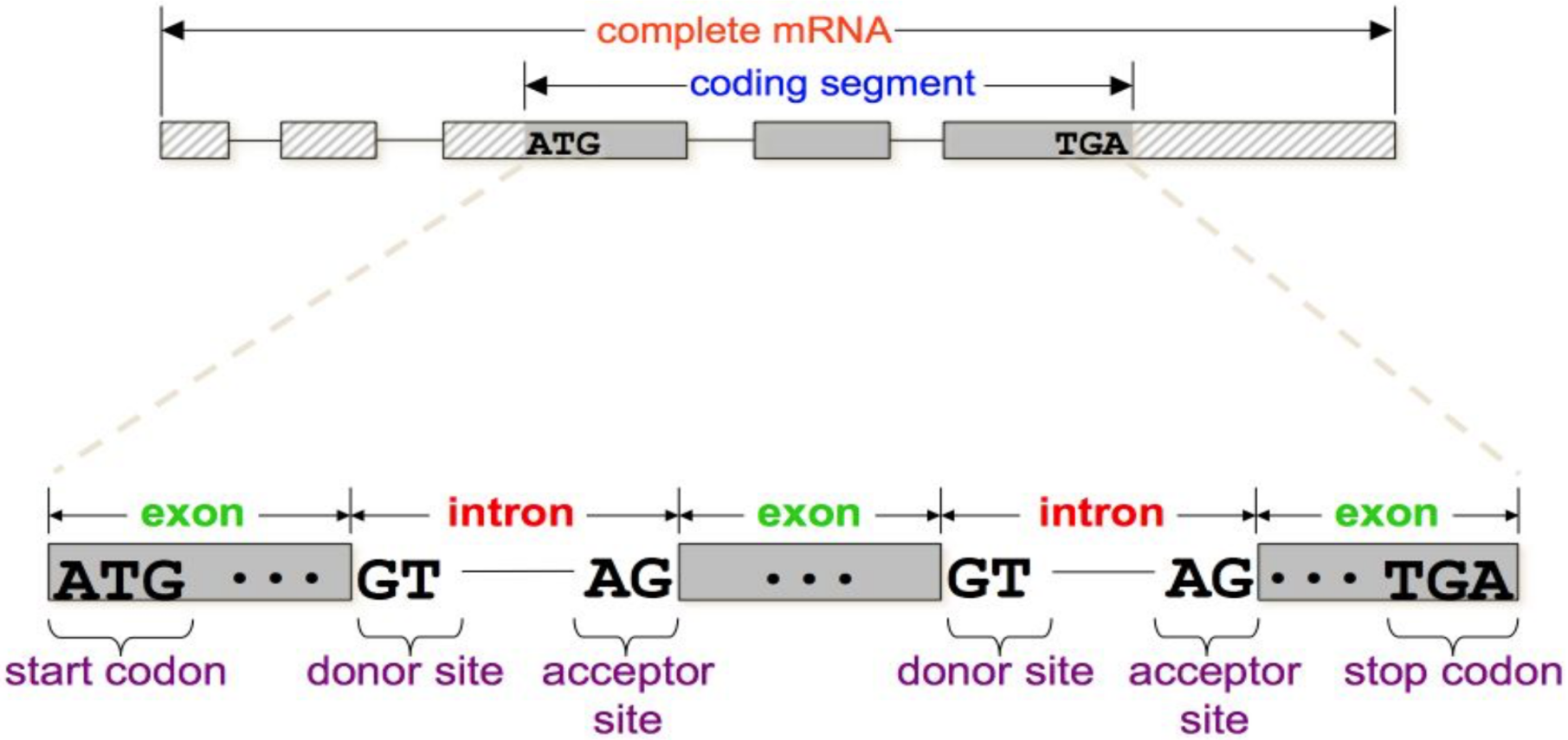
RNA-seq Introduction

Mikhail Dozmorov

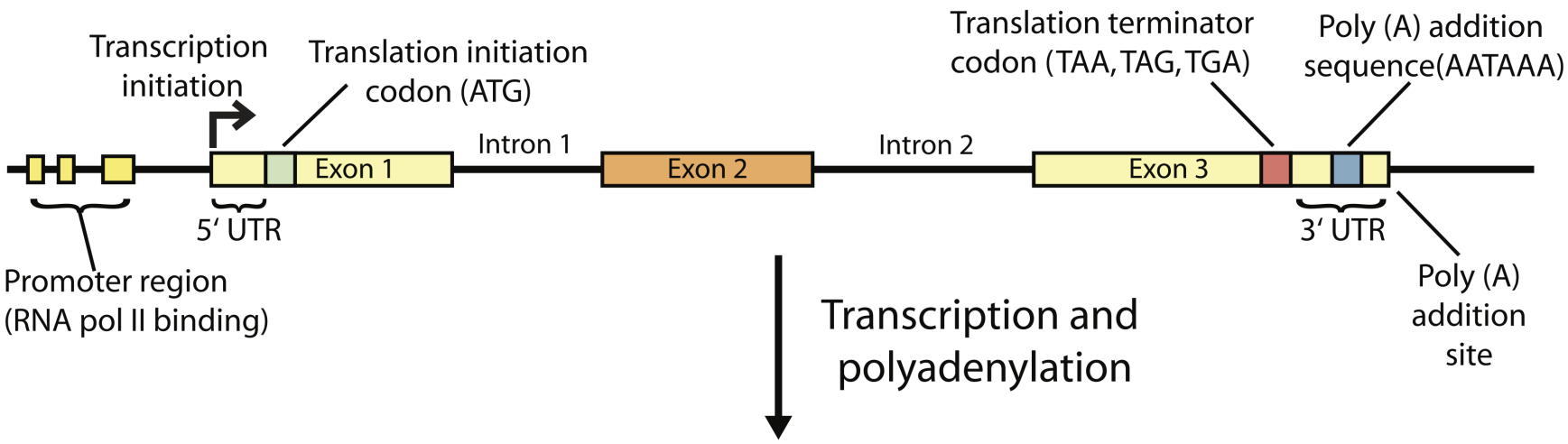
Spring 2018

Eukaryotic gene structure



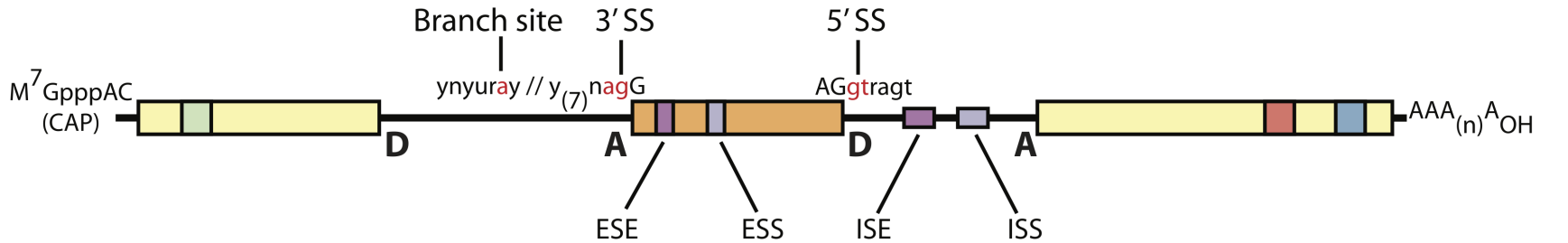
Gene expression

Double-stranded genomic DNA template



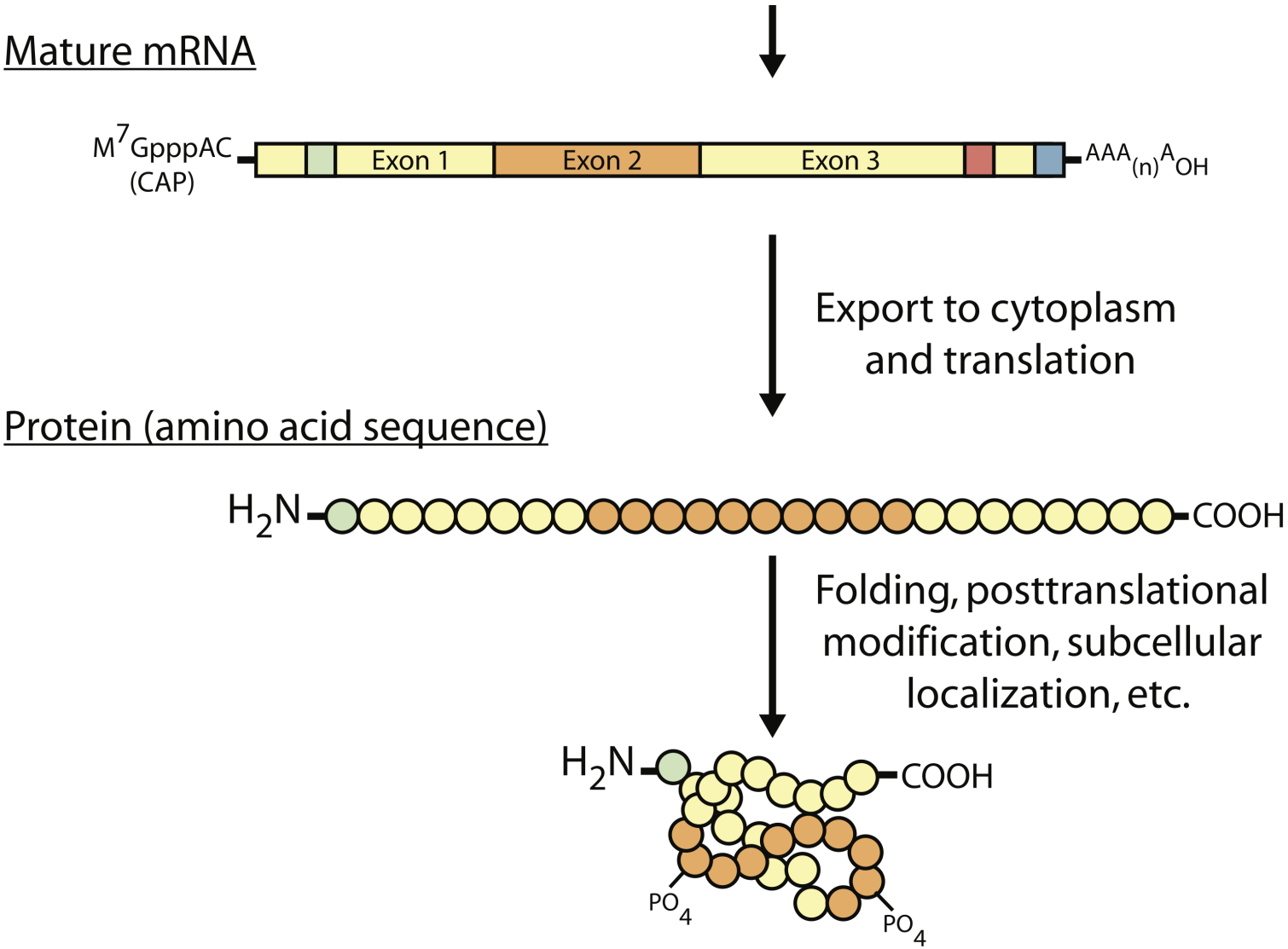
Transcription and polyadenylation

Single-stranded pre-mRNA (nuclear RNA)



RNA processing

Gene expression



What is RNA sequencing?

- Massive parallel sequencing to **characterize and quantify transcriptomes** (all actively transcribed genes)
- Detection of **differential gene expression**
- **Transcriptome reconstruction**, identification of **new transcripts**
- Detection of **alternative splicing events**
- Detection of **structural variants**, e.g., fusion transcripts
- **Allele-specific** gene expression measurements
- **Mutation analysis** – presence of genomic mutations and their effect on gene expression

http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s003

Sequencing technologies

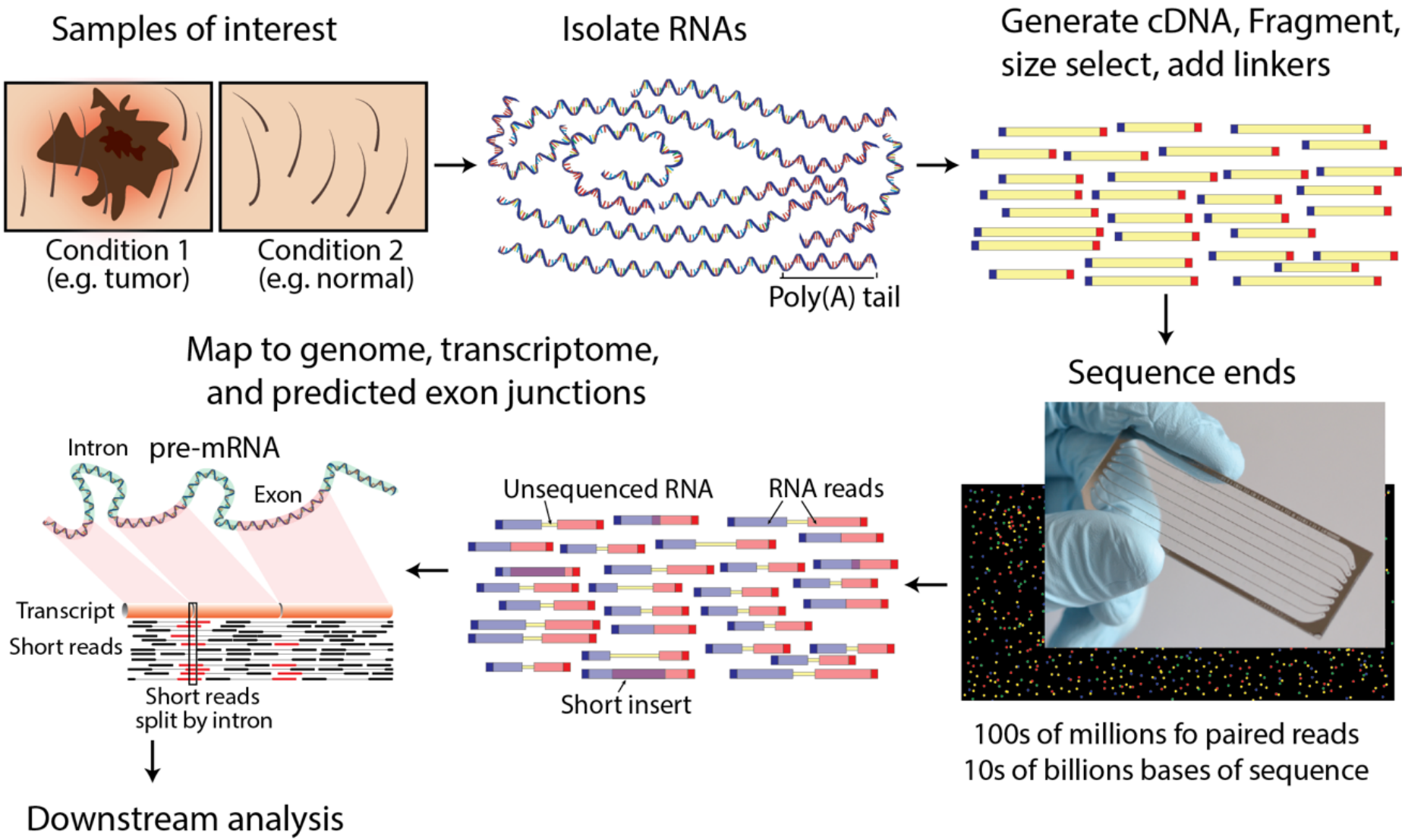
Commercially available

- **Illumina/Solexa** - short reads, sequencing-by-synthesis
- **Life Technologies Ion Torrent/Proton** - short reads, Ion Semiconductor sequencing
- **Pacific Biosciences** - long reads, Single Molecule Real Time sequencing

Experimental

- **Nanopore sequencing** - continuous sequencing (very long reads), fluctuations of the ionic current from nucleotides passing through the nanopore

Overview of RNA sequencing technology



Source: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

Advantage of RNA-Seq over Microarray

- Much richer information beyond quantitation
  - Boundary of gene transcripts: both 5' and 3' end, to nucleotide level
  - Alternative exon usage, novel splicing junction detection
  - SNP/indel discovery in transcripts: both coding and UTRs
  - Allele specific expression: critical in imprinting, cancer

- Not relying on gene annotation by mapping to the whole genome
  - No longer biased by probe design
  - Novel gene and exon discovery enabled

Advantage of RNA-Seq over Microarray

- Better performance at quantitation
  - Unlimited dynamic range: by increasing depth as needed
  - Higher specificity and accuracy: digital counts of transcript copies, very low background noise
  - Higher sensitivity: more transcripts and more differential genes detected

- Re-analysis easily done by computation, as gene annotation keeps evolving
- *De novo* assembly possible, not relying on reference genome sequence
- Comparable cost, continuing to drop

RNA-Seq Limitations

Quantitation influenced by many confounding factors

- "Sequenceability" - varying across genomic regions, local GC content and structure-related
- Varying length of gene transcripts and exons
- Bias in read ends due to reverse transcription, subtle but consistent
- Varying extent of PCR amplification artifacts
- Effect of RNA degradation in the real world
- Computational bias in aligning reads to genome due to aligners

RNA-Seq Limitations

SNP discovery in RNA-seq is more challenging than in DNA

- Varying levels of coverage depth
- False discovery around splicing junctions due to incorrect mapping

De novo assembly of transcripts without genome sequence: computationally intensive but possible, technical improvements will help

- Longer read length
- Lower error rate
- More uniform nucleotide coverage of transcripts - more equalized transcript abundance