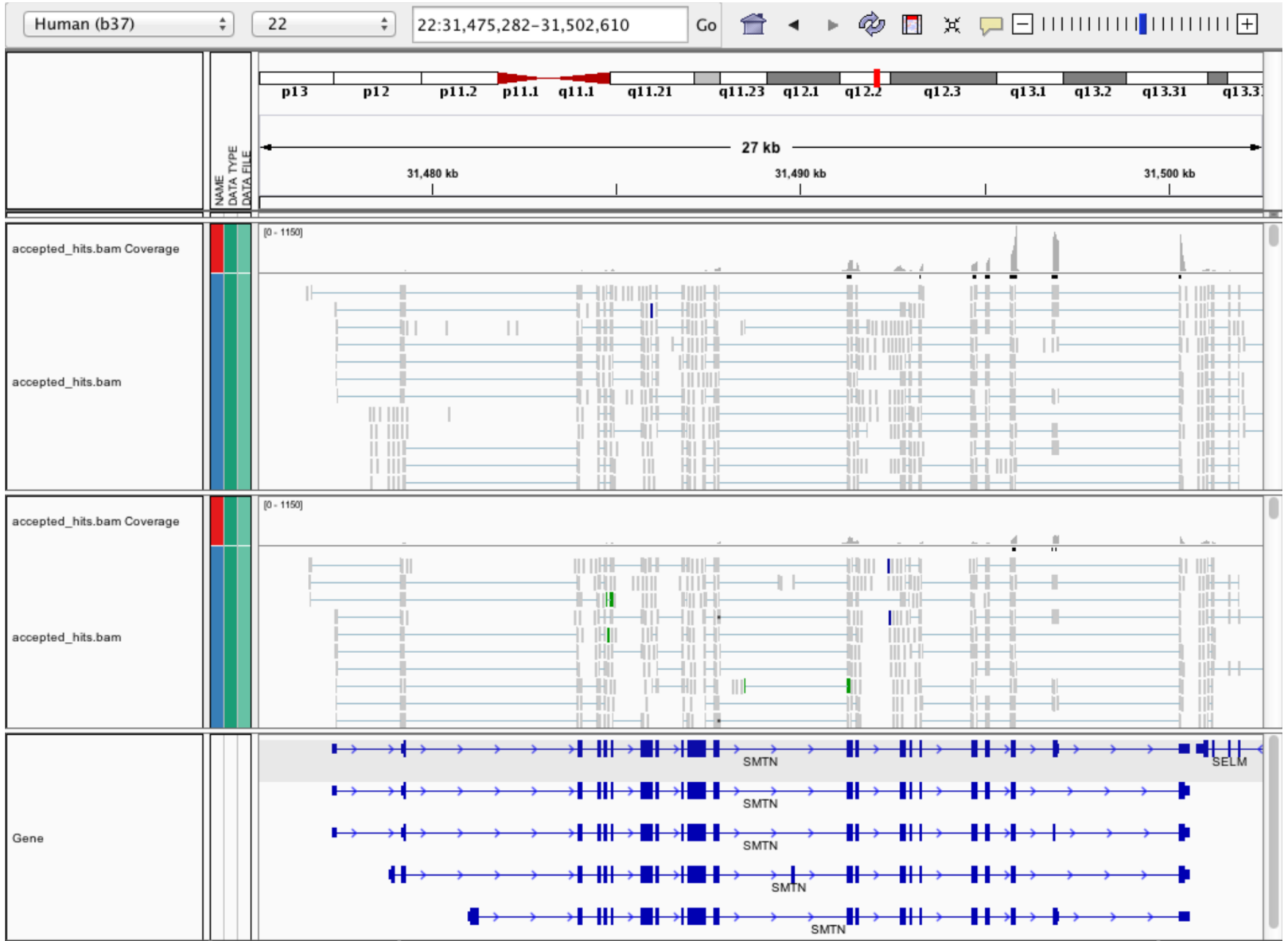
- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
  - Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

General themes of RNA-seq workflows

- Each type of RNA-seq analysis has distinct requirements and challenges but also a common theme:

1. Obtain raw data (convert format)
2. Align/assemble reads
3. Process alignment with a tool specific to the goal
    - e.g. ‘StringTie’ for expression analysis, ‘STARfusion’ for fusion detection, etc.
4. Post process
    - Import into downstream software (R, Python, etc.)
5. Summarize and visualize
    - Create gene lists, prioritize candidates for validation, etc.

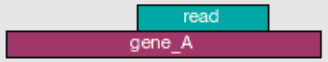
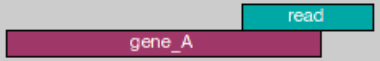
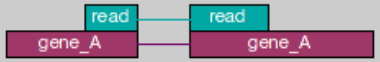
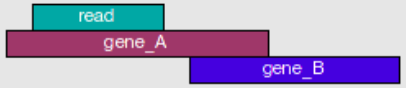
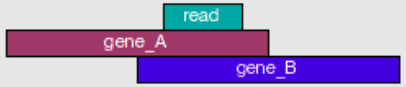
Expression estimation for known genes and transcripts



3' bias →

↓ Down-regulated

HTSeq-count basically counts reads supporting a feature (exon, gene) by assessing overlapping coordinates

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Note, if gene_A and gene_B on opposite strands, sequence data is stranded, and correct HTSeq parameter set then this read may not be ambiguous

Whether a read is counted depends on the nature of overlap and “mode” selected

What is FPKM?

- Why not just count reads in my RNAseq data? → **Fragments**

- The relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. However:

  - # fragments is biased towards larger genes → **Per Kilobase of transcript**

  - # fragments is related to total library depth → **per Million mapped reads.**

What is FPKM?

- FPKM attempts to normalize for gene size and library depth
    - remember – RPKM is essentially the same!

- C = number of mappable fragments for a gene (transcript)
- N = total number of mappable fragments in the library
- L = number of base pairs in the gene (transcript)
    - FPKM = (C / (N x L)) x 1,000 x 1,000,000
    - FPKM = (1,000,000,000 x C) / (N x L)
    - FPKM = (C / (N / 1,000,000)) / (L/1000)

- More reading:
    - http://www.biostars.org/p/11378/
    - http://www.biostars.org/p/68126/

How do FPKM and TPM differ?

- TPM: Transcript per Kilobase Million
- The difference is in the order of operations:

FPKM

1) Determine total fragment count, divide by 1,000,000 (per Million)

2) Divide each gene/transcript fragment count by #1 (Fragments Per Million)

3) Divide each FPM by length of each gene/transcript in kilobases (FPKM)

TPM

1) Divide each gene/transcript fragment count by length of the transcript in kilobases (Fragments Per Kilobase)

2) Sum all FPK values for the sample and divide by 1,000,000 (per Million)

3) Divide #1 by #2 (TPM)

- The sum of all TPMs in each sample is the same. Easier to compare across samples!

- http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/
- https://www.ncbi.nlm.nih.gov/pubmed/22872506

‘FPKM/TPM’ expression estimates vs. ‘raw’ counts

- Which should I use?
  - Long running debate, but the general consensus:

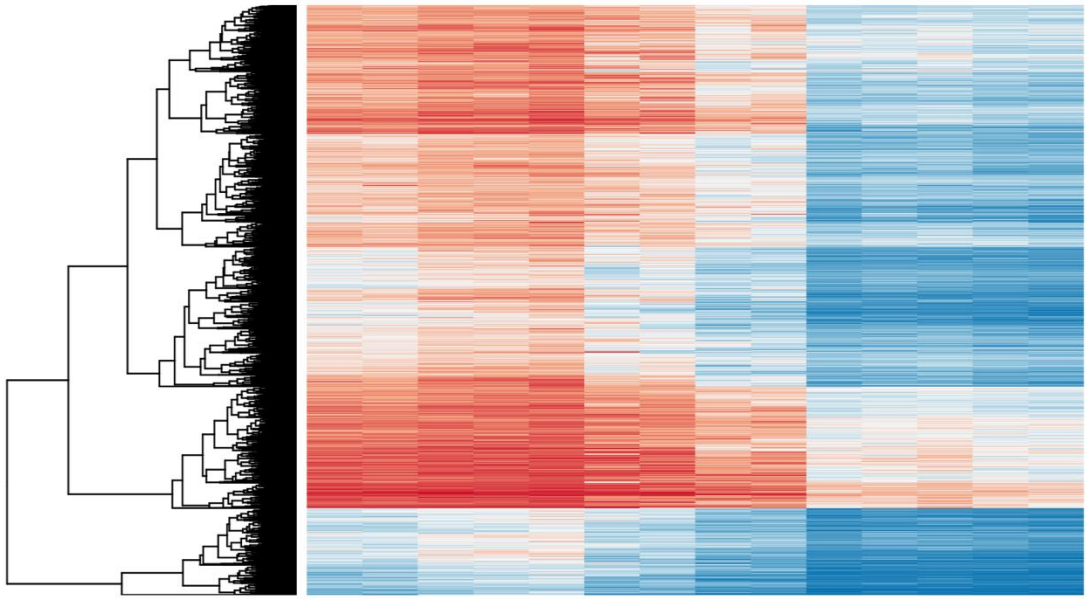
- FPKM/TPM
  - When you want to leverage more holistic approaches
    - Transcript/Isoform deconvolution
  - Good for visualization (e.g., heatmaps)
  - Calculating fold changes, etc.

- Counts
  - “More robust" statistical methods for differential expression
  - Accommodates more sophisticated experimental designs with appropriate statistical tests
  - Not suitable for transcript/isoform level estimation

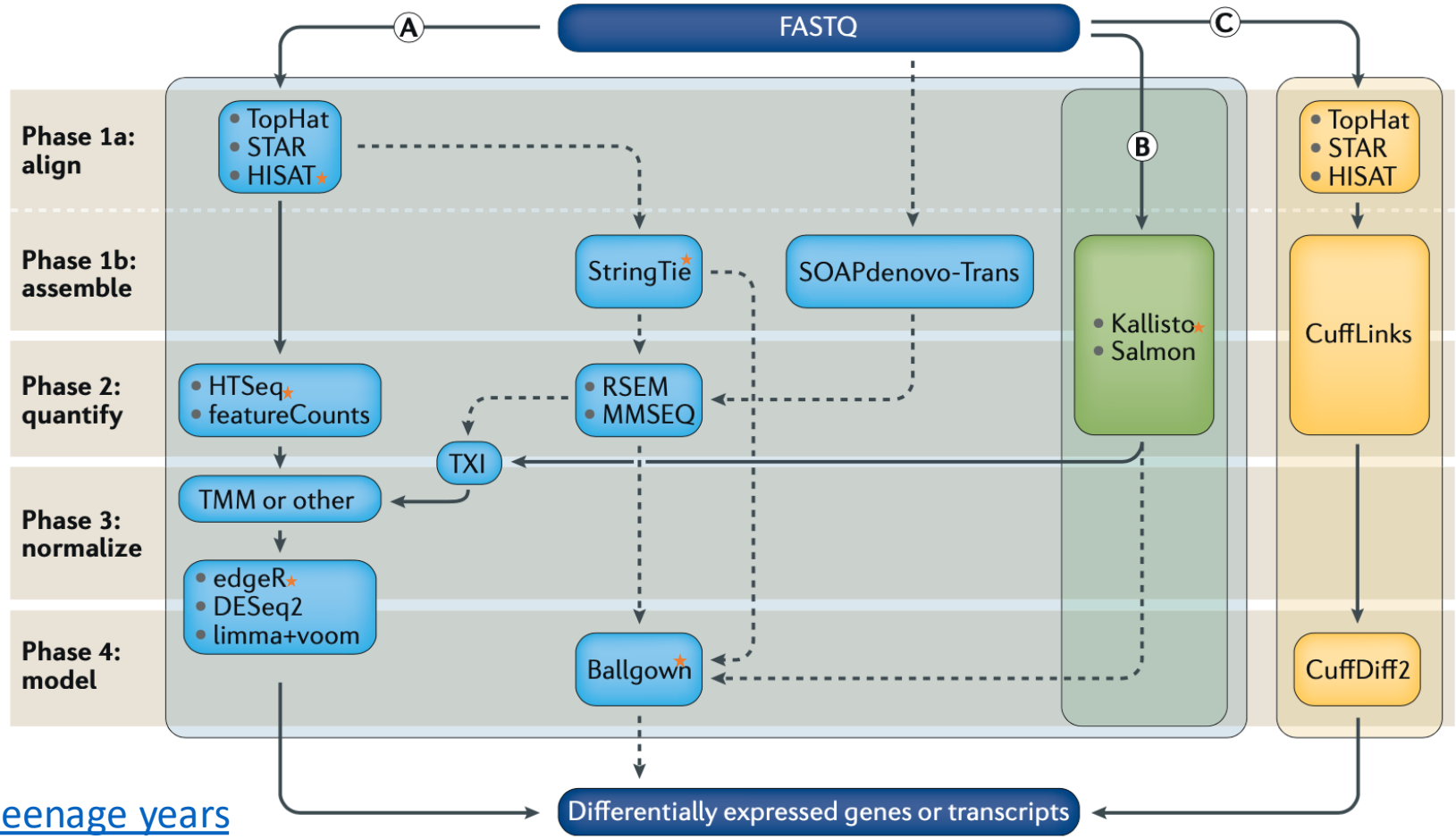
Differential Expression

- Tying gene expression back to genotype/phenotype

- What genes/transcripts are being expressed at higher/lower levels in different groups of samples?
  - Are these differences ‘significant’, accounting for variance/noise?



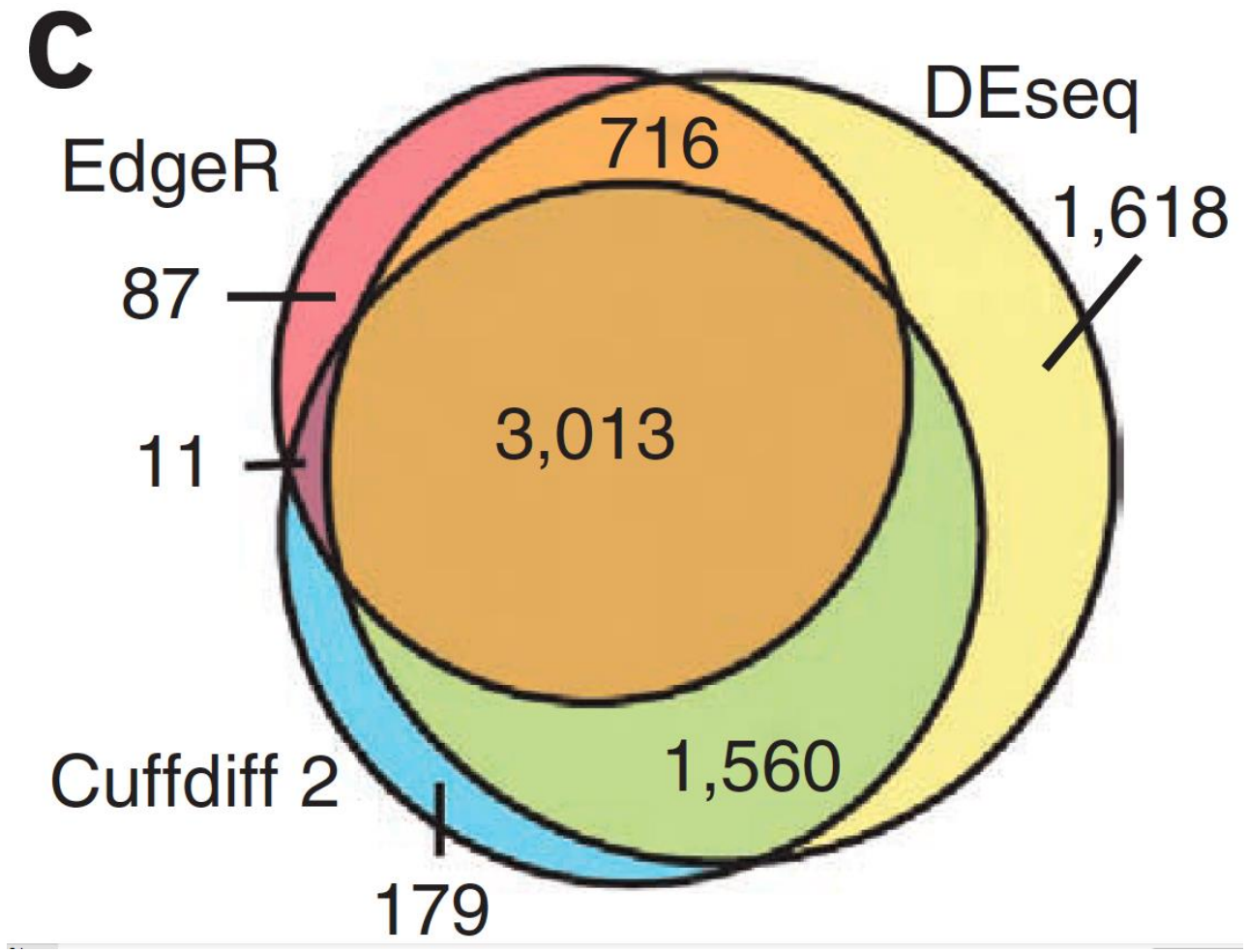
Examples of RNA-seq data analysis workflows for differential gene expression



[RNA sequencing: the teenage years]

★ Covered in rnabio.org

Multiple approaches advisable



Multiple testing correction

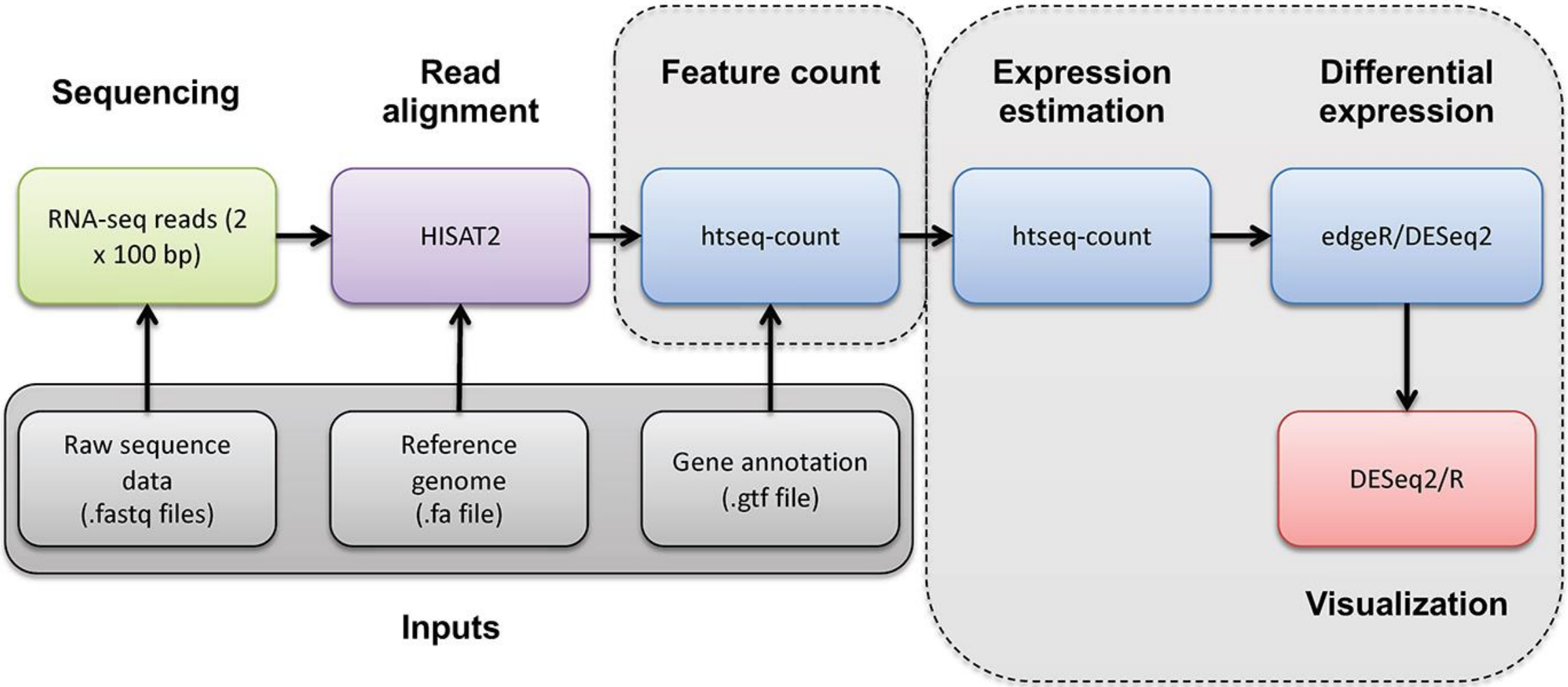
- As more attributes are compared, differences due solely to chance become more likely!

- Well known from array studies
  - 10,000s genes/transcripts
  - 100,000s exons

- With RNA-seq, more of a problem than ever
  - All the complexity of the transcriptome gives huge numbers of potential features
    - Genes, transcripts, exons, junctions, retained introns, microRNAs, lncRNAs, etc

- Bioconductor multtest
  - http://www.bioconductor.org/packages/release/bioc/html/multtest.html

Steps upstream of the hands-on exercises



Bioinformatics troubleshooting cheat sheet

- ❑ Check your inputs!
- ❑ Mix of incompatible reference genomes used (see this tutorial)
- ❑ Mix of incompatible gene/transcript identifiers
- ❑ Reference sequence names (e.g. “1” vs “chr1”)
- ❑ 1-based vs 0-based coordinates (see this tutorial)
- ❑ Computational tasks fail due to resource limitations (memory and storage)
- ❑ Dependency hell for bioinformatics tools. Learn to use containers (e.g. docker) or environment managers (e.g. conda)

Introduction to rnabio.org