RNA-seq

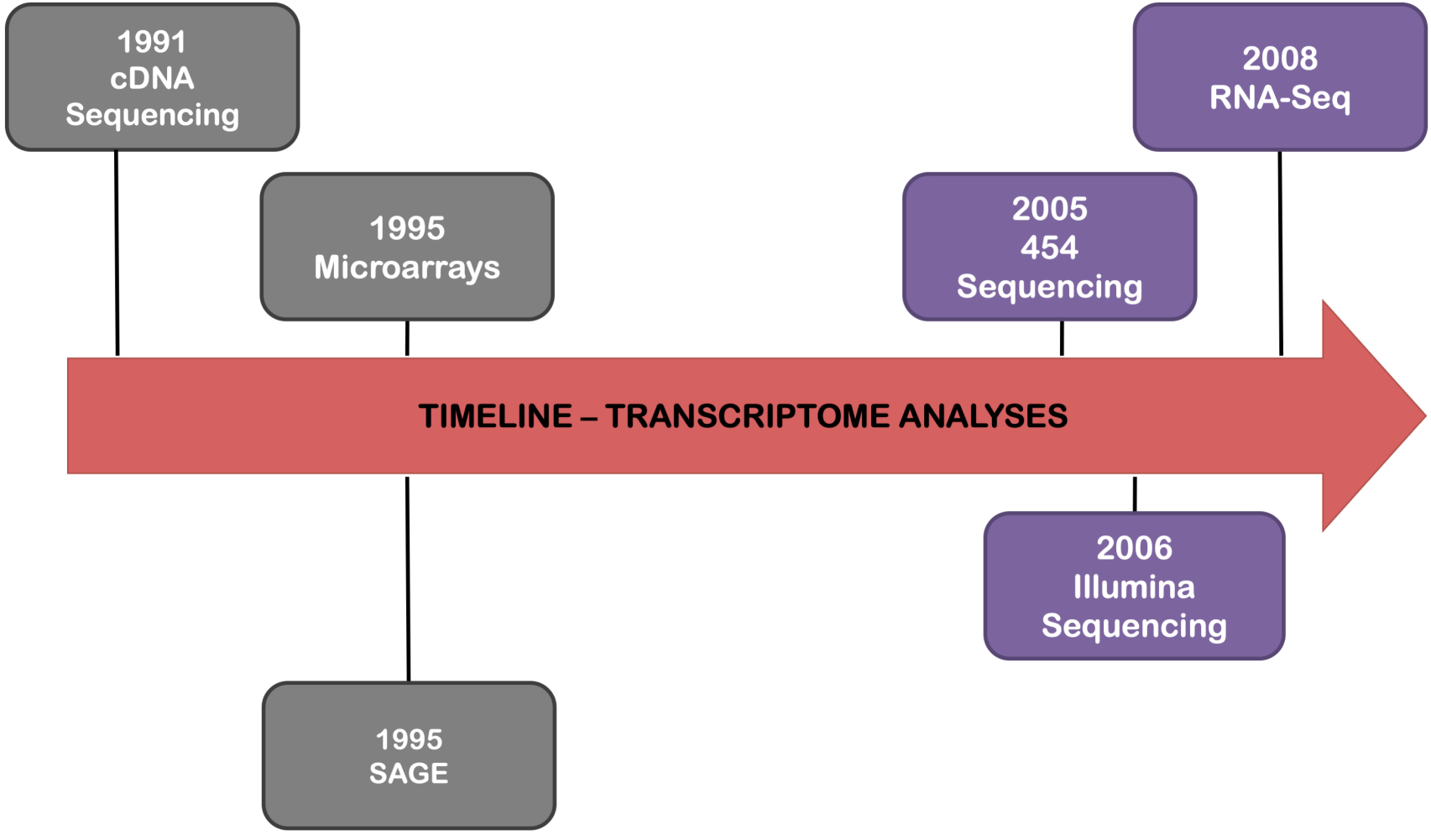
Introduction & Important Considerations

Day 02

https://tttorres.github.io/transcriptomics/

Transcriptomics

The evolution of the methods



Transcriptomics

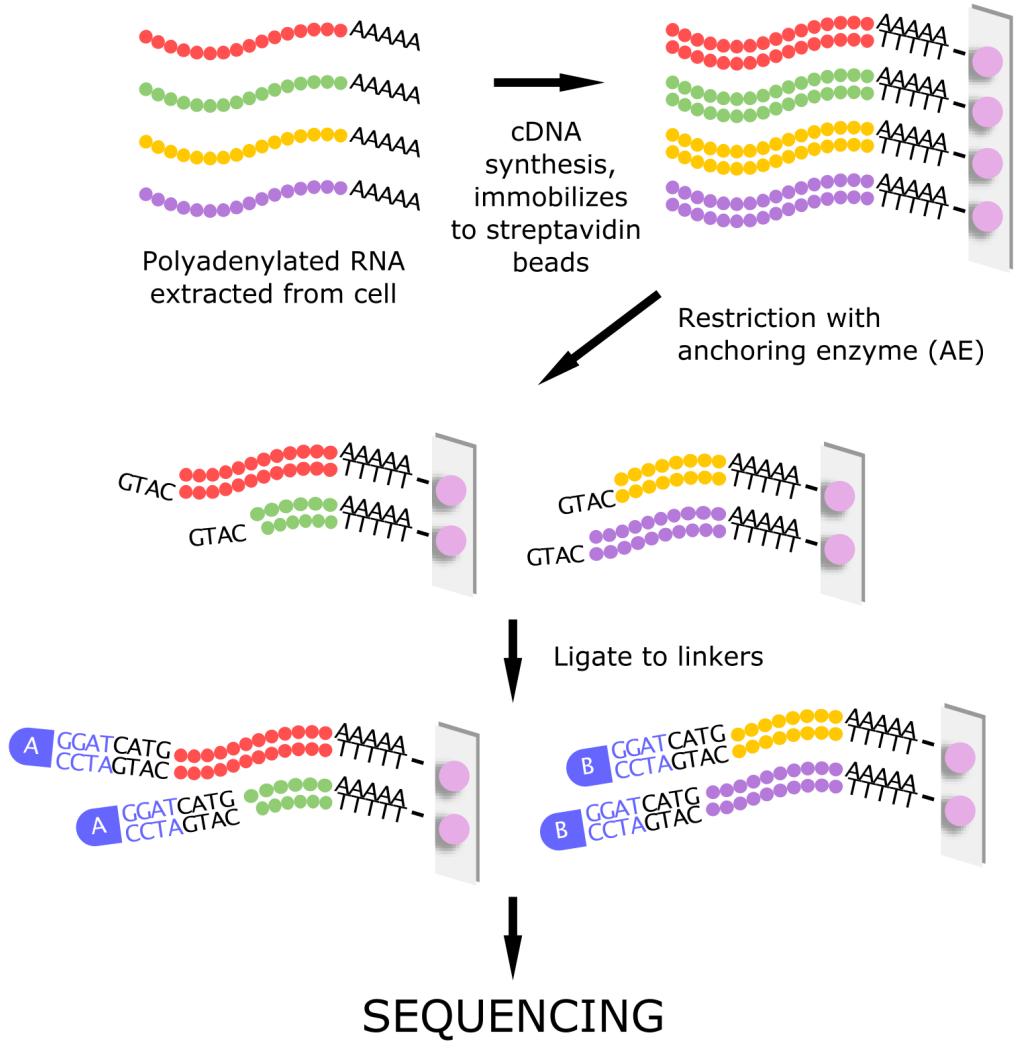
The evolution of the methods

454 / Roche

Solexa / Illumina

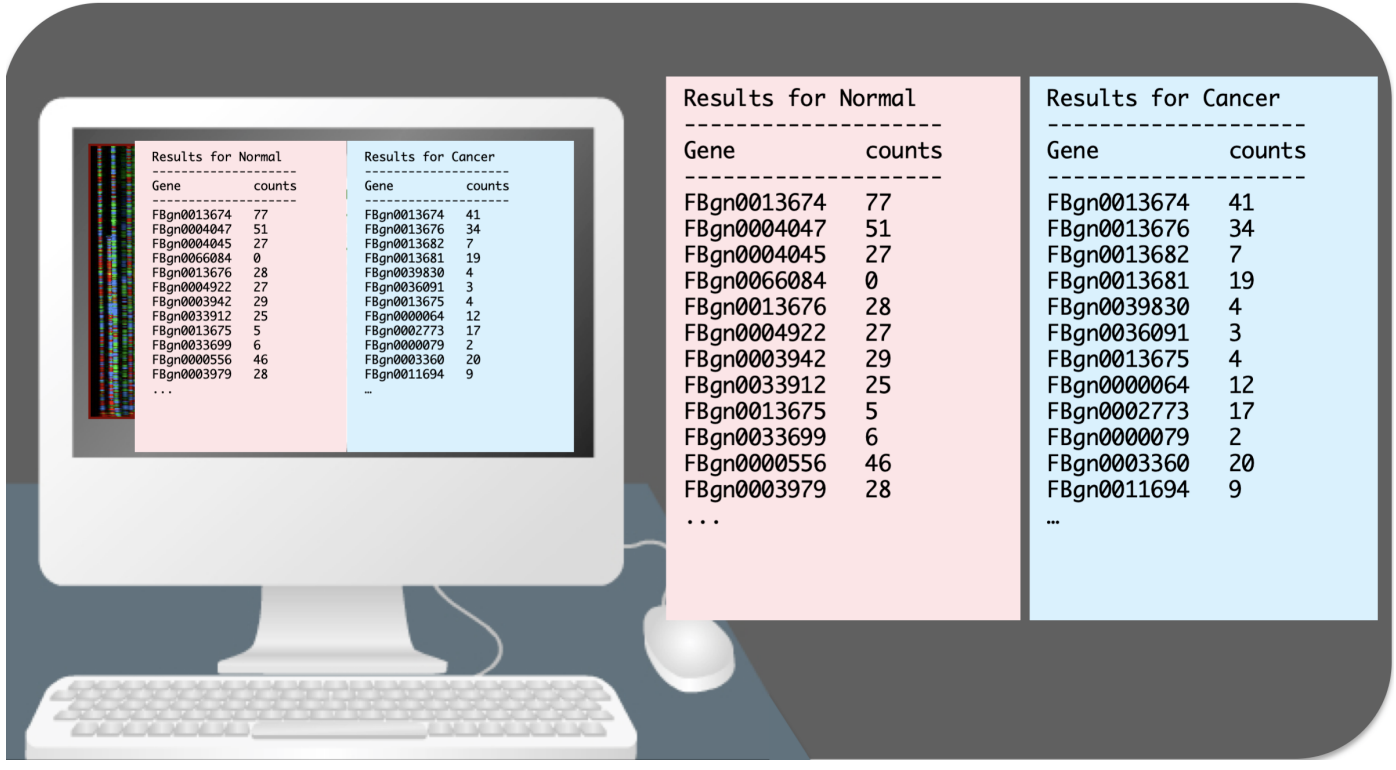
Transcriptomics

Massively Parallel Sequencing



Transcriptomics

Massively Parallel Sequencing



Torres et al. *Genome Res*. 2008 Jan;18(1):172-177.

Transcriptomics

Massively Parallel Sequencing

Methods

Gene expression profiling by massively parallel sequencing

Tatiana Teixeira Torres,[1] Muralidhar Metta,[1] Birgit Ottenwälder,[2] and Christian Schlötterer[1,3]

[1]*Institut für Tierzucht und Genetik, Veterinärmedizinische Universität Wien, 1210 Vienna, Austria;* [2]*Eurofins Medigenomix GmbH, 82152 Martinsried, Germany*

Massively parallel sequencing holds great promise for expression profiling, as it combines the high throughput of SAGE with the accuracy of EST sequencing. Nevertheless, until now only very limited information had been available on the suitability of the current technology to meet the requirements. Here, we evaluate the potential of 454 sequencing technology for expression profiling using *Drosophila melanogaster*. We show that short (< ~80 bp) and long (> ~300–400 bp) cDNA fragments are under-represented in 454 sequence reads. Nevertheless, sequencing of 3′ cDNA fragments generated by nebulization could be used to overcome the length bias of the 454 sequencing technology. Gene expression measurements generated by restriction analysis and nebulization for fragments within the 80- to 300-bp range showed correlations similar to those reported for replicated microarray experiments (0.83–0.91); 97% of the cDNA fragments could be unambiguously mapped to the genomic DNA, demonstrating the advantage of longer sequence reads. Our analyses suggest that the 454 technology has a large potential for expression profiling, and the high mapping accuracy indicates that it should be possible to compare expression profiles across species.

Torres et al. *Genome Res*. 2008 Jan;18(1):172-177.

Transcriptomics

RNA-Seq

The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

Ugrappa Nagalakshmi,[1]* Zhong Wang,[1]* Karl Waern,[1] Chong Shou,[2] Debasish Raha,[1] Mark Gerstein,[2,3] Michael Snyder[1,2,3]†

The identification of untranslated regions, introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method called RNA-Seq for mapping transcribed regions, in which complementary DNA fragments are subjected to high-throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames also were identified for many yeast genes. We also found unexpected 3′-end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.

Mapping and quantifying mammalian transcriptomes by RNA-Seq

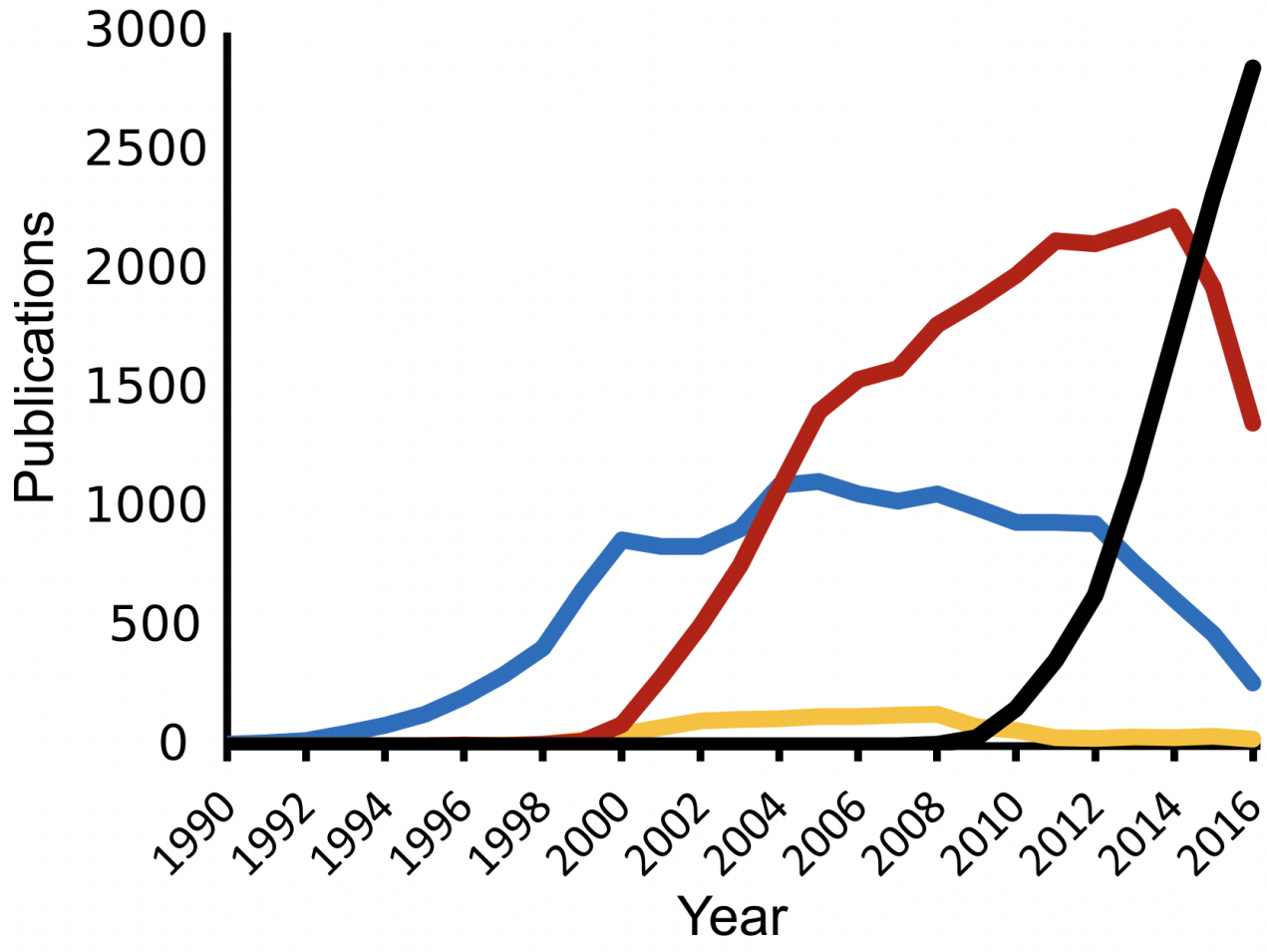
Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

We have mapped and quantified mouse transcriptomes by deeply sequencing them and recording how frequently each gene is represented in the sequence sample (RNA-Seq). This provides a digital measure of the presence and prevalence of transcripts from known and previously unknown genes. We report reference measurements composed of 41–52 million mapped 25-base-pair reads for poly(A)-selected RNA from adult mouse brain, liver and skeletal muscle tissues. We used RNA standards to quantify transcript prevalence and to test the linear range of transcript detection, which spanned five orders of magnitude. Although >90% of uniquely mapped reads fell within known exons, the remaining data suggest new and revised gene models, including changed or additional promoters, exons and 3′ untranscribed regions, as well as new candidate microRNA precursors. RNA splice events, which are

approaches to large-scale RNA analysis are serial analysis of gene expression (SAGE)[4,5] and related methods such as massively parallel signature sequencing (MPSS)[6], which use DNA sequencing of previously cloned tags 17–25 base pairs (bp) from terminal 3′ (or 5′) sequence tags. These sequence tags are then identified by informatic mapping to mRNA reference databases or, for longer tag lengths, to the source genome. A strength of SAGE and SAGE-like methods is that they produce digital counts of transcript abundance, in contrast to the analog-style signals obtained from fluorescent dye–based microarrays. However, SAGE-family assays provide no information about splice isoforms or new gene discovery, and fully comprehensive measurements of lower-abundance-class RNAs have not been achieved owing to cost and technology constraints. Expressed sequence tag (EST) sequencing of cloned cDNAs has long been the core method for reference transcript

Transcriptomics

The evolution of the methods



Lowe et al (2017) PLoS Comput Biol, 13, e1005457.

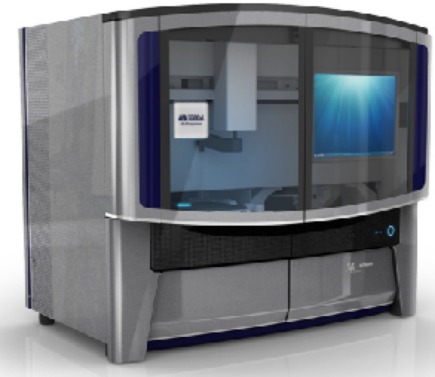
RNA-Seq

Sequencing platforms

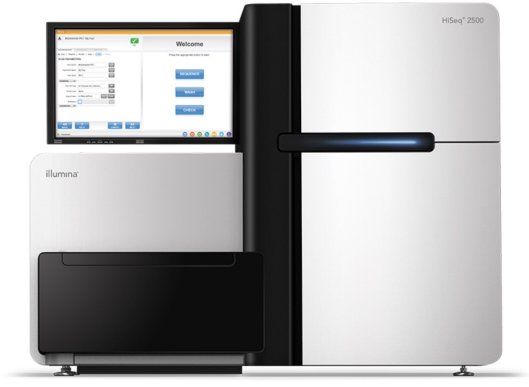
Roche 454 ‘pyrosequencing’

Life Technologies
Ion Torrent

Life Technologies SOLiD



Illumina



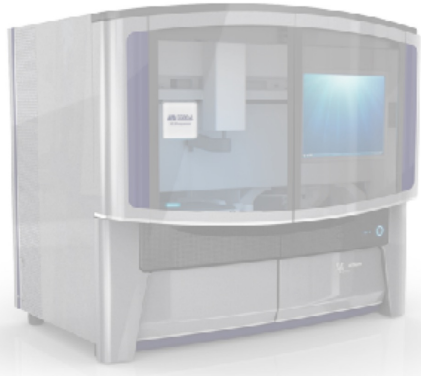
RNA-Seq

Sequencing platforms

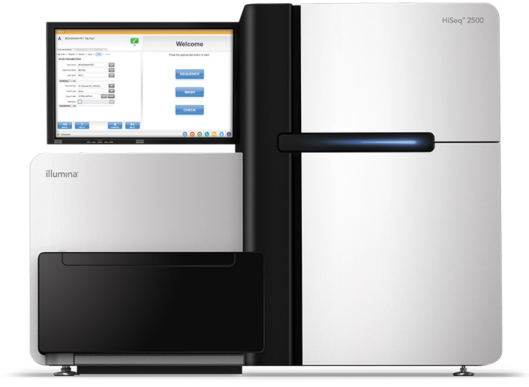
Roche 454 ‘pyrosequencing’

Life Technologies
Ion Torrent

Life Technologies SOLiD

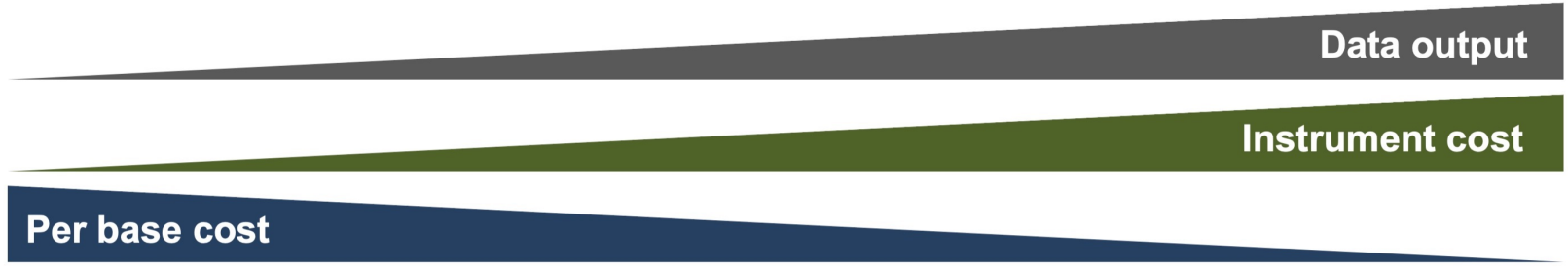


Illumina



RNA-Seq

Sequencing platforms: Illumina



RNA-Seq

Sequencing platforms: Illumina

MiniSeq

MiSeq

1-25 million reads
150 - 250bp paired-end
24 hr run time

NextSeq

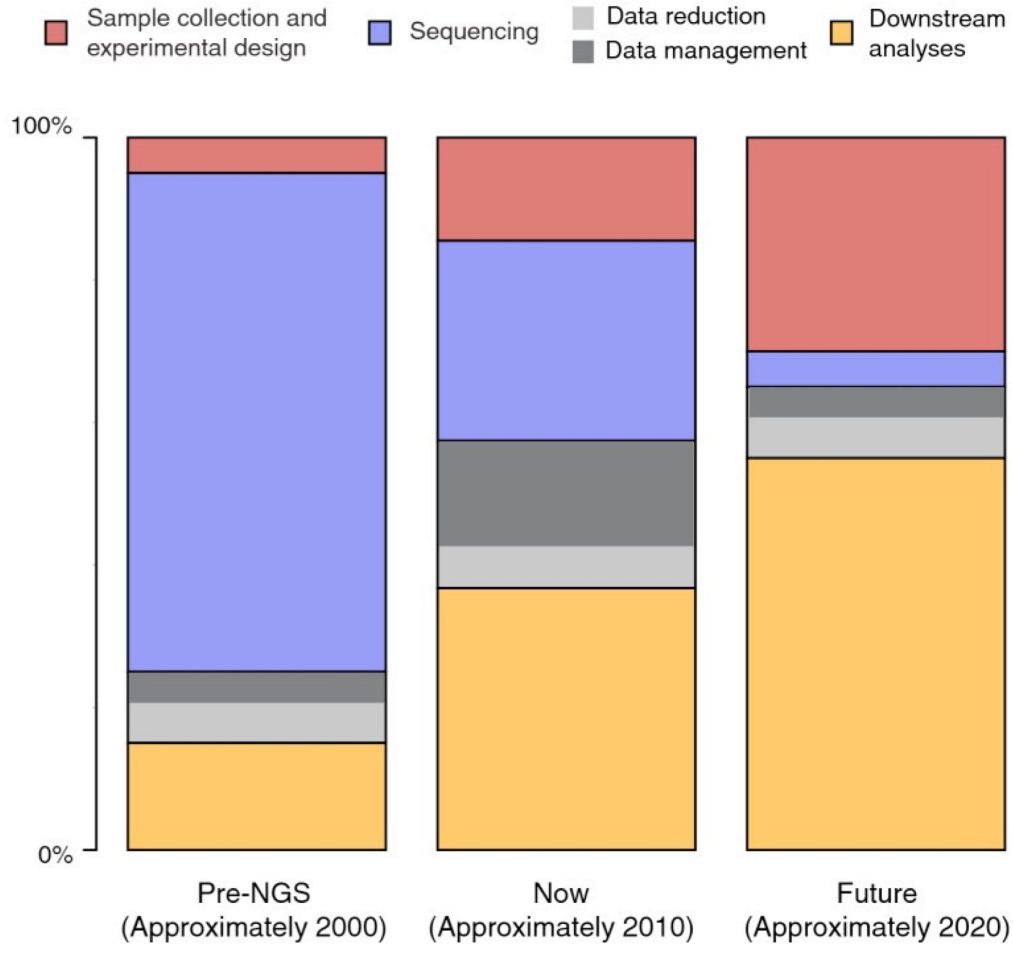
0.4 - 1.1 billion reads
150bp paired-end
24 hr run time

NovaSeq

0.6 - 10 billion reads
50-250bp paired-end
5 day run time

RNA-Seq

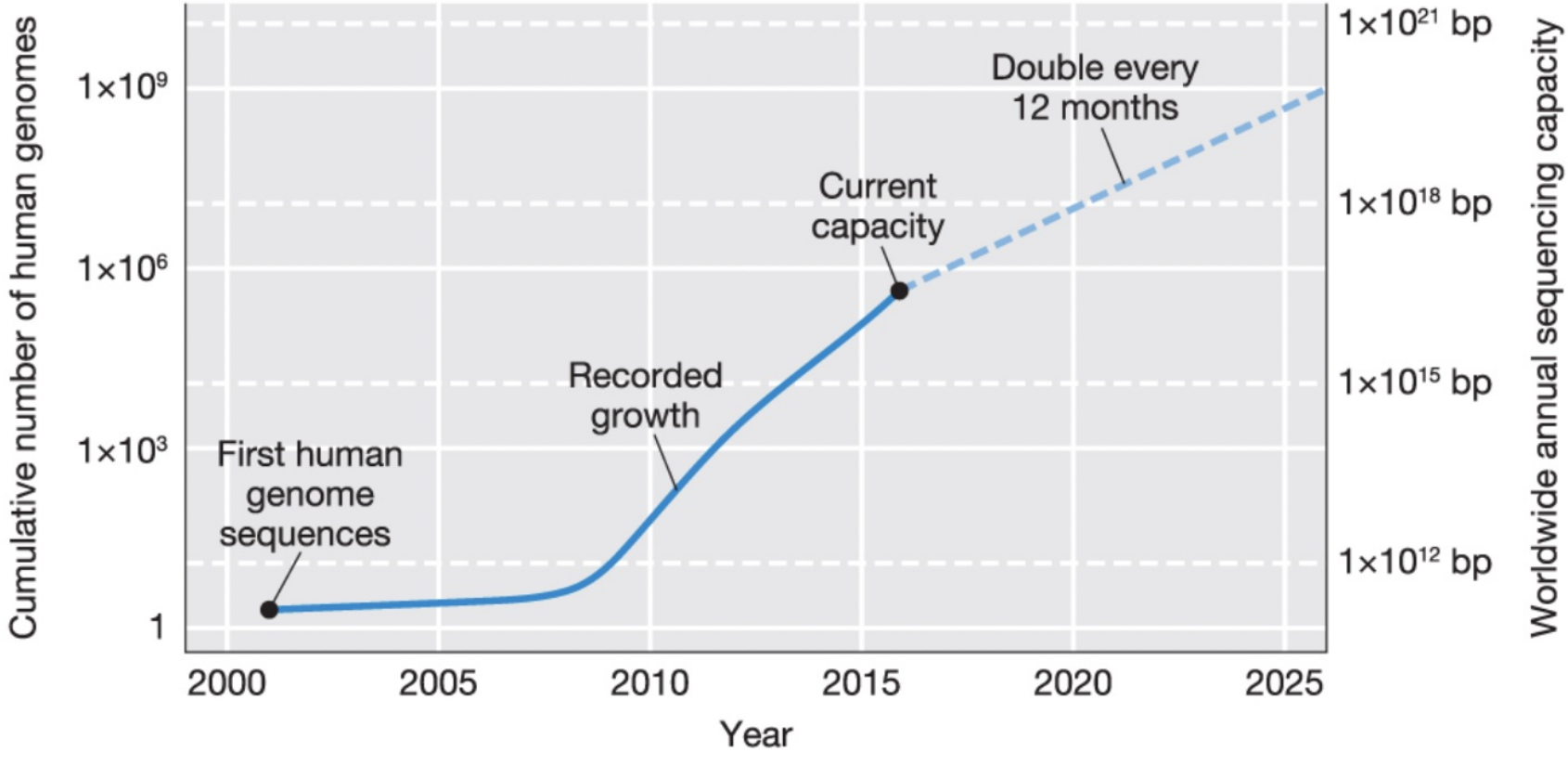
Sequencing platforms: Illumina



Sboner et al (2011) Genome Biology, 12, 125

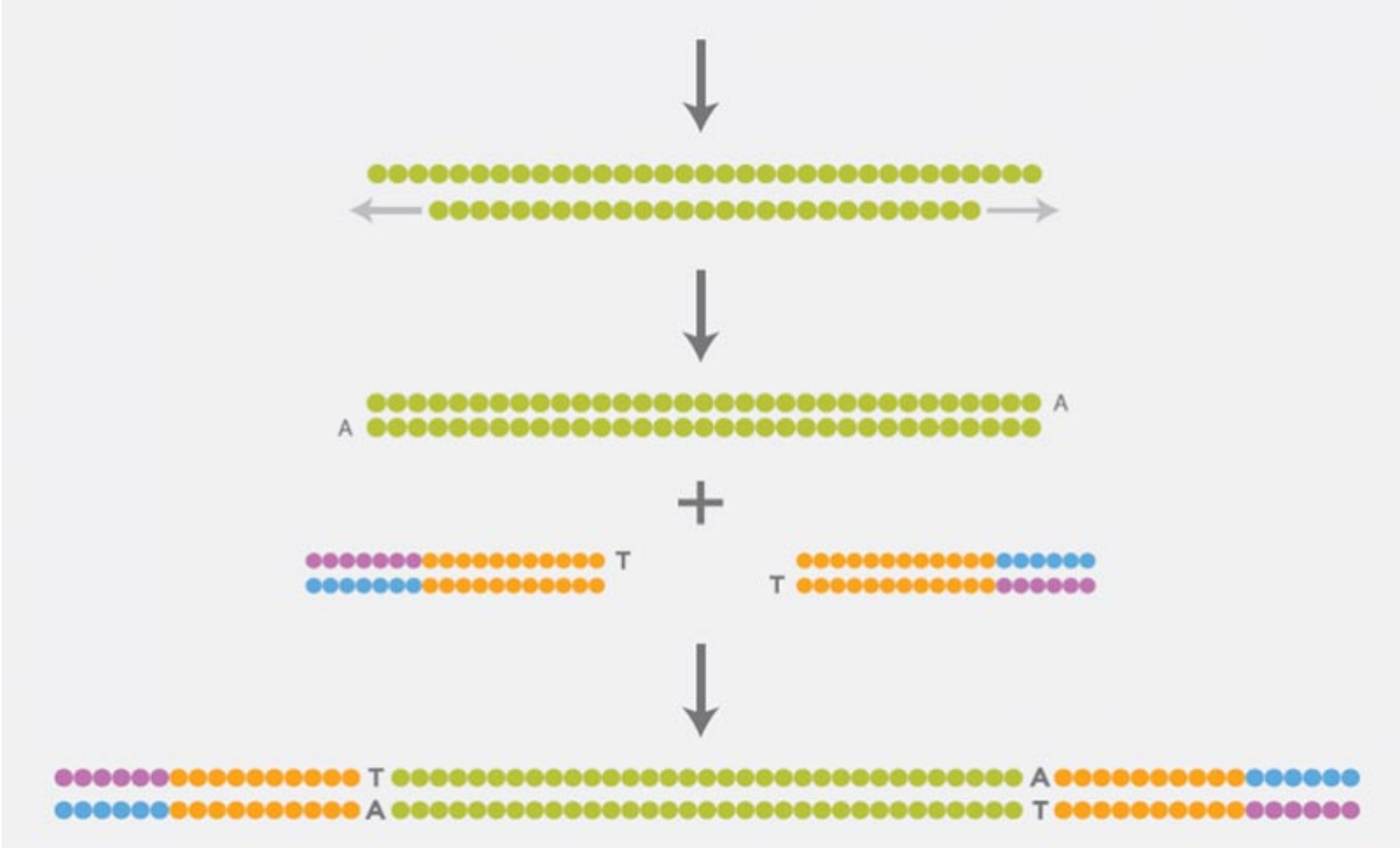
RNA-Seq

Sequencing platforms: Illumina



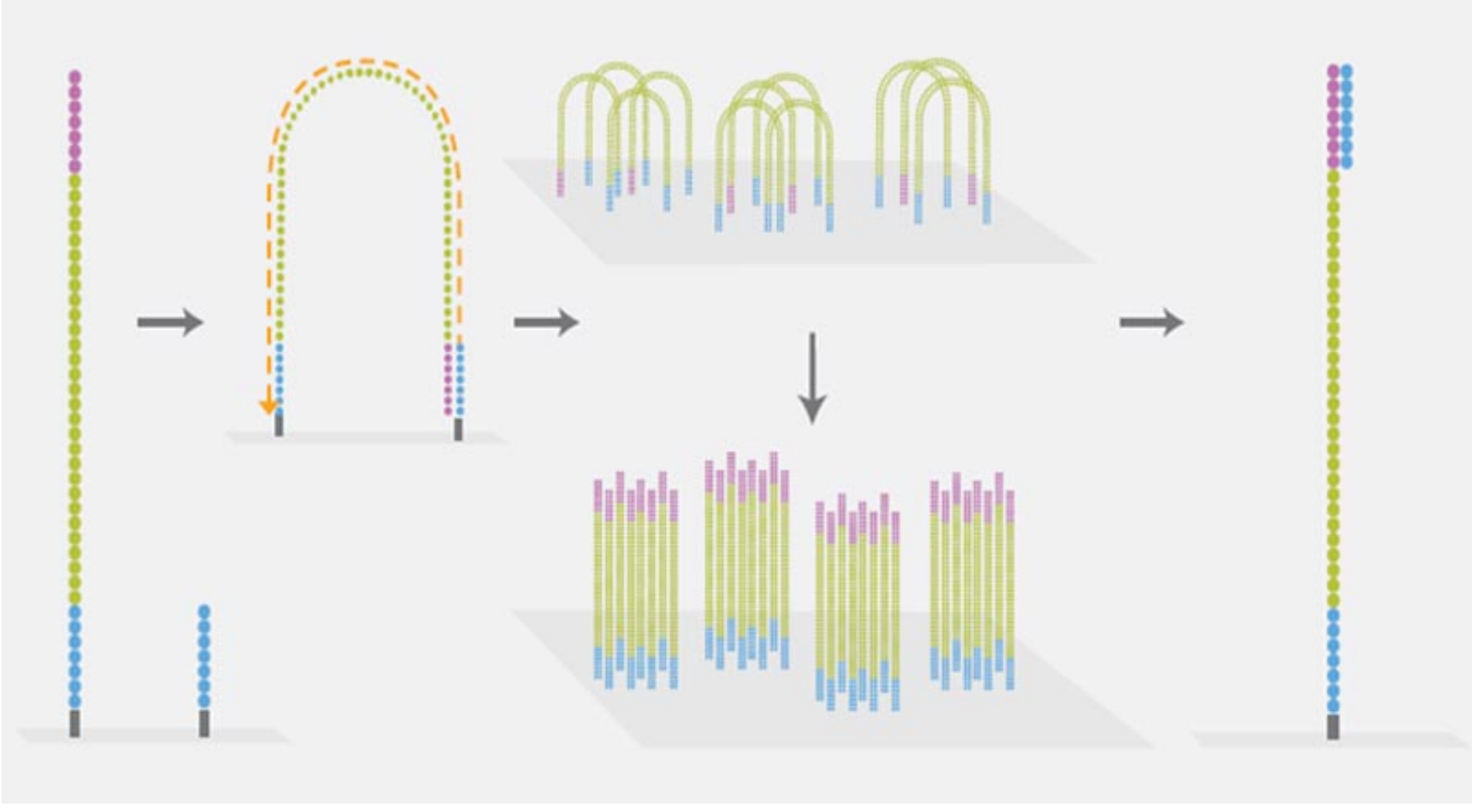
RNA-Seq

Illumina: Sequencing by Synthesis



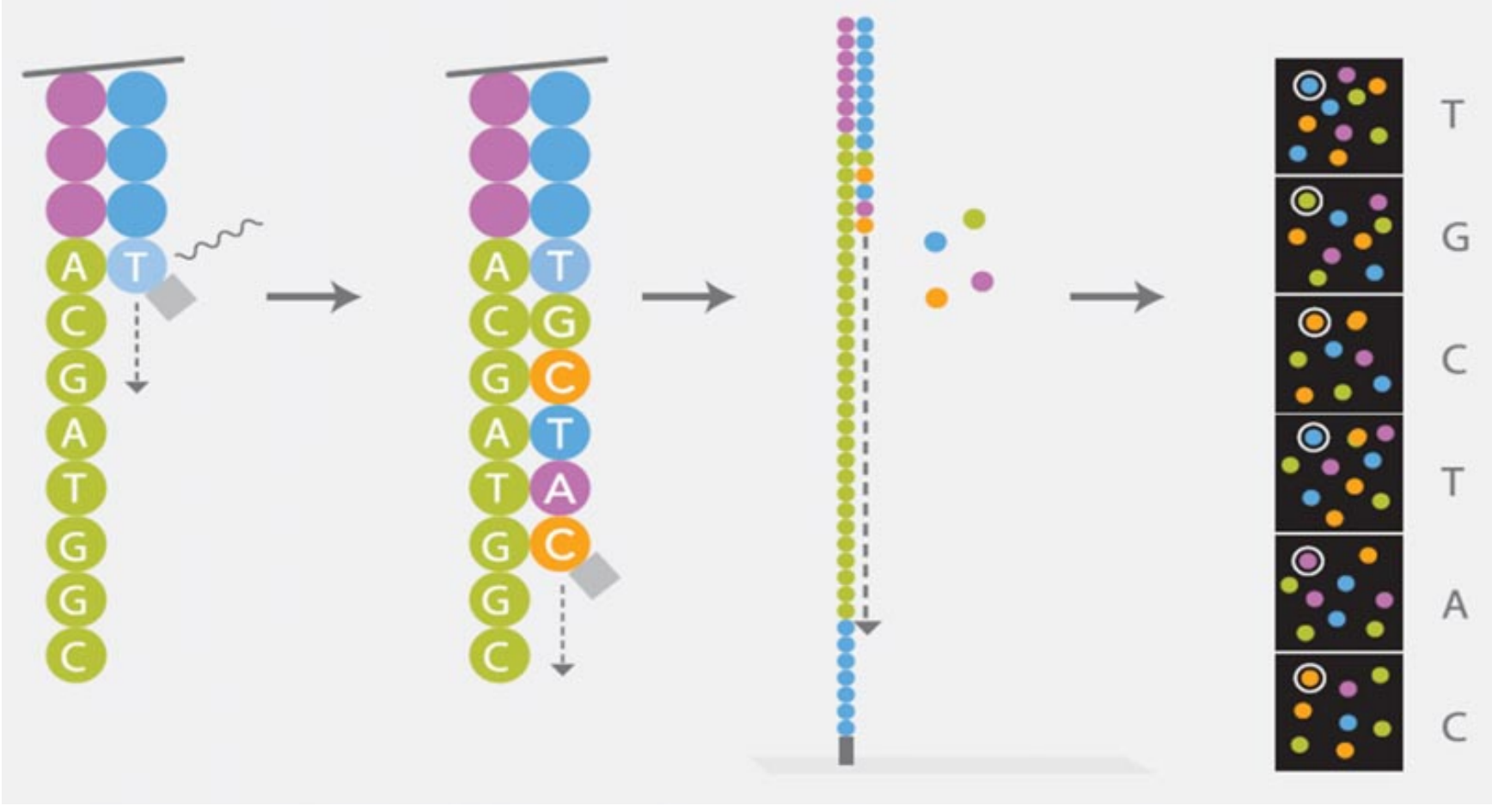
RNA-Seq

Illumina: Sequencing by Synthesis



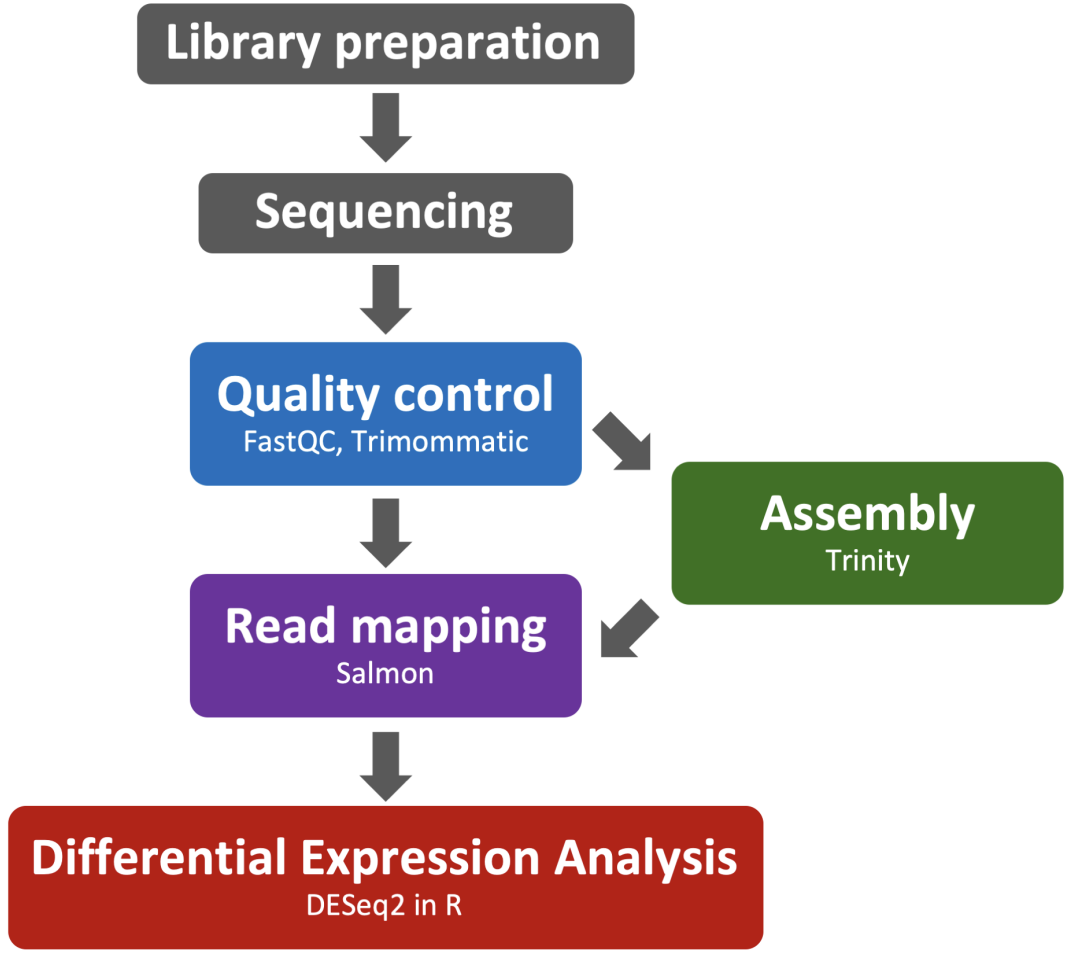
RNA-Seq

Illumina: Sequencing by Synthesis



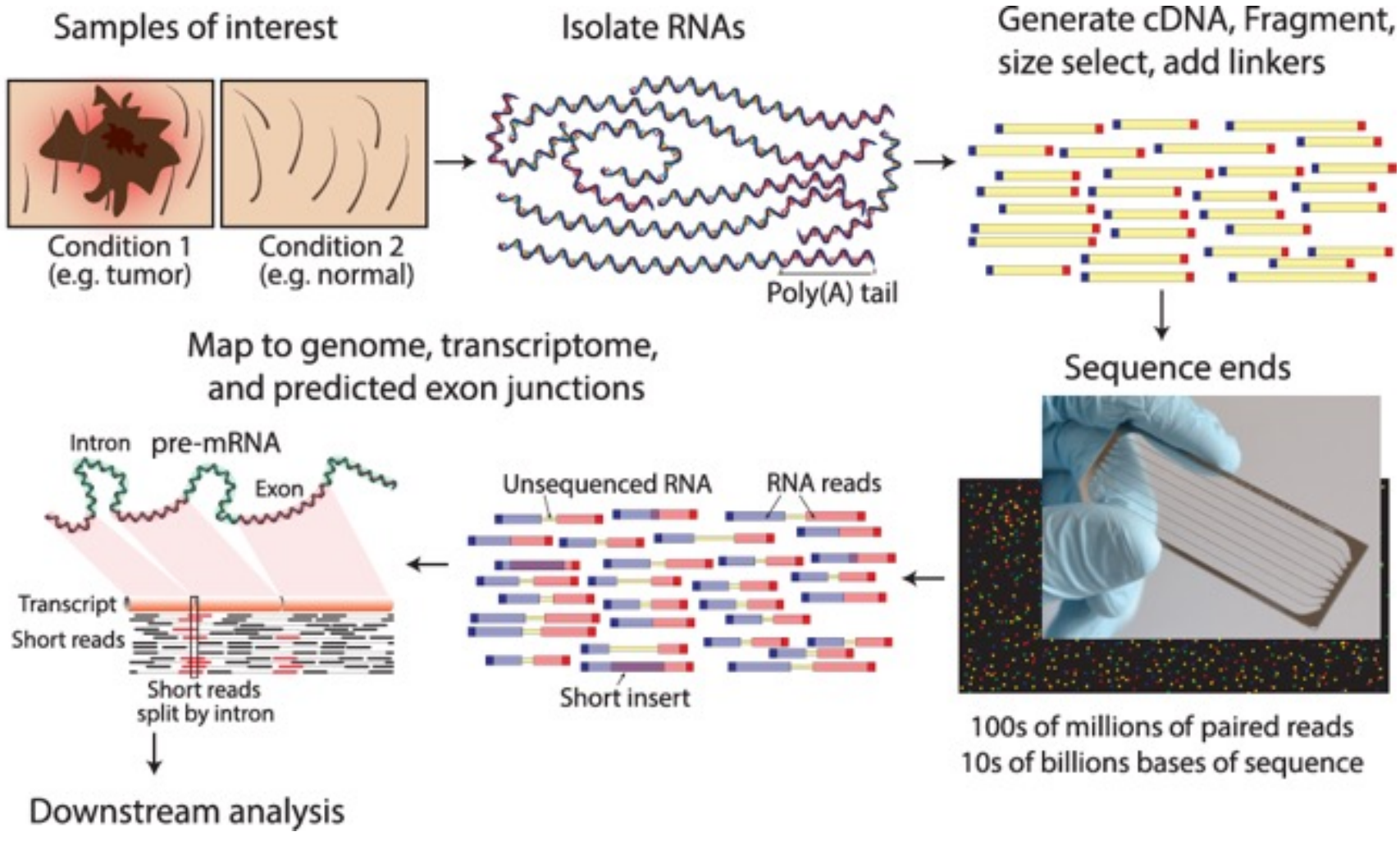
RNA-Seq

Illumina



RNA-Seq

Illumina



RNA-Seq

Before you start: Experimental Design

1. **Study design**
    - independent evaluation
    - appropriate controls
2. **Library preparation**
    - steps may differ depending on application
    - quality control
3. **Sequencing**
    - how many replicates?
    - How many reads?

RNA-Seq: Sample preparation

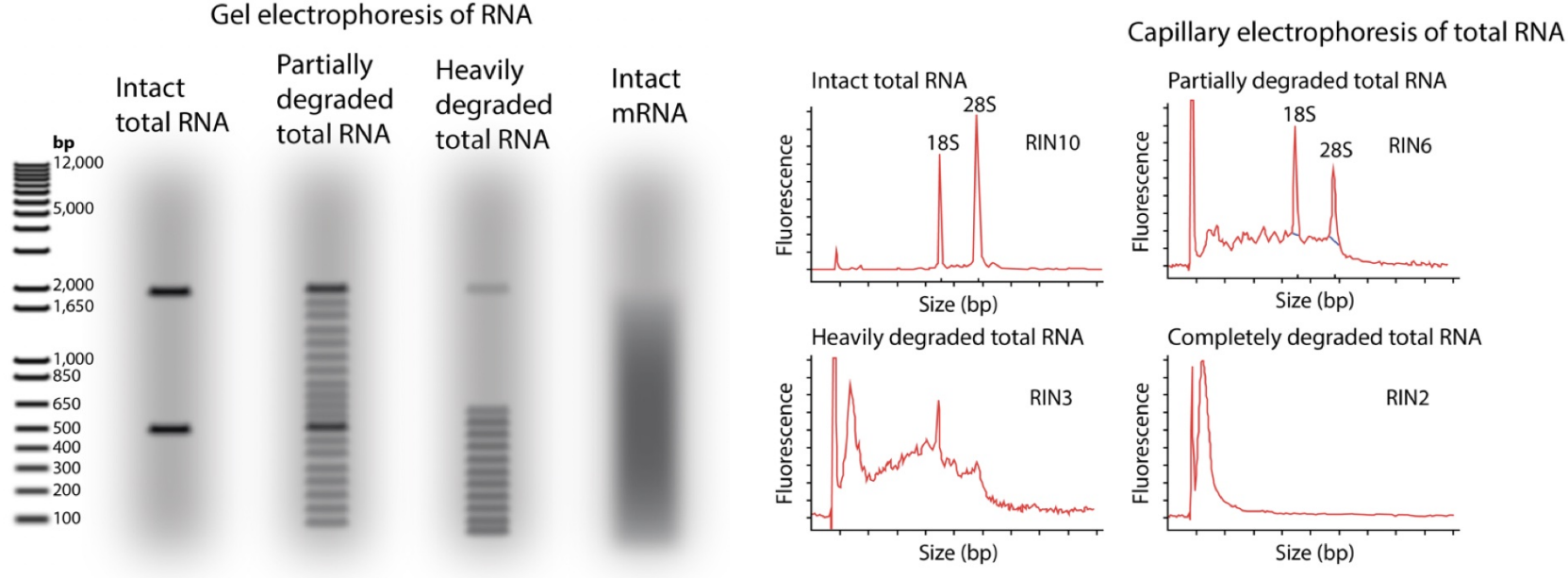
Challenges

- Sample quantity
- Sample quality
- RNA sequencing works by random sampling = a small fraction of highly expressed genes may consume the majority of reads
    - Ribosomal
    - Mitochondrial genes
- PolyA selection of large RNAs may result in 3′ end bias
- RNA is fragile compared to DNA (easily degraded)

RNA-Seq: Sample preparation

Agilent Bioanalyzer system

‘RIN’ = RNA integrity number: 0 (bad) to 10 (perfect)



RNA-seq: Library preparation

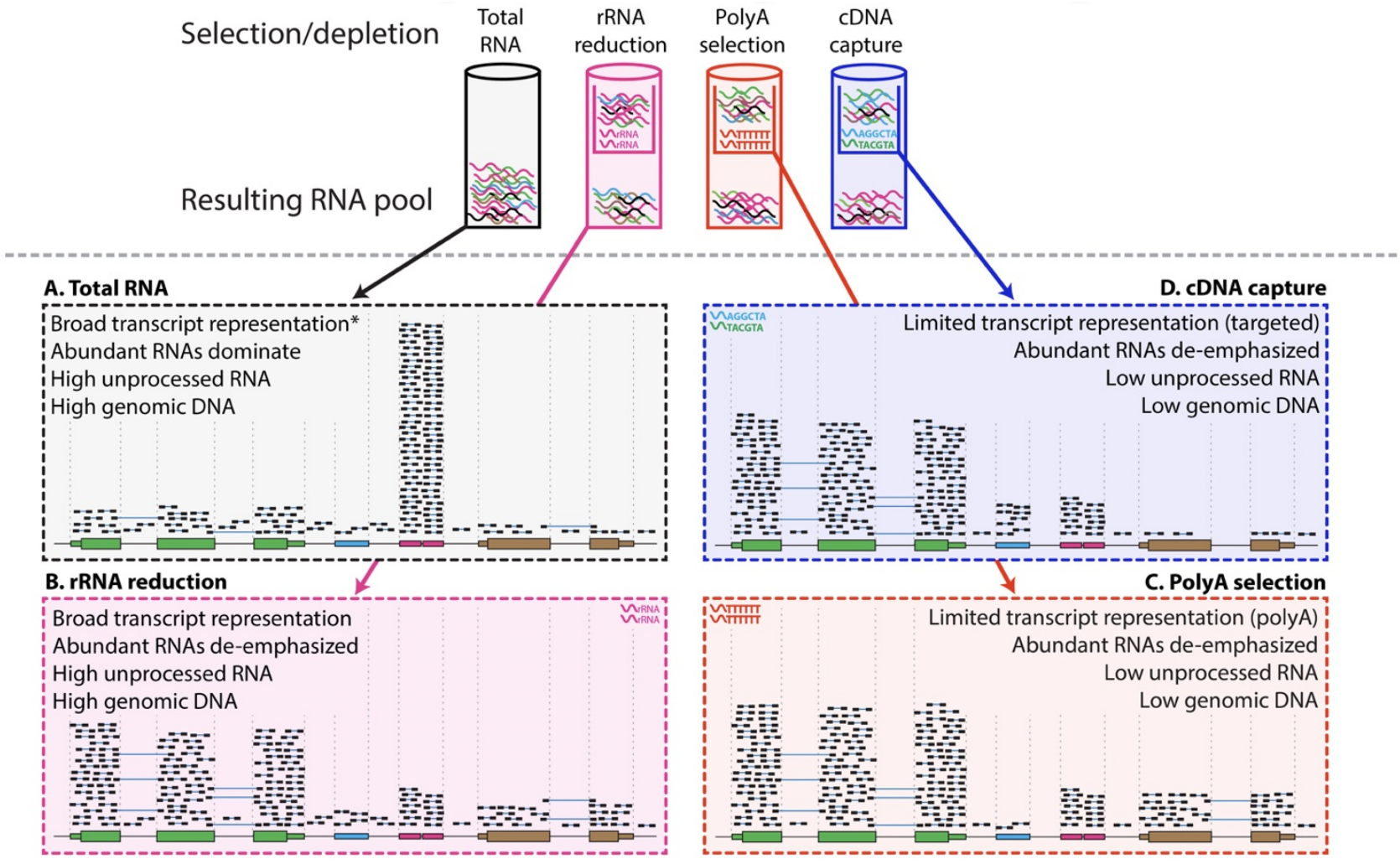
Strategies

A few examples:

- Total RNA versus polyA+ RNA
- Ribo-reduction
- Size selection (before and/or after cDNA synthesis)
- Small RNAs (microRNAs) vs. large RNAs
- Stranded vs. un-stranded libraries

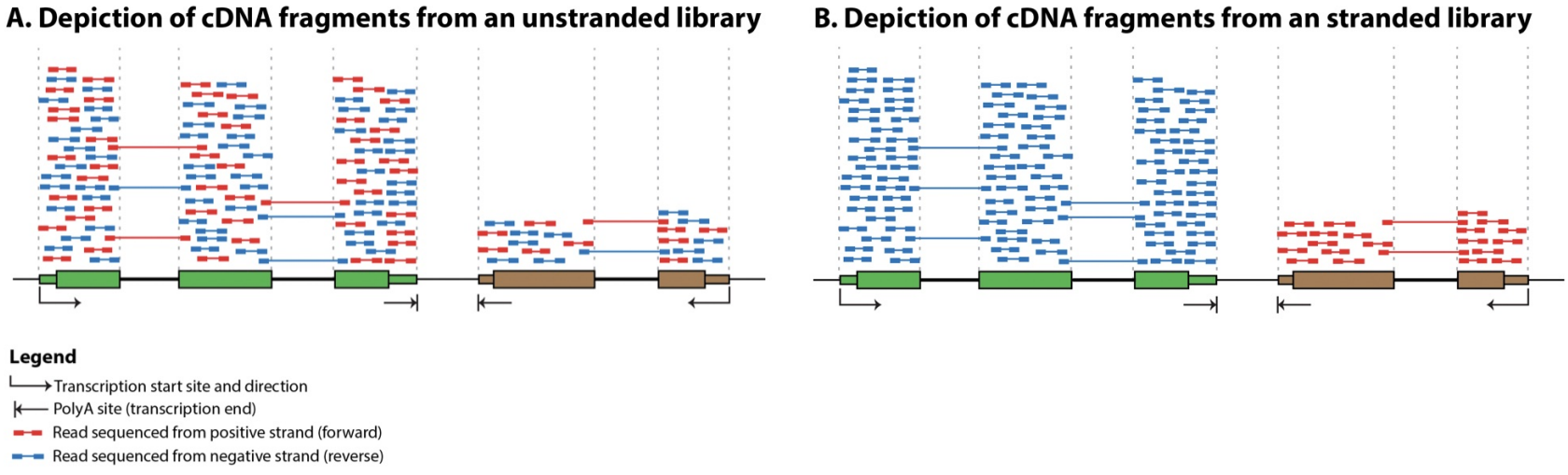
RNA-seq: Library preparation

Strategies



RNA-seq: Library preparation

Strategies



RNA-seq: Library preparation

Strategies

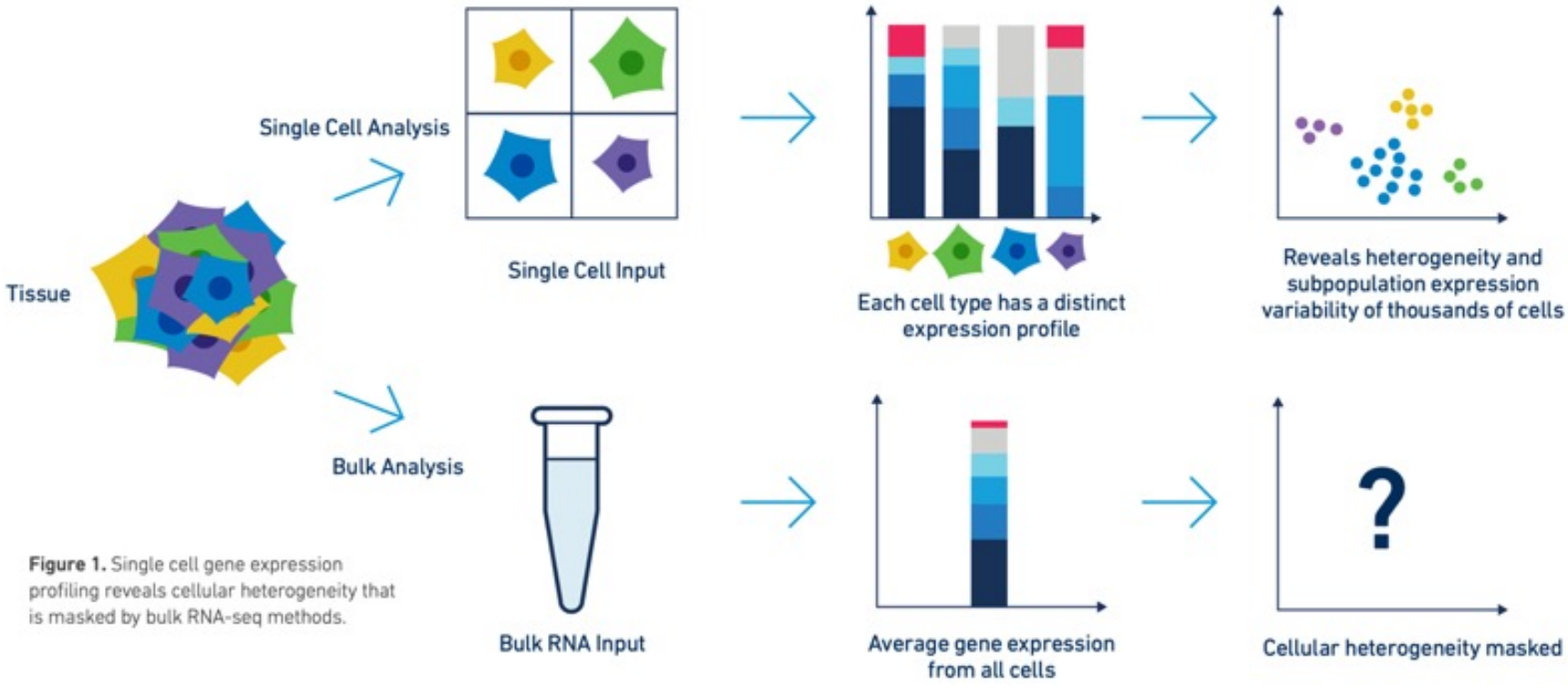
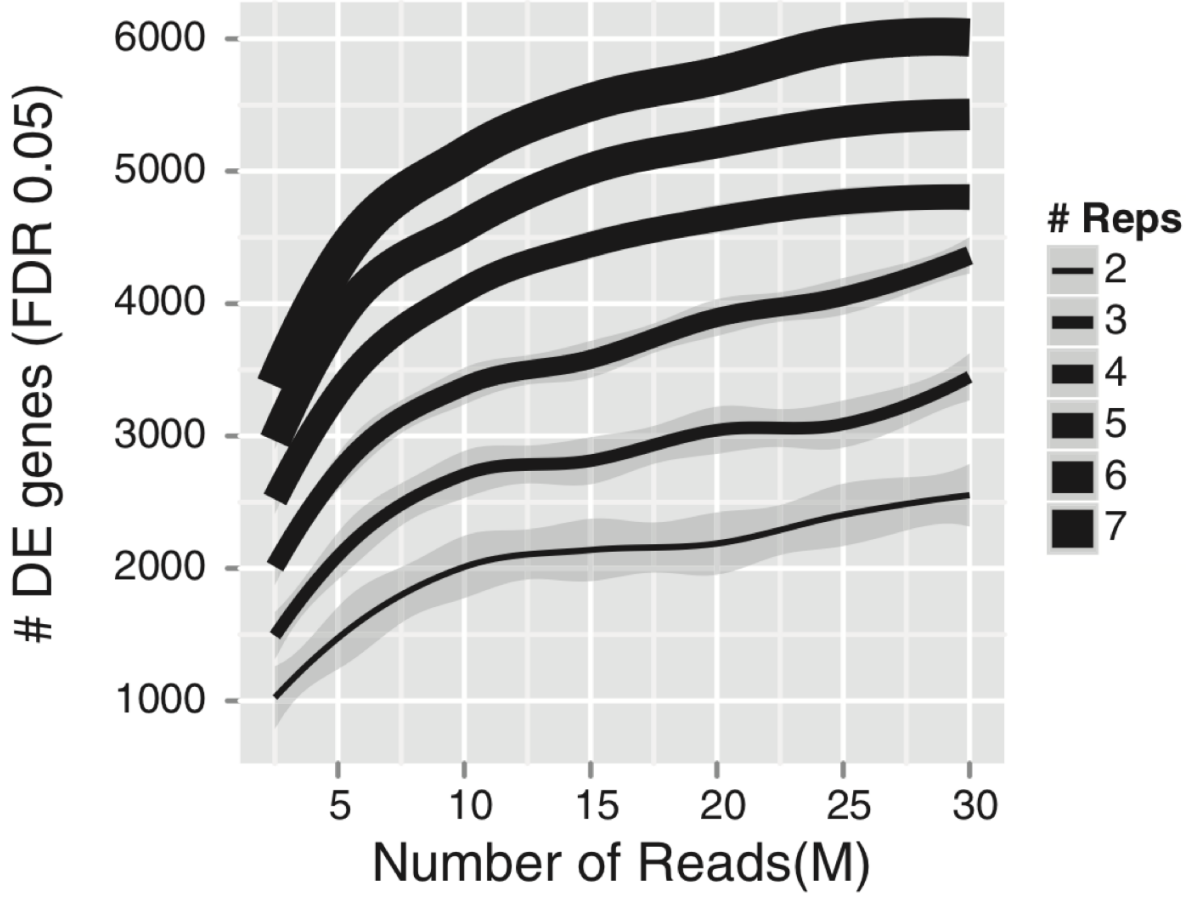


Figure 1. Single cell gene expression profiling reveals cellular heterogeneity that is masked by bulk RNA-seq methods.

RNA-seq: Sequencing

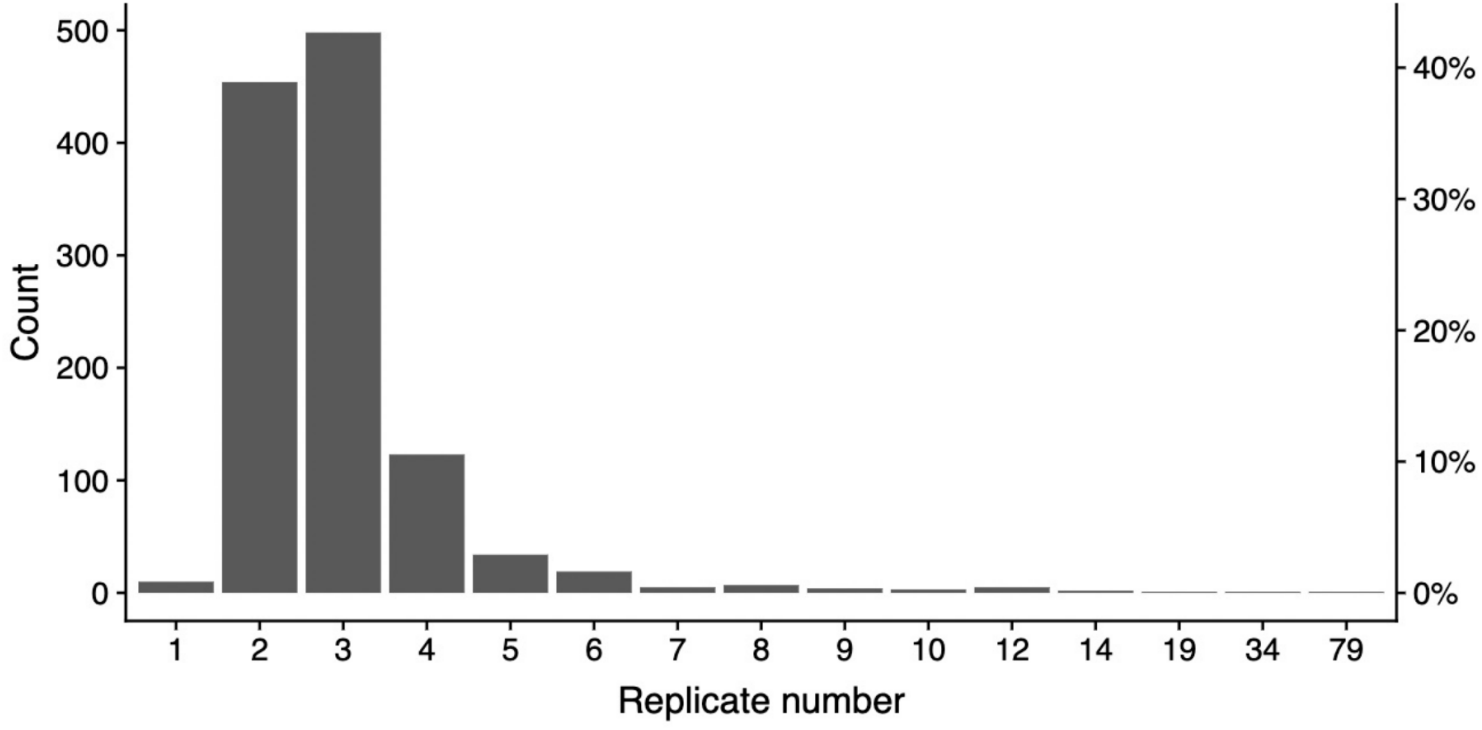
Number of replicates vs. number of reads (depth)



Liu et al (2014) Bioinformatics, 30, 301-304

RNA-seq: Sequencing

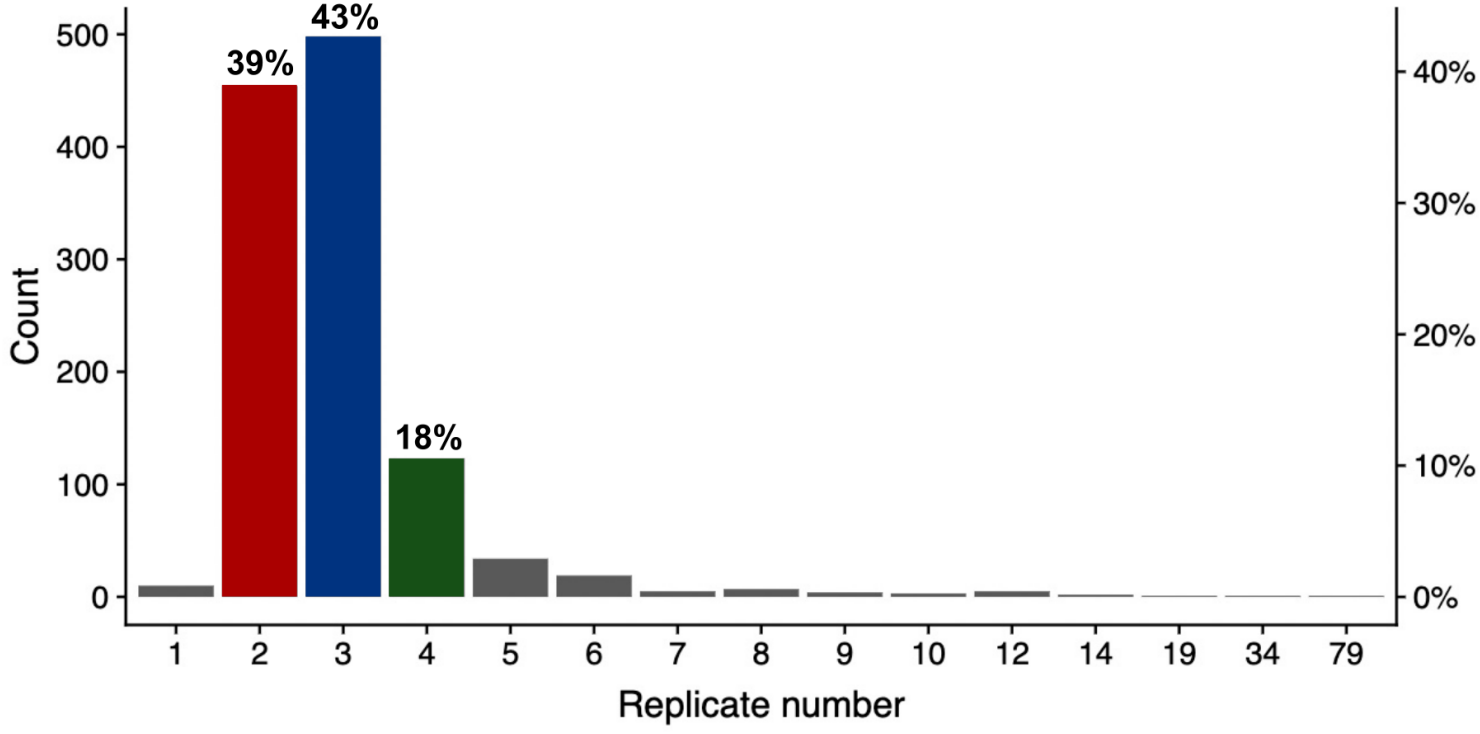
Number of replicates



Li et al (2020) Genome Biology, 21, 66

RNA-seq: Sequencing

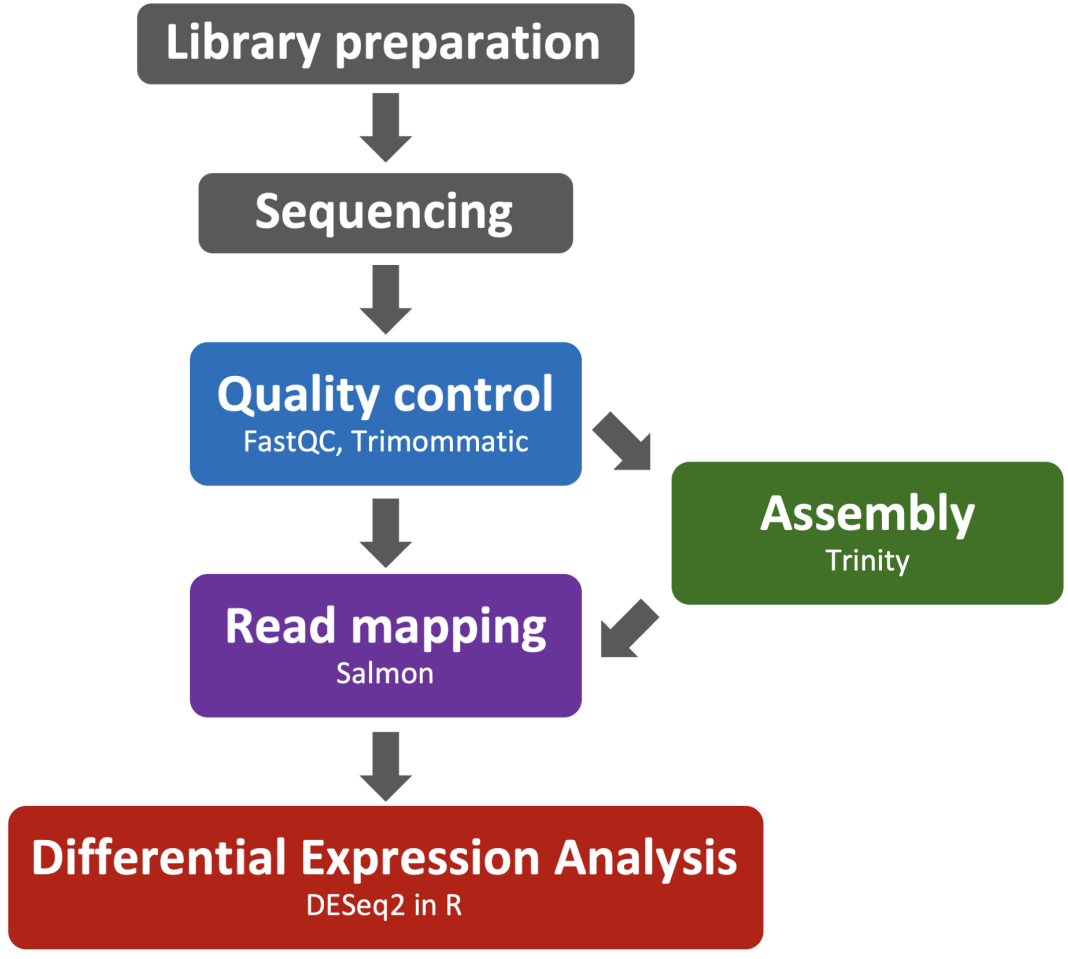
Number of replicates



Li et al (2020) Genome Biology, 21, 66

RNA-Seq

Workflow



RNA-Seq

Workflow