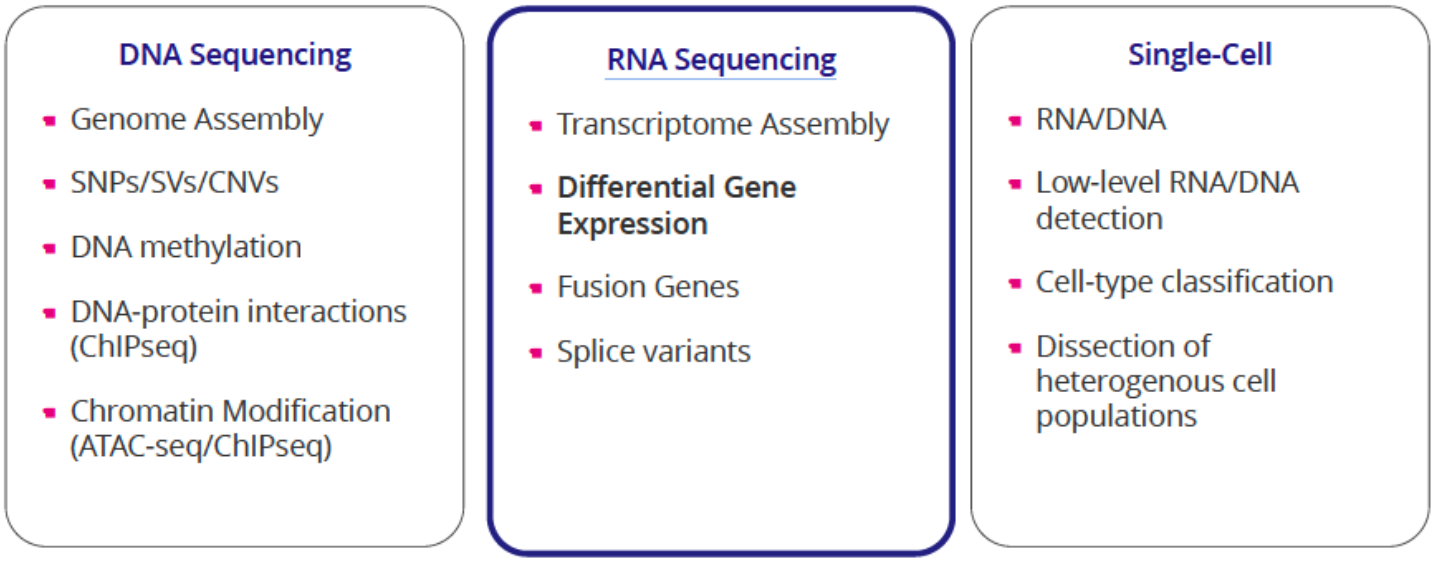
STAT718/BIOL703: Genomic Data Science
Introduction to RNAseq Methods & Experimental Design

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

High-throughput Sequencing Applications: An Overview

- Introduction
- Experiment Design
- RNAseq Workflows

High-throughput Sequencing Applications-Overview



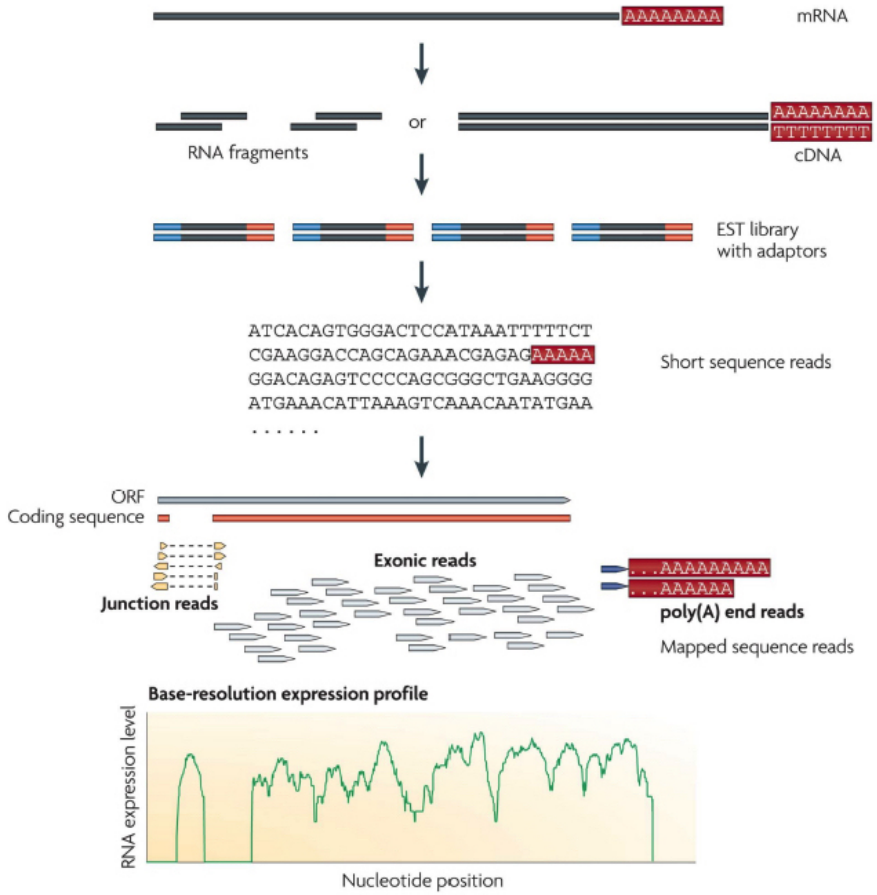
RNAseq Workflow[1]

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis



[1]Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

Designing the Right Experiment

A good experiment should:

- Have clear objectives
- Have sufficient power
- Be amenable to statistical analysis
- Be reproducible

Designing the Right Experiment

Practical considerations for RNAseq

- Coverage: how many reads?
- Read length & structure: Long or short reads? Paired or Single end?
- Controlling for batch effects
- Library preparation method: Poly-A, Ribominus, other?

Designing the Right Experiment: How Many Reads Do We Need?

The coverage is defined as:

$$\frac{\textit{Read Length} \times \textit{Number of Reads}}{\textit{Length of Target Sequence}}$$

- For a general view of differential expression: 5–25 million reads per sample
- For alternative splicing and lowly expressed genes: 30–60 million reads per sample.
- In-depth view of the transcriptome/assemble new transcripts: 100–200 million reads
- Targeted RNA expression requires fewer reads.
- miRNA-Seq or Small RNA Analysis require even fewer reads.

Designing the Right Experiment: Read Length

Long or short read? Paired or Single end? The answer depends

- Gene expression: typically just a short read e.g. 50/75 bp; SE or PE
- kmer-based quantification of Gene Expression (Salmon etc.) - benefits from PE.
- Transcriptome Analysis – longer paired-end reads (such as 2 x 75 bp).
- Small RNA Analysis – short single read, e.f. SE50 - will need trimming.

Designing the Right Experiment

Biological Replication

- Measures the biological variations between individuals
- Accounts for sampling bias

Technical Replication

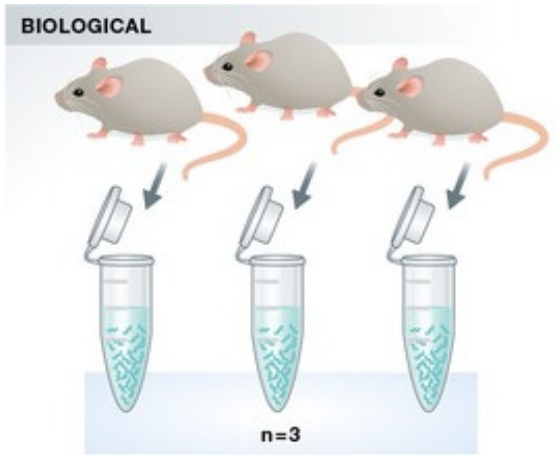
- Measures the variation in response quantification due to imprecision in the technique
- Accounts for technical noise

Designing the Right Experiment

Biological Replication

Each replicate is from an independent biological individual

- In Vivo
  - Patients
  - Mice
- In Vitro
  - Different cell lines
  - Different passages



Designing the Right Experiment

Technical Replication: replicates are from the same individual but processed separately

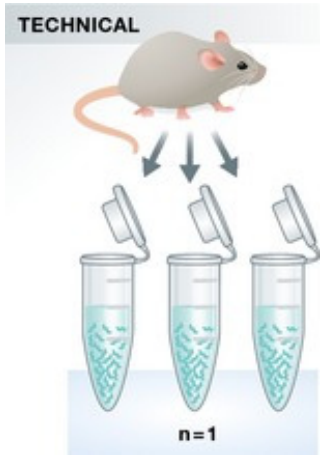
- Experimental protocol
- Measurement platform

RNA Sequencing

- Transcriptome Assembly
- **Differential Gene Expression**
- Fusion Genes
- Splice variants

Single-Cell

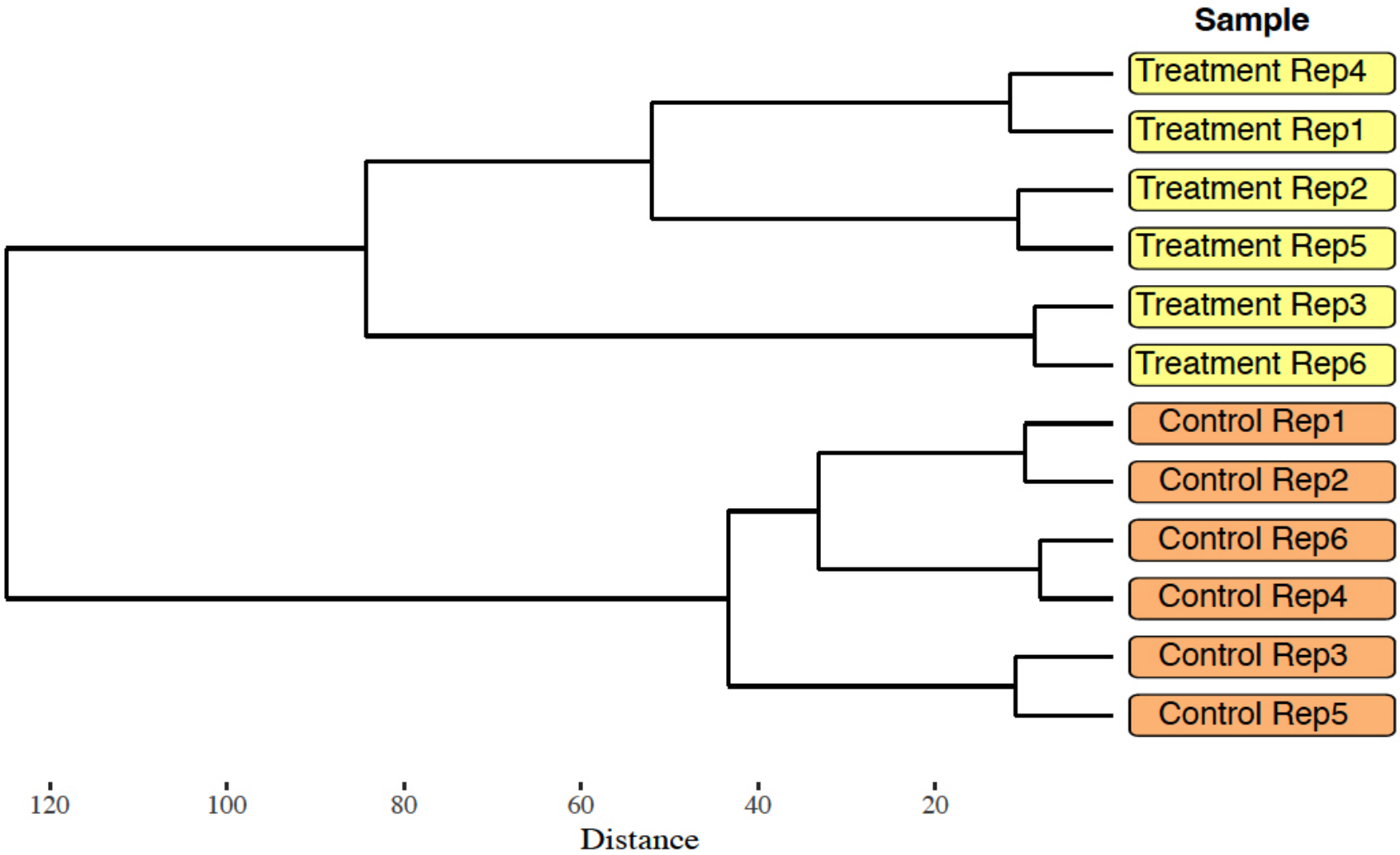
- RNA/DNA
- Low-level RNA/DNA detection
- Cell-type classification
- Dissection of heterogenous cell populations



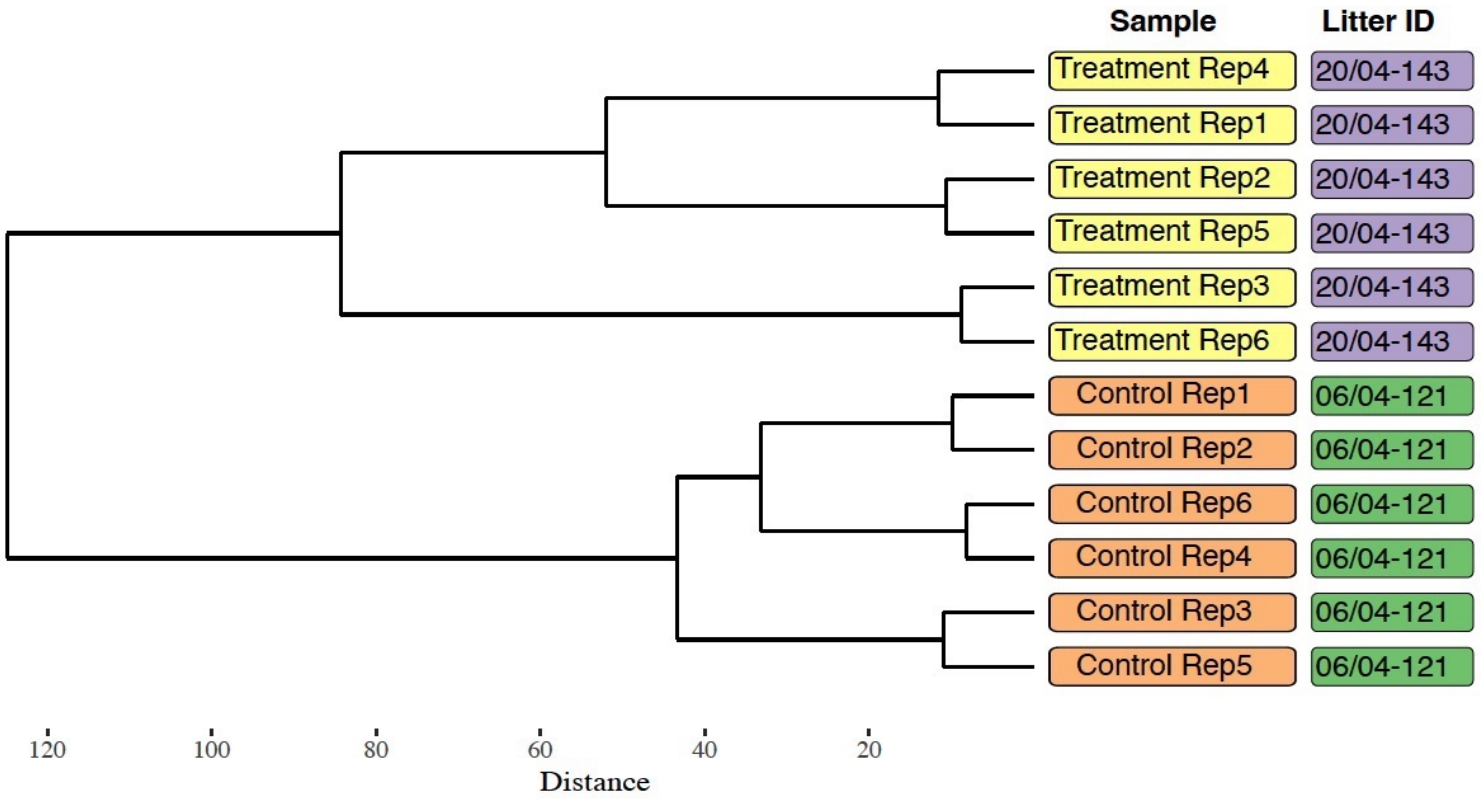
Designing the Right Experiment

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.

Designing the Right Experiment



Designing the Right Experiment

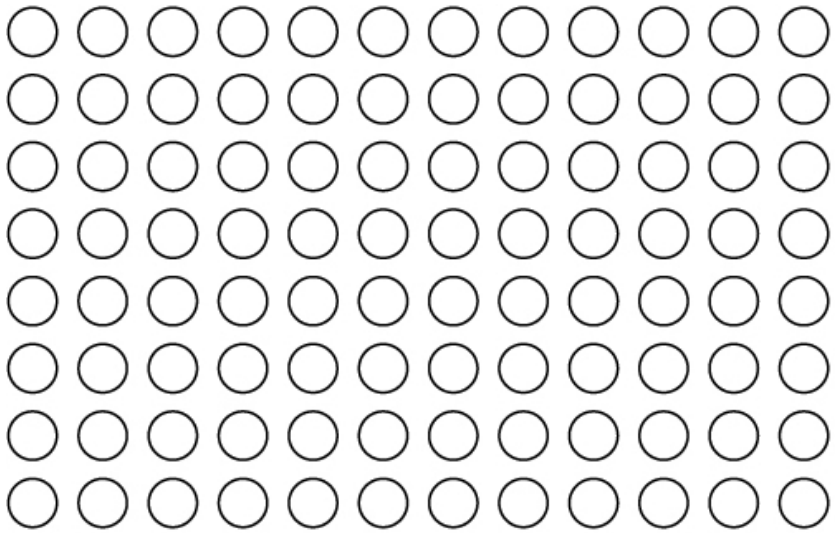


Designing the Right Experiment: Batch Effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.

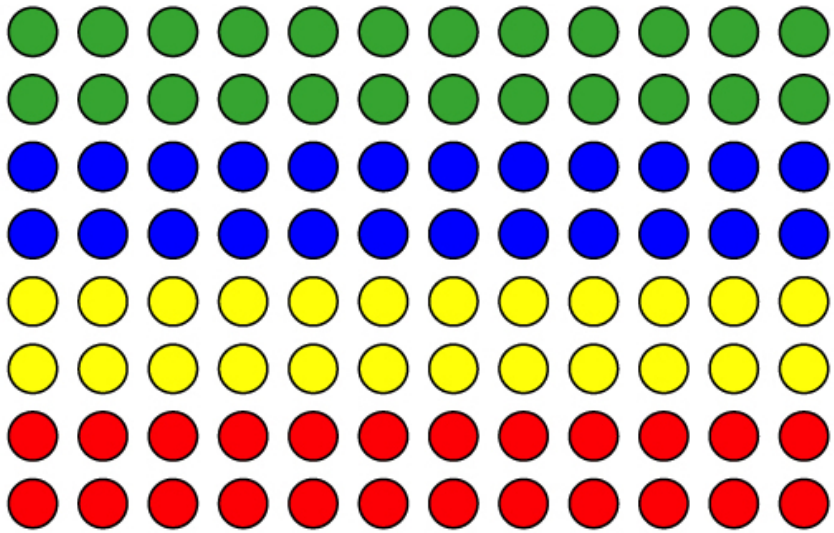
Designing the Right Experiment: Batch Effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.



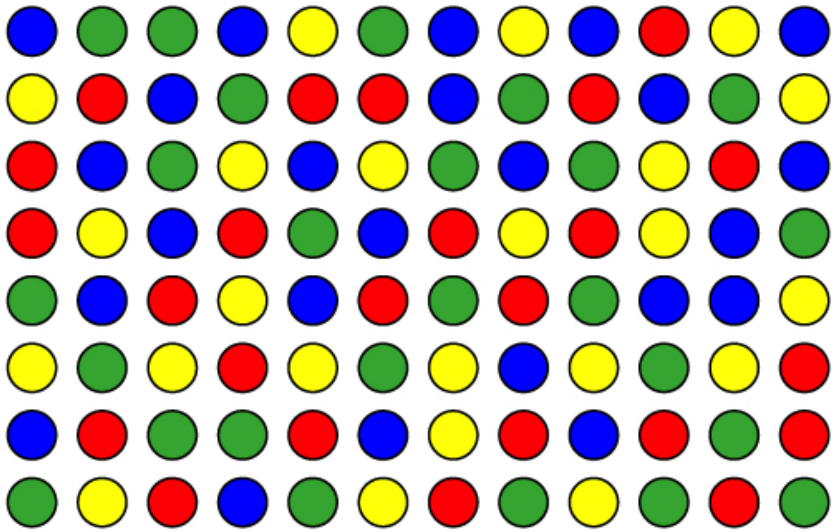
Designing the Right Experiment: Batch Effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.



Designing the Right Experiment: Batch Effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.



Designing the Right Experiment: Batch Effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.
- **Record everything**: Age, sex, litter, cell passage ..

RNAseq Workflow

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis

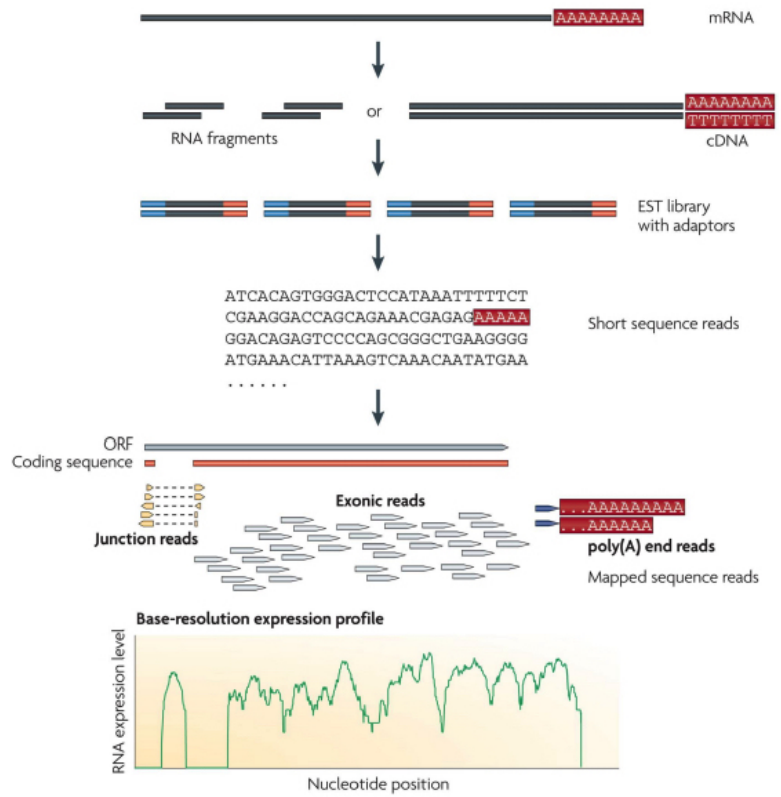
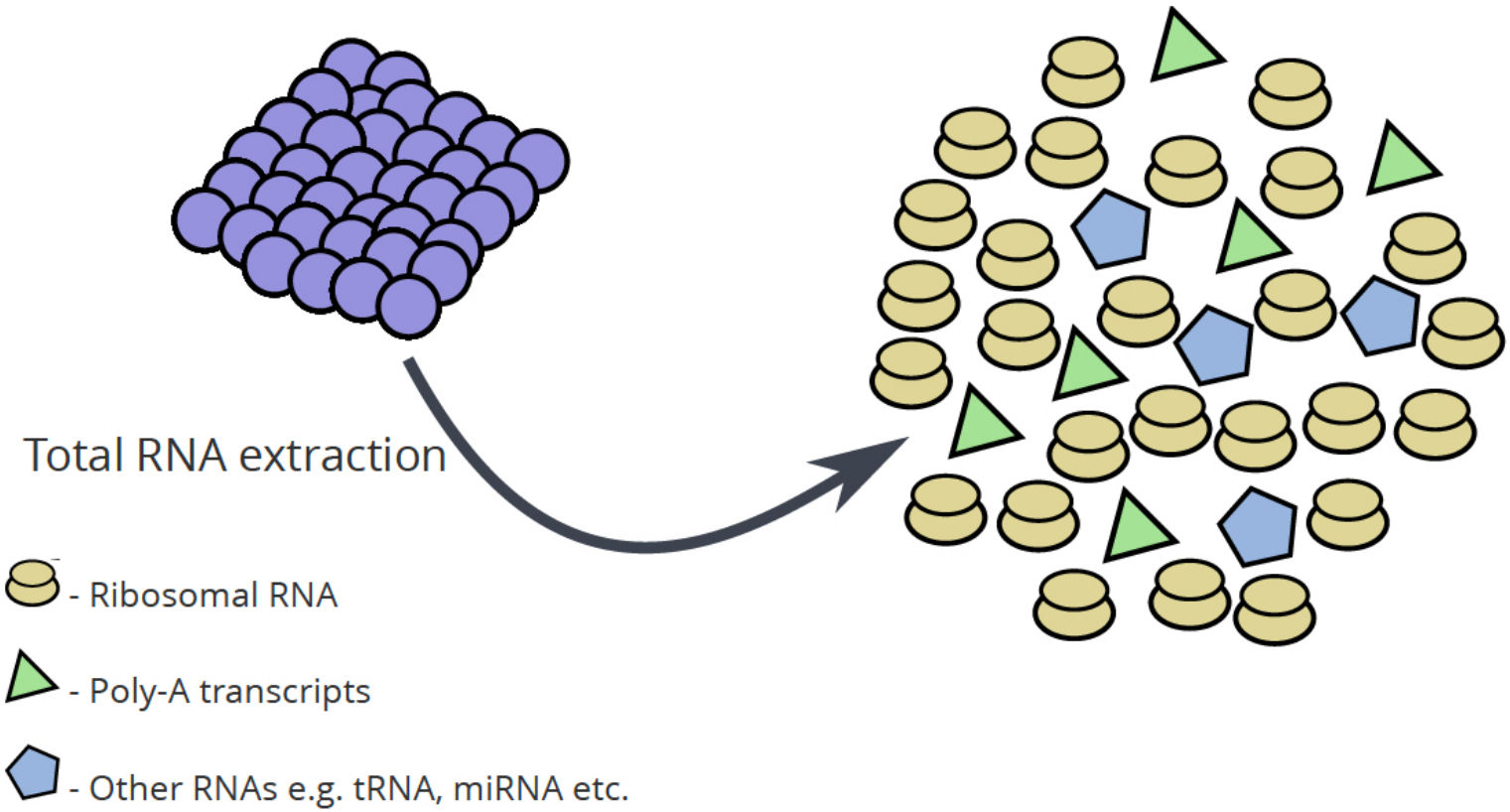


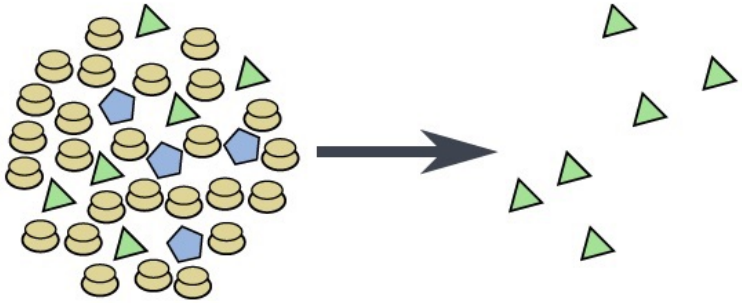
Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

Library Preparation



Library Preparation

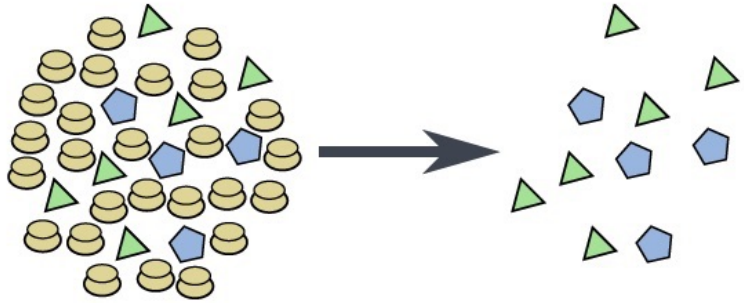
Poly-A Selection



Poly-A transcripts e.g.:

- mRNAs
- immature miRNAs
- snoRNA

Ribominus selection



Poly-A transcripts + Other mRNAs e.g.:

- tRNAs
- mature miRNAs
- piRNAs

RNAseq Workflow

- Experimental Design
- Library Preparation
- Sequencing
- Bioinformatics Analysis

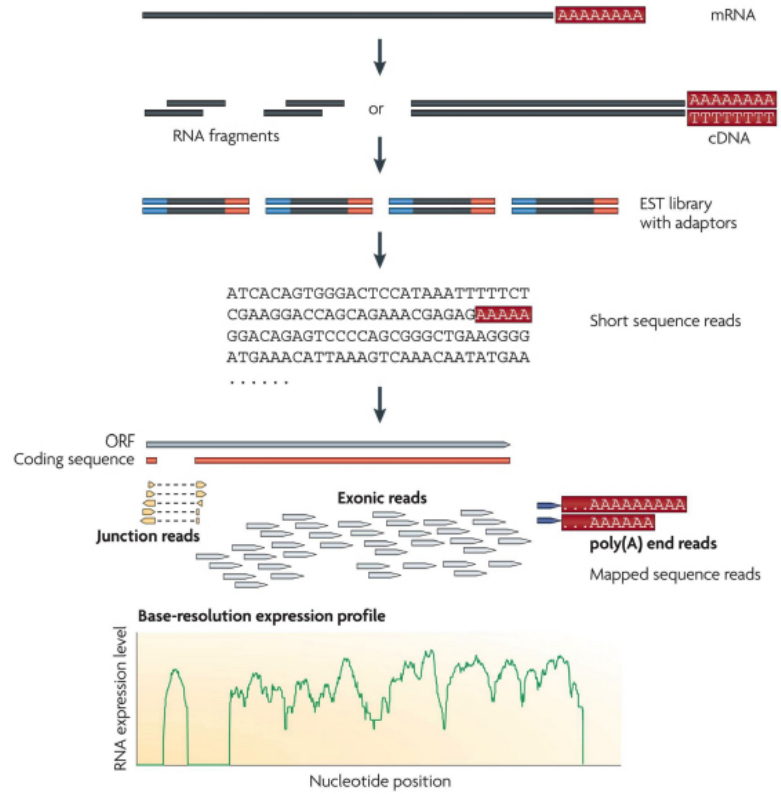


Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

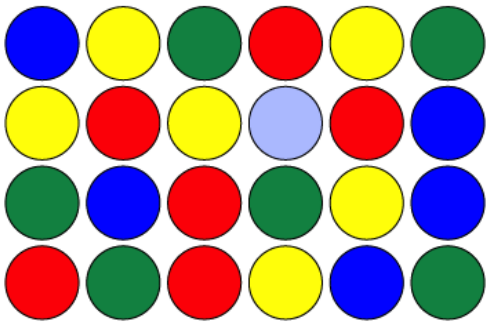
Sequencing by synthesis

- A complimentary strand is synthesized using the cDNA fragment as template.
- Each nucleotide includes a fluorescent tag and as the new strand is synthesized, the color of the fluorescence indicates which base is being added.
- The sequencer records the order of these flashes of light and translates them to a base sequence.

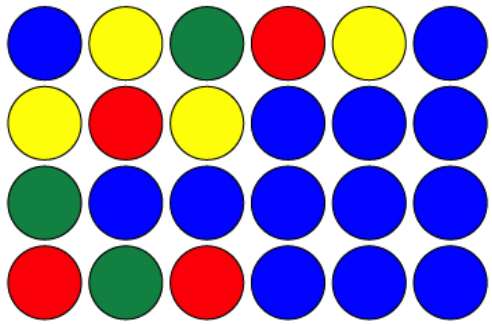
Sequencing by synthesis

Sequencing errors cause uncertainty in calling the nucleotide at a given location. These reductions in confidence would be reflected int he quality scores in your fastq output.

If a probe doesn't shine as bright as it should, the sequencer is less confident in calling that base.



If there are lots of probes the same colour in the same region the sequencer finds it harder to identify the individual reads.



Differential Gene Expression Analysis Workflow