SBCNY

NIGMS funded Center

Experimental Methods in Systems Biology

Part of the Coursera Certificate in Systems Biology

Marc Birtwistle, PhD
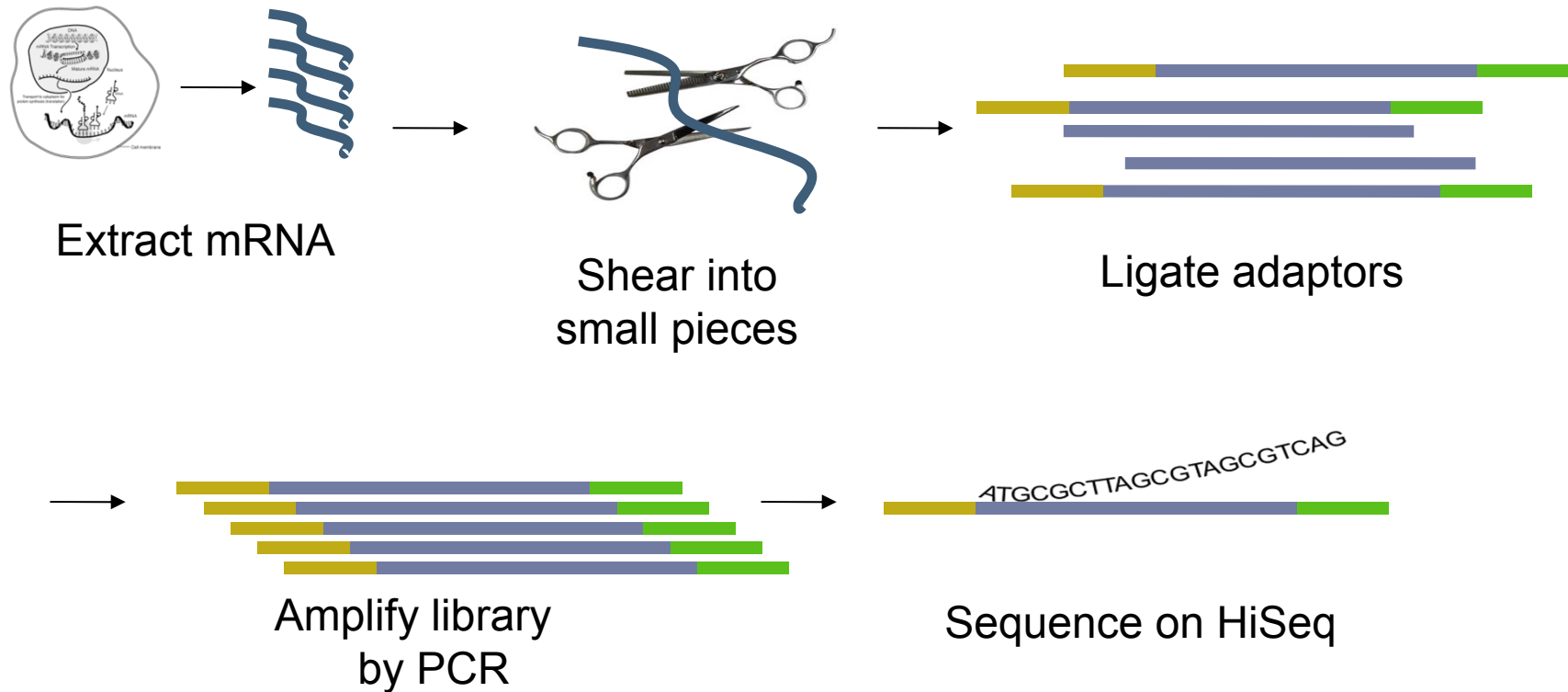Department of Pharmacology & Systems Therapeutics
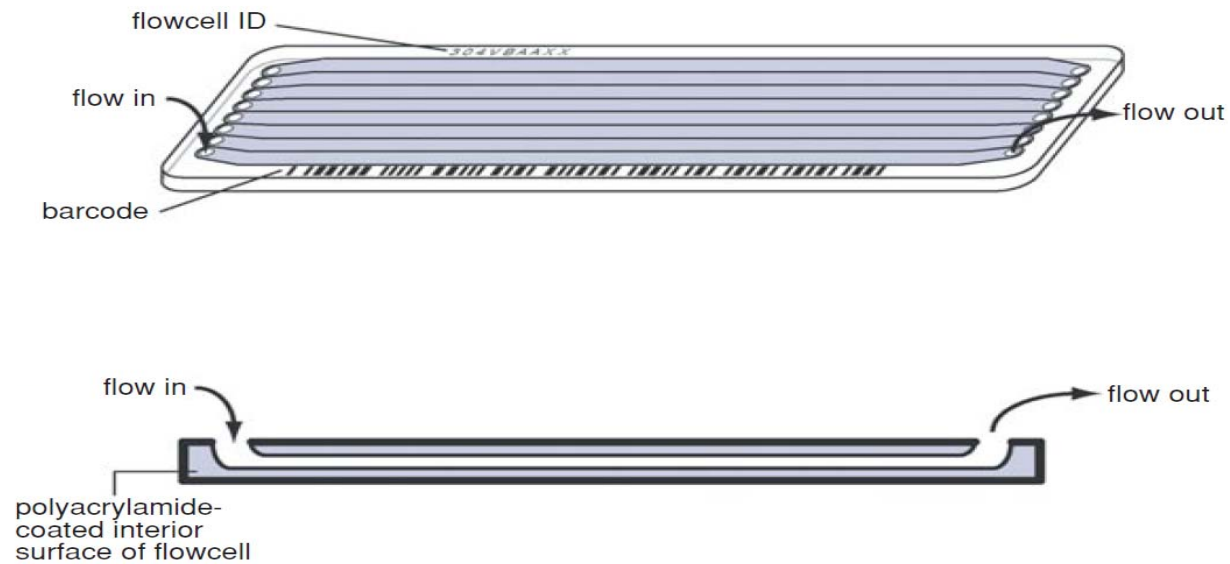Fall 2014, Week 2, Deep mRNA Sequencing

Icahn School
of Medicine at
Mount
Sinai

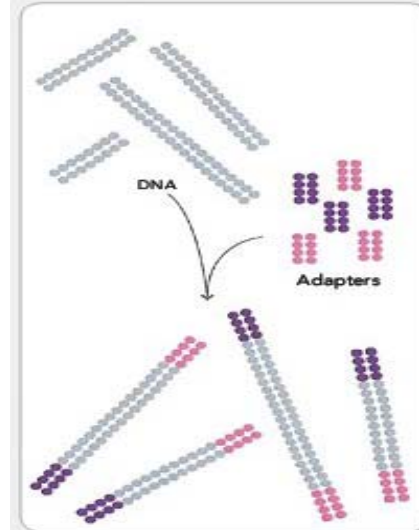# The Illumina Procotol

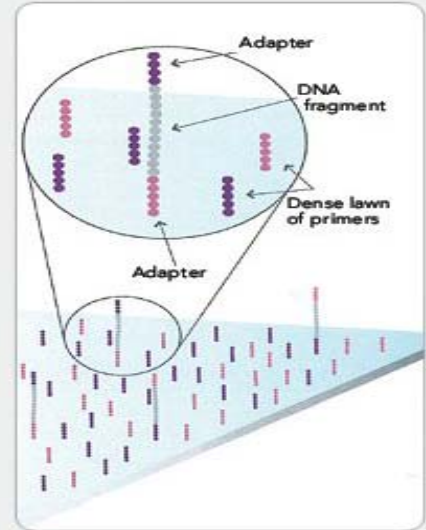# A typical Illumina mRNA sequencing process

Extract mRNA

Shear into
small pieces

Ligate adaptors

Amplify library
by PCR

ATGCGCTTAGCGTAGCGTCAG

Sequence on HiSeq

Illumina sequencing

Layering Seq library on the chip (flowcell) surface

Preparation of DNA on flowcell
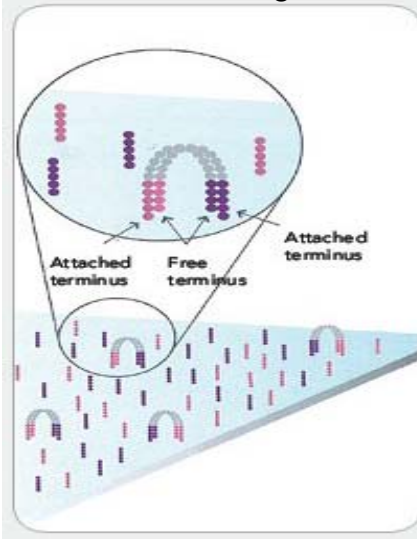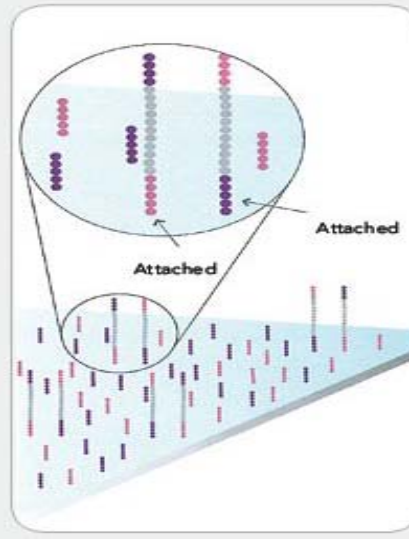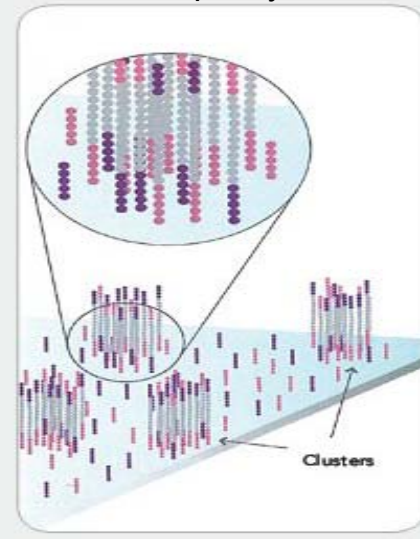
Isothermal "Bridge PCR" with immobilized primers and chemical denaturation

Visualization of clusters

Cluster growth

100um
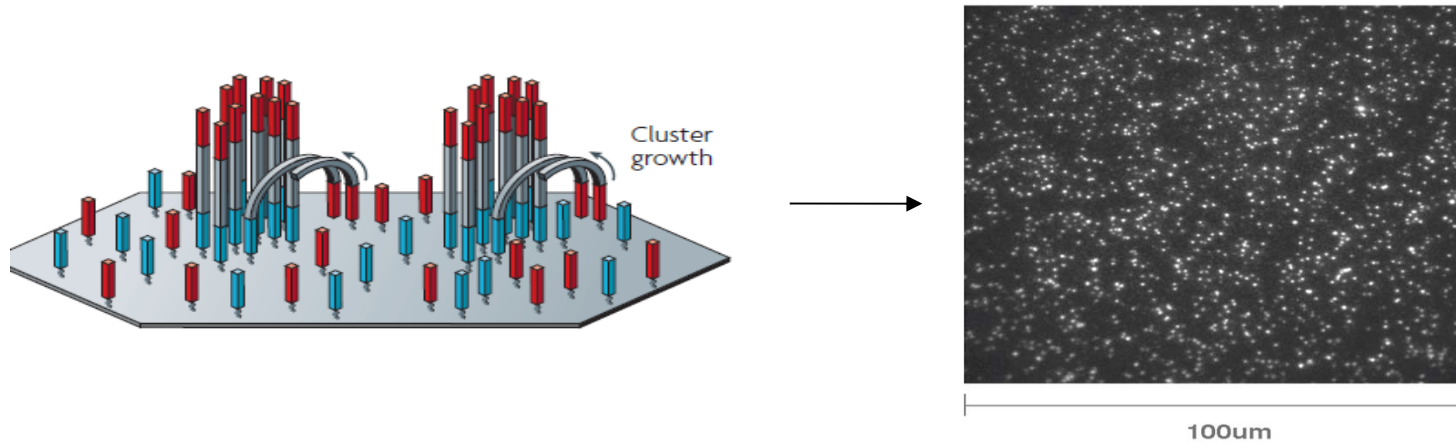
- Position of initial strand is random
- Must avoid overlapping bridges
- Length of library strand matters

Sequencing reaction

Incorporate

Rinse & Image

Cleave fluorophore, deblock
3' end & repeat cycle

Repeat cycles
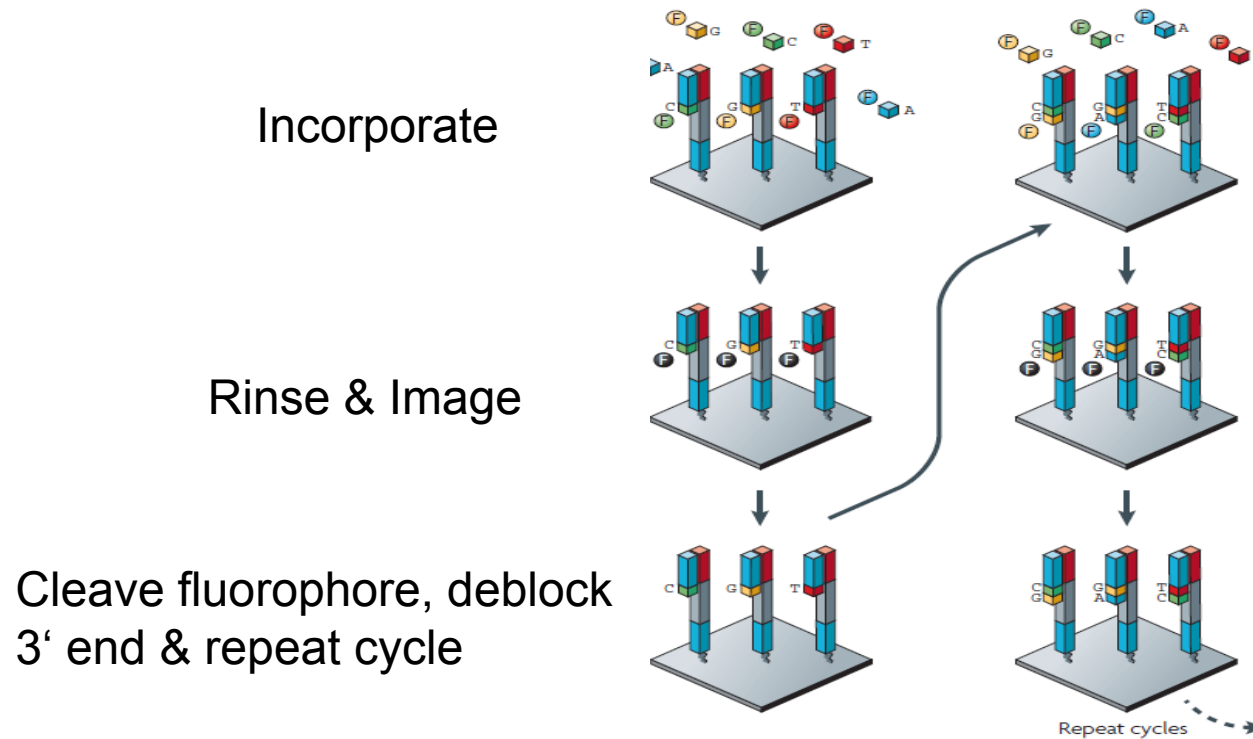
Read length is the
number of repeats.
The whole fragment
is usually not
sequenced
(fragment length >
read length)

8

Single-end vs. Paired-end

Reference genome

Read 1

Read 2

Paired end gives a constraint on length & position
For mRNAseq gives great advantage for alternative splicing analysis
However it doubles the cost and running time
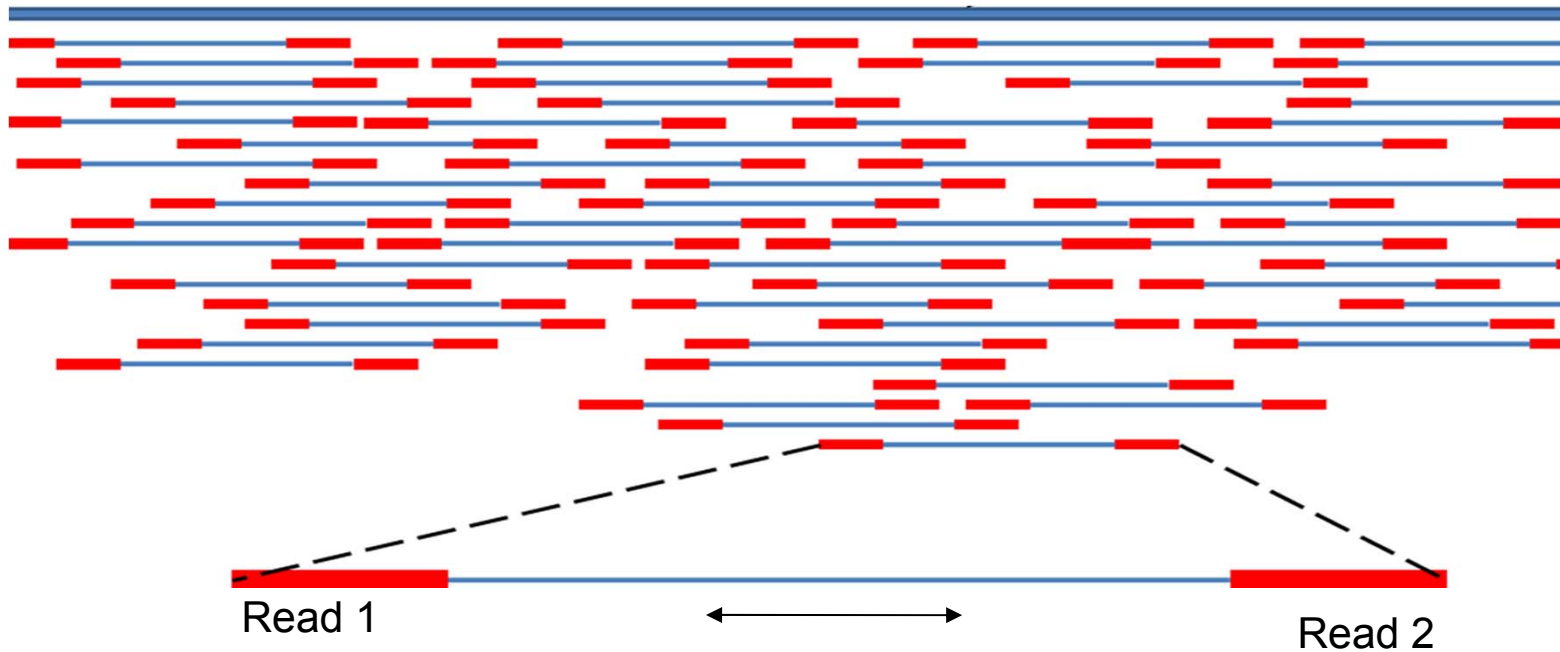
Illumina Seq Data analysis: Overview

Data collection images → **Image analysis intensities** → **Base Calling** → **Alignment/counting** → **Biologically Meaningful Results**

Usually done with platform-specific software

Usually done with open source academic software: e.g. tophat/cufflinks/cuffdiff
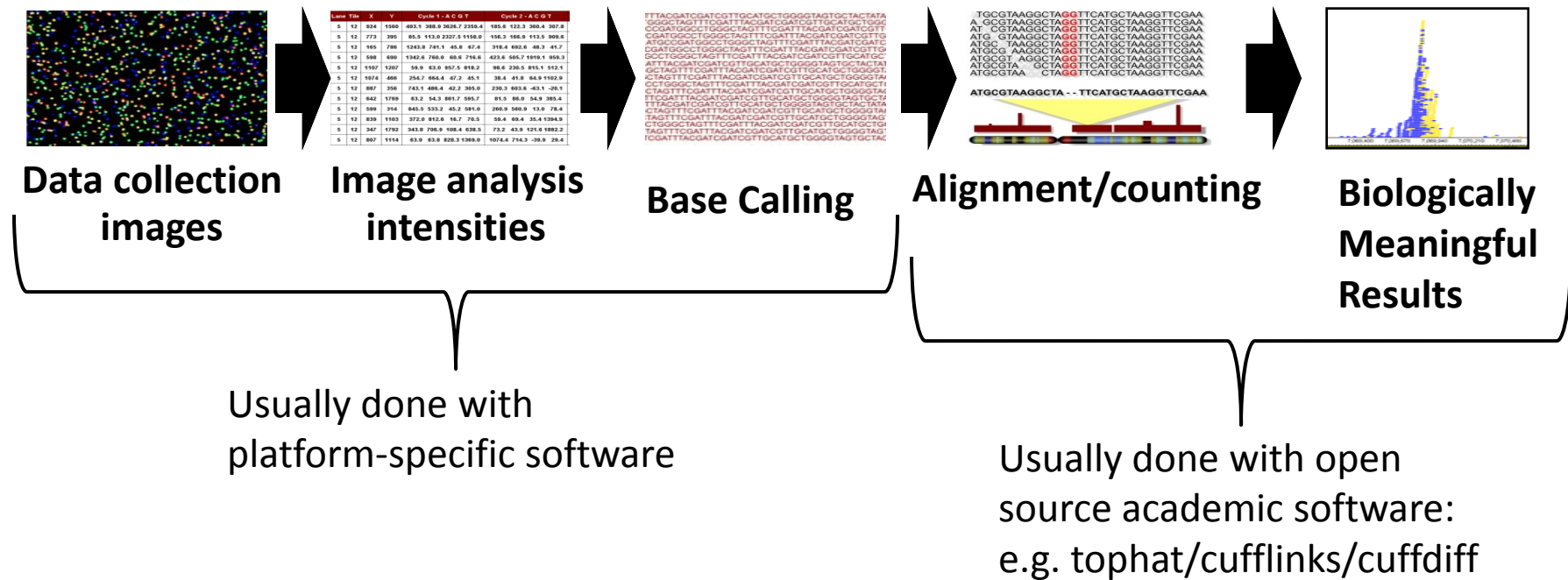
mRNA Seq Quantification

- Counts
  - Simply count the number of reads that align to a transcript
  - Biased by transcript length
  - Not comparable between experiments
    - Different sequencing depth (number of clusters identified, sequenced and mapped)
- RPKM—Reads per kilobase of transcript length per million mapped reads
  - Addresses the above problems with pure count data
- FPKM—Fragments per kilobase of transcript length per million mapped reads
  - Applicable for paired end experiments
  - Sometimes only one end of a read is uniquely assignable; FPKM corrects for this
- Cufflinks software will calculate transcript abundances with uncertainties
  - Uncertainties arise from randomness of where reads fall in relation to splicing junctions
  - The mathematical model underlying cufflinks calculations is well-built and an excellent read but is outside the scope of this course (buried in the supplement of Trapnell et al., Nat Biotech, 2008)
  - They have updated it to attempt to also correct for biases in the PCR steps and capture steps
- There are many statistical nuances of differential expression analysis which are also outside the scope of this course but Cuffdiff software addresses many of them.