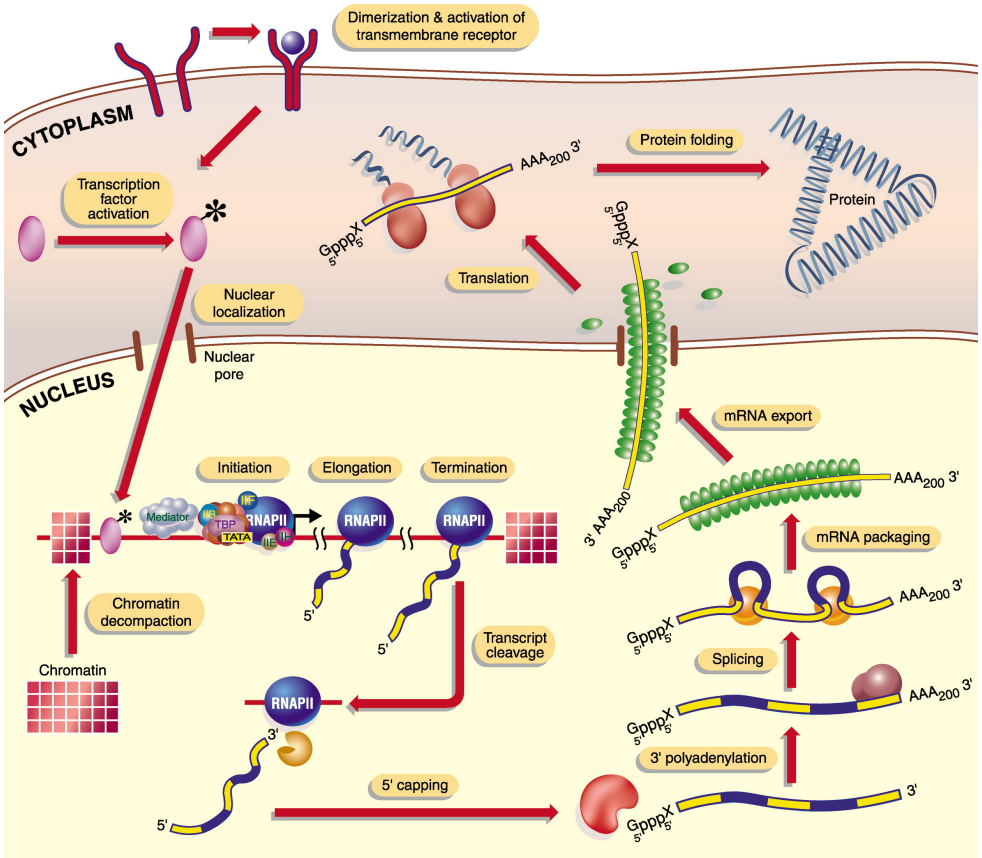
RNA SEQ ANALYSIS

Algorithms in Sequence Analysis

GENE EXPRESSION – THE FOUNDATION



3 main questions

- What is expressed?
- When is it expressed?
- In what amounts is it expressed?

Source: Orphanides et al. Cell VOLUME 108, ISSUE 4, P439-451,

TRANSCRIPTION – FROM DNA TO RNA

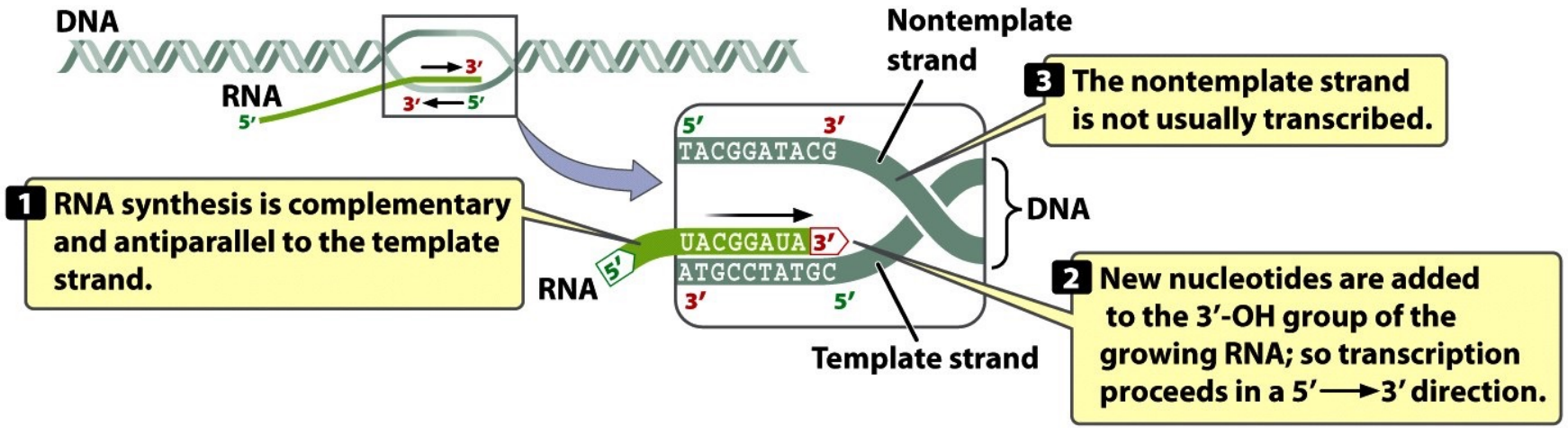
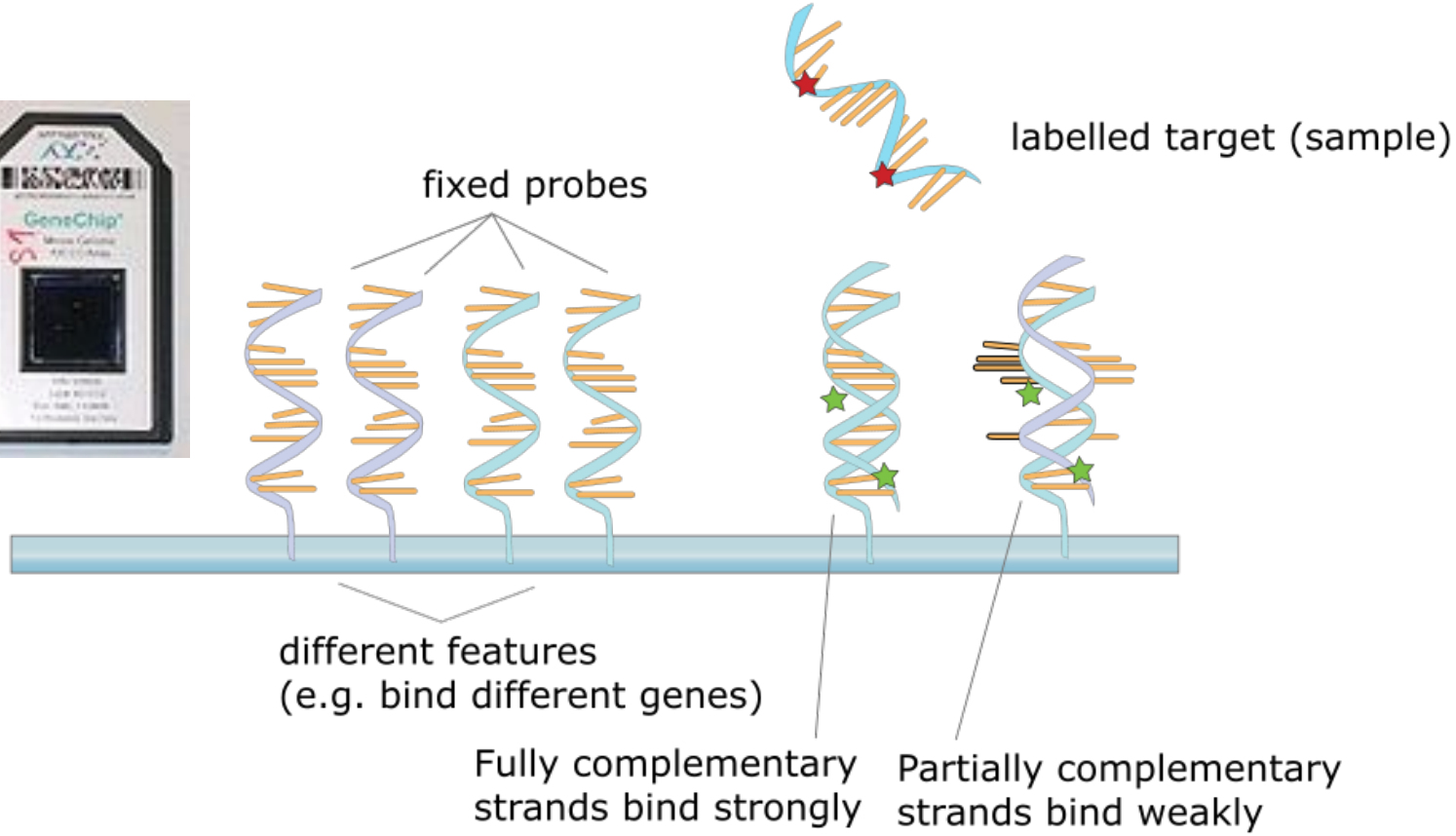


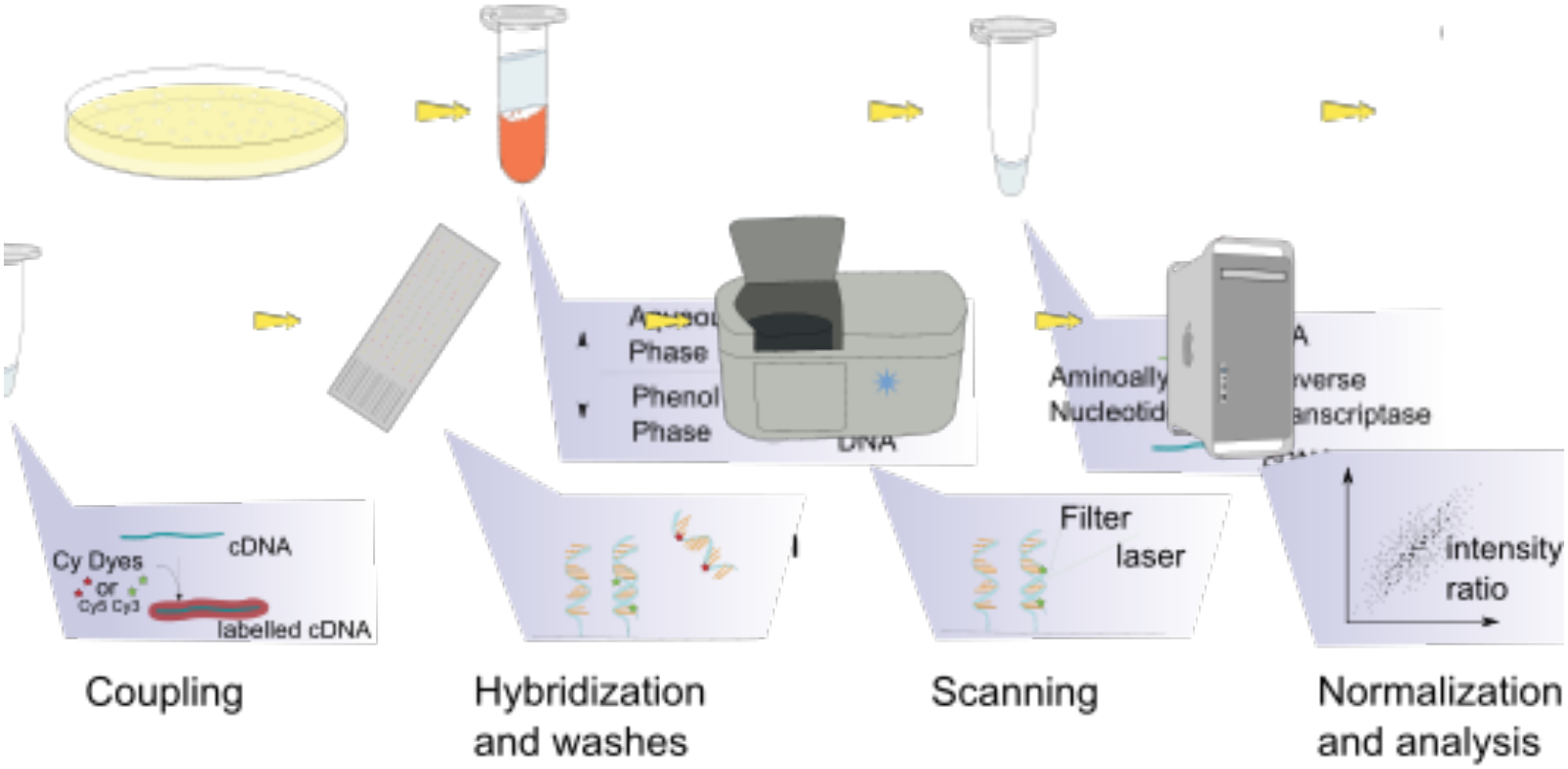
Figure 13-4
Genetics: A Conceptual Approach, Third Edition

If we could measure the amount of RNA produced from a given gene, we can determine its expression level

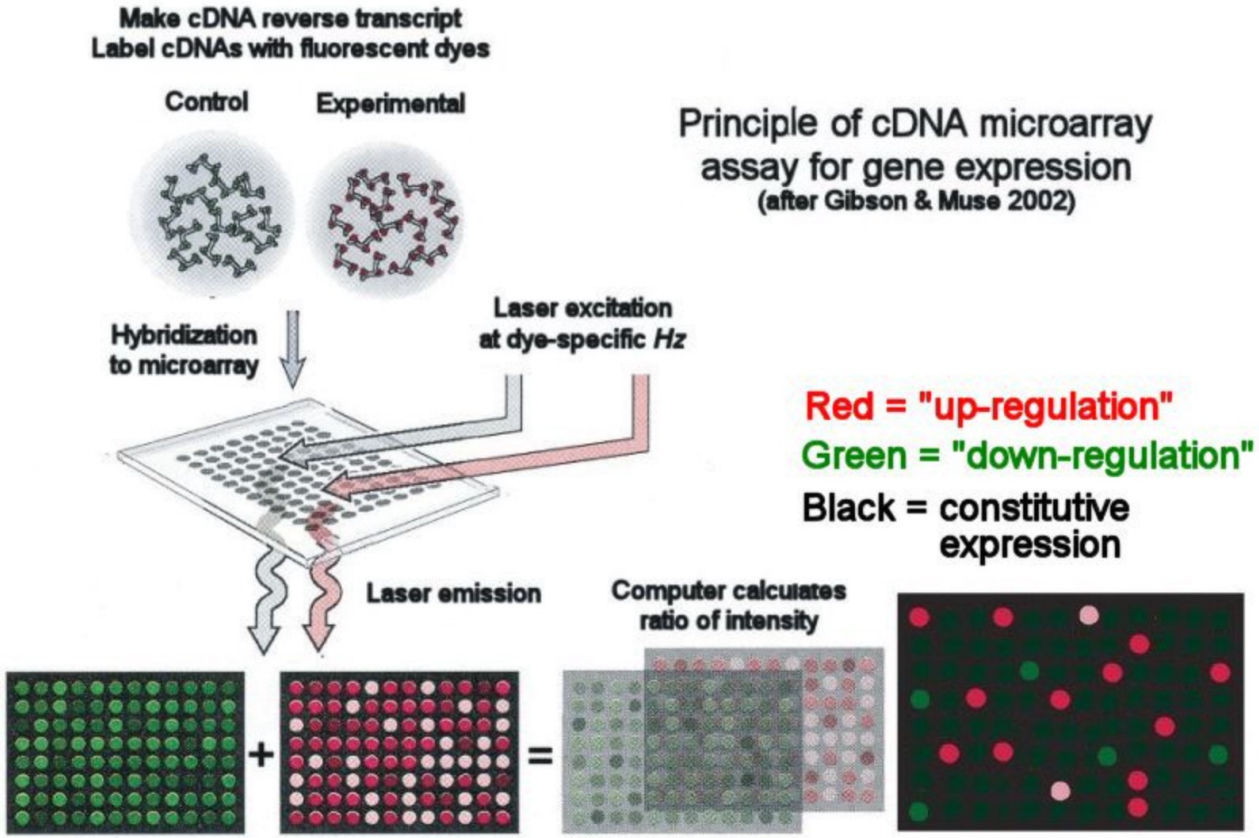
EARLY METHODS: MICROARRAYS DETECT TRANSCRIPTS EXPLOITING THEIR HYBRIDIZATION TO COMPLEMENTARY PROBE SETS



EARLY METHODS: MICROARRAYS WORKFLOW



DIFFERENTIALLY LABELLED CDNAS HELP TO DETERMINE DIFFERENTIAL GENE EXPRESSION ANALYSIS



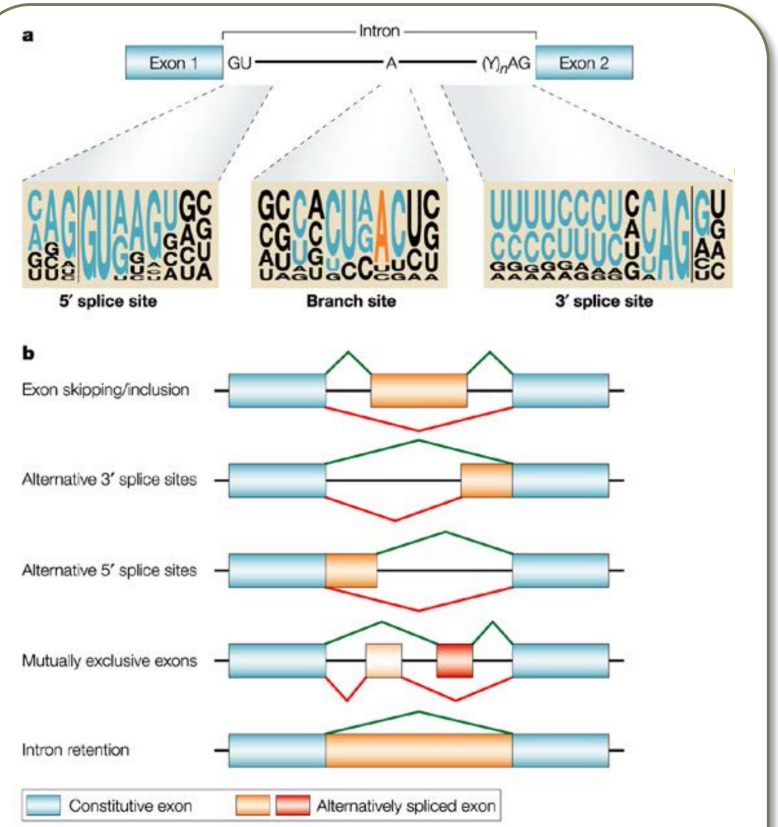
Advantages

1. fast
2. Intuitive

Disadvantages

1. Difficult to quantify
2. Differences in probe binding efficiacy
3. Probe design not straightforward
4. Cross-hybridization adds noise
5. expenisve

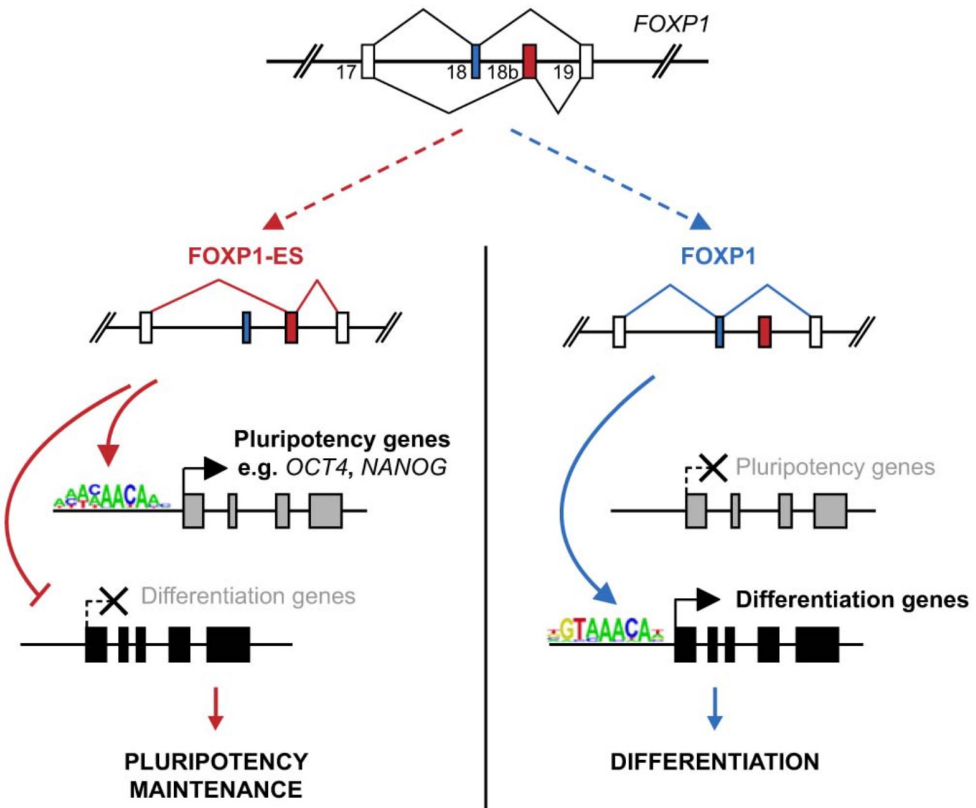
ALTERNATIVE SPLICING



Listening to silence and understanding nonsense: exonic mutations that affect splicing. Cartegni et al. 2002

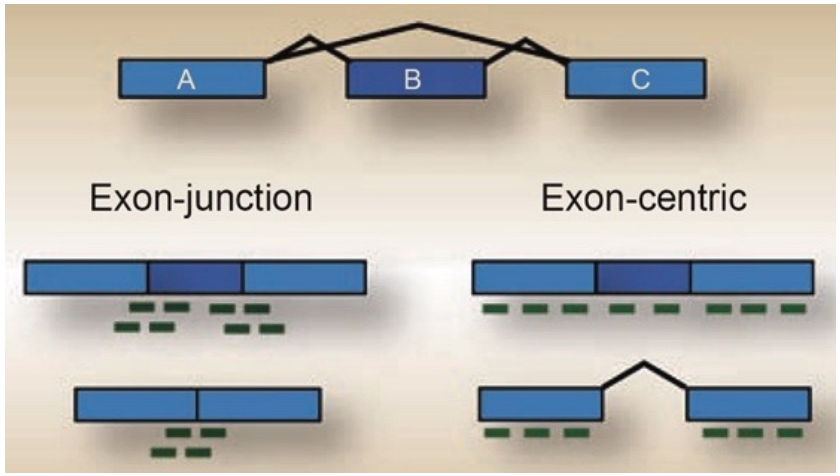
Nature Reviews | Genetics

Cell fate determination in human stem cells



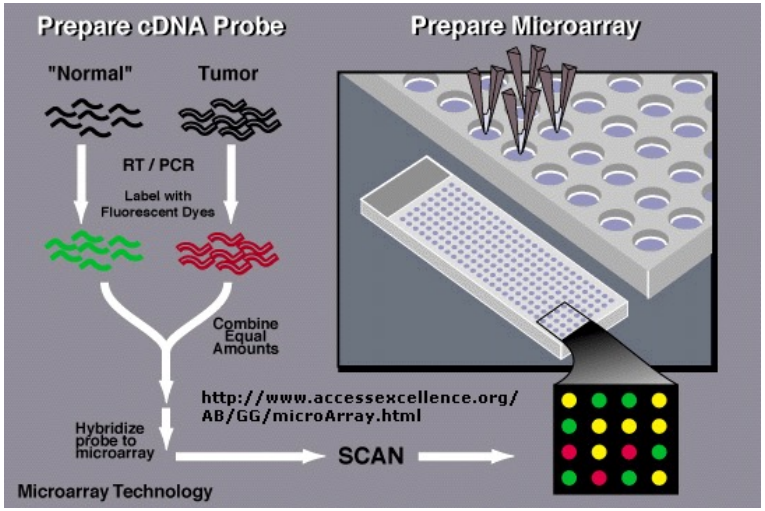
Gabut et al. (2011) Cell 147, 132-146

MICROARRAY APPLICATIONS – IT ALL DEPENDS ON THE PROBE DESIGN



Probe sets of different conditions, specifically measuring the occurrence of both samples

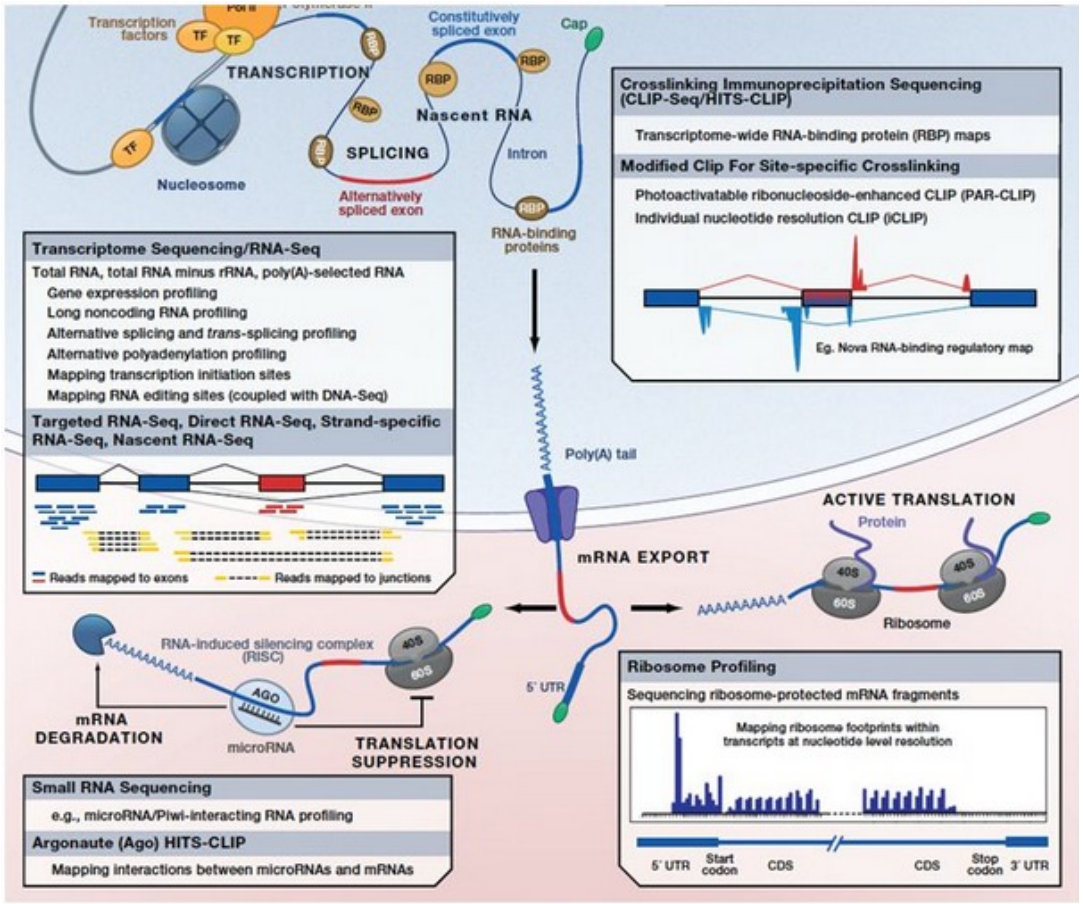
Reciprocal behaviour of splice probe sets, specifically measuring the inclusion or exclusion of AS



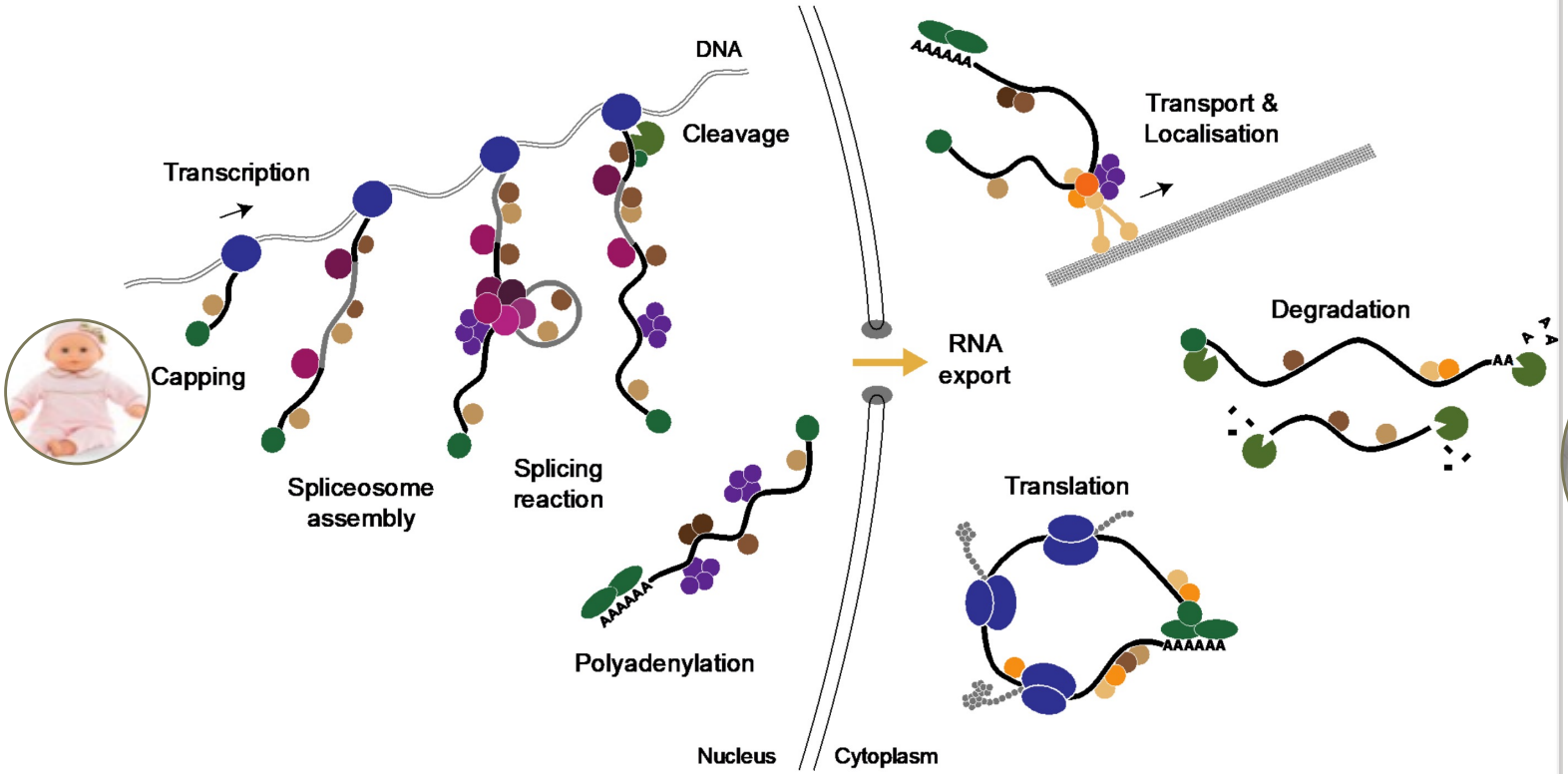
WHY BOTHER SEQUENCING RNAS?

- **Identification of genes**
  - Build new or improved profile of transcribed regions ("gene models") of an uncharacterized genome
  - Rapid access to (protein-coding) genes without bothering with genome assembly and gene prediction
- **Differential Gene Expression (DGE)**
  - Quantitative evaluation and comparison of transcript levels, usually between different groups
  - Vast majority of RNA-Seq is for DGE
- **Metatranscriptomics**
  - Transcriptome analysis of a community of different species (e.g., gut bacteria, hot springs, soil)
  - Gain insights on the functioning and activity rather than just who is present
- **Study RNA-Protein interaction**
  - Gain insights into regulatory networks controlling gene expression

THE END OF MICROARRAYS – RNA SEQUENCING



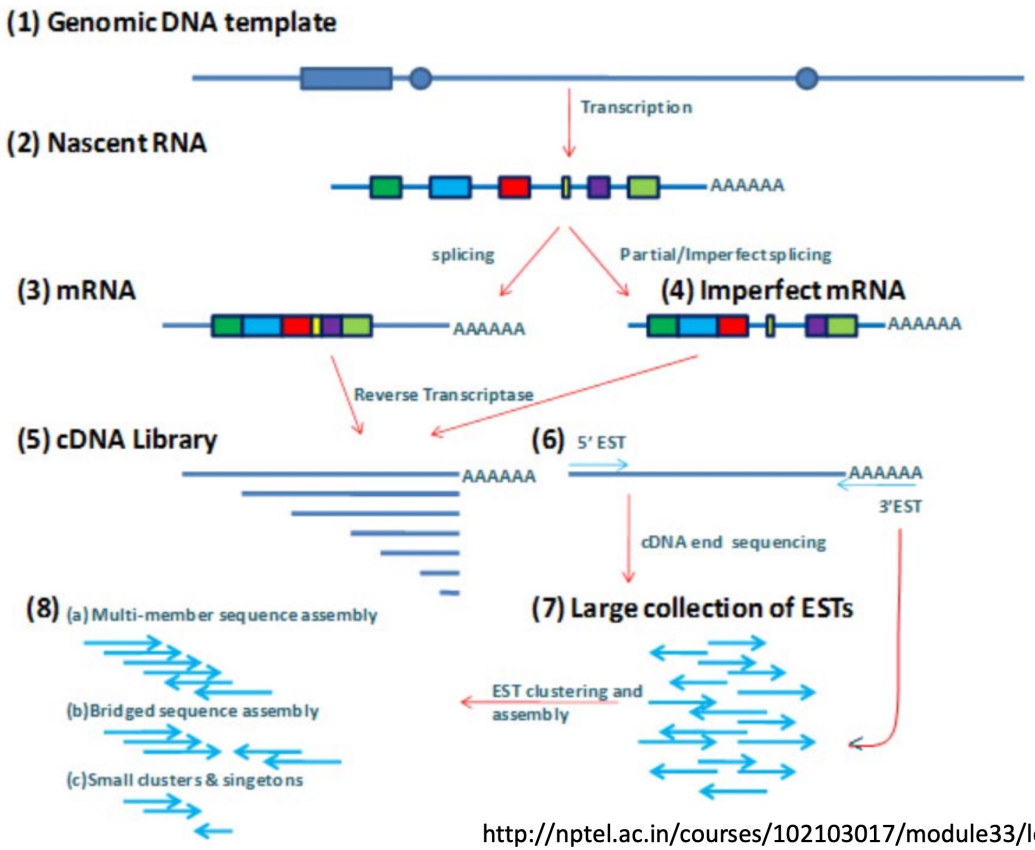
AT A GIVEN TIME IN A CELL, WE FIND A GENE'S TRANSCRIPTS IN ALL POSSIBLE STAGES OF ITS LIFE[1]



modified from
McKee and Silver, Cell Res. 2007

1 this becomes relevant for the interpretation of RNA seq data

FIRST STEPS IN RNASEQ – EST[1] SEQUENCING



Transcription of Genomic DNA:

Genomic DNA is first transcribed to generate Nascent mRNA followed by splicing of synthesize perfect mRNA.

Reverse transcription of mRNA:

mRNA can also be directly isolated from the species by using different kits (e.g. RNAgent Promega). mRNA synthesized undergoes reverse transcription to form cDNA library.

Generation of ESTs:

From the cDNA library 5' or 3'-ESTs are generated by cDNA end sequencing. 5' EST is formed from a region of transcript which forms protein whereas the ending portion of cDNA forms 3'EST.

Assembly and organization of ESTs:

The constructed ESTs can then be assembled separately in multimember sequence assembly, Bridged sequence assembly and small clusters on the basis of size of ESTs.

http://nptel.ac.in/courses/102103017/module33/lec33_slide2.htm

1 Expressed sequence tags

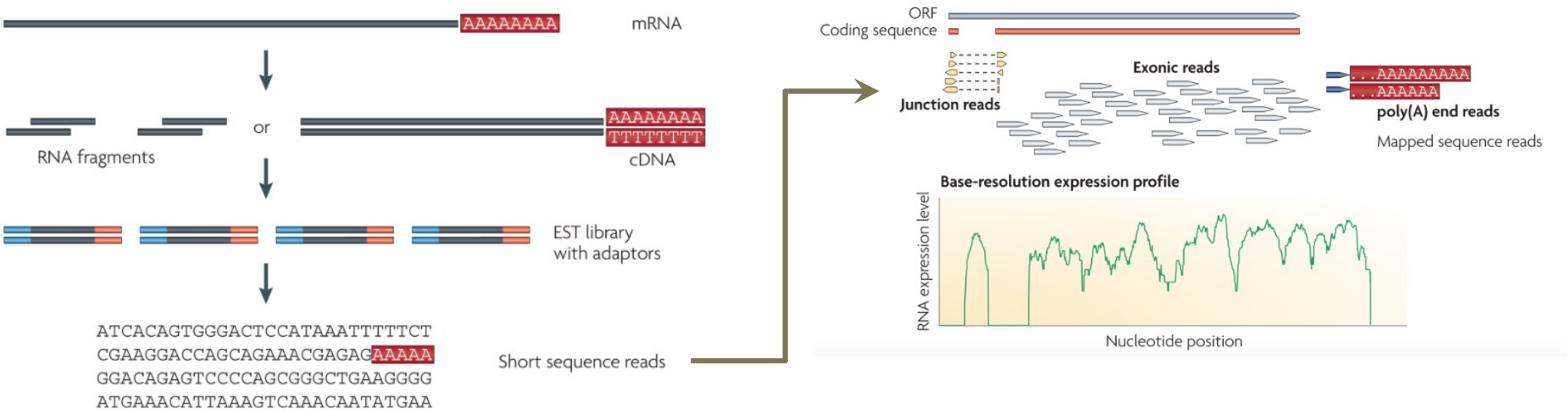
BEYOND EST SEQUENCING – THE QUEST FOR THE FULL TRANSCRIPT TO BROADEN THE SCOPE OF ANALYSES (E.G. ALTERNATIVE SPLICING)

- once: Cloning and sequencing of long cDNA-fragments (300 – 400 nt)
- Now: shotgun RNA sequencing and asseembly
- Even newer: PacBio Iso-Seq – single molecule sequencing of full RNAs
- Analysis: mapping to reference genome, e.g. with BLAT (Blast like alignment tool) or with splice-site aware mappers, e.g. GMAP

1 Kent (2002) Genome Res. 12: 656-664

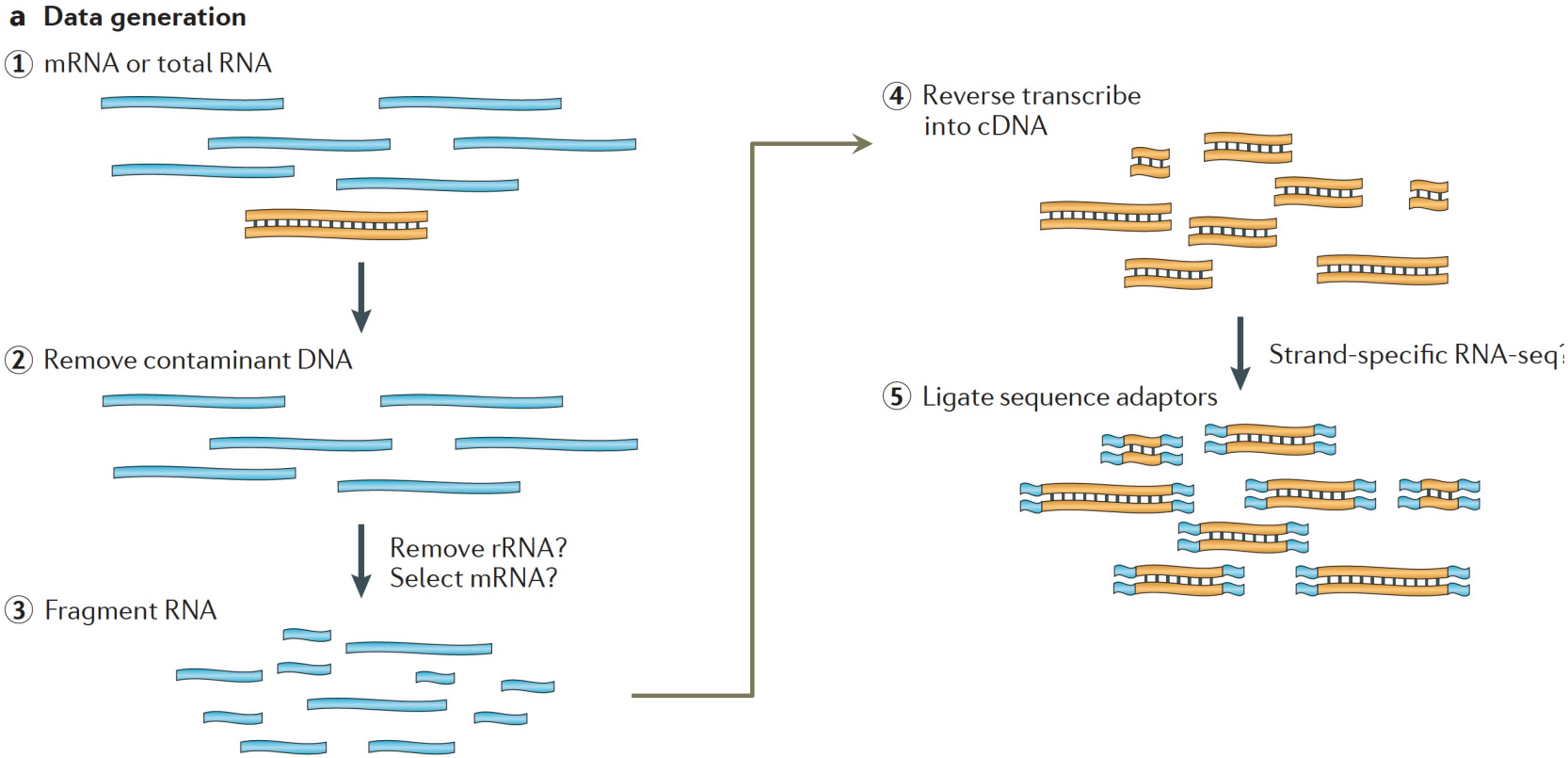
2 Wu and Watanabe (2005) Bioinformatics 21:1859-1875

THE GENEREAL WORKFLOW OF RNASEQ



Wang et al. 2009 Nat Rev Genet. 10(1): 57–63

FROM RNA TO SEQUENCE DATA



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

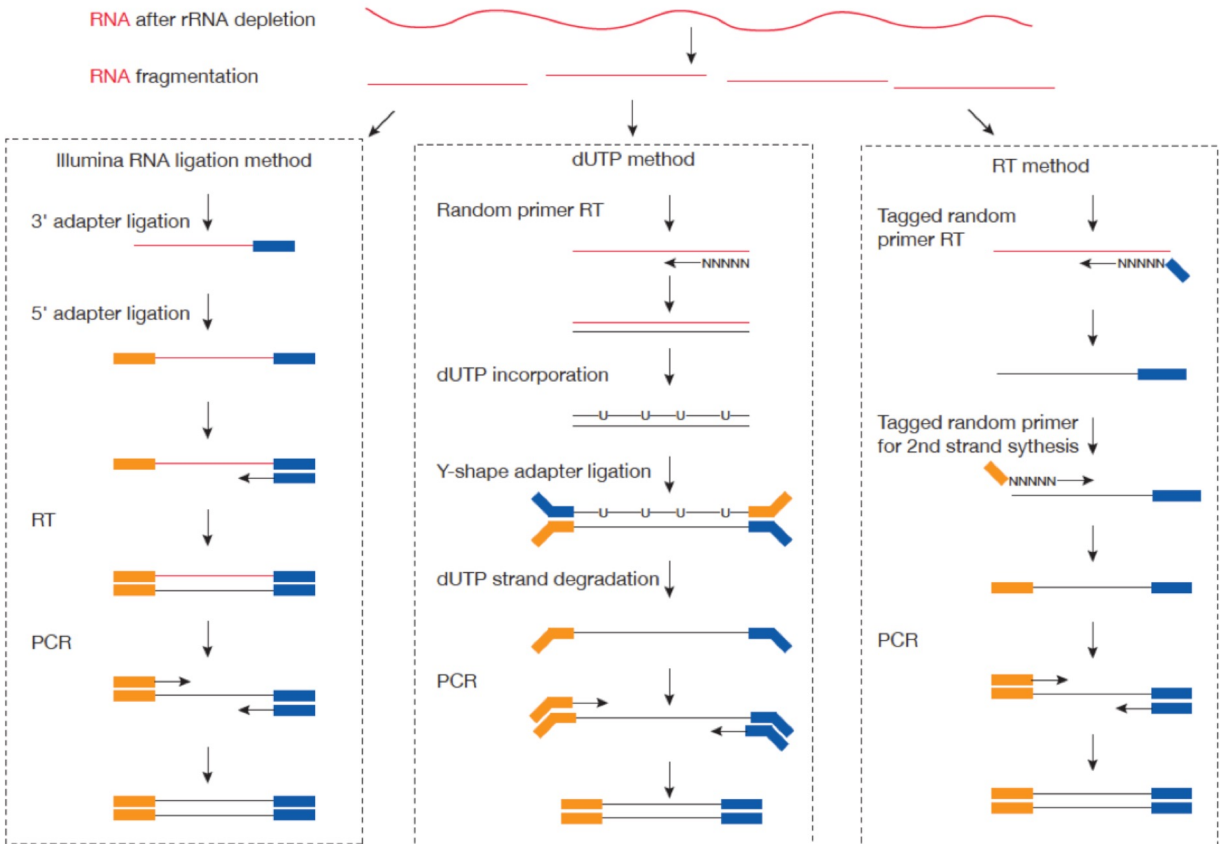
DIFFERENT WAYS TO CREATE RNASEQ LIBRARIES

• **Library preparation kits:**
– standard Illumina Tru-seq kit (small RNA or mRNA)
– average insert size of about 200 bp

• **Regular libraries:**
– prokaryotes: rRNA depletion or synthetic A-tailing purification
– eukaryotes: Poly-A purification or rRNA depletion

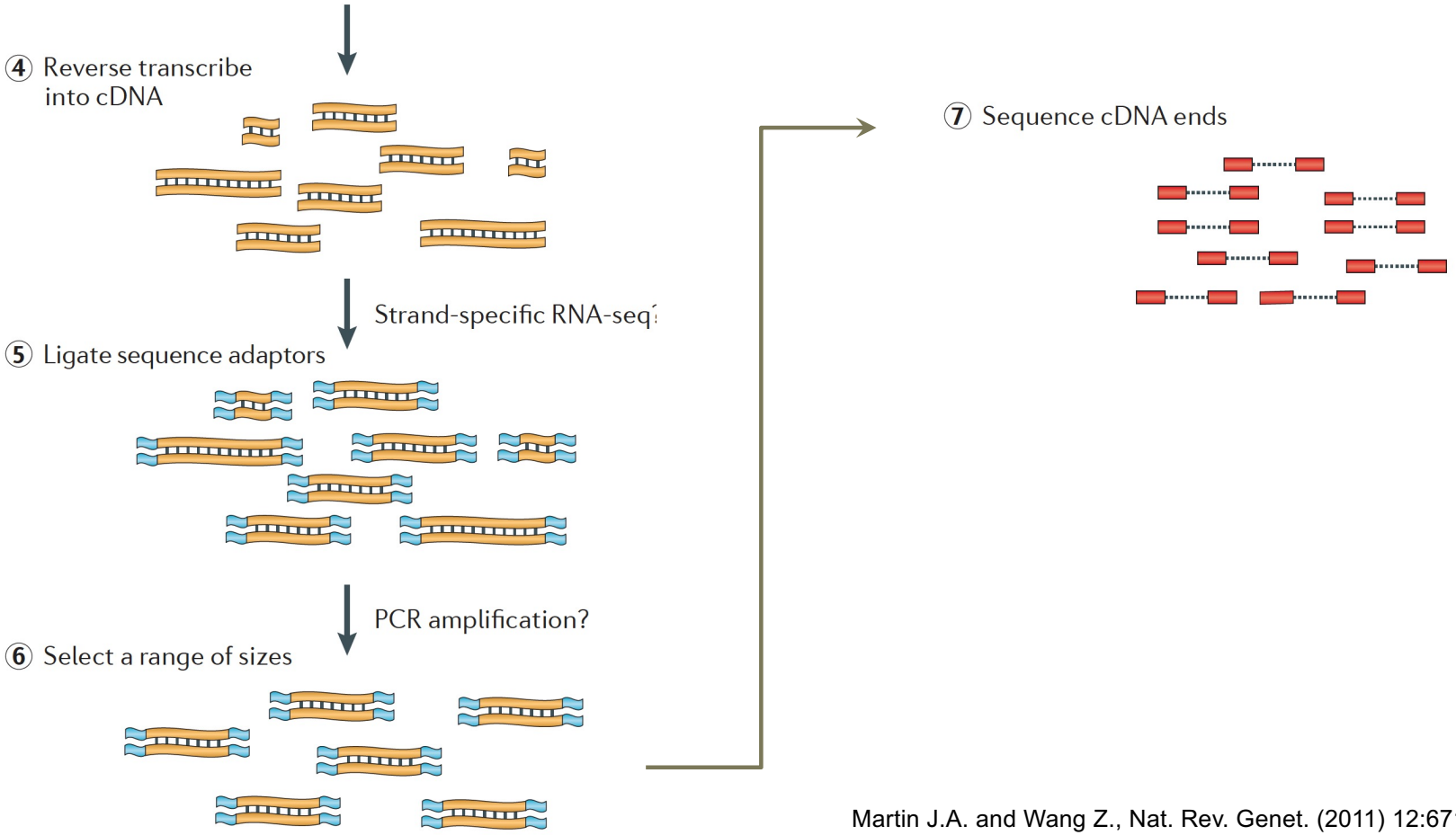
• **Normalized libraries (single cell sequencing, low abundant RNAs):**
– double strand nuclease normalization:
* denature double-stranded RNAs
* re-hybridization
* cleavage with double strand-specific nuclease -> abundant RNAs are more likely to re-hybridize

RNASEQ LIBRARIES ARE OFTEN ‚STRANDED'



Stranded libraries are, for example, relevant for the detection of anti-sense transcripts, which often have regulatory functions

FROM RNA TO SEQUENCE DATA

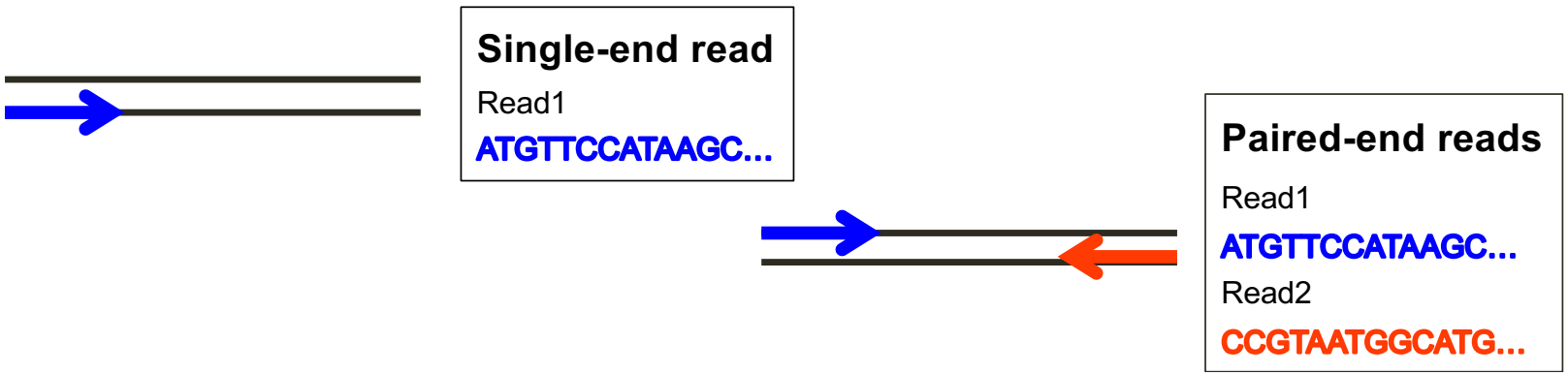


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

RNA-SEQ - EXPERIMENTAL AND PRACTICAL CONSIDERATIONS

Single-end (SE) or Paired end (PE)?

- SE is most common for DGE analysis
- PE is best for assemblies, for isoform differentiation, and for paralogous & orthologous gene differentiation (i.e. high-ploidy genomes & metatransciptomes)



IRRESPECTIVE OF WHAT YOU DO: RNA QUALITY IS CRUCIAL!

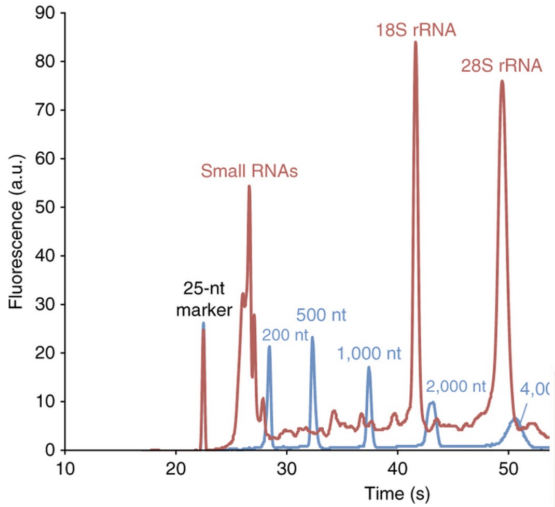
Take home message

Garbage in garbage out…

Thus, make sure you start your analysis with high quality RNA…

But how to assess extraction quality, if gene length, and thus transcript length is not uniform?

RNA EXTRACTION – RIBOSOMAL RNAS HELP TO DETERMINE RNA QUALITY

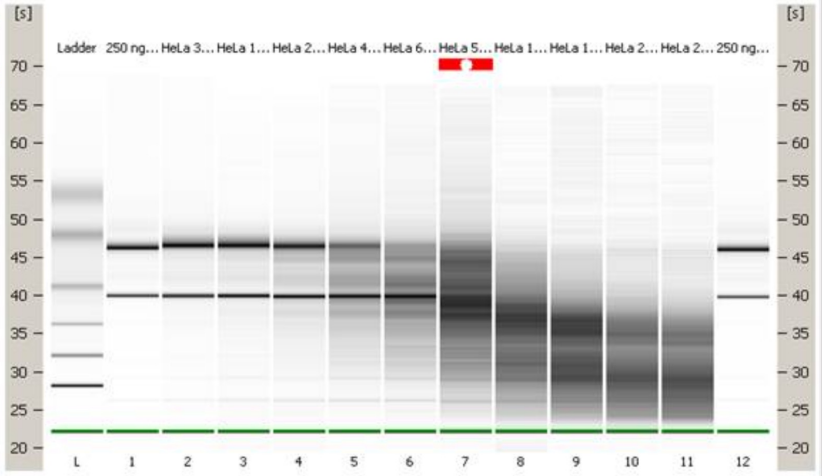


The idea

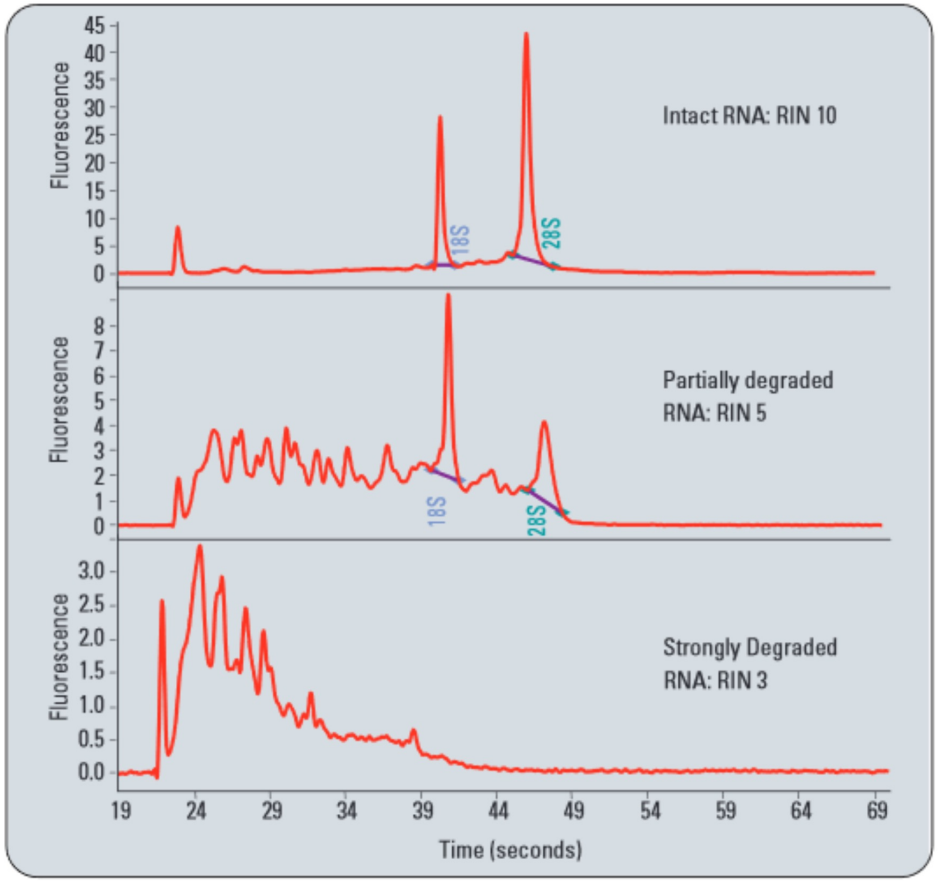
- rRNA is the most abundant RNA species in a cell
- The two large rRNAs, 18S and 28S, form two prominent peaks in a size separation of total RNA on a gel
- The more smear on a gel, the more the two rRNAs (but also the other RNA molecules) are degrated

The application

- Checking RNA integrity by using the two large rRNAs, 18S and 28S, as marker
- Categorization to RIN factors
- RIN10 – perfect quality; RIN1 – fully degrated

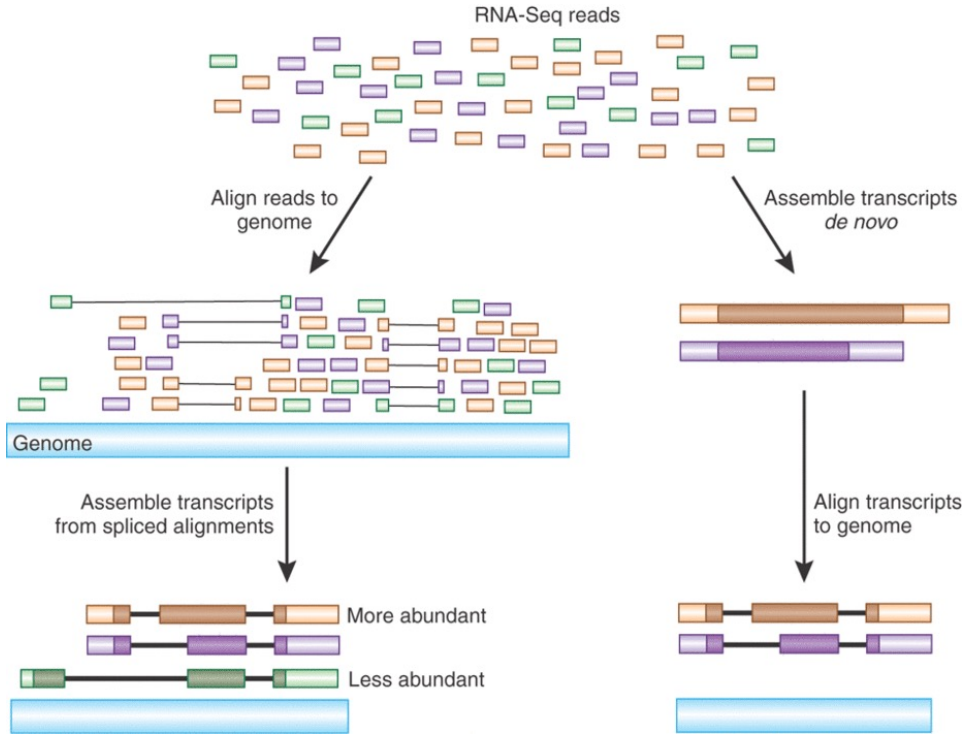


RNA EXTRACTION – QUALITY CHECKING



TWO FLAVORS OF RNASEQ ASSEMBLY: REFERENCE BASED AND DE-NOVO…OR GAPPED VS. UNGAPPED ANALYSIS

Reference based assembly - RNA seq reads are mapped against the corresponding genomic position. In a perfect world, reads map only to exonic regions and *split reads* identify exon-intron boundaries. Thus, splice-aware mappers are required



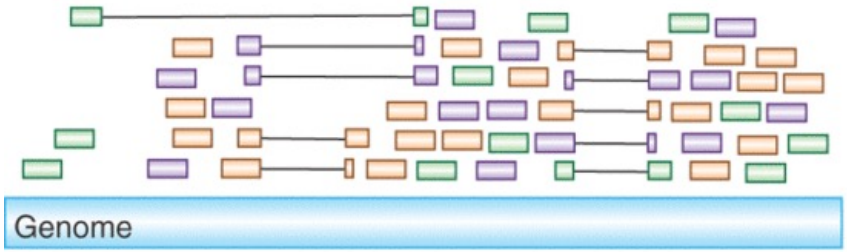
De-novo assembly – Overlapping sequence reads with sufficient sequence similarity are collapsed into longer sequences (aka *contigs*). The contigs serve as reconstructions of the original transcript. The assembly of RNA seq reads does not require the consideration of gaps due to exon-intron-boundaries

Haas and Zody (2010) *Nature Biotechnology* **28,** 421–423

ASSUMING THAT WE CAN PERFECTLY MAP, WHAT DO READ COUNTS TO A GENE/TRANSCRIPT TELL US ABOUT EXPRESSION LEVEL[1]?

We can, in principle, directly compare expression of a gene between different replicates and/or experimental conditions:

- library size & differences in sample pools
- experimental biases introduced at any stage



We have to solve the following problems

- longer mRNAs → sampled more frequently
- different sequences may be preferentially enriched during the sample preparation

Difficulties to keep genuine biological RNA abundance:

- small subsets of the total dataset → overfitting
- expression between samples → significant differences

1 Transcript abundance

COMPARISON OF NORMALIZATION METHODS

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

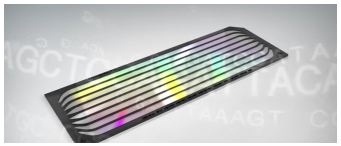
Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle ... Show more

Author Notes

Briefings in Bioinformatics, Volume 14, Issue 6, November 2013, Pages 671–683,
https://doi.org/10.1093/bib/bbs046

Published: 15 September 2012 **Article history** ▾

COMPARISON OF NORMALIZATION METHODS



- **Definitions**
- Sequencing technology: Illumina sequencing machines, data sets differ in their read length and overall throughput but share the same sequencing technology (flow cell).
- A flow cell is made up of eight independent sequencing areas, or ‘lanes’.
- A library contains cDNAs representative of the RNA molecules that are extracted from a given culture or tissue
- Libraries are deposited on these lanes in order to be sequenced.
- Similarly to microarrays, the library composition reflects the RNA repertoire expressed in the corresponding culture or tissue.
- ‘library size’ refers to the number of mapped short reads obtained from the sequencing process of the library.
- In this study (Dillies et al. 2013), a single library was sequenced in each lane.

RNA-SEQ NORMALIZATION METHODS (A SELECTION)

Because the most obvious source of variation between lanes is the differences in library size (i.e. sequencing depth), the simplest form of inter-sample normalization is achieved by scaling raw read counts in each lane by a single lane-specific factor reflecting its library size.

Total count (TC)

Gene counts are divided by the total number of mapped reads (or library size) associated with their lane and multiplied by the mean total count across all the samples of the dataset

Upper Quartile (UQ)

Like TC but total counts are replaced by the upper quartile of counts different from 0

Median (Med)

Like TC but total counts are replaced by the median counts different from 0

See Dillies et al. https://doi.org/10.1093/bib/bbs046 for further information

RNA-SEQ NORMALIZATION METHODS (A SELECTION)

Quantile (Q)

normalization method consists in matching distributions of gene counts across lanes

RPKM

re-scales gene counts to correct for differences in both library sizes and gene length

DESeq

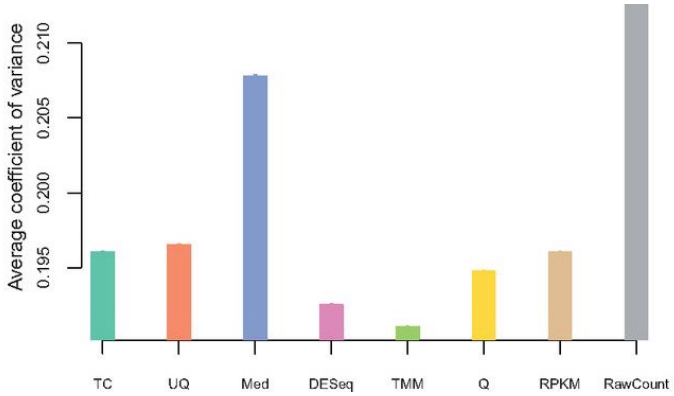
DESeq scaling factor for a given lane is computed as the median of the ratio, for each gene, of its read count over its geometric mean across all lanes. The underlying idea is that non-DE genes should have similar read counts across samples, leading to a ratio of 1

IMPORTANT RNA-SEQ NORMALIZATION

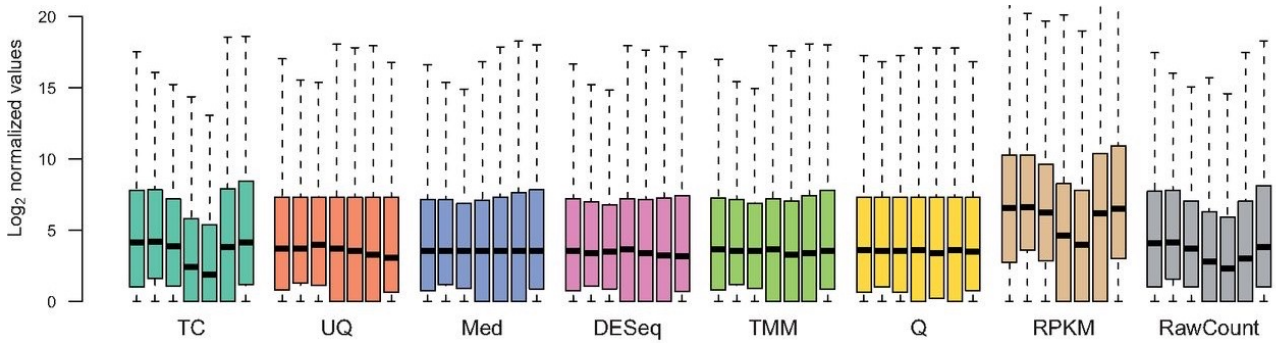
	Normalization Method	Reference
Library size	RPM (Reads Per Million) 1. Includes library size normalization	Tarazona et al. 2011
Gene length	TPM (Transcripts Per Million) 1. Average length of transcripts + library size	Wagner et al. 2012
Scaling factors	DESeq/DESeq2 1. Included in the DESeq Bioconductor package (version 1.6.0)	Anders & Huber 2010

COMPARISON OF NORMALIZATION METHODS

Variation in expression among a set of 30 housekeeping genes in the human data, which may be assumed to be similarly expressed across samples.

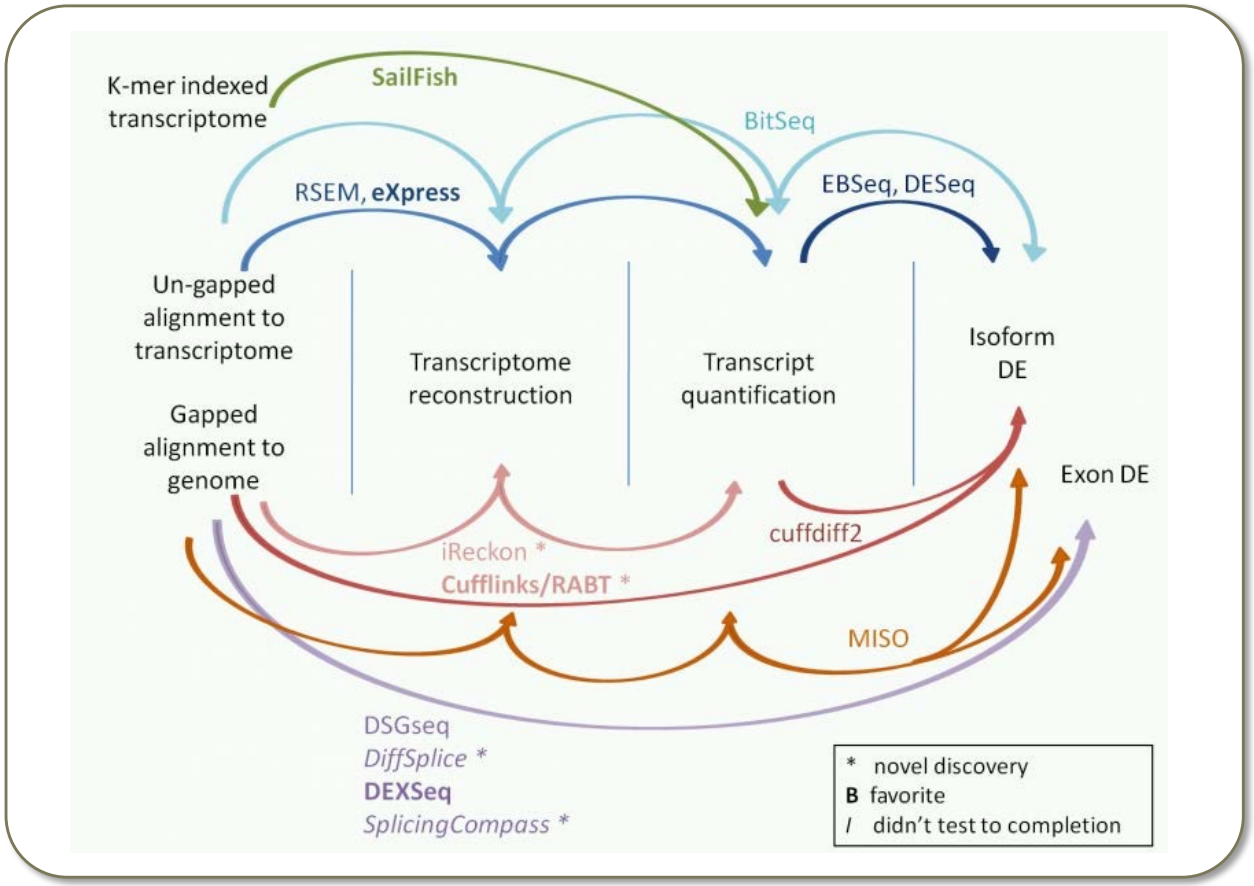


in typical DE analyses the majority of genes under consideration are assumed to be non-differentially expressed between conditions. For this reason, it is useful to examine boxplots of counts across samples in each dataset, both before and after normalization; an effective normalization scheme should result in a stabilization of read count distributions across replicates.

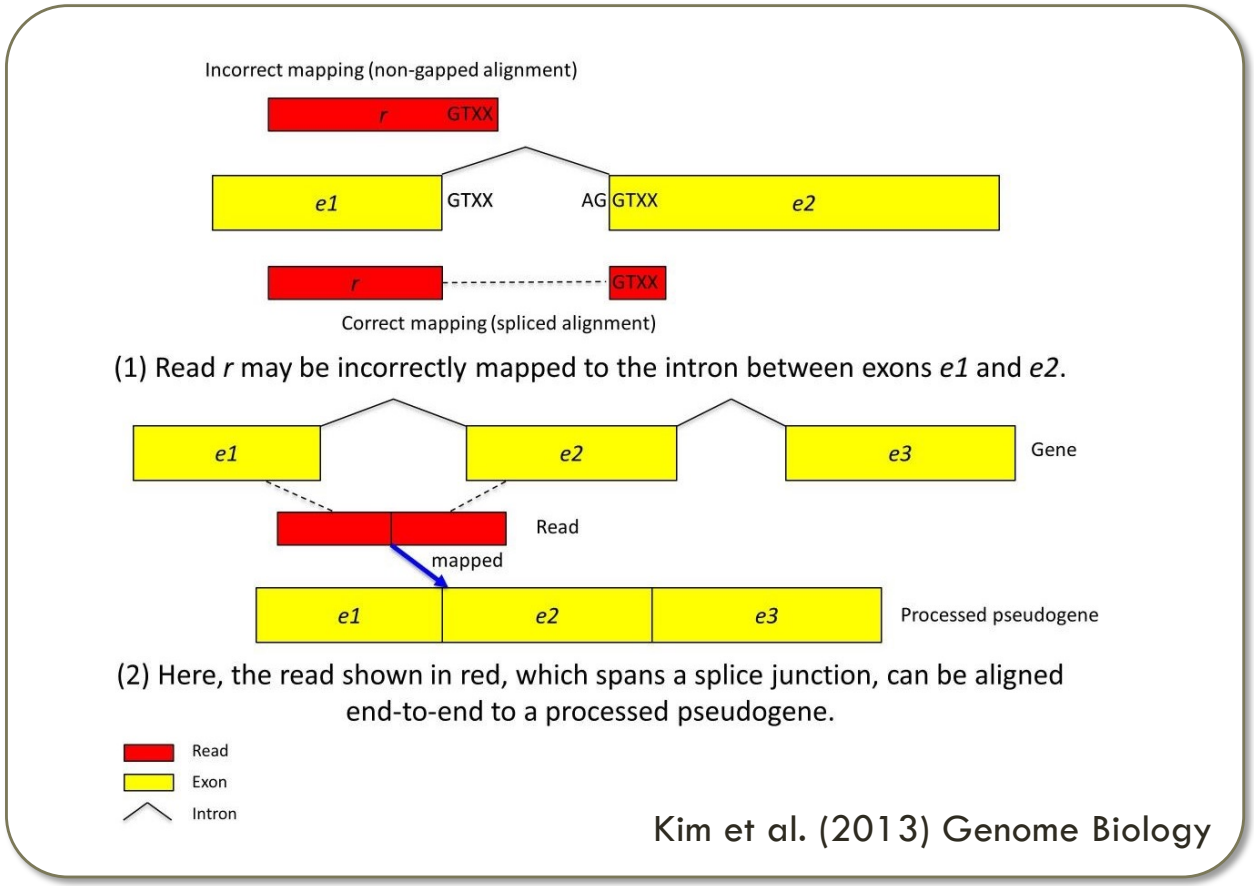


Marie-Agnès Dillies et al. Brief Bioinform 2012;bib.bbs046

COMING BACK TO THE MAPPING PROBLEM – TWO MAIN FLAVORS OF RNASEQ MAPPING – GAPPED VS UNGAPPED



MAPPING RNASEQ TO A GENOME – WHERE ARE THE ISSUES?



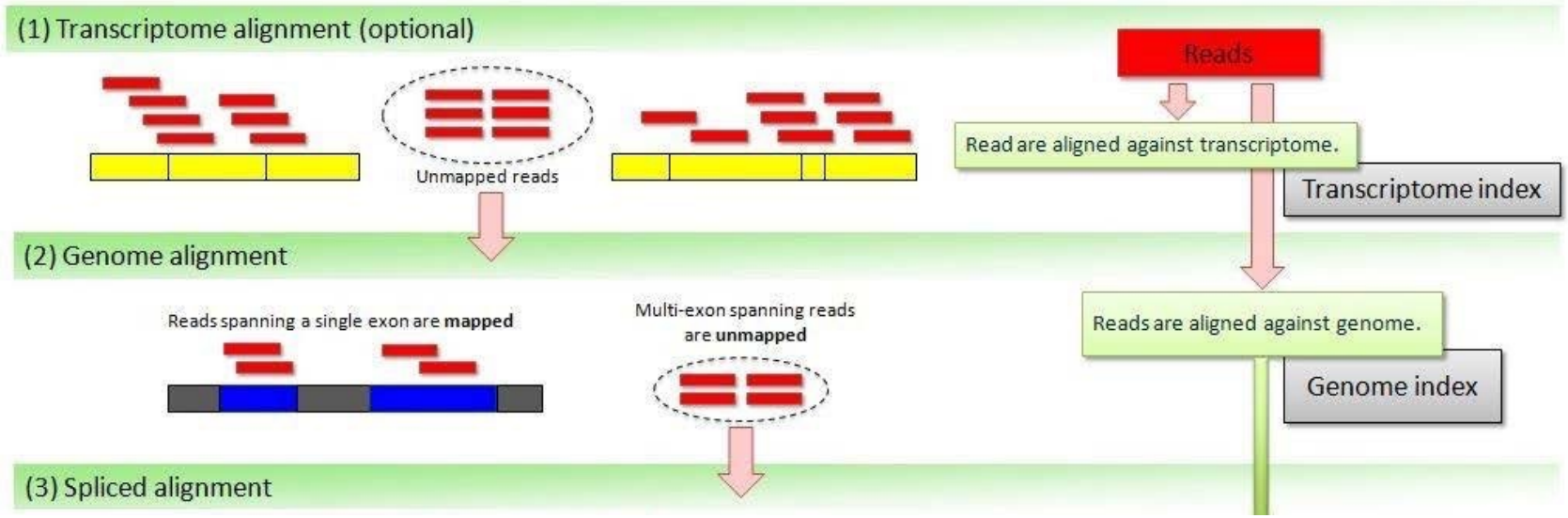
SPLICE AWARE ALIGNMENTS

***Splice-aware* alignment algorithms**

- TopHat1/2[1]
- HiSat2
- SpliceMap[2]
- GSNAP[3]
- STAR[4]

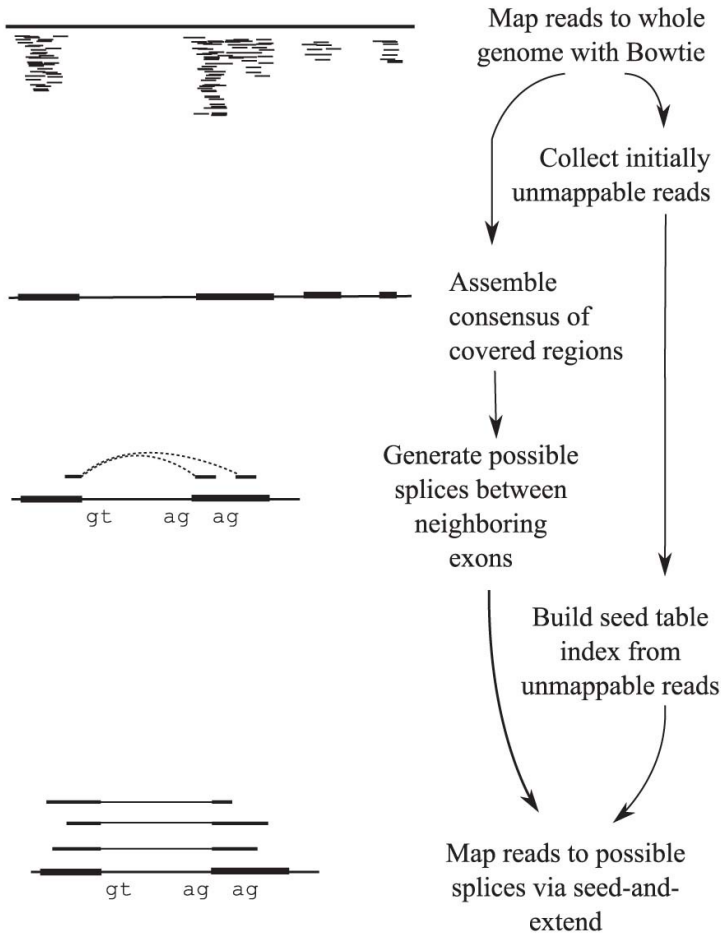
1 Wang et al. (2008) Nature 456: 470-6; 3 Robertson et al. (2010) Nat Methods 7: 909-12.
2 Wu and Nacu (2010) Bioinformatics 26: 873-81. 4 Dobin et al. (2013) Bioinformatics 29: 15-21.

THE TASK – SPLICE-AWARE MAPPING OF RNASEQ READS



- known junction signals (GT-AG, GC-AG, and AT-AC)
- Re-map due to edit distance
- split into smaller non-overlapping segments (25 bp)
- left and right segments of the same read are mapped
- within a user-defined maximum intron size

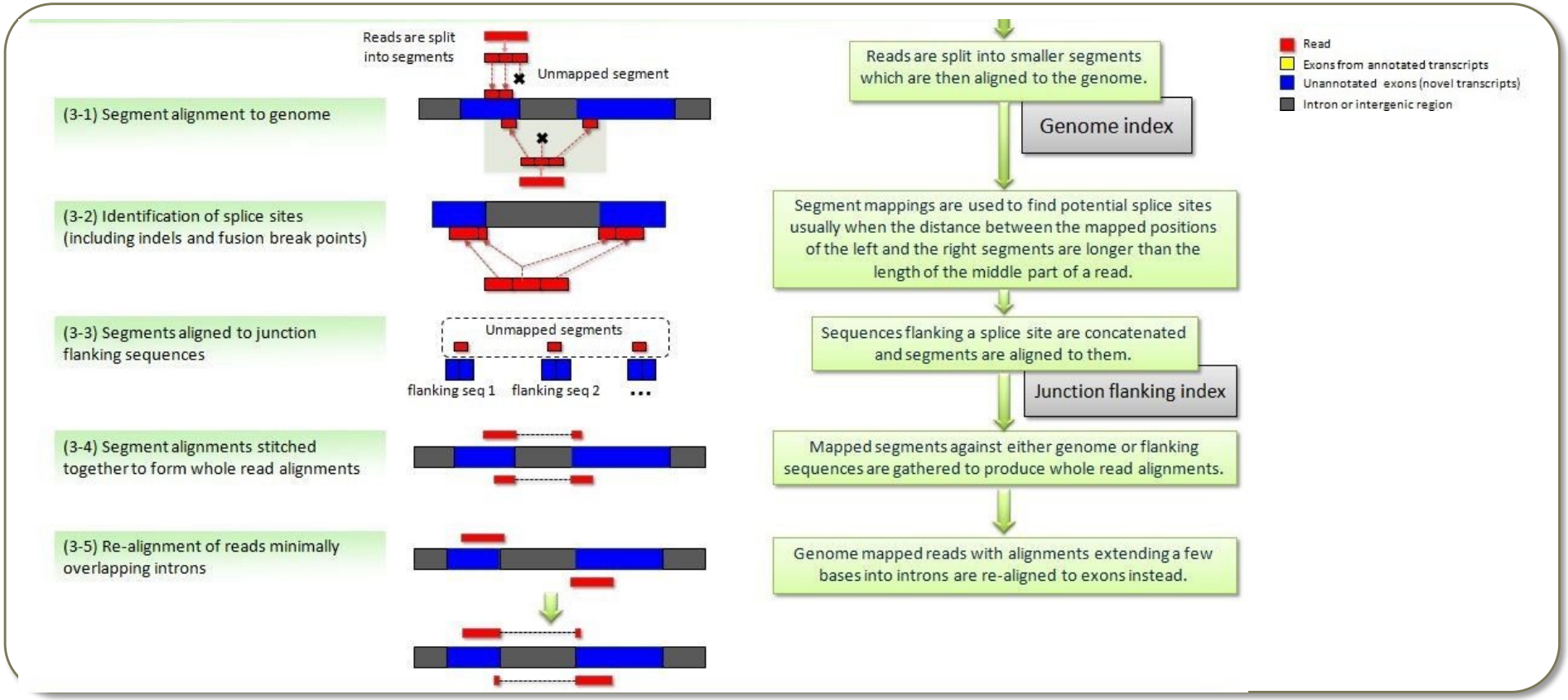
SPLICE-AWARE ALIGNMENT I - TOPHAT



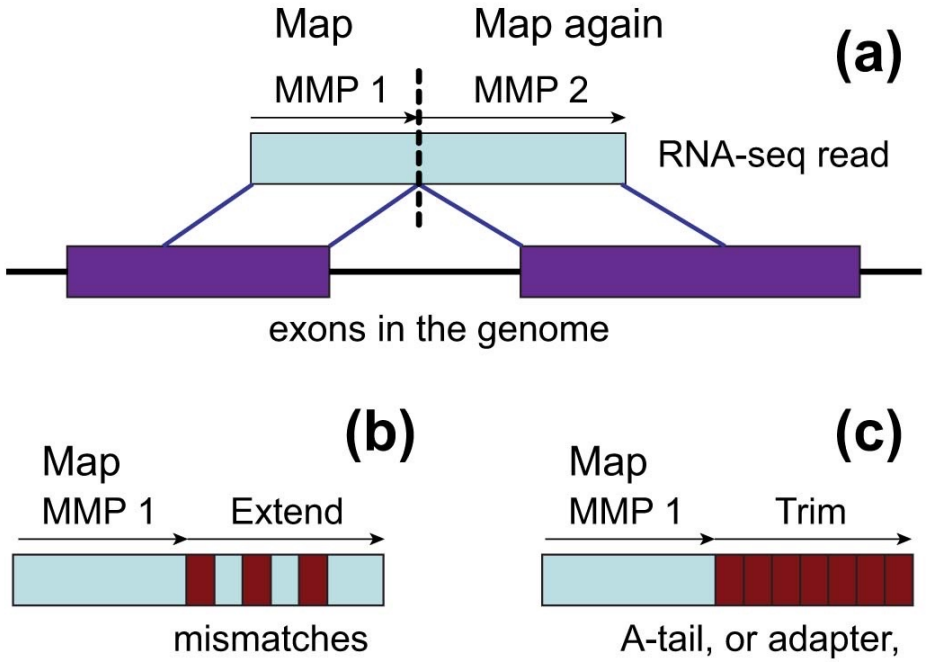
1. Genomic alignment via Bowtie
2. 'Islands of reads' show probable Exonic regions
3. Connection of exons via spliced reads using seed-and-extend strategy

Wang et al. (2008) Nature 456: 470-6

SPLICE-AWARE ALIGNMENT – TOPHAT2 ALIGNMENTS OF READS



SPLICE-AWARE ALIGNMENT II - STAR



a) Maximum Mappable Prefix

1. Look for the longest substring of a read that can be continuously aligned to a genome
2. Repeat with the remaining part of the read

b) Connect all seeds that could be placed in a genome via an extend step

c) Trim terminal poorly mapping regions

Dobin et al. (2013) Bioinformatics 29: 15-21

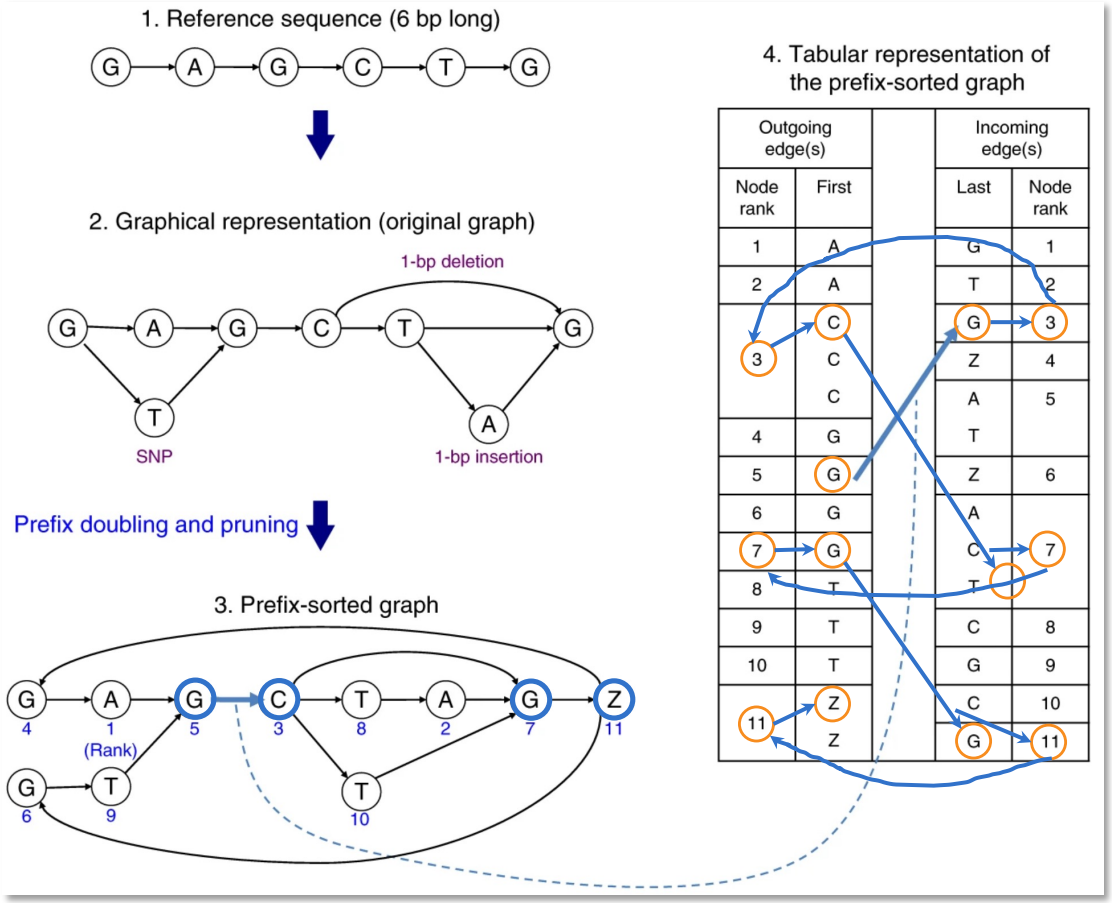
HISAT2

Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype

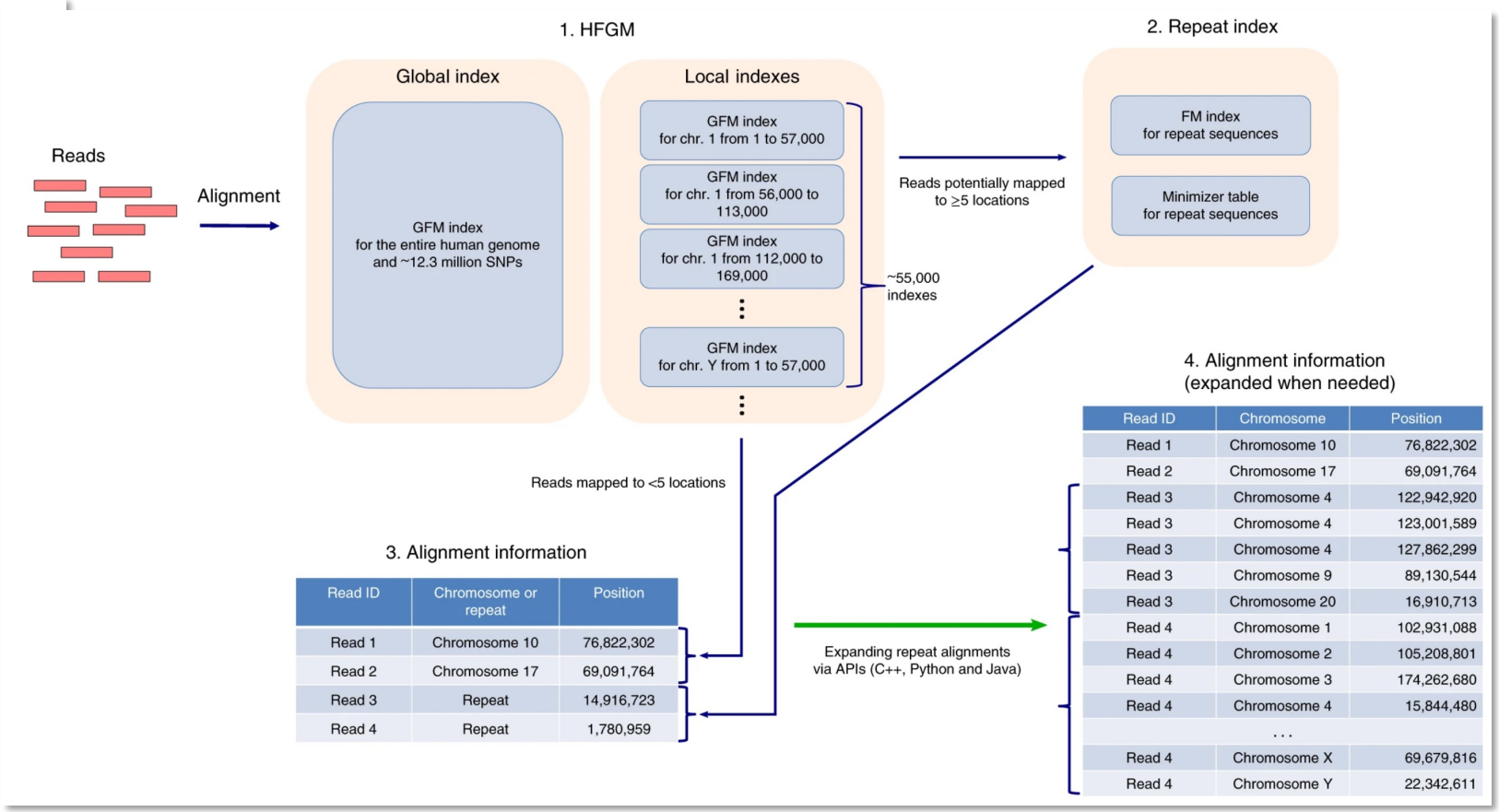
Daehwan Kim[1*], Joseph M. Paggi[2], Chanhee Park[1], Christopher Bennett[1] and Steven L. Salzberg[3,4]

The human reference genome represents only a small number of individuals, which limits its usefulness for genotyping. We present a method named HISAT2 (hierarchical indexing for spliced alignment of transcripts 2) that can align both DNA and RNA sequences using a graph Ferragina Manzini index. We use HISAT2 to represent and search an expanded model of the human reference genome in which over 14.5 million genomic variants in combination with haplotypes are incorporated into the data structure used for searching and alignment. We benchmark HISAT2 using simulated and real datasets to demonstrate that our strategy of representing a population of genomes, together with a fast, memory-efficient search algorithm, provides more detailed and accurate variant analyses than other methods. We apply HISAT2 for HLA typing and DNA fingerprinting; both applications form part of the HISAT-genotype software that enables analysis of haplotype-resolved genes or genomic regions. HISAT-genotype outperforms other computational methods and matches or exceeds the performance of laboratory-based assays.

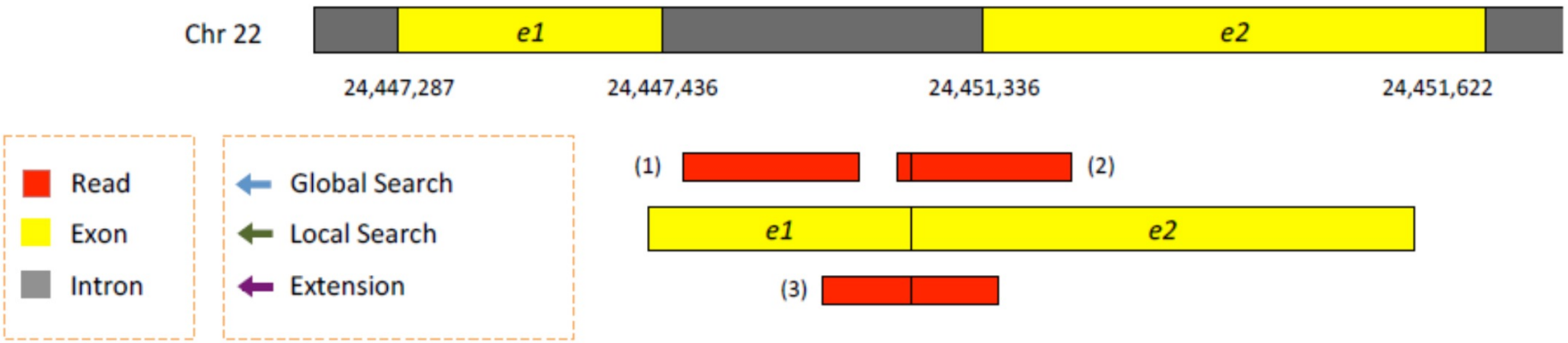
HISAT2 - INDEXING



HISAT2 – WORKING WITH DIFFERENT INDICES: GLOBAL, LOCAL & REPEAT



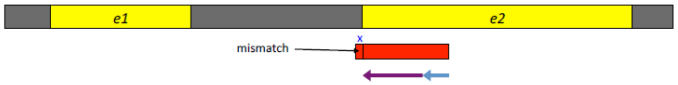
THE GENERAL IDEA OF HISAT2



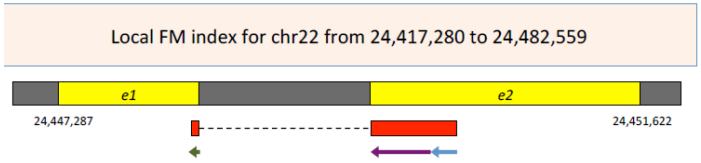
Global Search: a global FM index that represents the entire genome
Local Search: numerous small FM indexes for regions that collectively cover the genome, where each index represents 57,000 bp

SPLICE-AWARE ALIGNMENT – HISAT2 ALIGNMENT

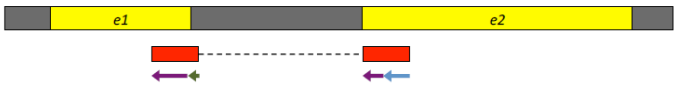
1. reads that map within an exon



2. … across two exons with 8–15 bp mapping to one exon

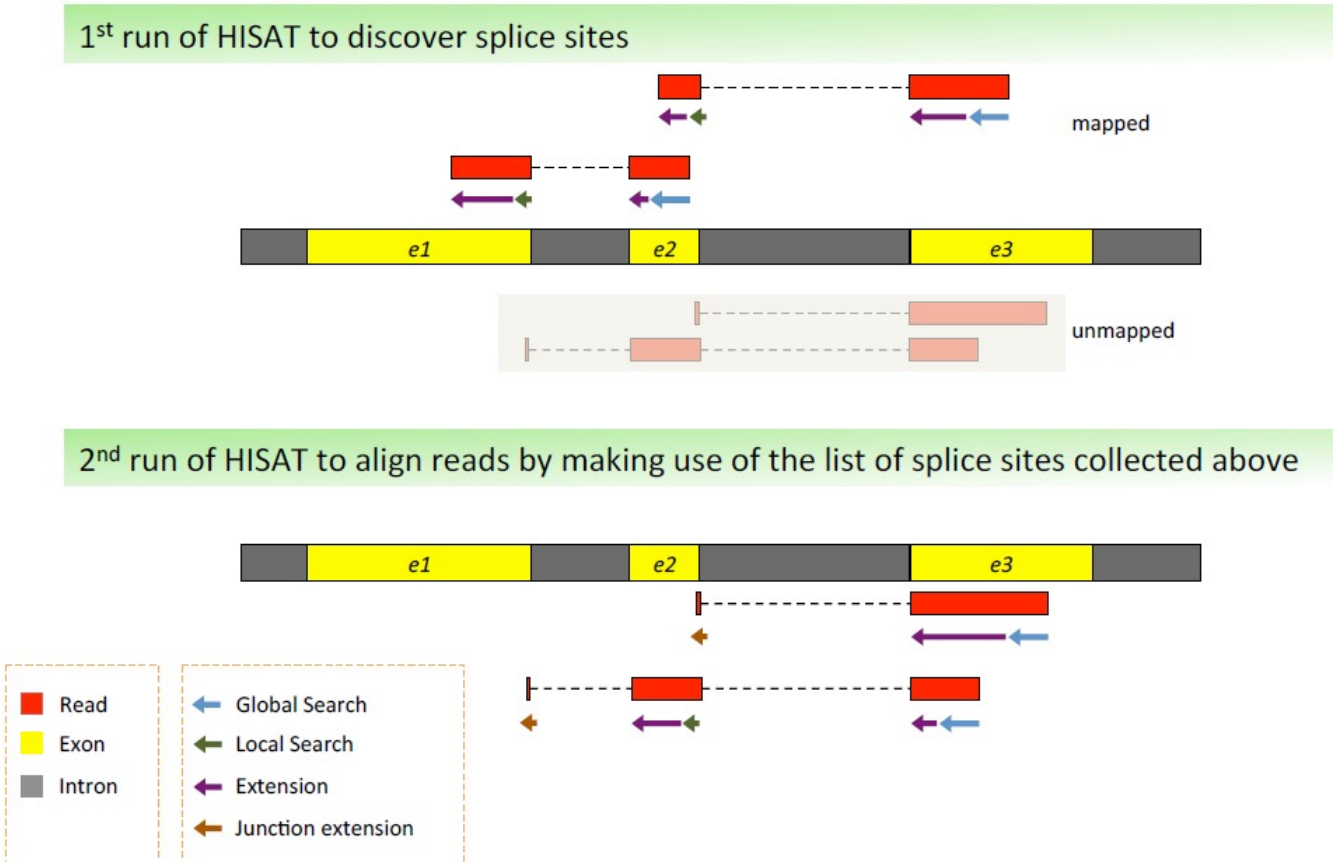


3. across two exons with at least 15 bp in each exon



(a) it is at least 28 bp long and (b) it maps onto exactly one location

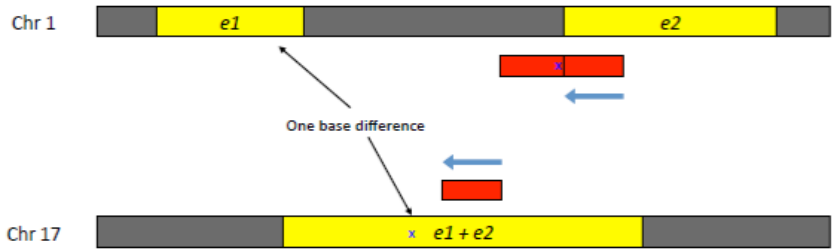
HISAT RUNS IN TWO PHASES



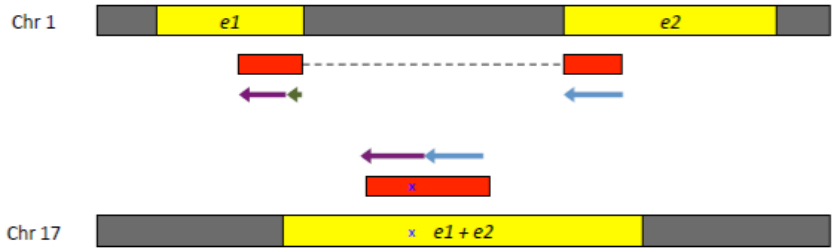
Phase 1 requires a sufficient coverage to detect splice junctions / splice sites

Phase 2 makes use of splice sites to map reads only barely overlapping an exon

SPLICE-AWARE ALIGNMENT – HISAT2 HANDLING PSEUDOGENES

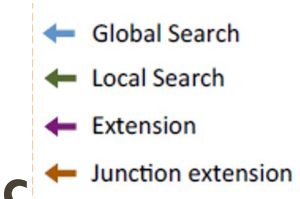


(i) Mapping of both locations long enough for partial mapping

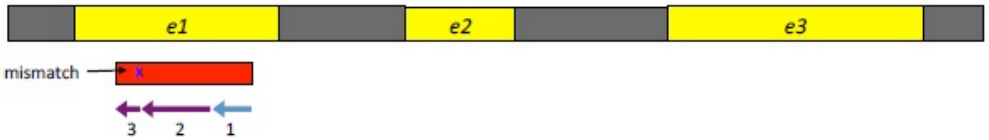


(ii) One locationby local indexing, the other by extension with one mismatch

SPLICE-AWARE ALIGNMENT – HISAT2 ERRORS IN READS

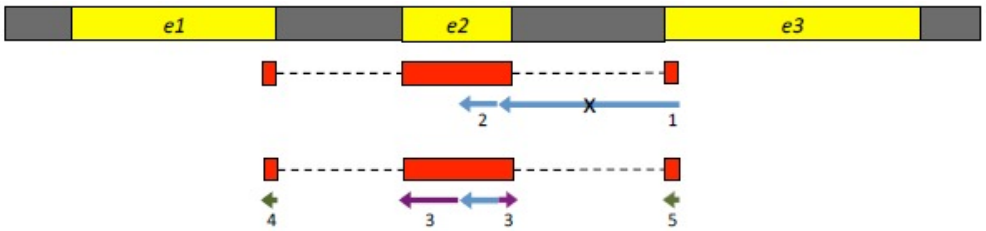


(1)

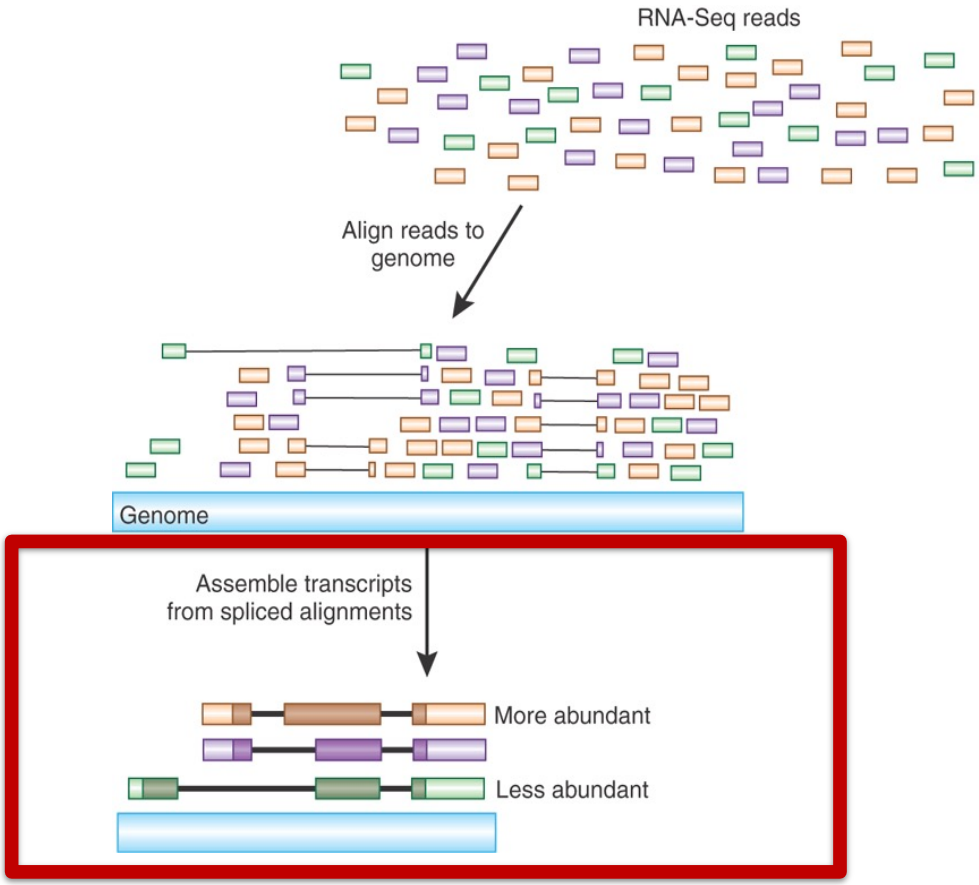


(2)

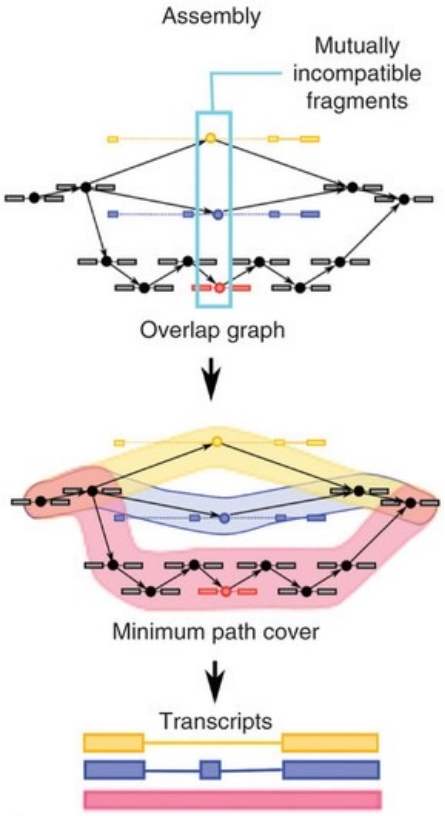
(3)



AFTER MAPPING COMES TRANSCRIPT RECONSTRUCTION



TRANSCRIPT RECONSTRUCTION



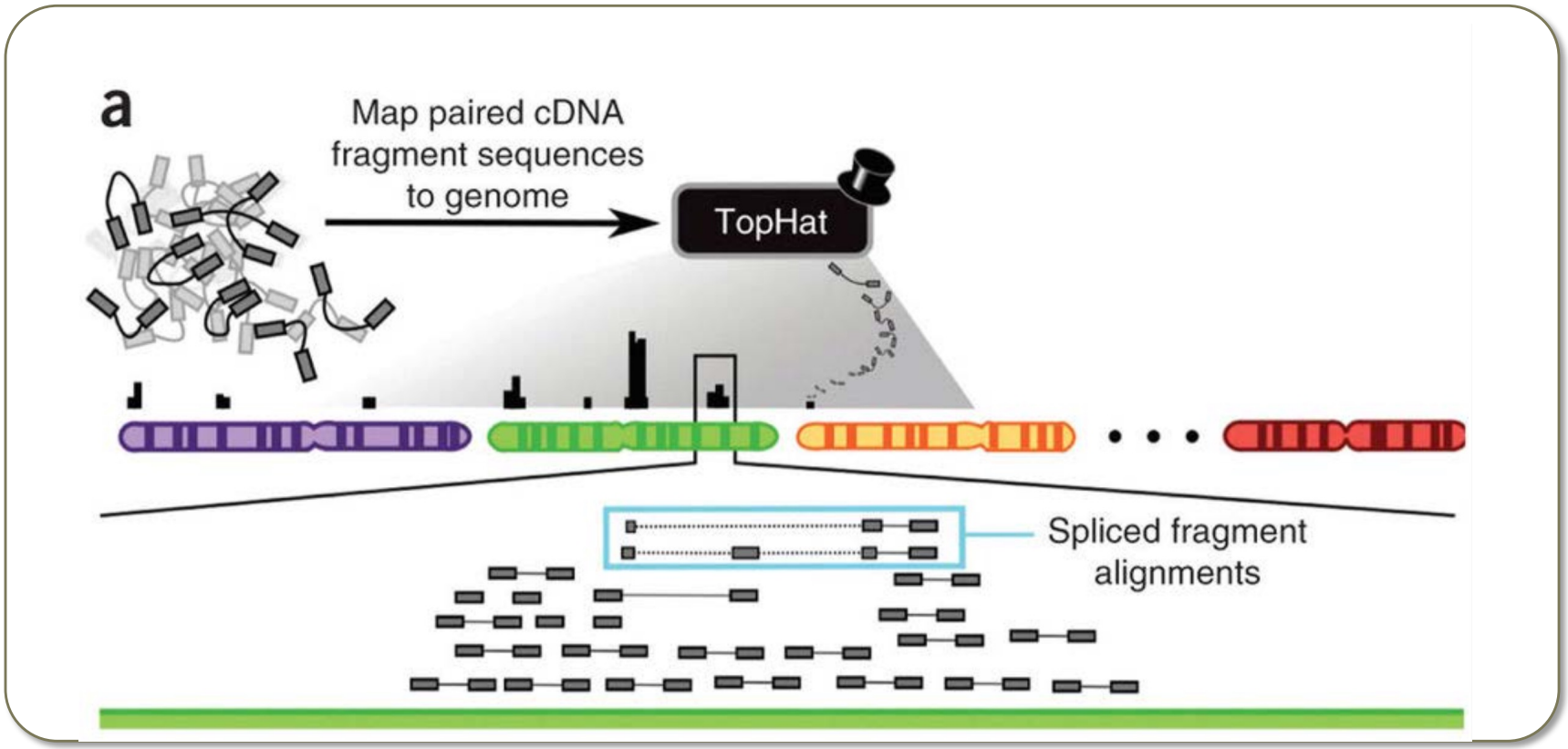
Cufflinks[1]: minimal set of compatible isoforms (maximum precision)

Scripture[2]: all isoforms that are compatible with the read data (maximum sensitivity)

MISO[3]: estimate expression of alternatively spliced exons and isoforms

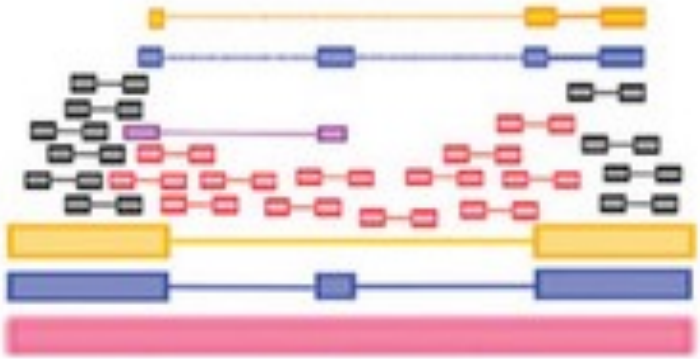
1 Trapnell et al. (2010) Nat Biotechnol 28: 511-15
2 Guttman et al. (2010) Nat Biotechnol 28: 503-10
3 Katz et al. (2010) Nat Methods 7: 1009-15

CUFFLINKS FOLLOWS TOPHAT



Trapnell et al. (2010) Nat Biotechnol 28: 511-15

CUFFLINKS – DEFINE POSSIBILE TRANSCRIPTS



1. With paired-end RNA-Seq, Cufflinks treats each pair of **fragment reads** as a single alignment. The algorithm assembles overlapping 'bundles' of fragment alignments
2. Every fragment is consistent with at least one assembled transcript
3. Every transcript is tiled by reads
4. The number of transcripts is the smallest required to satisfy requirement 1

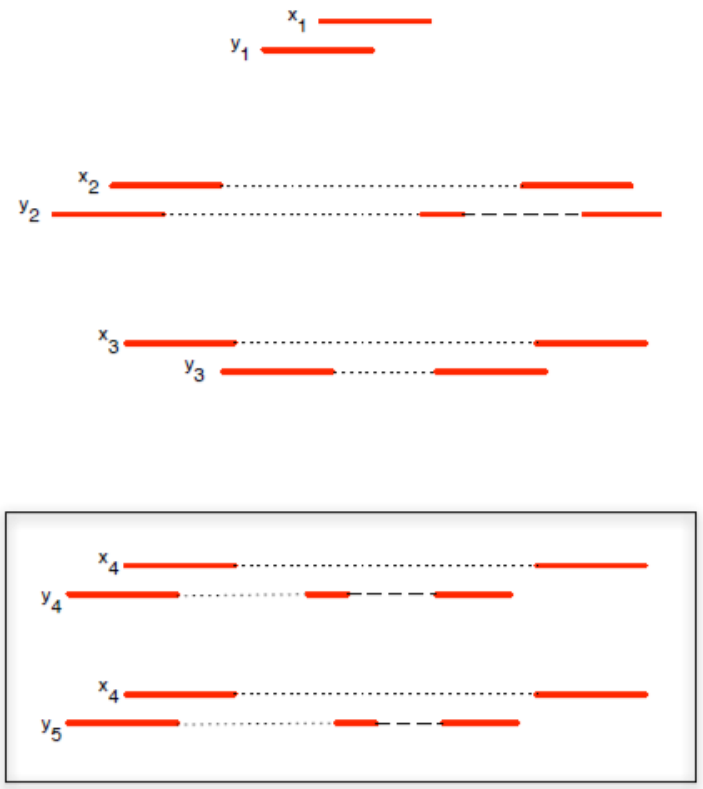
Bundle: Set of overlapping fragments representing splice variants of only one or few genes

CUFFLINKS – TRANSCRIPT ASSEMBLY

- directed acyclic graph (DAG):
  - with one node for each fragment
  - Each fragment→aligned pair of mated reads
- Fragment (paired-end) alignments are of two types:
  - those where reads align in their entirety to the genome
  - reads which have a split alignment (due to an implied intron)
- Single reads checking for compatibility:
  - Two reads are compatible if their overlap contains the exact
  - same implied introns (or none)

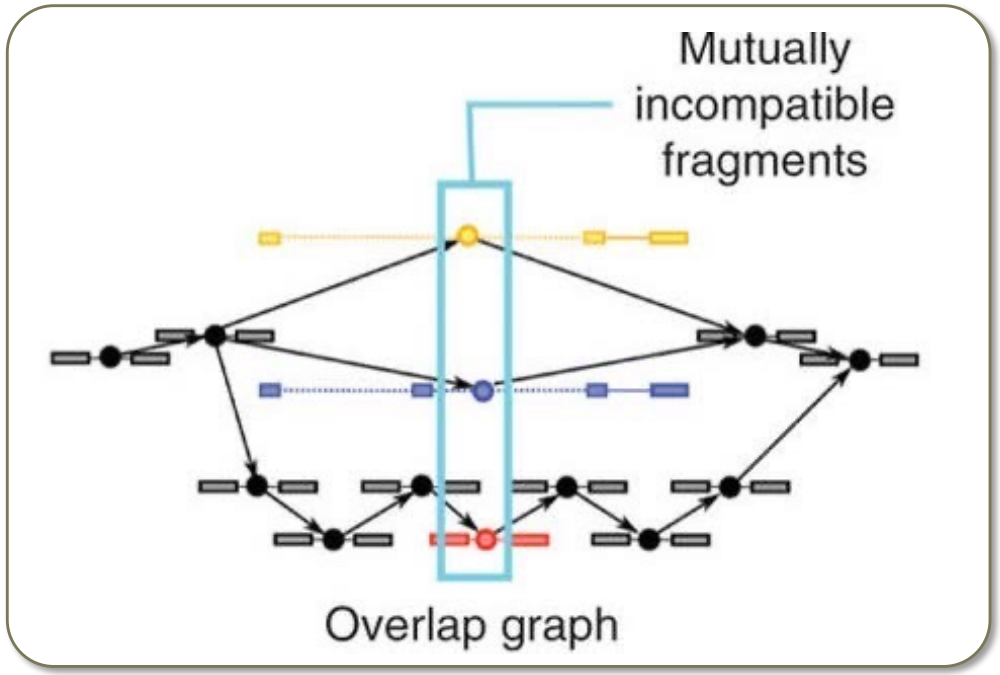
Trapnell et al. (2010) Nat Biotechnol 28: 511-15

CUFFLINKS – TRANSCRIPT ASSEMBLY CASES



- Compatible
- Incompatible
- Nested
- Uncertainty of x4
- Y4 & y5 incompatible

CUFFLINKS – CREATE GRAPH OF PUTATIVE TRANSCRIPTS

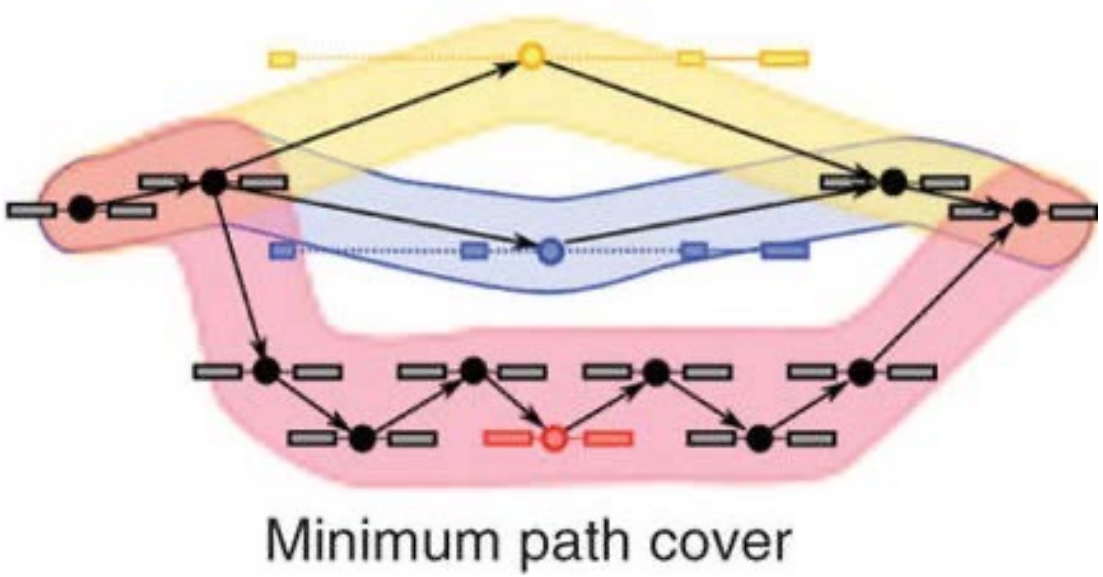


Within a ‚bundle‘
Identify pairs of ‘incompatible’ fragments that must have originated from distinct spliced mRNA isoforms. Fragments are connected in an ‘overlap graph’ when they are compatible and their alignments overlap in the genome. Each fragment has one node in the graph, and an edge, directed from left to right along the genome, is placed between each pair of compatible fragments.

Example:
yellow, blue, and red fragments must have originated from separate isoforms, but any other fragment could have come from the same transcript as one of these three

Bundle: Set of overlapping fragments representing splice variants of only one or few genes

CUFFLINKS – DEFINE POSSIBILE PATH WITHIN THE GRAPH

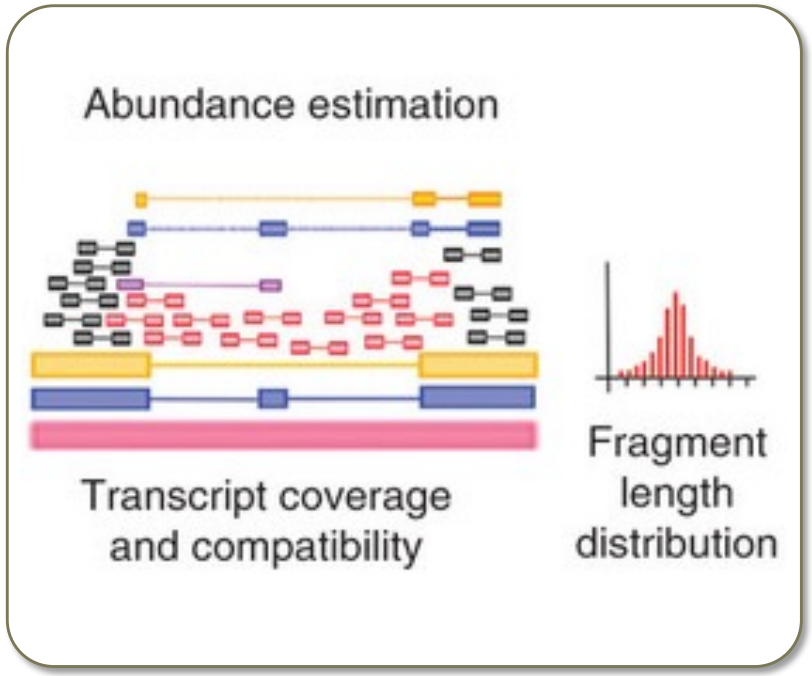


Minimum path cover

Minimal Path coverage to determine number of alternative transcripts

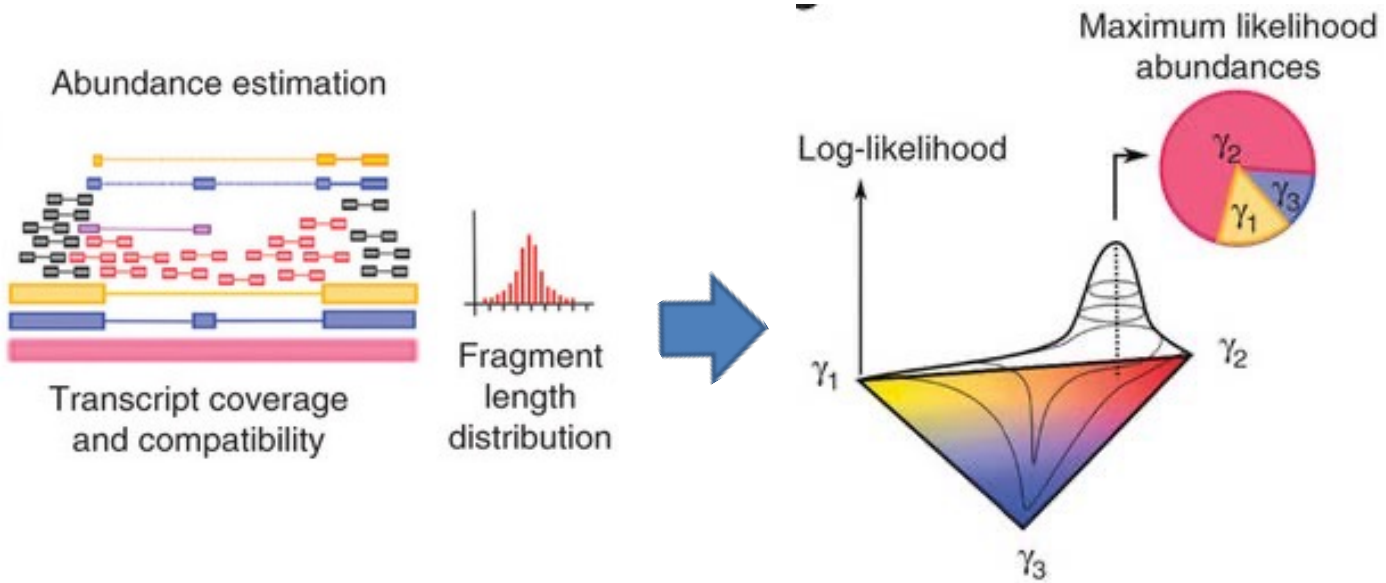
The overlap graph here can be minimally ‘covered’ by three paths, each representing a different isoform. Dilworth's Theorem states that the number of mutually incompatible reads is the same as the minimum number of transcripts needed to “explain” all the fragments.

CUFFLINKS - TRANSCRIPT ABUNDANCE AND PROBABILITY



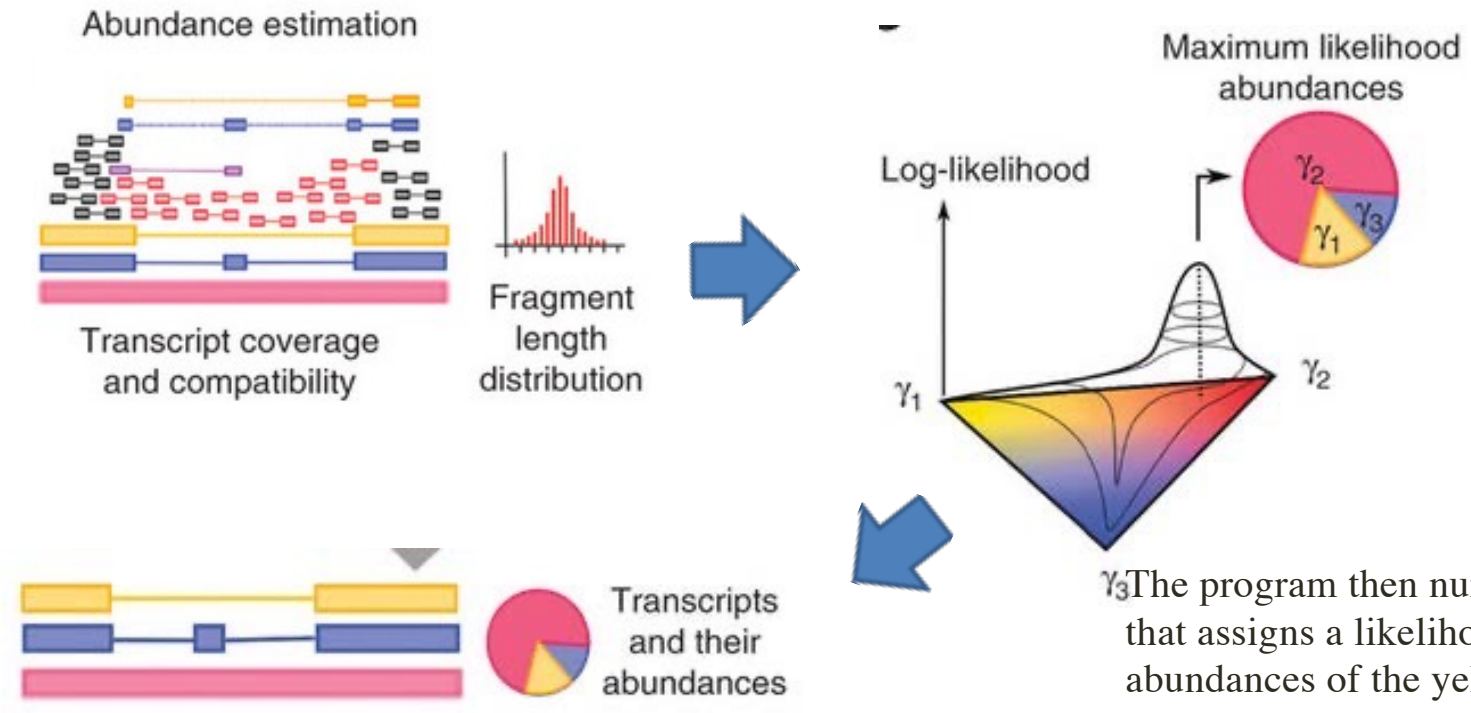
Fragments are matched (denoted here using color) to the transcripts from which they could have originated. The violet fragment could have originated from the blue or red isoform. Gray fragments could have come from any of the three shown.

CUFFLINKS - TRANSCRIPT ABUNDANCE AND PROBABILITY



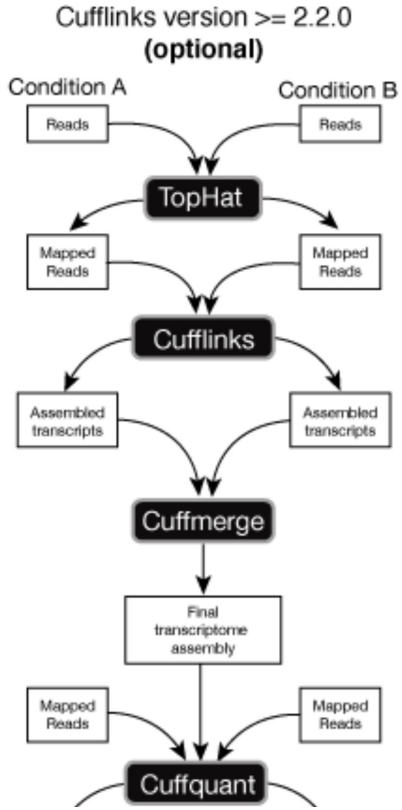
- Cufflinks estimates transcript abundances using a statistical model in which the probability of observing each fragment is a linear function of the abundances of the transcripts from which it could have originated.
- Because only the ends of each fragment are sequenced, the length of each may be unknown. Assigning a fragment to different isoforms often implies a different length for it. Cufflinks can incorporate the distribution of fragment lengths to help assign fragments to isoforms. For example, the violet fragment would be much longer, and very improbable according to Cufflinks' model, if it were to come from the red isoform instead of the blue isoform.

CUFFLINKS - TRANSCRIPT ABUNDANCE AND PROBABILITY



γ_3The program then numerically maximizes a function that assigns a likelihood to all possible sets of relative abundances of the yellow, red and blue isoforms (γ_1,γ_2,γ_3), producing the abundances that best explain the observed fragments, shown as a pie chart.

THE WORKFLOW OF TRANSCRIPT RECONSTRUCTION FROM ASSEMBLED READS



Cufflinks → Workflow

TopHat / HISAT:
Splice-aware aligner for read mapping

Cufflinks:
Assemble transcripts from spliced reads

Cuffmerge:
Assembled transcriptomes are used for a master transcriptome

Cuffquant:
Quantifying the transcript expression as profiles

Trapnell et al. (2010) Nat Biotechnol 28: 511-15
http://cole-trapnell-lab.github.io/cufflinks/manual/