RNA-Seq

Harm Nijveen

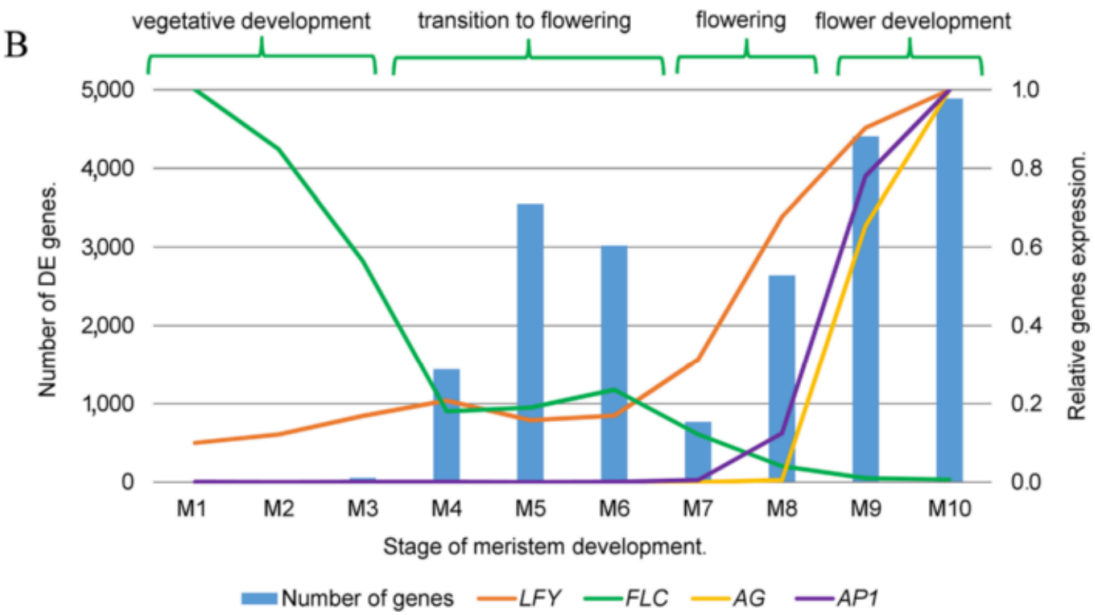
Klepikova *et al. BMC Genomics* (2015) 16:466
DOI 10.1186/s12864-015-1688-9

BMC Genomics

RESEARCH ARTICLE Open Access

RNA-seq analysis of an apical meristem time series reveals a critical point in *Arabidopsis thaliana* flower initiation

Anna V. Klepikova[1], Maria D. Logacheva[1,2,3], Sergey E. Dmitriev[2,4] and Aleksey A. Penin[1,2,5*]



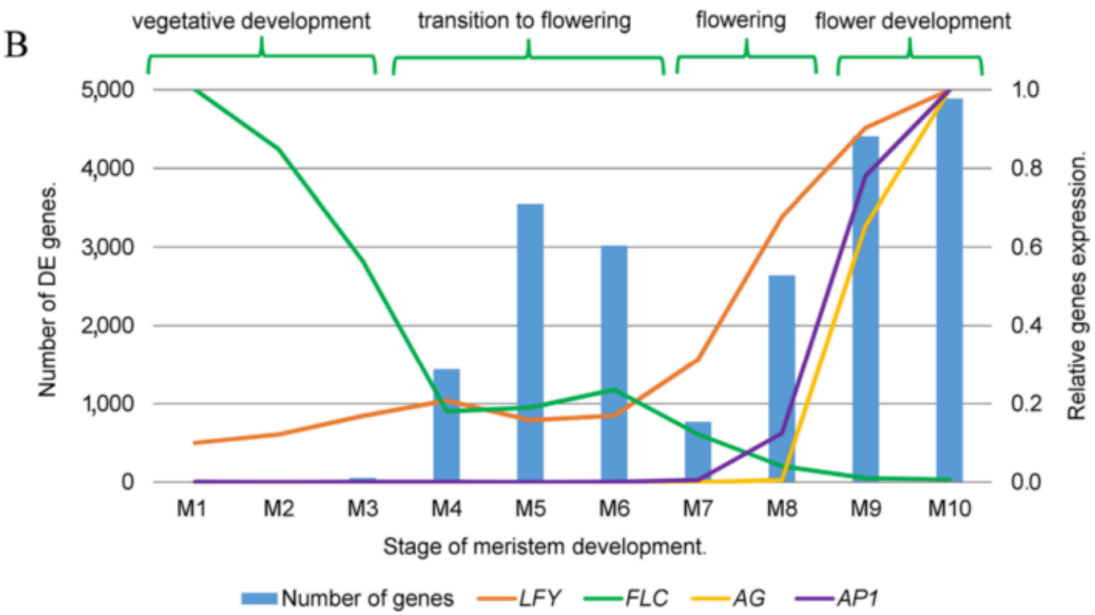
Klepikova *et al. BMC Genomics* (2015) 16:466
DOI 10.1186/s12864-015-1688-9

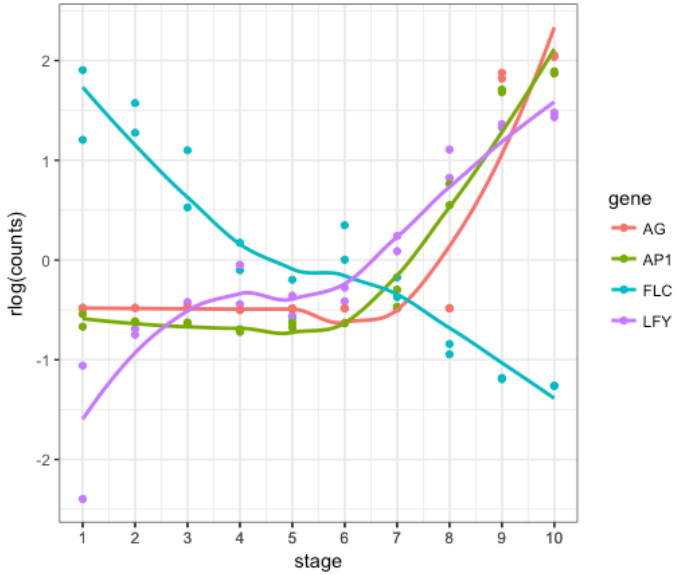
BMC Genomics

RESEARCH ARTICLE Open Access

RNA-seq analysis of an apical meristem time series reveals a critical point in *Arabidopsis thaliana* flower initiation

Anna V. Klepikova[1], Maria D. Logacheva[1,2,3], Sergey E. Dmitriev[2,4] and Aleksey A. Penin[1,2,5*]





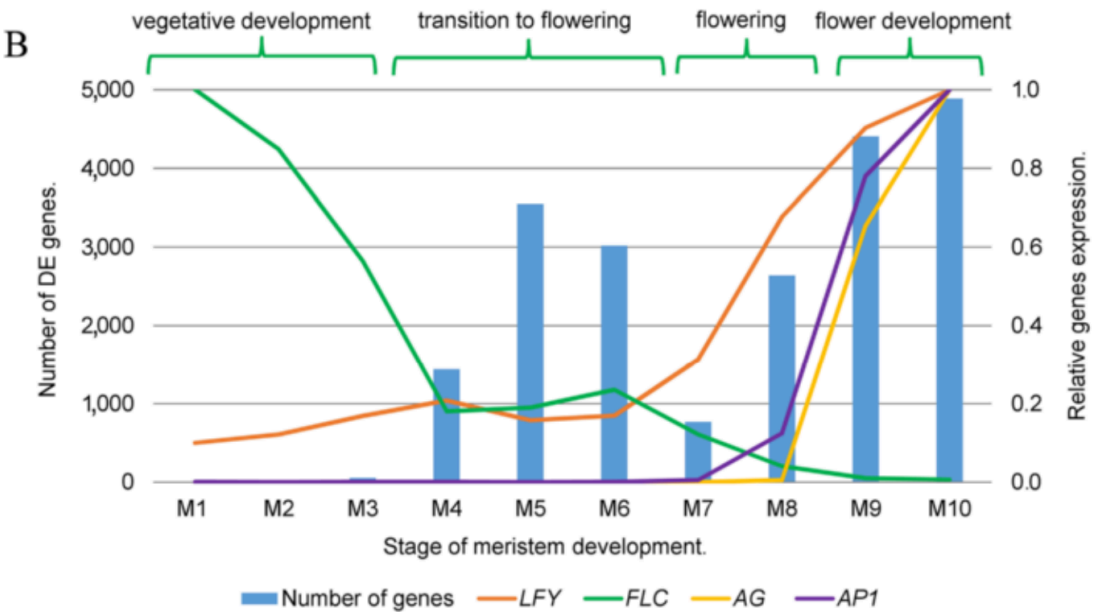
Klepikova *et al. BMC Genomics* (2015) 16:466
DOI 10.1186/s12864-015-1688-9

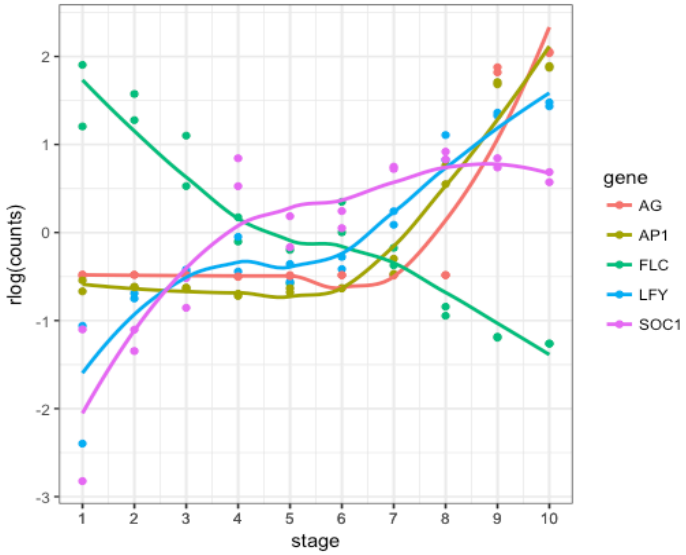
BMC Genomics

RESEARCH ARTICLE Open Access

RNA-seq analysis of an apical meristem time series reveals a critical point in *Arabidopsis thaliana* flower initiation

Anna V. Klepikova[1], Maria D. Logacheva[1,2,3], Sergey E. Dmitriev[2,4] and Aleksey A. Penin[1,2,5*]





mRNA levels …

differ between genes

differ between isoforms

differ between tissues

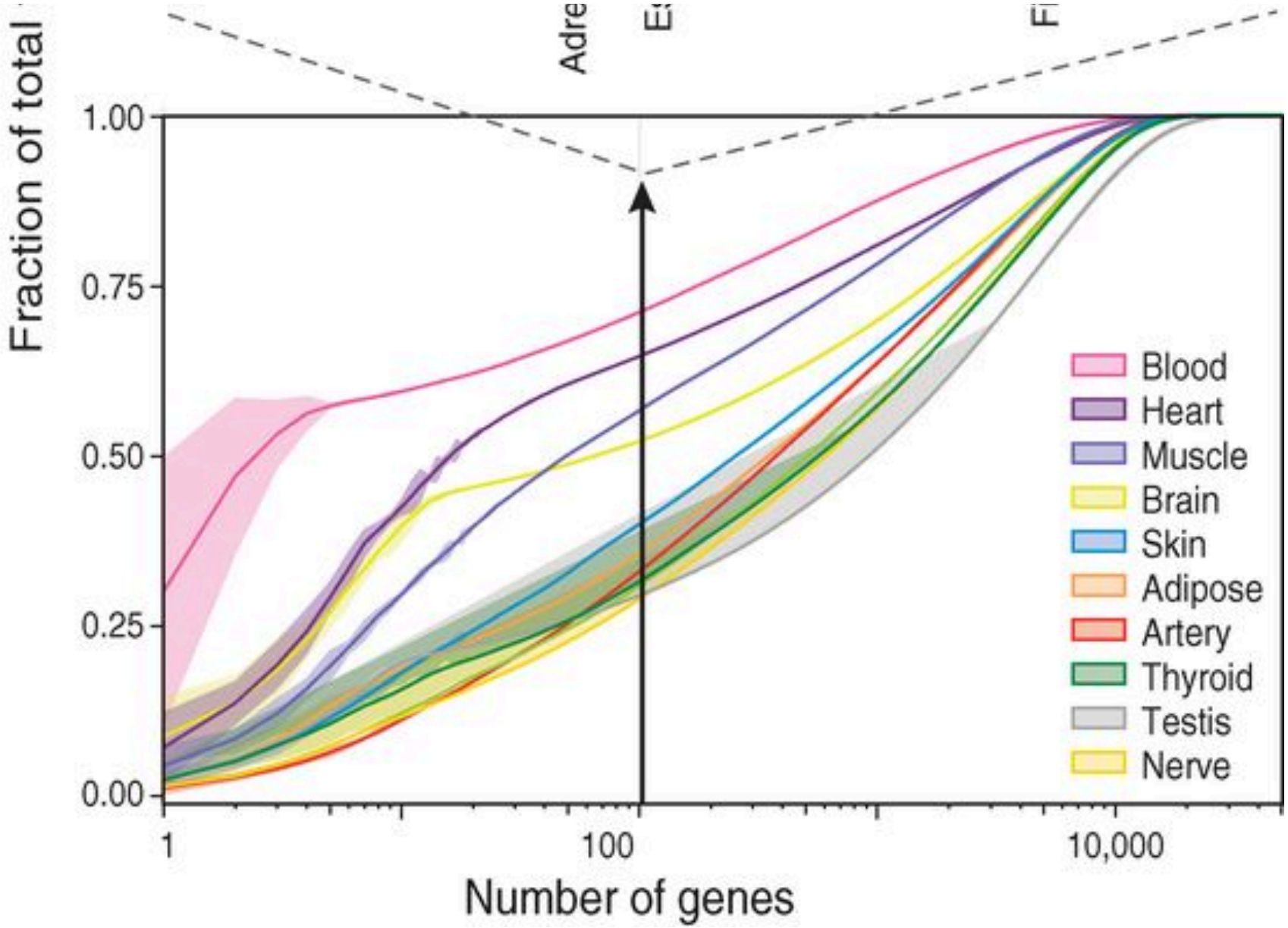
differ between developmental stages

vary during the day (circadian rhythm)

differ between individual cells

depend on the environment

are the result of mRNA synthesis and mRNA decay



Melé, M et al (2015). "The human transcriptome across tissues and individuals." Science **348**(6235): 660-665.

Transcriptomics

Qualitative

- Which genes are expressed?

Quantitative

- What are the expression levels of the genes?

Gene count and species complexity

A.thaliana
26.000

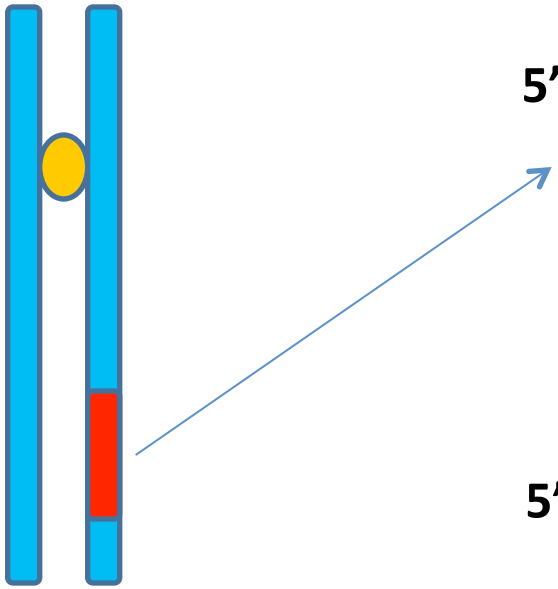
C. elegans
19.000

H. sapiens
20.000

O. sativa
41.000

D. melanogaster
13.600

Splicing



Pre-mRNA

5'- -3'

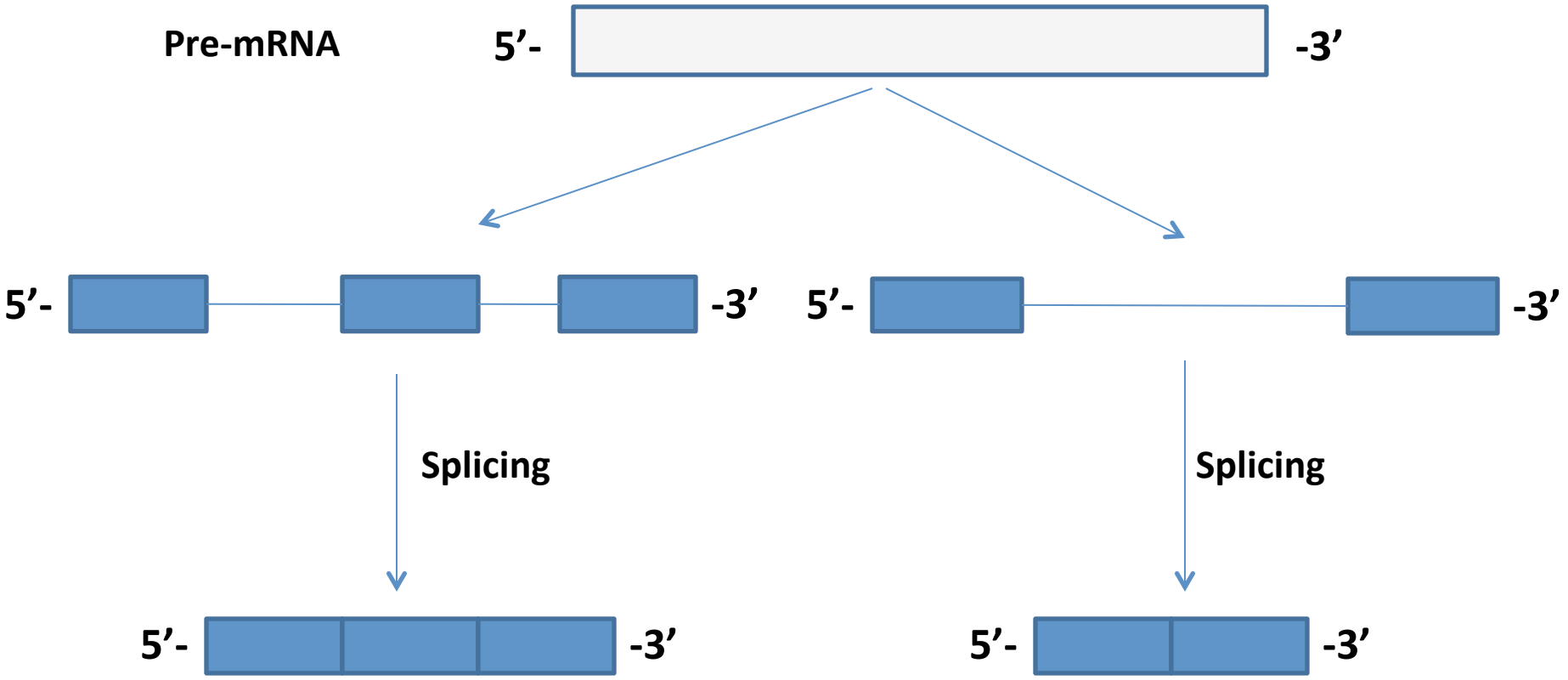
5'- Exon Intron Exon Intron Exon -3'

Splicing

mRNA

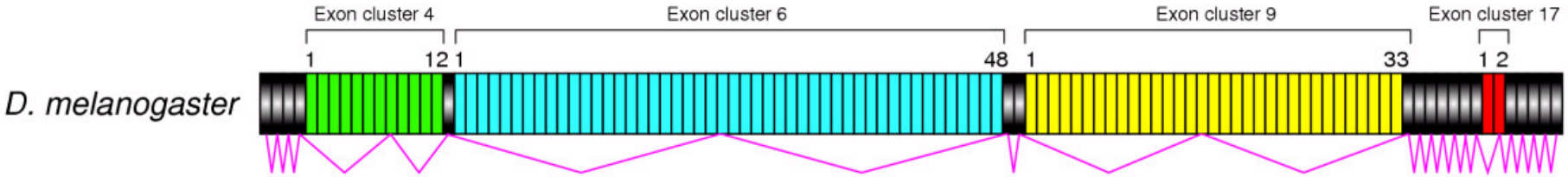
Cap 5'- Exon Exon Exon -3' AAAAAAA

Alternative splicing



Extreme case

***Dscam* gene produces over 38,000 different transcripts**



Genes producing multiple transcripts through alternative splicing

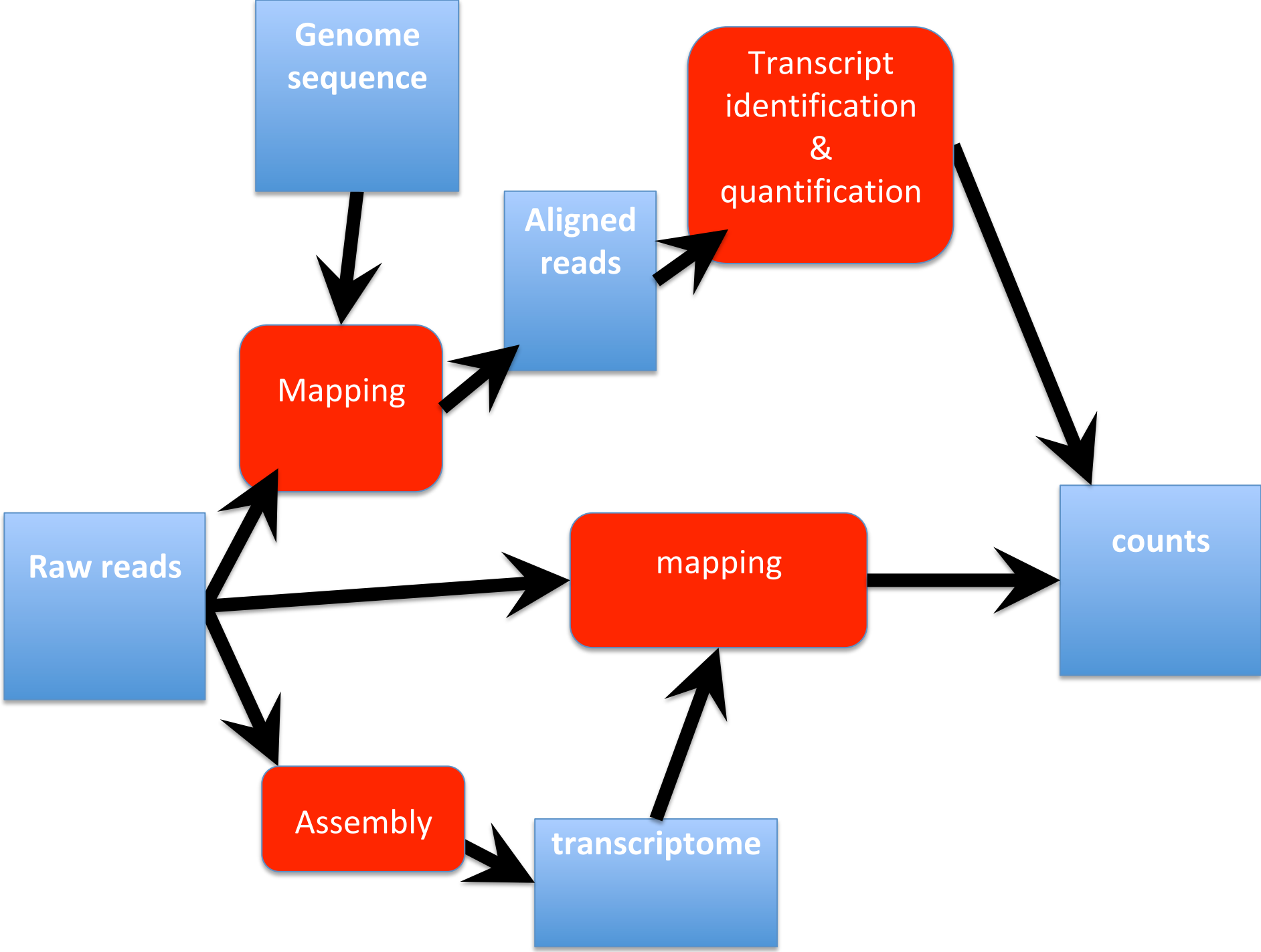
~60%

~50%

~95%

~25%

~31%


Genome sequence
Transcript identification & quantification
Aligned reads
Mapping
Raw reads
mapping
counts
Assembly
transcriptome

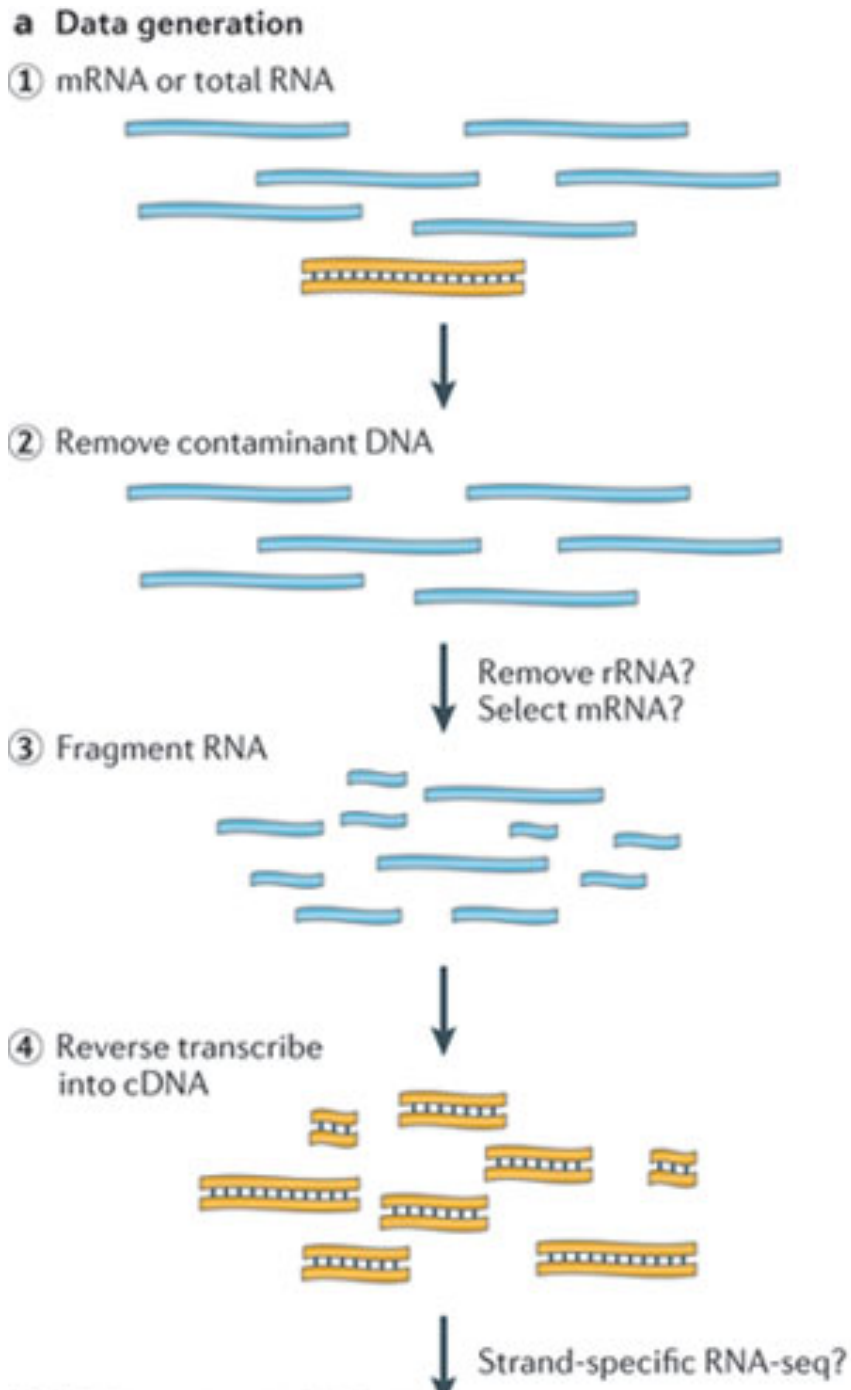
STUDY DESIGNS

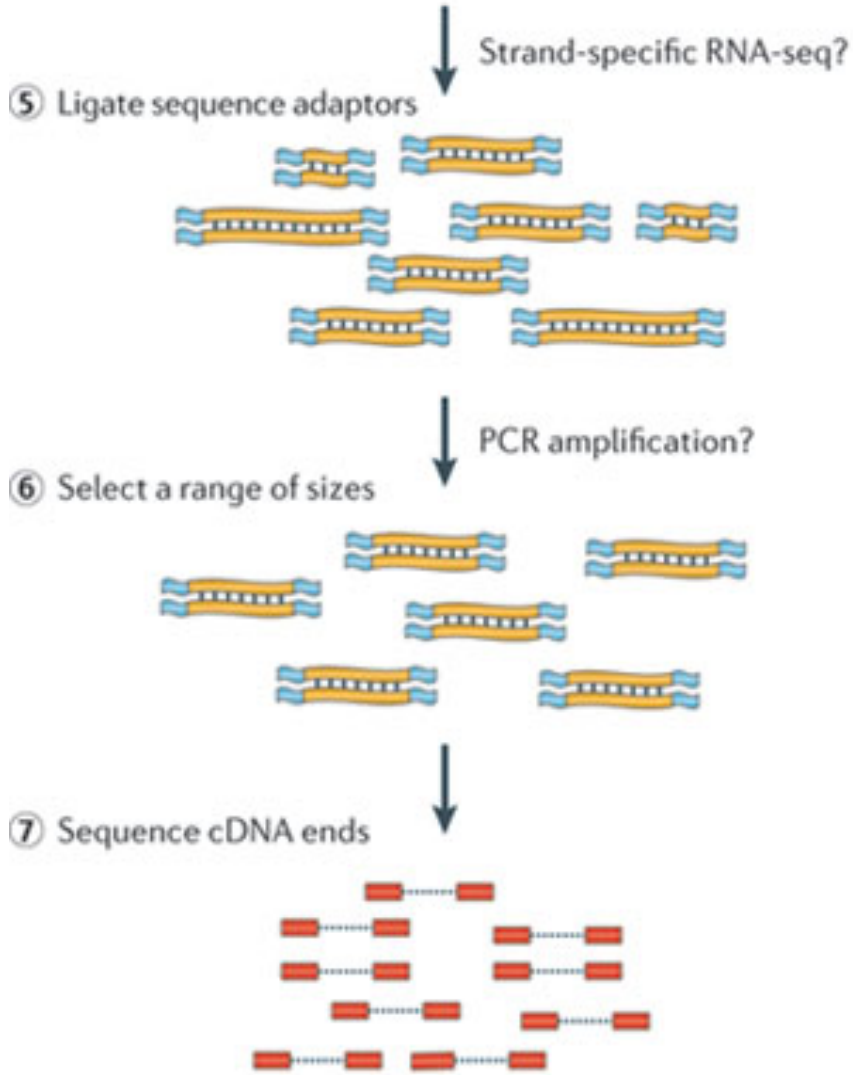
Next-generation transcriptome assembly

Jeffrey A. Martin and Zhong Wang

Abstract | Transcriptomics studies often rely on partial reference transcriptomes that fail to capture the full catalogue of transcripts and their variations. Recent advances in sequencing technologies and assembly algorithms have facilitated the reconstruction of the entire transcriptome by deep RNA sequencing (RNA-seq), even without a reference genome. However, transcriptome assembly from billions of RNA-seq reads, which are often very short, poses a significant informatics challenge. This Review summarizes the recent developments in transcriptome assembly approaches — reference-based, *de novo* and combined strategies — along with some perspectives on transcriptome assembly in the near future.

Nat Rev Genet **12**(10): 671-682.


a Data generation
1 mRNA or total RNA
2 Remove contaminant DNA
Remove rRNA?
Select mRNA?
3 Fragment RNA
4 Reverse transcribe
into cDNA
Strand-specific RNA-seq?


Strand-specific RNA-seq?
5 Ligate sequence adaptors
PCR amplification?
6 Select a range of sizes
7 Sequence cDNA ends

Sequencing by Synthesis

http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html

Questions

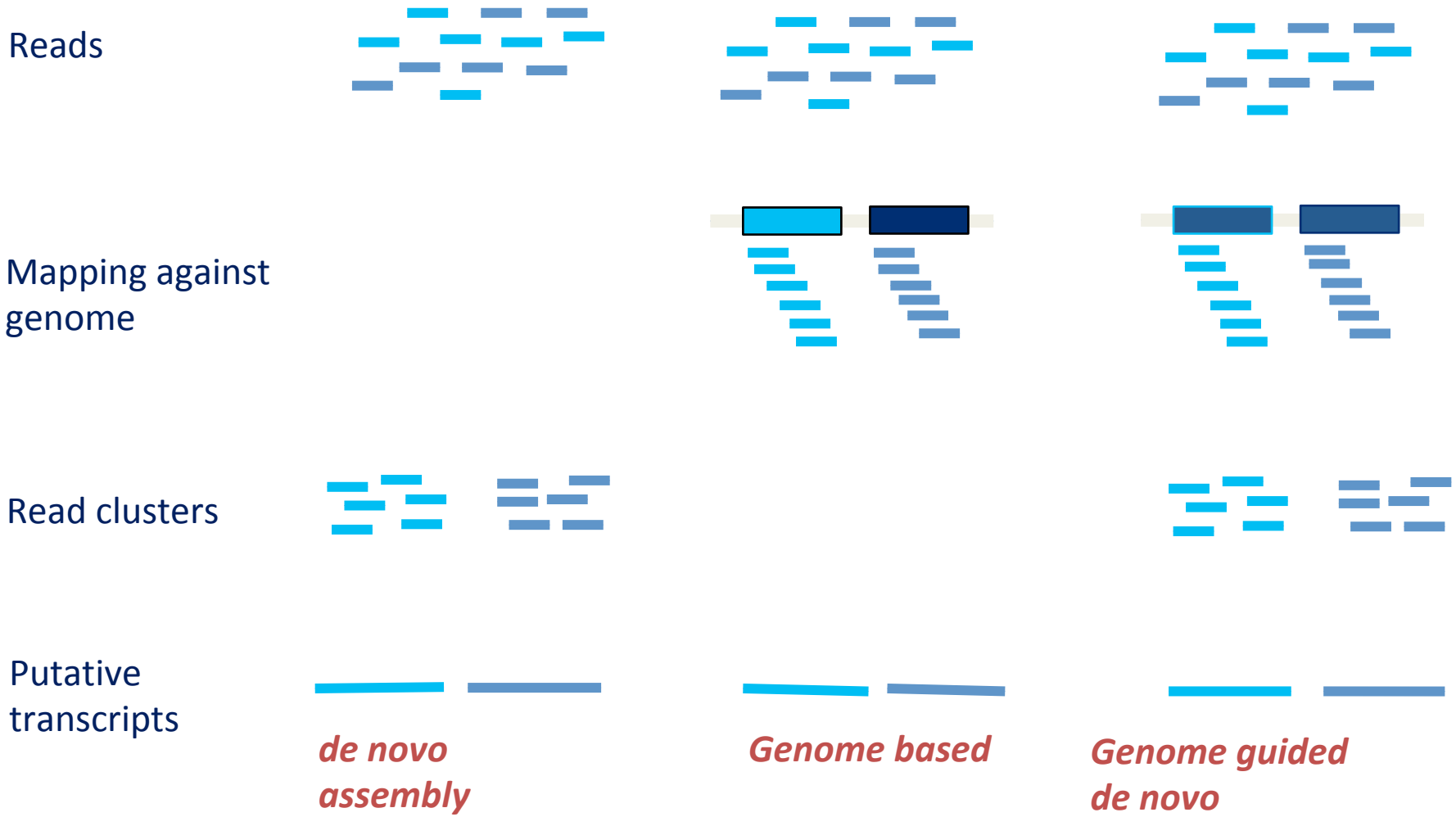
With read lengths of 100 nt, why do we sequence fragments of mRNAs and not whole mRNAs?

With paired-end sequencing both ends of a fragment are sequenced. Why would you do that?

When would you use these sequencing strategies?

- Paired-end with reads of 150 nt
- single-end with reads of 75 nt

From reads to transcripts



Challenges

Different isoforms (splice variants)

Paralogous sequences

Exon/intron structure

Sequencing errors

SNPs

Ever growing data volumes

(millions of reads per sample)

Read mapping

Map the reads on a genome

- Software tools: Blast?

Read mapping

Map the reads on a genome

Software tools: ~~Blast?~~

Too slow….

Read mapping

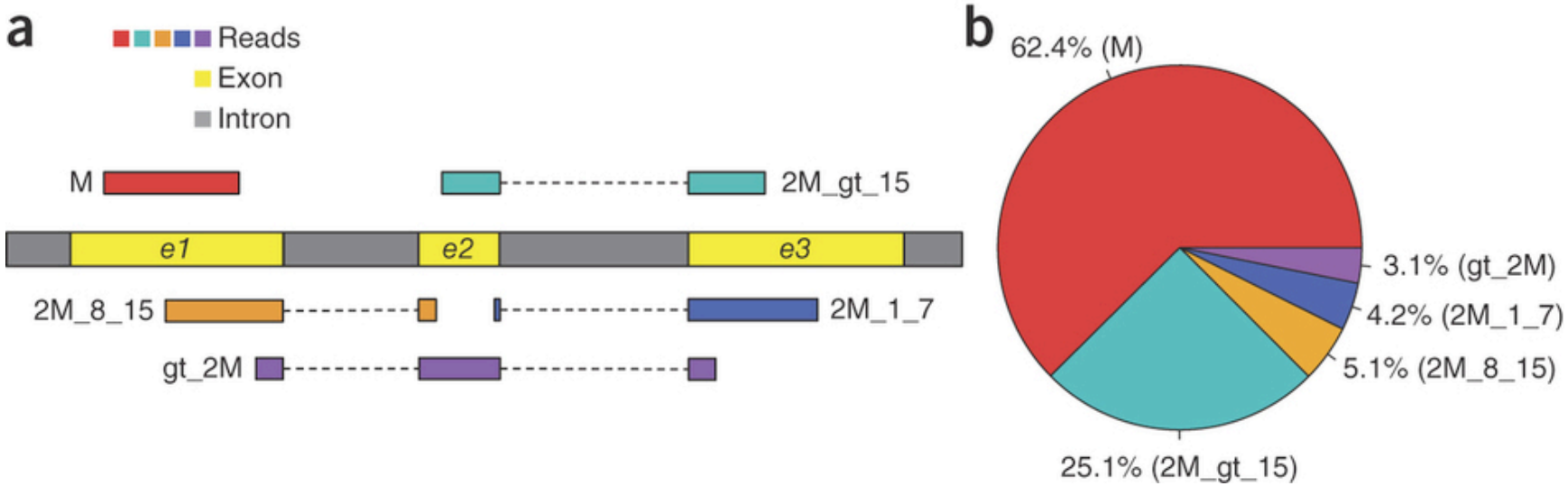
Map the reads on a genome

- Software tools: Bowtie, BWA

- use an **index** to rapidly find a sequence
- FM-index based on Burrows-Wheeler Transform

But what about introns…

- Bowtie and BWA do not allow many mismatches or large indels



(a) Five types of RNA-seq reads: (i) M, exonic read; (ii) 2M_gt_15, junction reads with long, >15-bp anchors in both exons; (iii) 2M_8_15, junction reads with intermediate, 8- to 15-bp anchors; (iv) 2M_1_7, junction reads with short, 1- to 7-bp, anchors; and (v) gt_2M, junction reads spanning more than two exons.

(b) (b) Relative proportions of different types of reads in the 20 million 100-bp simulated read data.

Kim, D., B. Langmead and S. L. Salzberg (2015). "HISAT: a fast spliced aligner with low memory requirements." Nat Meth **12**(4): 357-360.

Enter: splice junction mappers

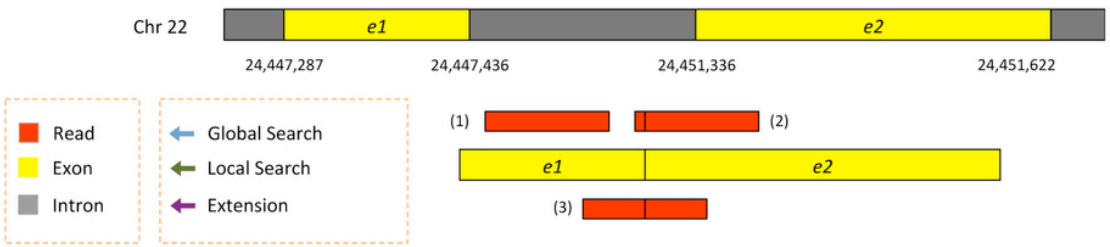
Tools: Tophat, Star, GSnap, HISAT2

Allow large gaps in the mapping process

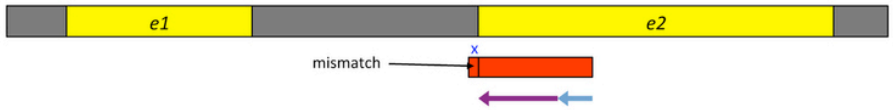
Differ in sensitivity, memory usage, speed

Produce alignment files, i.e. in **SAM** format

HISAT

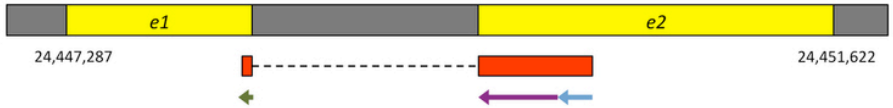


a

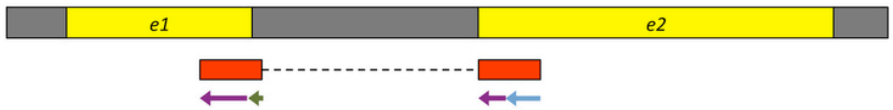


b

Local FM index for chr22 from 24,417,280 to 24,482,559



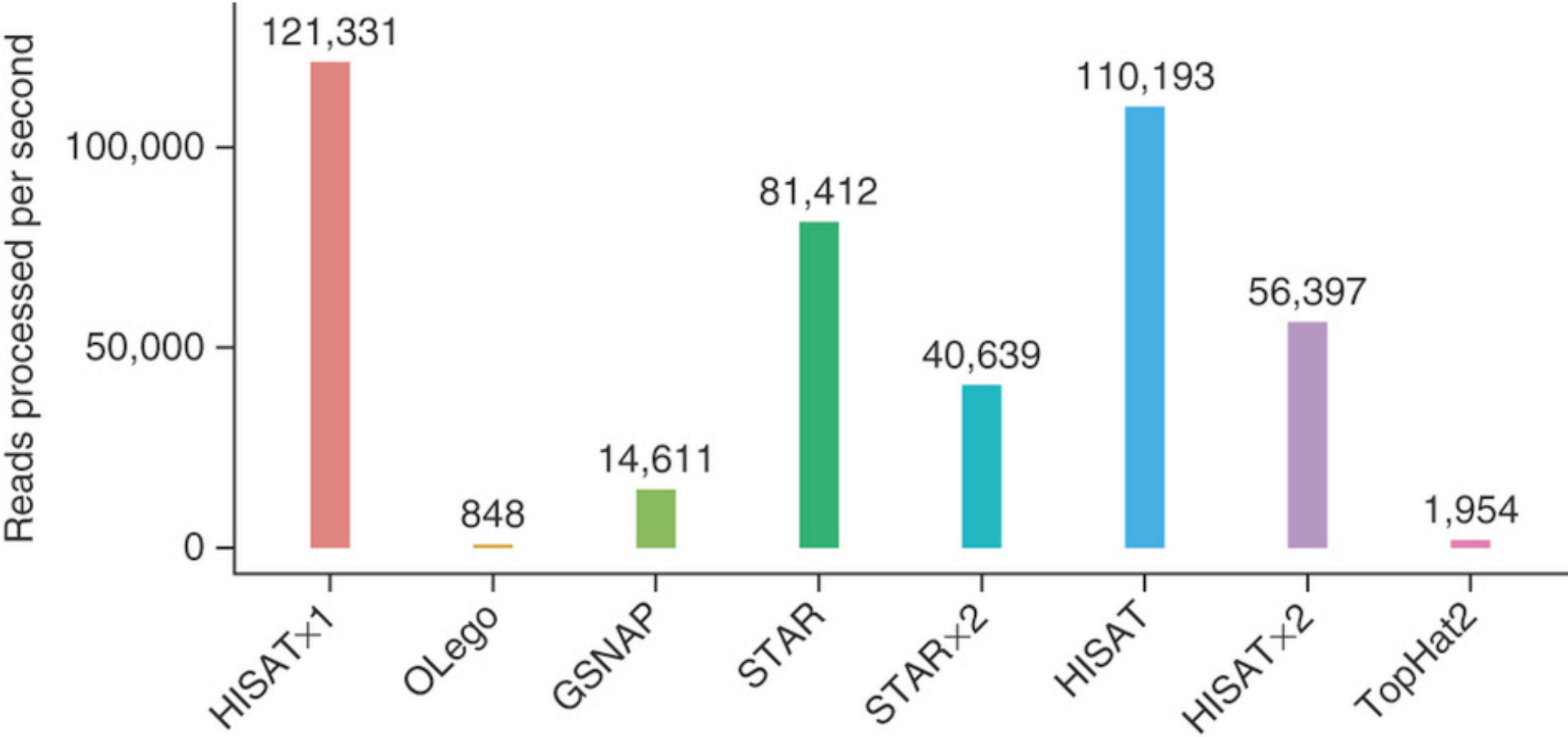
c

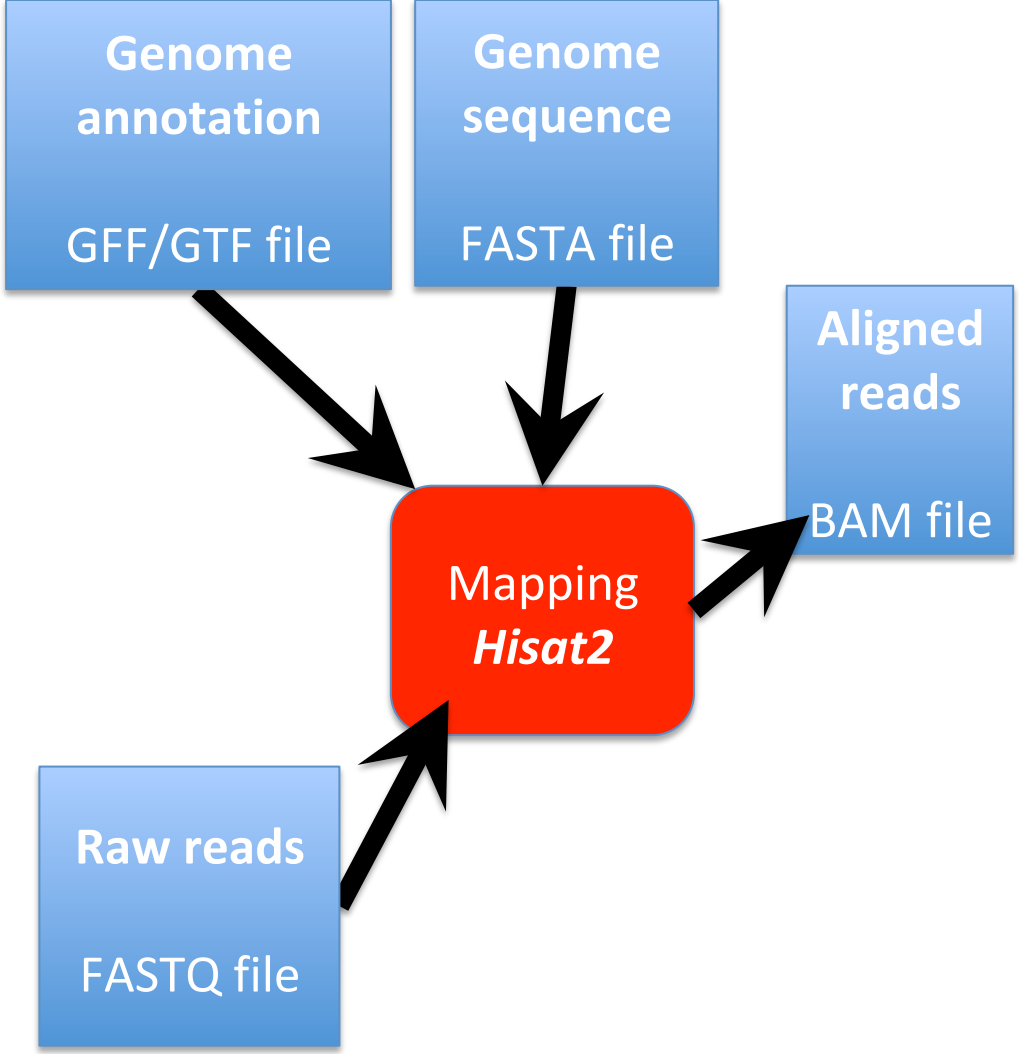


Nature Methods 12, 357–360 (2015) doi:10.1038/nmeth.3317

HISAT: a fast spliced aligner with low memory requirements

Daehwan Kim[1,2], Ben Langmead[1–3] & Steven L Salzberg[1–3]




Genome annotation
GFF/GTF file
Genome sequence
FASTA file
Aligned reads
BAM file
Mapping
Hisat2
Raw reads
FASTQ file

Hands on