Transcriptomics (I)

BBMS 3009: Genome Science (First Semester, 2021)

Dr. Yuanhua Huang
School of Biomedical Sciences &
Department of Statistics and Actuarial Science

yuanhua@hku.hk
https://web.hku.hk/~yuanhua/

香 港 大 學
THE UNIVERSITY OF HONG KONG

Today's learning objectives

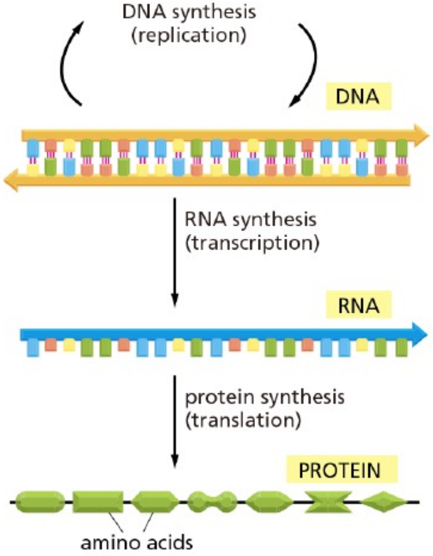
1. Transcriptome: what and why?
2. Additional layer of complexity: RNA splicing
3. Technologies to measure transcriptome? pros vs cons
4. RNA-seq: QC, alignment, assembly / quantification

Reading list

1) Stark, Grzelak, Hadfield. RNA sequencing: the teenage years. Nat Rev Gen, 2019
2) A survey of best practices for RNA-seq data analysis, Genome Biology, 2016
3) Alberts et al. Molecular Biology of The Cell (Chapter 6 & 7): https://www.ncbi.nlm.nih.gov/books/NBK21054/
4) Wikipedia: https://en.wikipedia.org/wiki/Transcriptomics_technologies

What is transcriptome?

- Central dogma of molecular biology & information flow via RNAs
- Transcriptome: all RNA transcripts, including coding and non-coding, in an individual or a population of cells



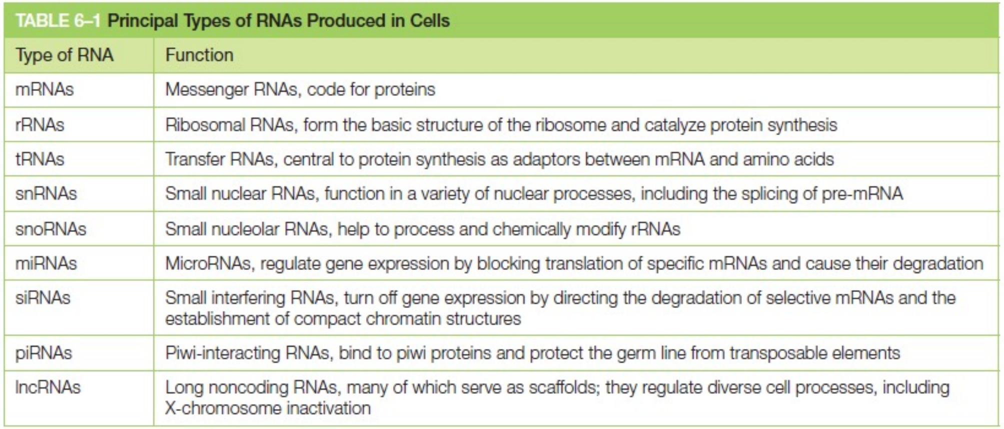
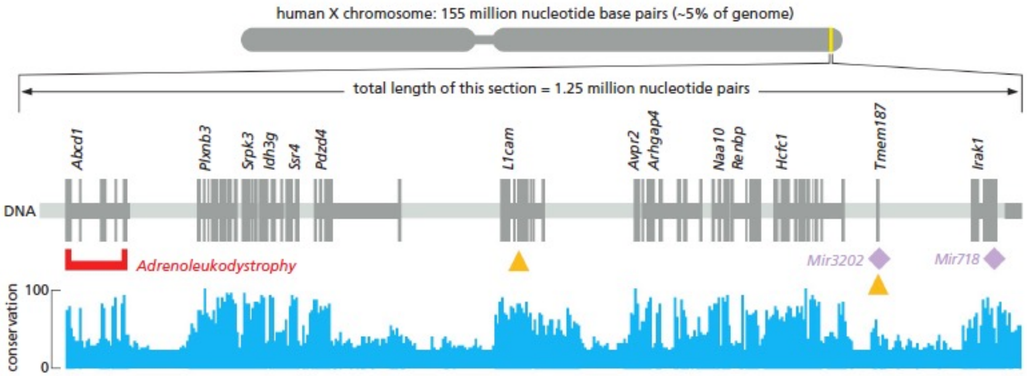


TABLE 6–1 Principal Types of RNAs Produced in Cells

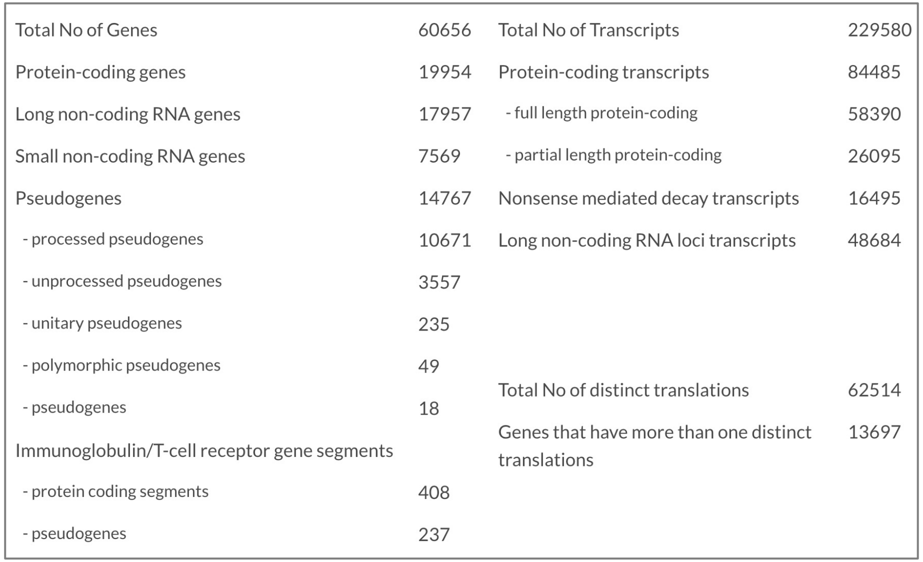
Type of RNA	Function
mRNAs	Messenger RNAs, code for proteins
rRNAs	Ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
tRNAs	Transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
snRNAs	Small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA
snoRNAs	Small nucleolar RNAs, help to process and chemically modify rRNAs
miRNAs	MicroRNAs, regulate gene expression by blocking translation of specific mRNAs and cause their degradation
siRNAs	Small interfering RNAs, turn off gene expression by directing the degradation of selective mRNAs and the establishment of compact chromatin structures
piRNAs	Piwi-interacting RNAs, bind to piwi proteins and protect the germ line from transposable elements
lncRNAs	Long noncoding RNAs, many of which serve as scaffolds; they regulate diverse cell processes, including X-chromosome inactivation

Alberts et al. Molecular Biology of The Cell. Six Edition

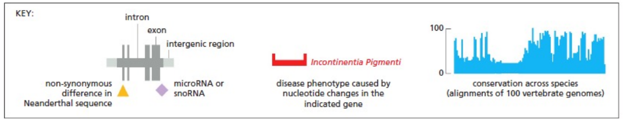
Gene annotation


human X chromosome: 155 million nucleotide base pairs (~5% of genome)
total length of this section = 1.25 million nucleotide pairs
Abcd1
Plxnb3
Srpk3
Idh3g
Ssr4
Pdzd4
L1cam
Avpr2
Arhgap4
Naa10
Renbp
Hcfc1
Tmem187
Irak1
DNA
Adrenoleukodystrophy
Mir3202
Mir718
conservation
100
0

GENCODE annotation v35 on human transcriptome
https://www.gencodegenes.org



Total No of Genes	60656	Total No of Transcripts	229580
Protein-coding genes	19954	Protein-coding transcripts	84485
Long non-coding RNA genes	17957	- full length protein-coding	58390
Small non-coding RNA genes	7569	- partial length protein-coding	26095
Pseudogenes	14767	Nonsense mediated decay transcripts	16495
- processed pseudogenes	10671	Long non-coding RNA loci transcripts	48684
- unprocessed pseudogenes	3557		
- unitary pseudogenes	235		
- polymorphic pseudogenes	49		
- pseudogenes	18	Total No of distinct translations	62514
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13697
- protein coding segments	408		
- pseudogenes	237		

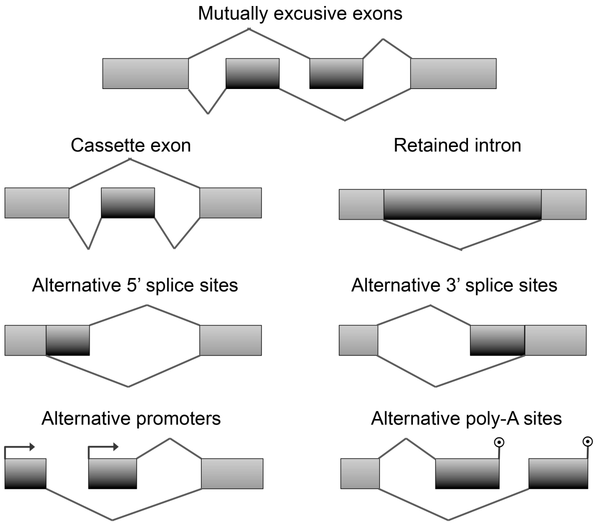

KEY:
intron
exon
intergenic region
non-synonymous difference in Neanderthal sequence
microRNA or snoRNA
Incontinentia Pigmenti
disease phenotype caused by nucleotide changes in the indicated gene
100
0
conservation across species (alignments of 100 vertebrate genomes)

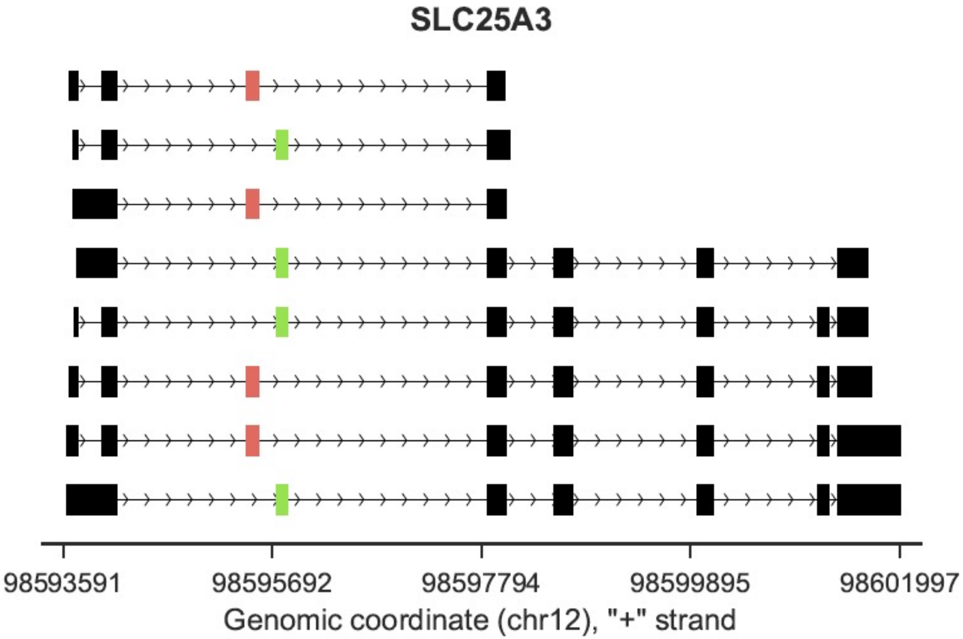
GENCODE annotation since 2003;
Transcriptome still not perfect even on human and human
Many more species: no good gene annotations

Alberts et al. Molecular Biology of The Cell. Six Edition

Alternative splicing

- One gene produces multiple transcripts (i.e., splicing isoforms)
- May increase the complexity in analysis

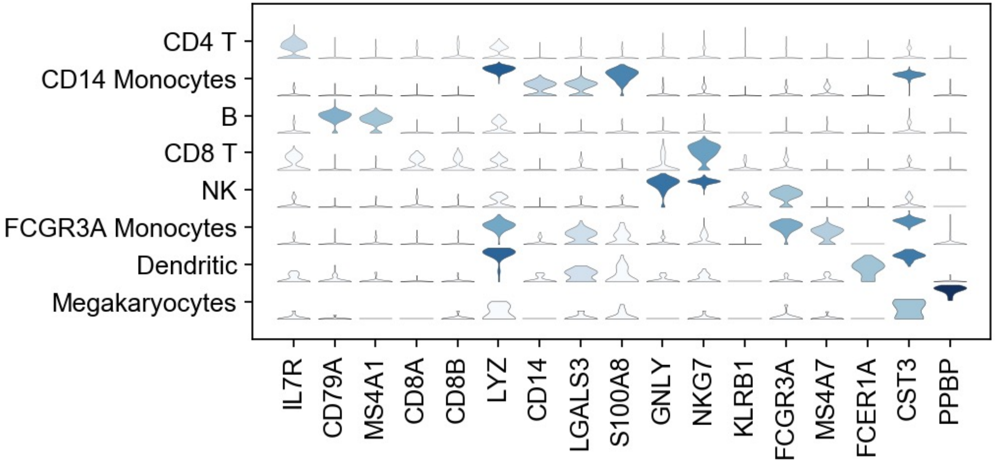




Transcriptome: cell differentiation & cell type

Identical DNA but completely different cellular functions and morphology

Immune cell expression profiles (subset genes)



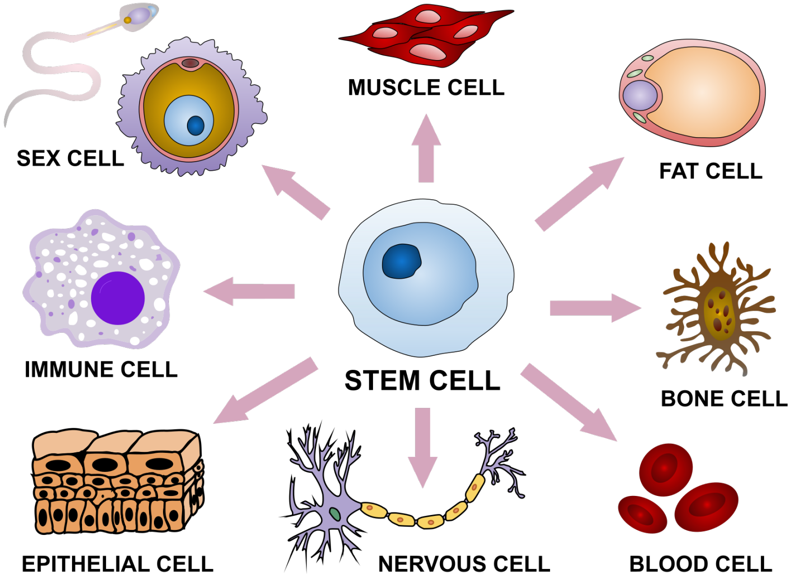


Fig from Scanpy tutorial on PBMC

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

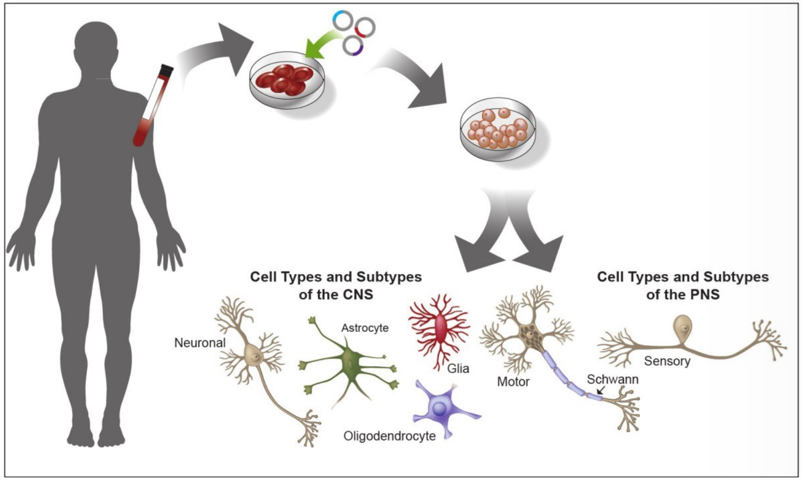
Transcriptome: cell reprogramming?

Change key regulatory genes may reprogram the cells

Key factors

- Oct3/4, Sox2, c-Myc, and Klf4
- Takahashi & Yamanaka, 2006, Cell

Transcriptome is precisely regulated in a complex way.


Cell Types and Subtypes of the CNS
Neuronal
Astrocyte
Glia
Oligodendrocyte
Motor
Cell Types and Subtypes of the PNS
Sensory
Schwann

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

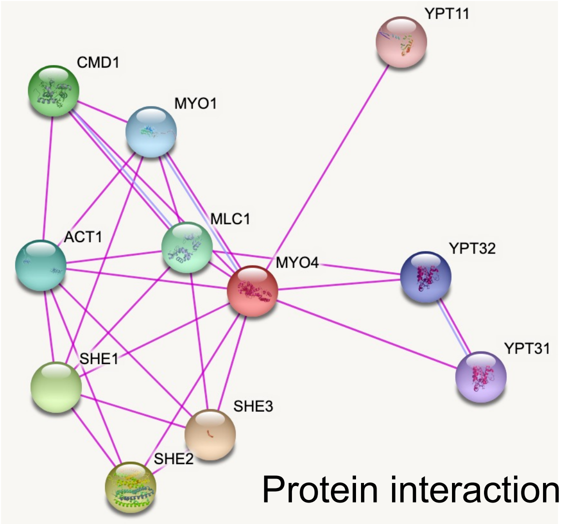
Transcriptomics: biomarker discovery

- Pair-wise comparison *MYCN*-amplified and single-copy tumours
  - 223 genes significantly differentially expressed
  - Schramm et al., 2013
- Transcriptomics analysis in multiple ASD mouse models
  - several recurrent target genes associated with Autism Spectrum Disorder
  - Duan et al. Autism Research, 2020
- Time-series transcriptomics across 48h
  - over 3000 circadian genes in liver
  - Zhang, et al. PNAS, 2014

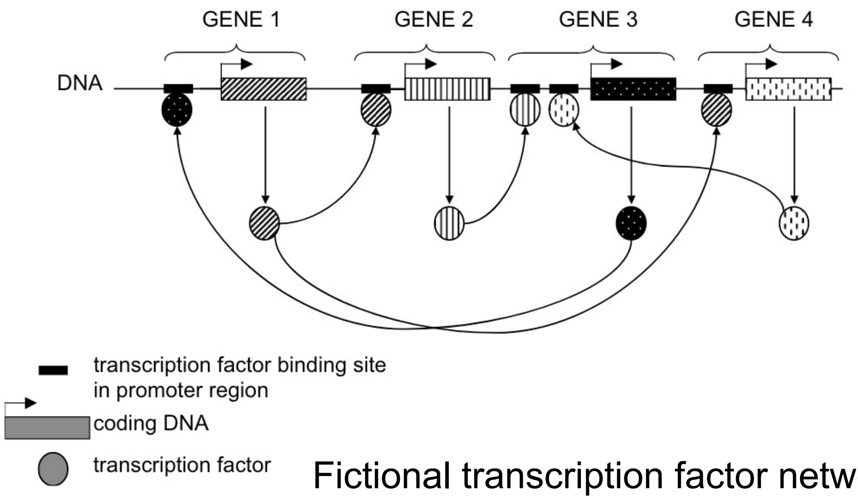
- Hypothesis minimal (or free) discovery?

Gene regulatory network

- Transcription factor: sequence-specific DNA-binding factor
- Gene pairs: co-expression; exclusive expression
- Protein physical interaction
  - Human: 365,000 edges across ~20,000 genes
  - Yeast: 131,000 edges across ~6,000 genes



Protein interaction



Fictional transcription factor network

Protein-protein interaction: https://string-db.org;
BioGRID: :https://academic.oup.com/nar/article/47/D1/D529/5204333
Schlitt & Brazma, BMC Bioinfo, 2007

Today's learning objectives

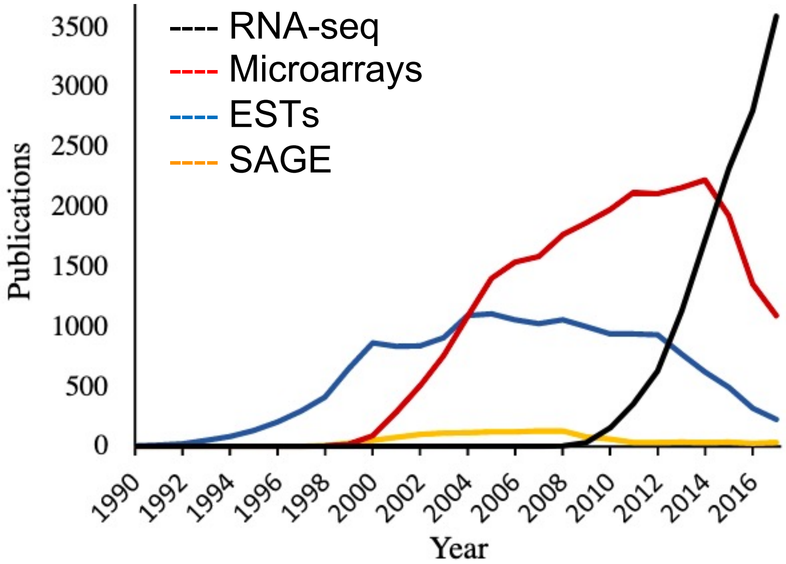
1. Transcriptome: what and why?
   - All RNAs in a cell: mRNA, rRNA, etc
   - Alternative splicing: additional layer of complexity
   - Characters of cells: cell types, states, tissues
   - Hypothesis free discovery: marker genes for disease

2. Technologies to measure transcriptome? Pros vs cons

3. RNA-seq and computational process: challenges and solutions
   - QC, alignment, assembly / quantification

Technology history on gene expression

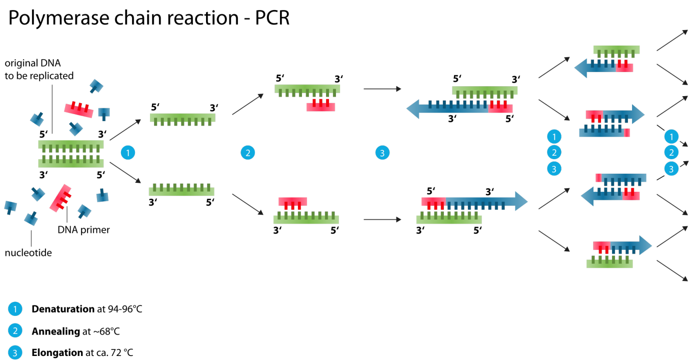
- Before transcriptomics
  - Sanger sequencing (popular in 1980s): ESTs, SAGE
  - Individual transcripts: RT-qPCR, Northern blotting, etc.
- cDNA microarrays
- **RNA-seq**
- Long reads: PacBio / Nanopore

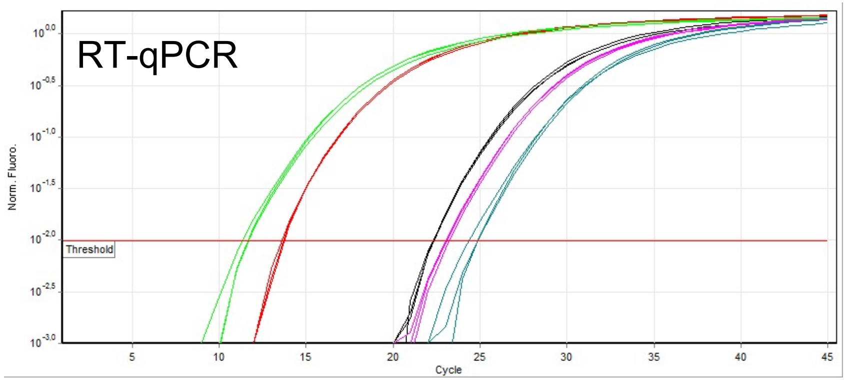


ESTs: expressed sequence tags
SAGE: serial analysis of gene expression
https://en.wikipedia.org/wiki/Transcriptomics_technologies

Before transcriptomics

- RT-qPCR (qPRC)
  - Gold standard measurement
  - laborious & usually a tiny subsection of a transcriptome
  - viral RNAs, e.g., HBV, SARS-CoV-19
- Sanger sequencing
  - First generation sequencing
  - Invented in 1977 by Frederick Sanger and colleagues
  - ESTs, SAGE

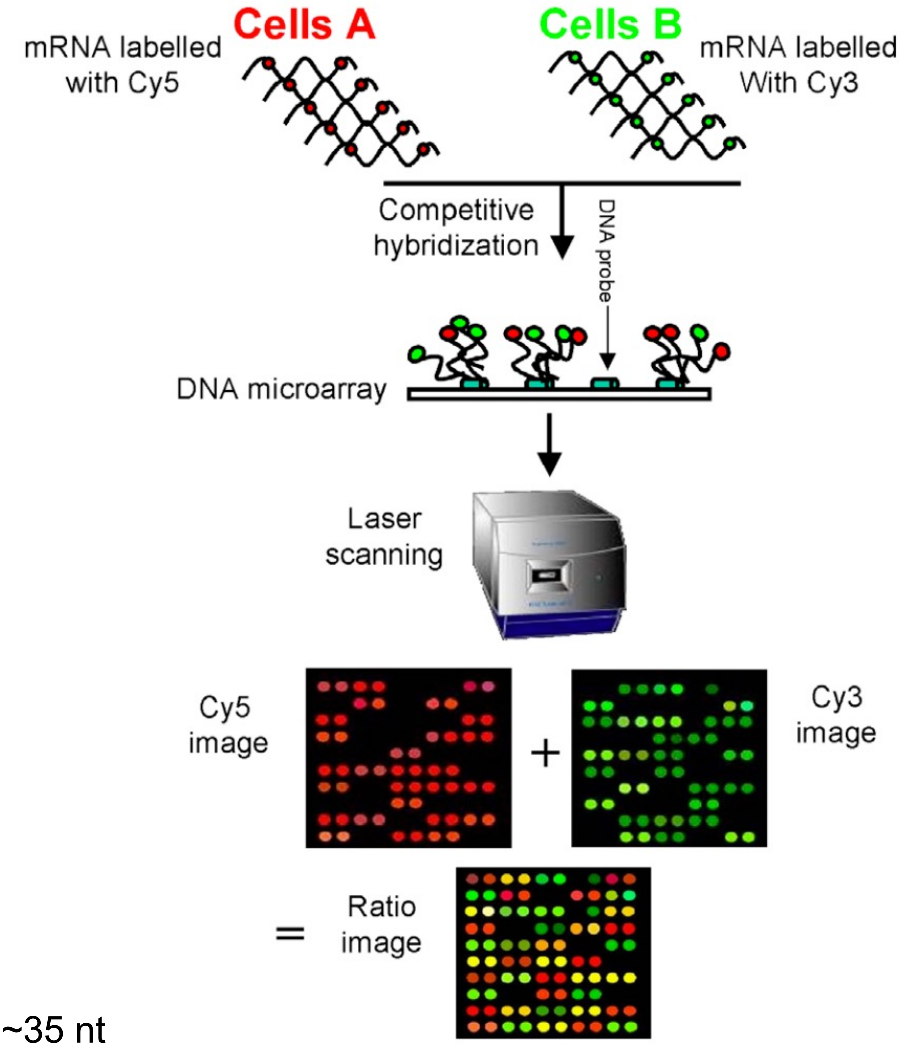




cDNA microarrays

- Since mid 1990s
- A set of transcripts in a sample, followed by fluorescent labelling
- Hybridization to an array of complementary probes

- Require known sequences first
- Common probes:
  - spotted oligonucleotide arrays
  - Affymetrix high-density arrays

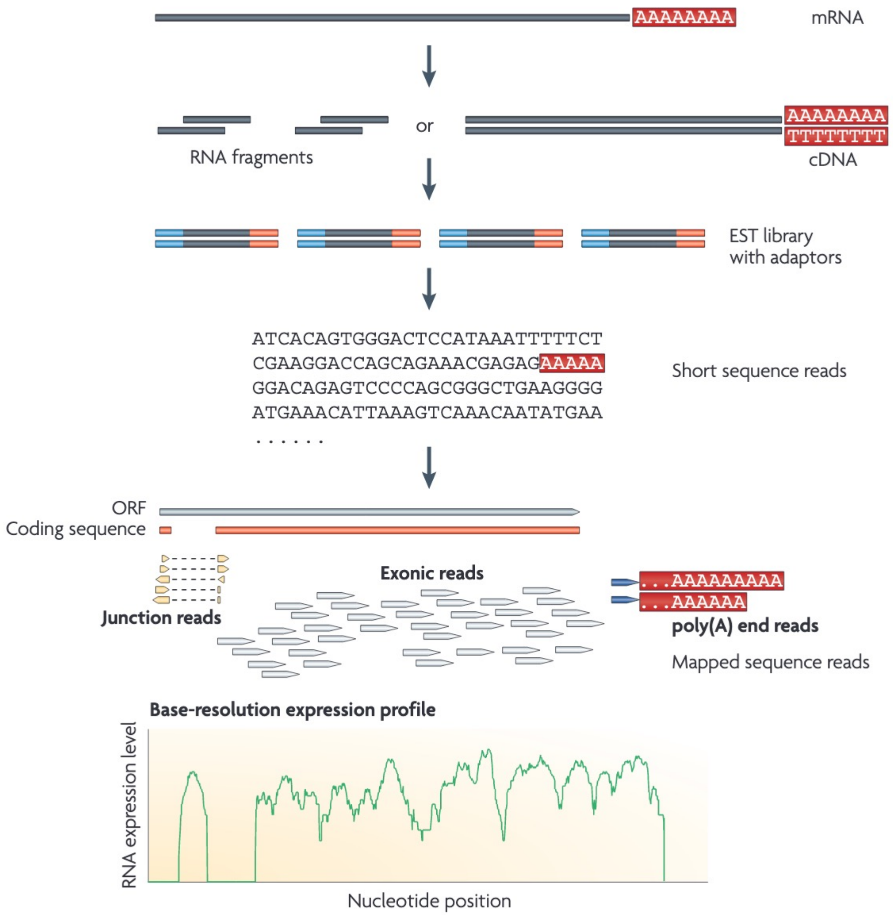


Oligonucleotides: short nucleic acid polymers, often 13~35 nt

Lamartine, Materials Science and Engineering: C, 2006

RNA-seq

- Next generation sequencing
- Library preparation
  - RNA extraction
  - Enrichment or depletion (rRNA)
  - cDNA synthesis and preparation
- Sequencing
  - Library size: 10-100 million reads
  - Single-end vs paired-end
  - RNA Fragmentation
- Computational processing
  - Multiple steps
  - Depending on the purpose



Wang, Gerstein, Snyder, 2009, Nat Rev Gen

RNA-seq parameters (w/ budget constraint)

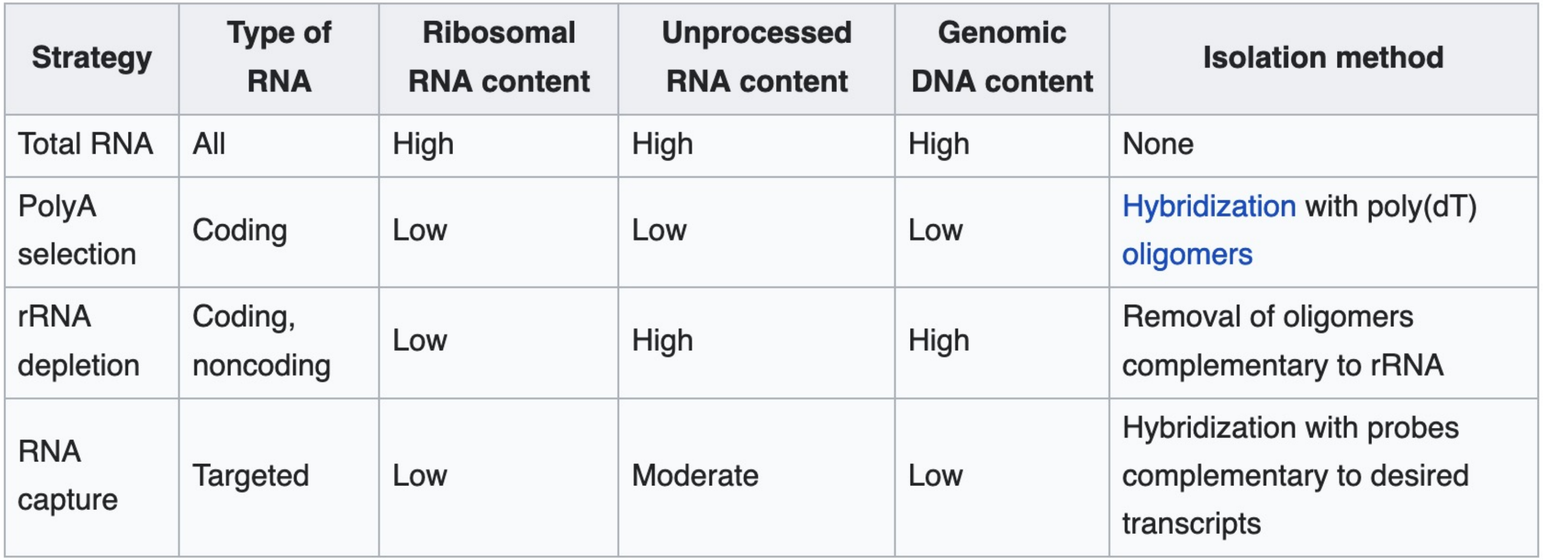
- Paired-end (vs single-end)
  - Pro: longer range to better cover splicing junctions
  - Con: waste half of the reads if only caring about gene level

- Sequencing depths (vs number of samples)
  - Read length: 75bp, 100bp, 150bp, (possibly) 250 bp
  - Rough pricing: 2x150 bp & 300 million reads: 2,500 USD
  - Balance between number of samples and depths
    - 2x150 bp: 100 million x 3 samples
    - 2x150 bp: 25 millions x 12 samples
    - 1x150 bp: 50 millions x 12 samples

Example pricing: http://med.stanford.edu/gssc/rates.html

RNA-seq variates

- 4tU- or 4sU labelling for nascent RNAs (usually time-series)
- Poly-A selection, rRNA depletion, specific targeted
- UPF1 depletion to protect mis-spliced transcripts from NMD



Strategy	Type of RNA	Ribosomal RNA content	Unprocessed RNA content	Genomic DNA content	Isolation method
Total RNA	All	High	High	High	None
PolyA selection	Coding	Low	Low	Low	Hybridization with poly(dT) oligomers
rRNA depletion	Coding, noncoding	Low	High	High	Removal of oligomers complementary to rRNA
RNA capture	Targeted	Low	Moderate	Low	Hybridization with probes complementary to desired transcripts

Aslanzadeh, et al. Genome Res, 2018; Barrass, et al. Genome Biology, 2015
https://en.wikipedia.org/wiki/RNA-Seq

Long reads: PacBio and nanopore

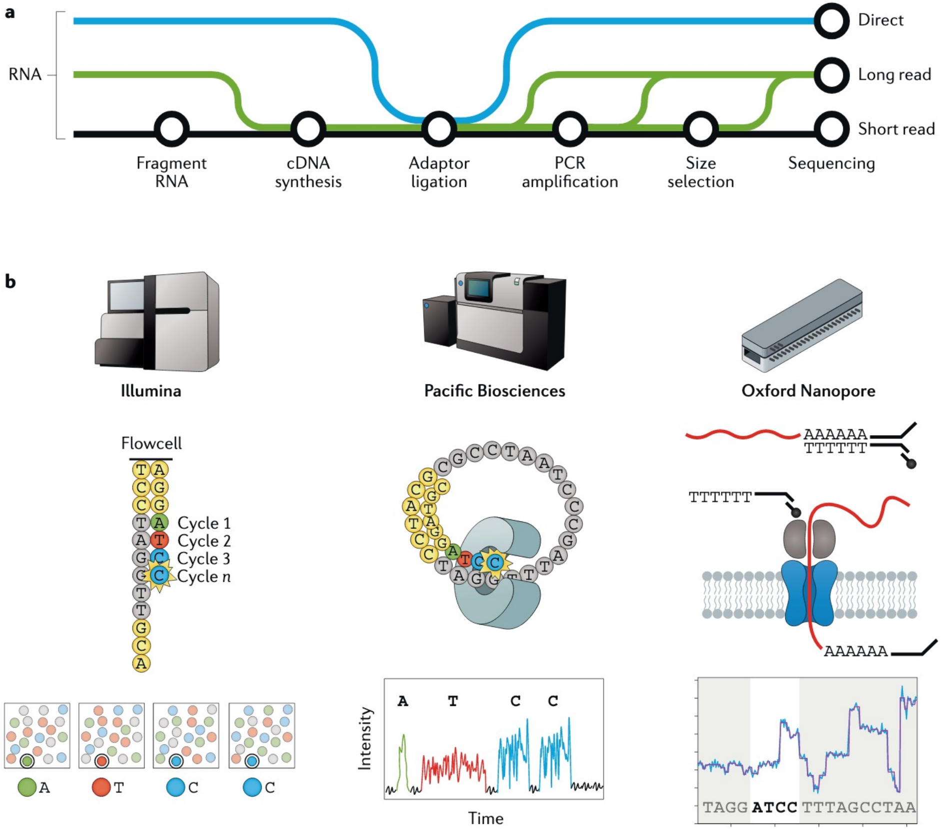
Long-read cDNA or direct RNA

Pros

- Long reads: 1-50 kb
- May capture full transcript
- Good for transcript assembly

Cons

- High error rate: 1~10%
- Low throughput: not sensitive to detect lowly expressed genes



Stark, Grzelak, Hadfield. RNA sequencing: the teenage years. Nat Rev Gen, 2019

Weirather, et al. F1000Research 2017

Technology comparison

- RT-qPCR (not really transcriptome)
  - A handful of transcripts; known sequence required (primer, Oligonucleotides)
  - Gold standard in terms of accuracy: validation & viral RNAs
- cDNA Microarrays (less popular now)
  - Thousands of RNAs; known sequences required (probes, Oligonucleotides)
- **RNA-seq** (versatile)
  - High throughput; return whole transcriptome in principle (can be enriched)
  - Experiment design requires optimization: paired- / single-end; depths, etc.
  - Computational analysis can be complex (industrial-standard software exist)
- Long-reads sequencing
  - Benefits in genome and transcriptome assembly
  - High error rate and low throughput (not sensitive to lowly expressed genes)

Today’s learning objectives

1. Transcriptome: what and why?

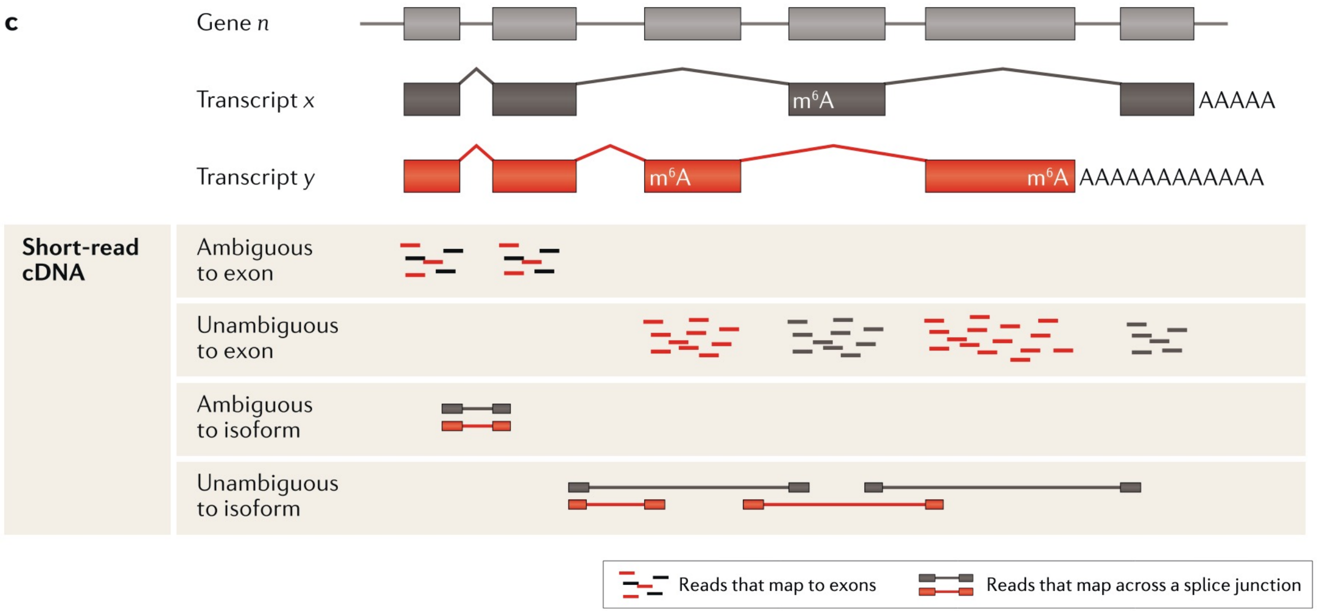
2. Technologies to measure transcriptome? Pros vs cons
   - Before transcriptomics: RT-qPCR, Sanger sequencing
   - cDNA Microarrays
   - RNA-seq
   - Long reads sequencing

3. RNA-seq & computational process: challenges and solutions
   - QC, alignment, assembly / quantification

Complexity from (alternative) RNA splicing

Gene level quantification vs transcript level quantification

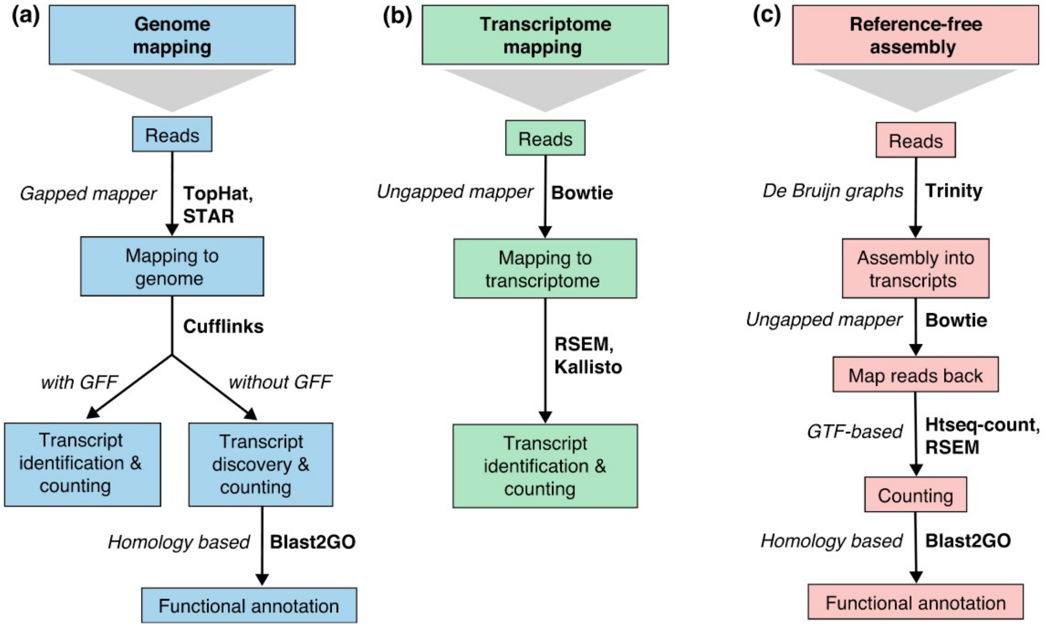
- Transcript level: detect differential transcript usage between conditions
- Gene level: simplify the analysis, but may miss information



Stark, Grzelak, Hadfield. RNA sequencing: the teenage years. Nat Rev Gen, 2019

RNA-seq analysis options

- Map to genome
  - Mostly used
  - Gene discovery
  - Novel splicing variants
- Map to transcriptome
  - Transcriptome available
  - Mouse and human
  - Faster
- Transcriptome assembly
  - with genome or without genome reference (de novo)
  - Challenging



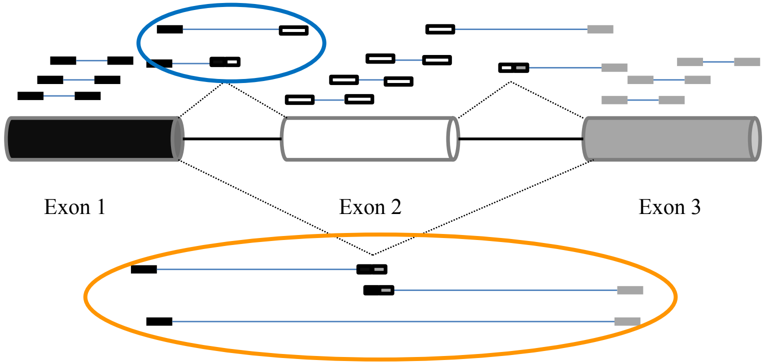
Conesa et al. A survey of best practices for RNA-seq data analysis, Genome Biology, 2016

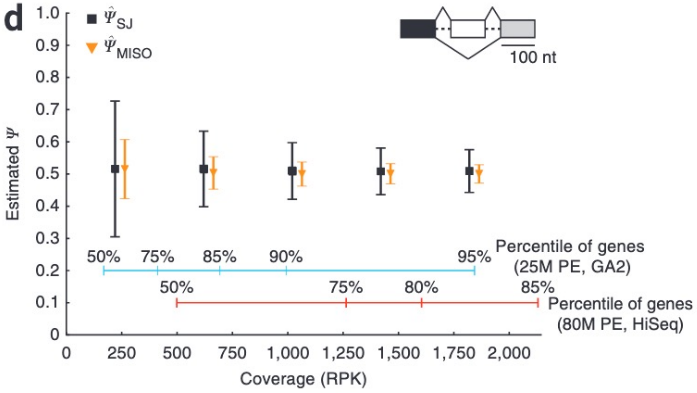
Option 1: Read alignment to genome

- Genome reference: each chromosome is a sequence
- Reads aligner: gap aware (mature & industrial standard now)
  - STAR, HISAT, and others
- Gene level counting (straightforward)
  - Feature-count: http://bioinf.wehi.edu.au/featureCounts/
  - HTseq-count: https://htseq.readthedocs.io

- Transcript level quantification: ambiguous reads
  - MISO: https://miso.readthedocs.io
  - Cufflinks: http://cole-trapnell-lab.github.io/cufflinks/
  - DICE-seq / BRIE (myself): http://diceseq.sf.net, https://brie.readthedocs.io
  - Mixture model: EM algorithm, MCMC sampling

Splicing quantification (MISO/BRIE model)





Estimate the proportions for 2 isoforms

- ✓ Direct method: count junction reads

  ψ = exon1_exon2 / (exon1_exon2 + exon1_exon3) = 2 / 5

- ✓ Probabilistic method: identity I_n for each read

  $L(\mathrm{R}_{1:\mathrm{N}} | \Psi) = \prod_{n=1}^{N} P(R_n|\Psi) = \prod_{n=1}^{N} \sum_{I_n=1}^{2} \{P(R_n|I_n)P(I_n|\Psi)\}$

  Maximize the likelihood on Ψ (mixture model)

- ✓ Bayesian method (posterior distribution)

  $$P(\Psi|\mathrm{R}_{1:\mathrm{N}}) \sim P(\Psi|\pi) \times L(\mathrm{R}_{1:\mathrm{N}} | \Psi)$$

Psi, ψ: probability of spliced exon inclusion (i.e., the fraction of isoform 2)

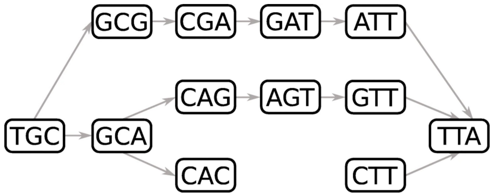
Option 2: Reads align to transcriptome

- Transcriptome reference: each transcript is a sequence
- Reads aligner: no require on gap (mature & industrial standard now)
  - Bowtie2, and others
  - Large fraction of reads have multiple alignment (multiple transcripts share exons)
- Statistical quantifications (relatively mature now)
  - BitSeq, RSEM, and many more
  - Mixture model: EM algorithm, MCMC sampling, variational inference
- Alternative strategy: combine alignment and quantification
  - Pseudo-alignment
  - Kallisto & Salmon

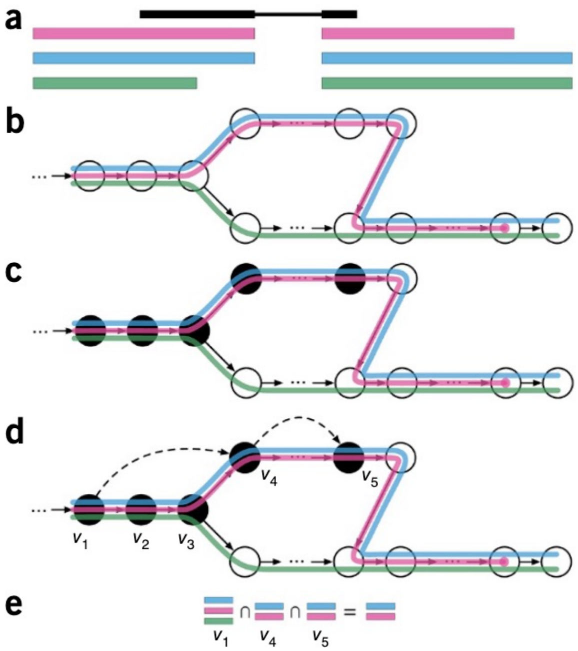
Kallisto: Bray, et al, Nat Biotech, 2016;
Salmon: Patro, et al. Nature Methods, 2017

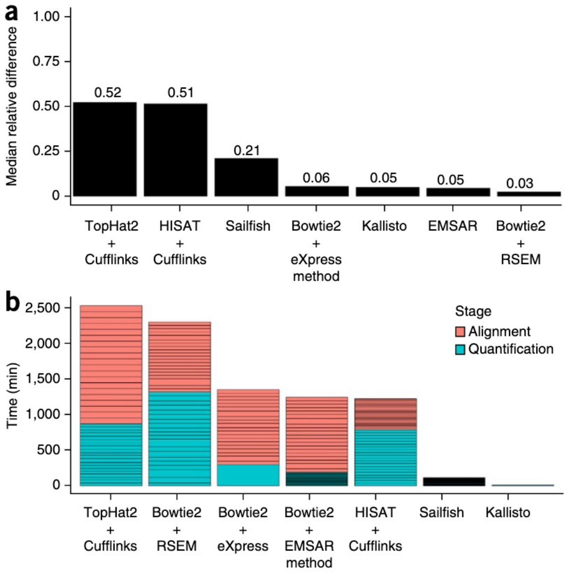
Kallisto: pseudo-alignment (de Bruijn graph)

- Not where in transcript the read comes from
- But whether it can come from the transcript



de Bruijn graph





Largely speed up

Kallisto: Bray, et al, Nat Biotech, 2016;
Salmon: Patro, et al. Nature Methods, 2017

Option 3: Transcriptome assembly

- De-novo or reference based
  - De-novo method based on de Bruijn graph
- Reference based is generally more accurate
  - Aligning reads to known genome reference first
  - Large assembly --> many smaller assembly
  - Often starting from generating splicing graph (with junction reads)
  - Whole genome sequencing to make genome reference first

Experiment designs

- High coverage and paired-end help
- Benefits from long-read sequencing from PacBio or Nanopore

Martin & Wang, Nat. Rev. Gen., 2011
Scallop: Shao & Kingsford, Nature biotech, 2017

Questions

1. Transcriptome: what and why?
2. Additional layer of complexity: RNA splicing
3. Technologies to measure transcriptome? pros vs cons
4. RNA-seq: QC, alignment, assembly / quantification

Reading list

1) Stark, Grzelak, Hadfield. RNA sequencing: the teenage years. Nat Rev Gen, 2019
2) A survey of best practices for RNA-seq data analysis, Genome Biology, 2016
3) Alberts et al. Molecular Biology of The Cell (Chapter 6 & 7): https://www.ncbi.nlm.nih.gov/books/NBK21054/
4) Wikipedia: https://en.wikipedia.org/wiki/Transcriptomics_technologies