UiO : Department of Biosciences
University of Oslo

MBV4410/9410 Fall 2016

Dec. 5 - Analysing transcriptome data (using R) – part 1

Outline

Monday

Before lunch:

- Transcriptomics (lectures/practical)
  - Sequencing technologies
  - Transcriptome assembly
  - Gene expression

After lunch:

- Basic R/RStudio (lecture)
- Installing/setting up R/RStudio
- Basic R (practical)

Tuesday

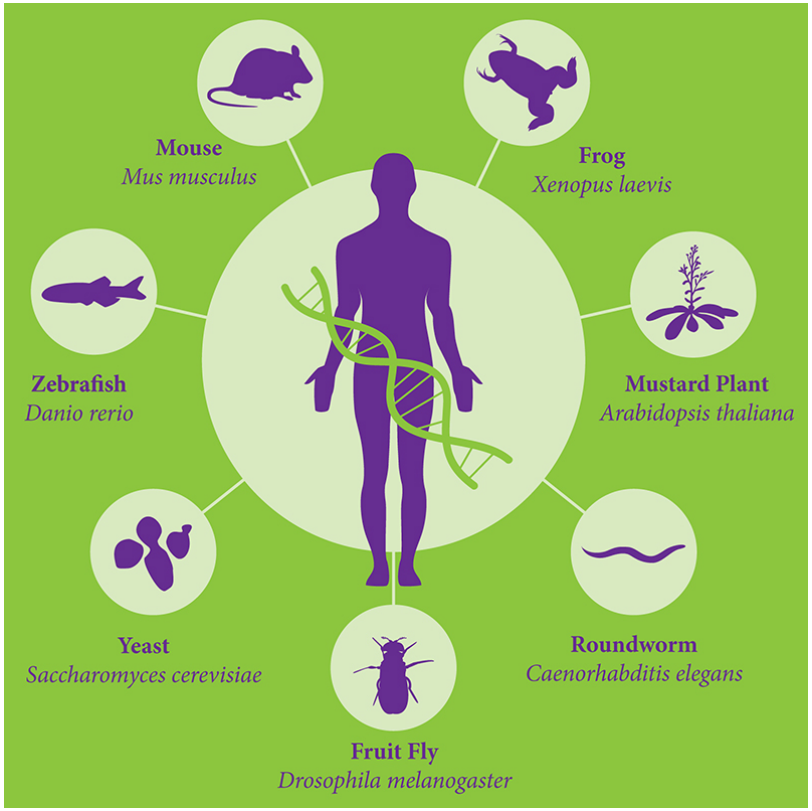
Before lunch:

- Transcriptomics (lectures/practical)
  - Experimental design
  - Quality assessement
  - Differential gene expression

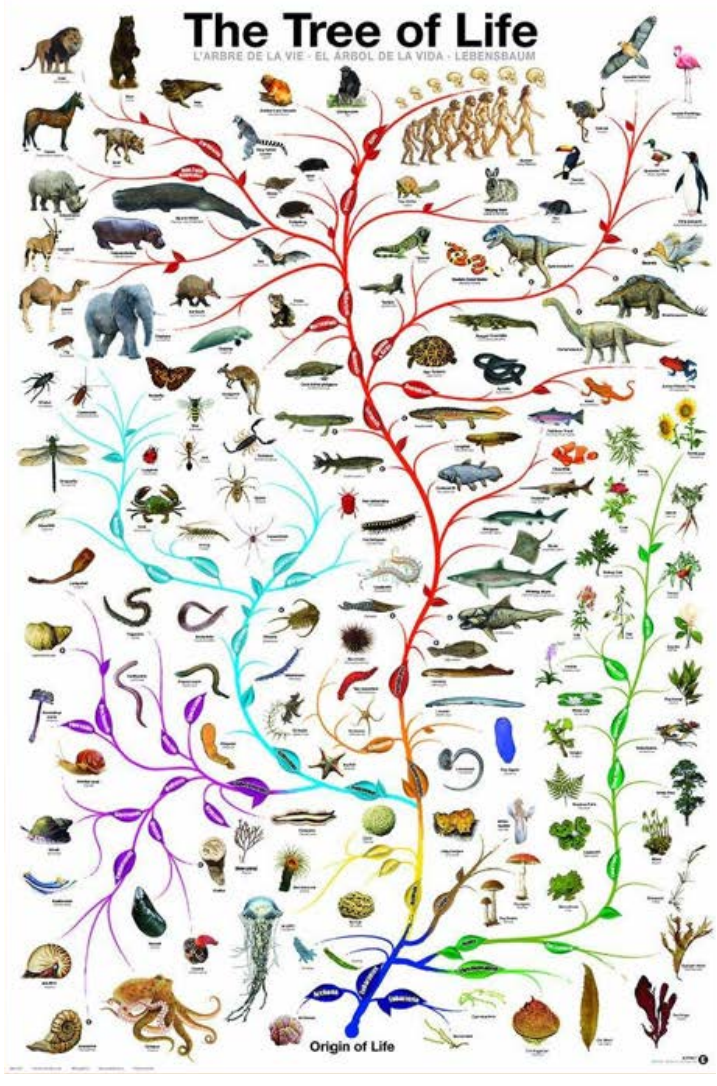
After lunch:

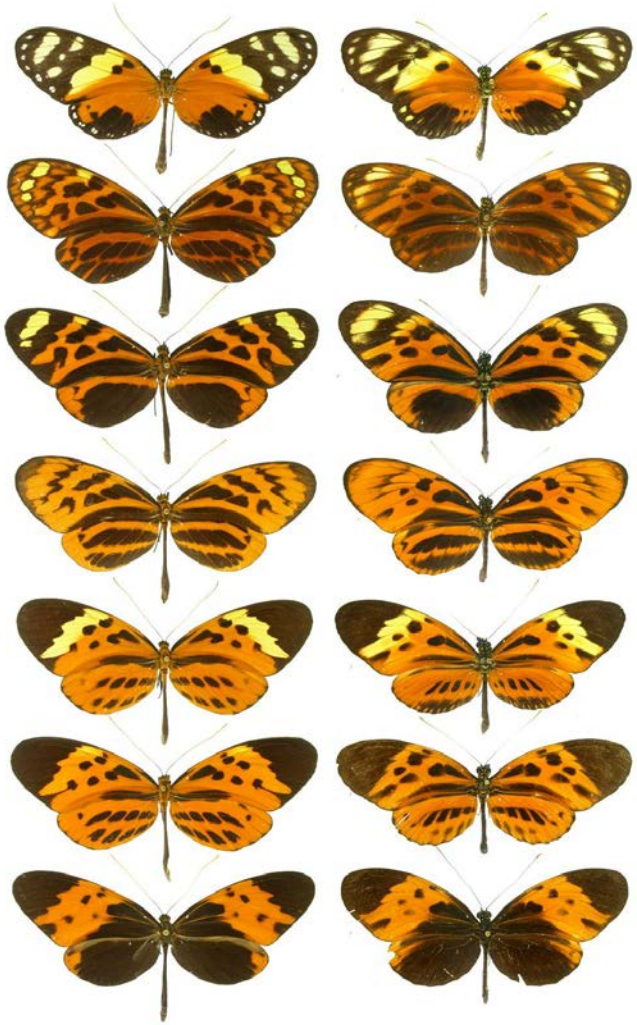
- Bioconductor (lecture)
- Transcriptomics/DE-test (lecture/practical)

Moving away from model organisms



https://biology.uiowa.edu/model-organisms


The Tree of Life
L'ARBRE DE LA VIE · EL ÁRBOL DE LA VIDA · LEBENSBAUM
Origin of Life

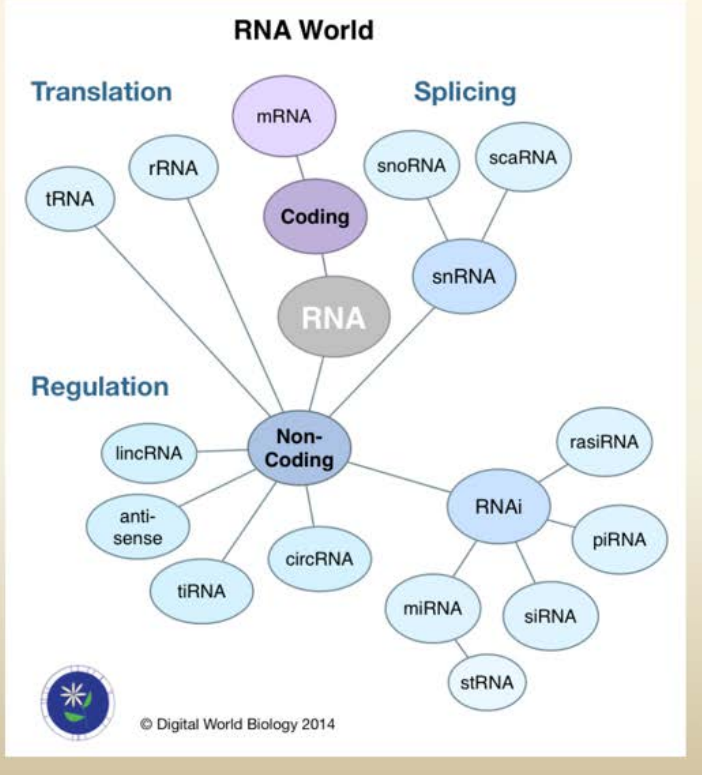


***Melinaea* spp. models**
(Nymphalidae: Ithomiinae)
M. ludovica ludovica, M. satevis cydon
M. marsaeus rileyi, M. marsaeus phasiana
M. mnemphilus ssp. nov., *M. mnemphilus hiatus*
M. ieconoma simulator

***Heliconius numata* mimetic forms**
(Nymphalidae: Heliconiinae)
H. n. f. silvana, H. n. f. elegans
H. n. f. aurora, H. n. f. arcuella
H. n. f. tarapotensis, H. n. f. timaeus
H. n. f. bicoloratus

Transcription – all the RNAs

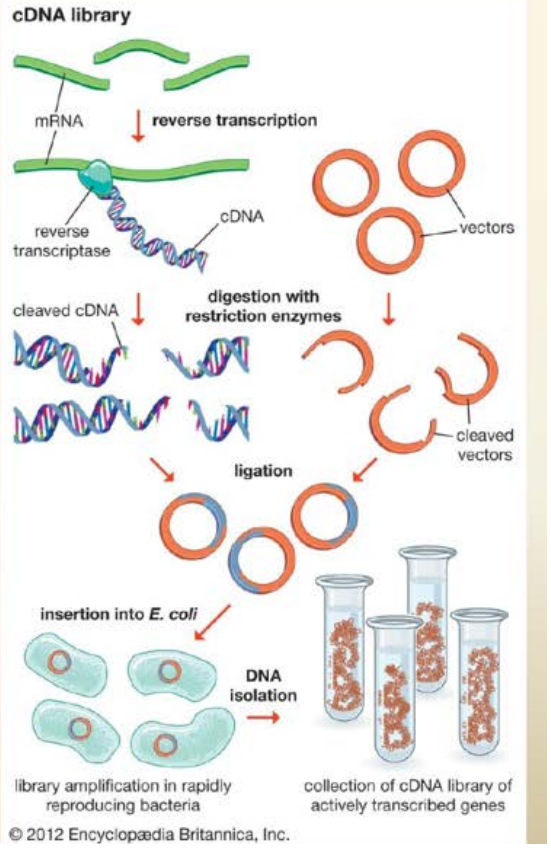
A transcriptome is a snapshot in time of **all RNAs** present in a sample isolated from a given cell, tissue or organism



Obtaining the transcriptome (not many years ago)

Sanger cDNA library sequencing

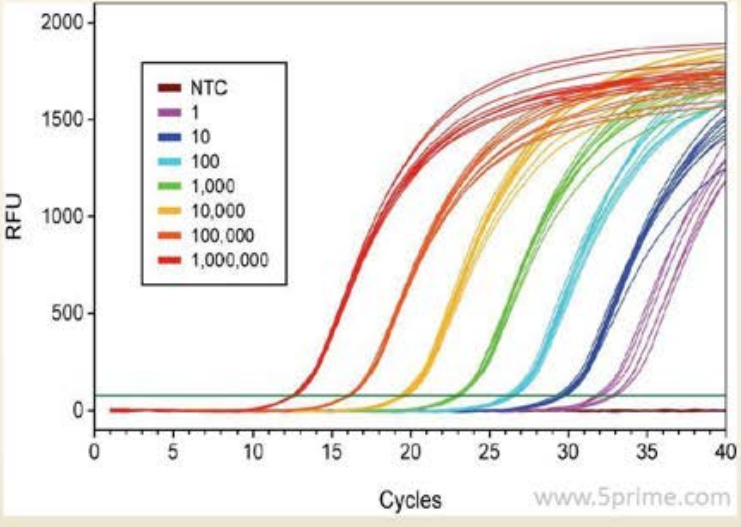
- mRNA converted to the more stable cDNA
- cDNA cleaved and ligated into vectors
- Vectors amplified (cloned) in *E. coli*
- DNA isolated = cDNA library
- Sequenced on Sanger
- Low throughput
- High accuracy



Obtaining the transcriptome (not many years ago)

Quantitative RT-PCR

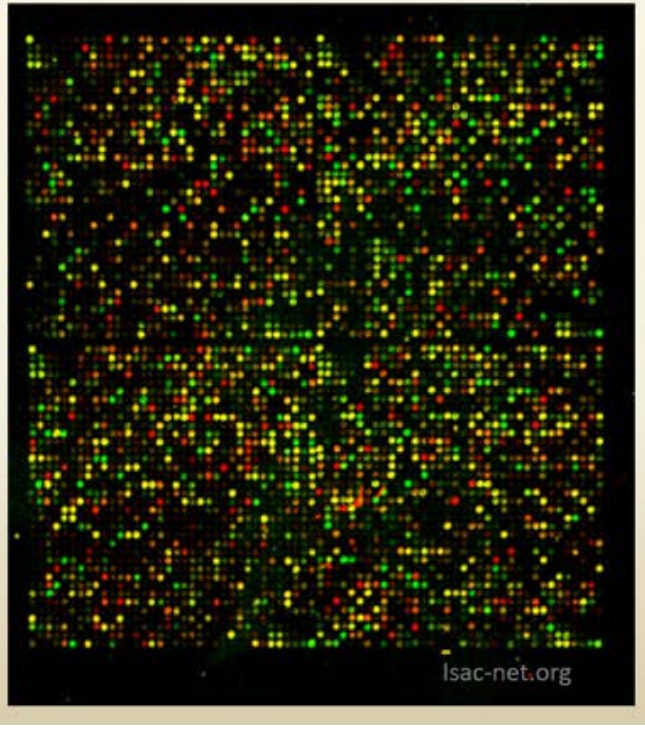
- qRT-PCR requires knowledge of gene sequence
- Hard manual work
- Low throughput
- Expression level relative to control (house-keeping gene)



Obtaining the transcriptome (not many years ago)

Microarray - expression determination

- Requires gene sequences for probe design
- High throughput compared to qRT-PCR
- Possibility of outsourcing
- Expression results relative to all probes

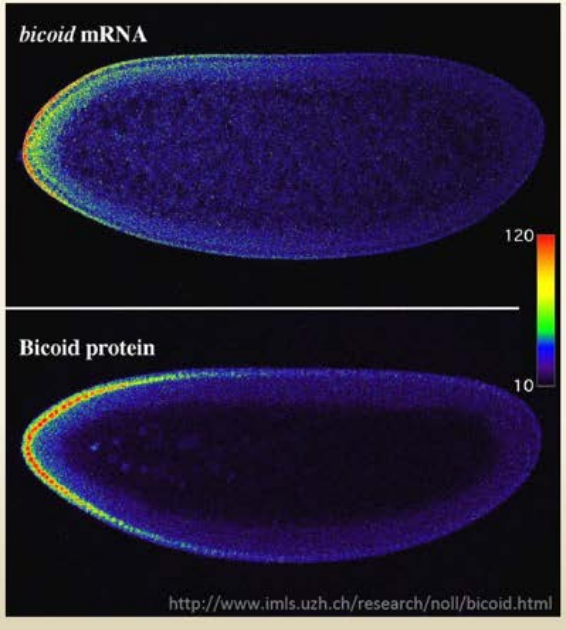


A word of caution…

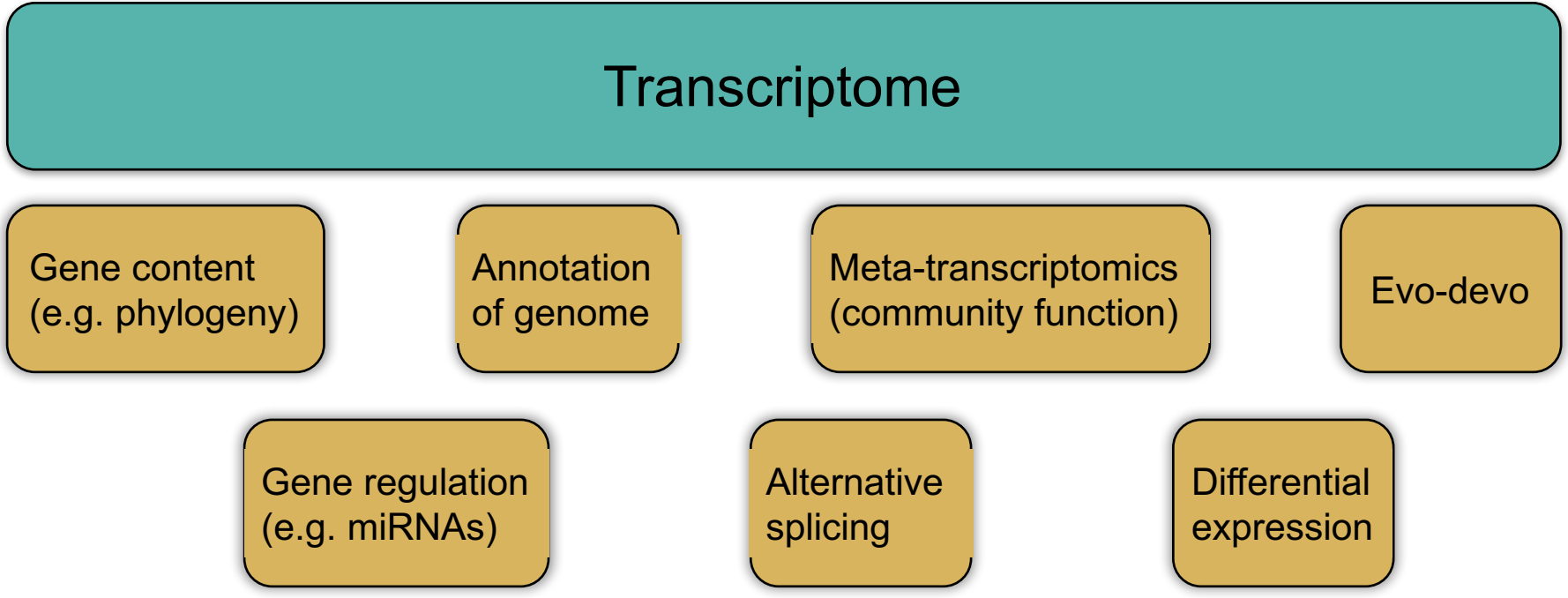
We tend to think…

- Transcriptome = mRNA
- mRNA = Protein
- Protein = Biological relevance

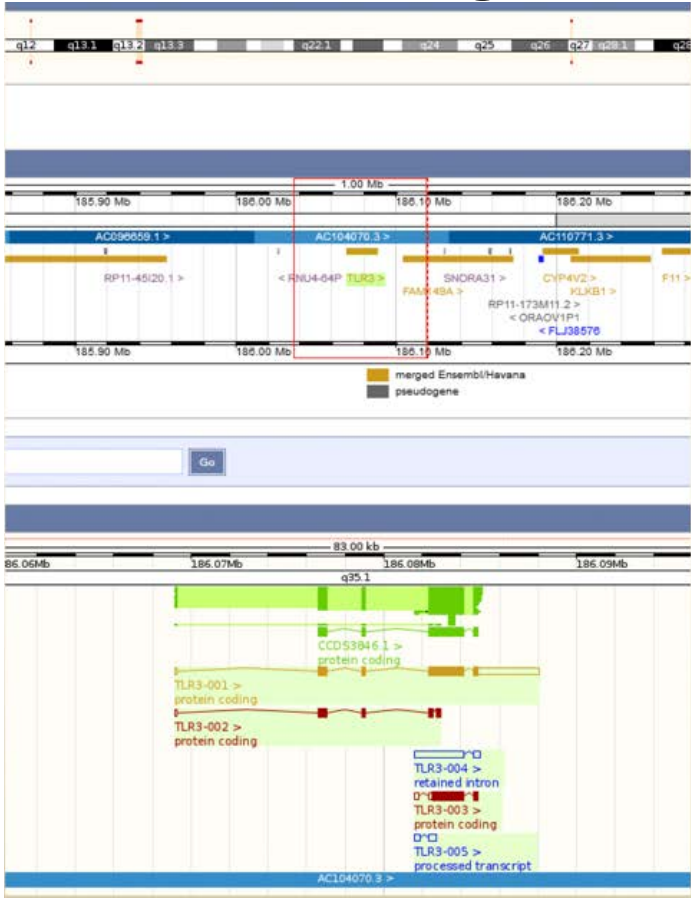
- Things are seldom as simple as clear cut…



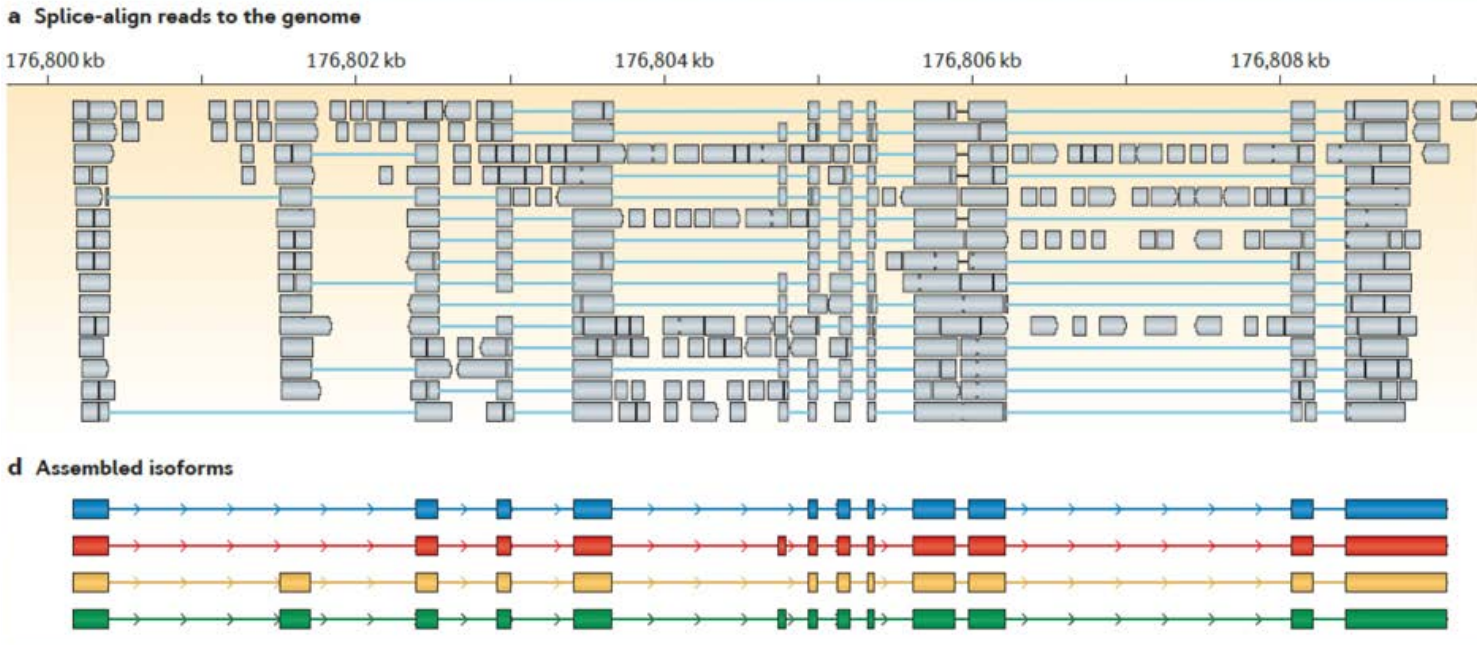
Uses of RNA-seq



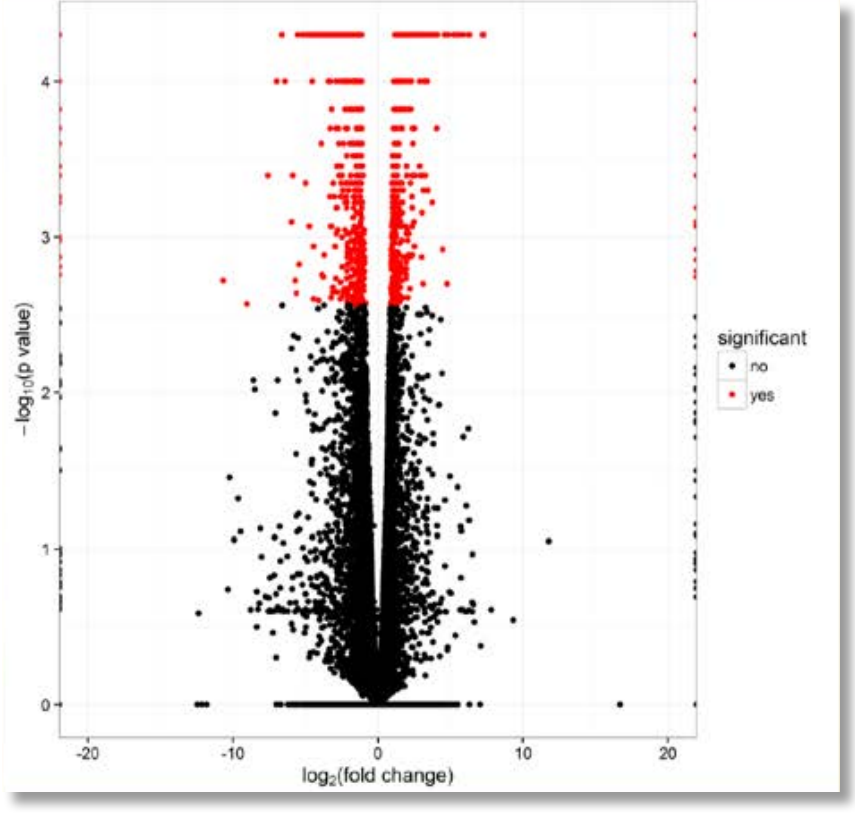
Genome annotation – better gene prediction



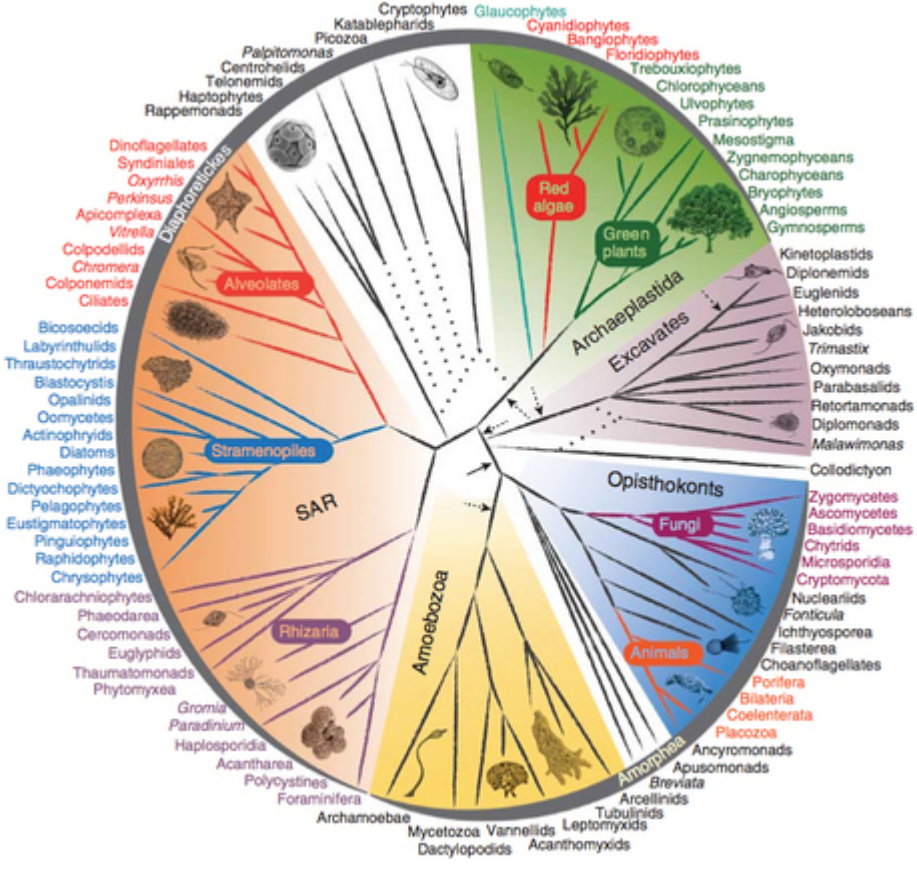
Isoforms – alternative splicing



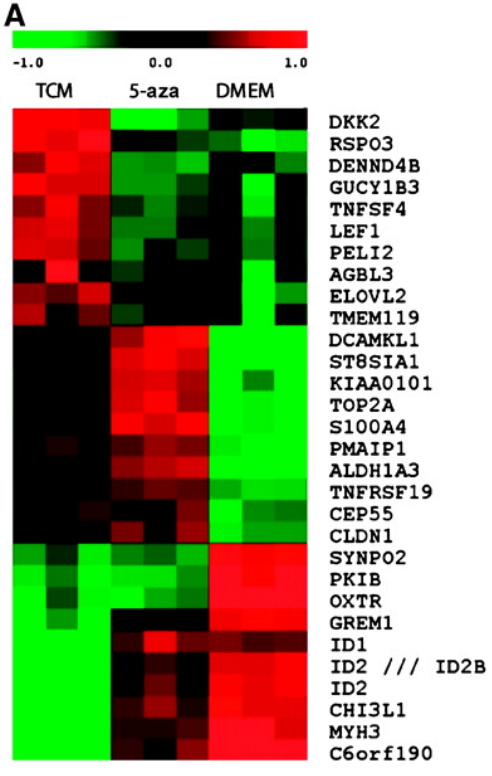
Differential gene expression - functional studies

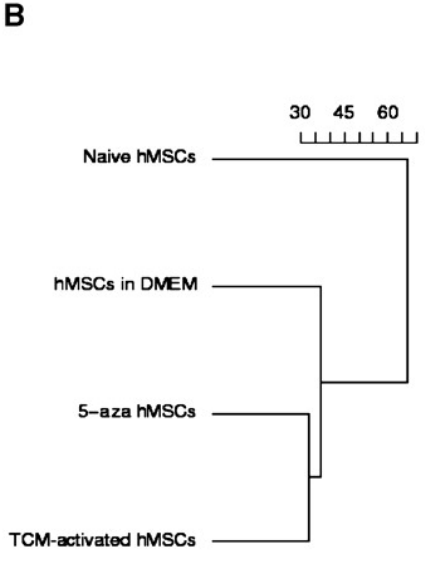


Phylogenomics

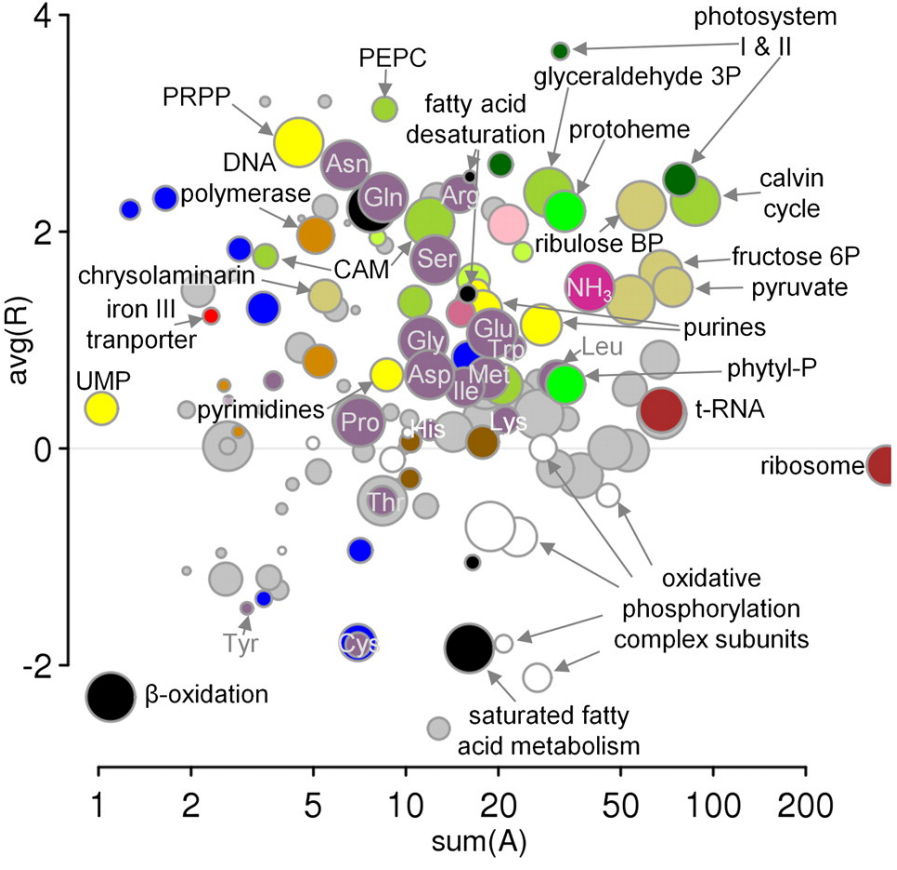


Comparative gene expression (between tissues or between species – or both)



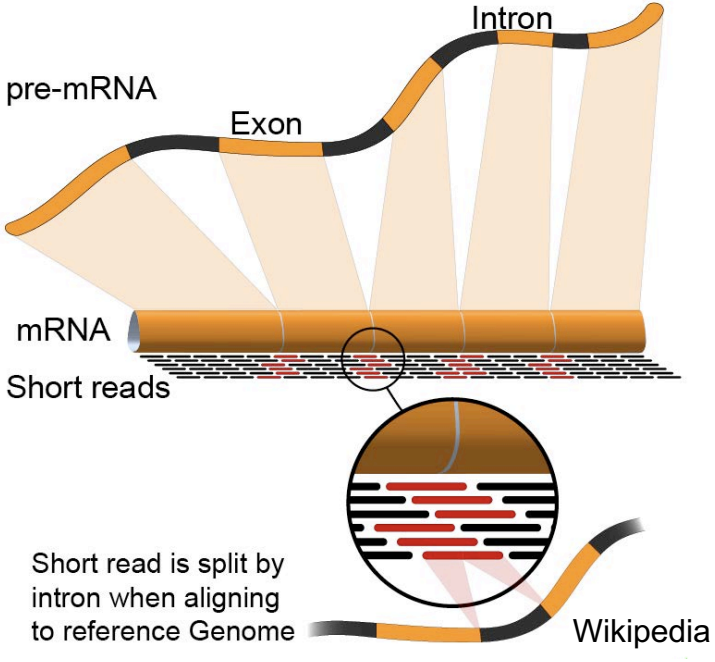


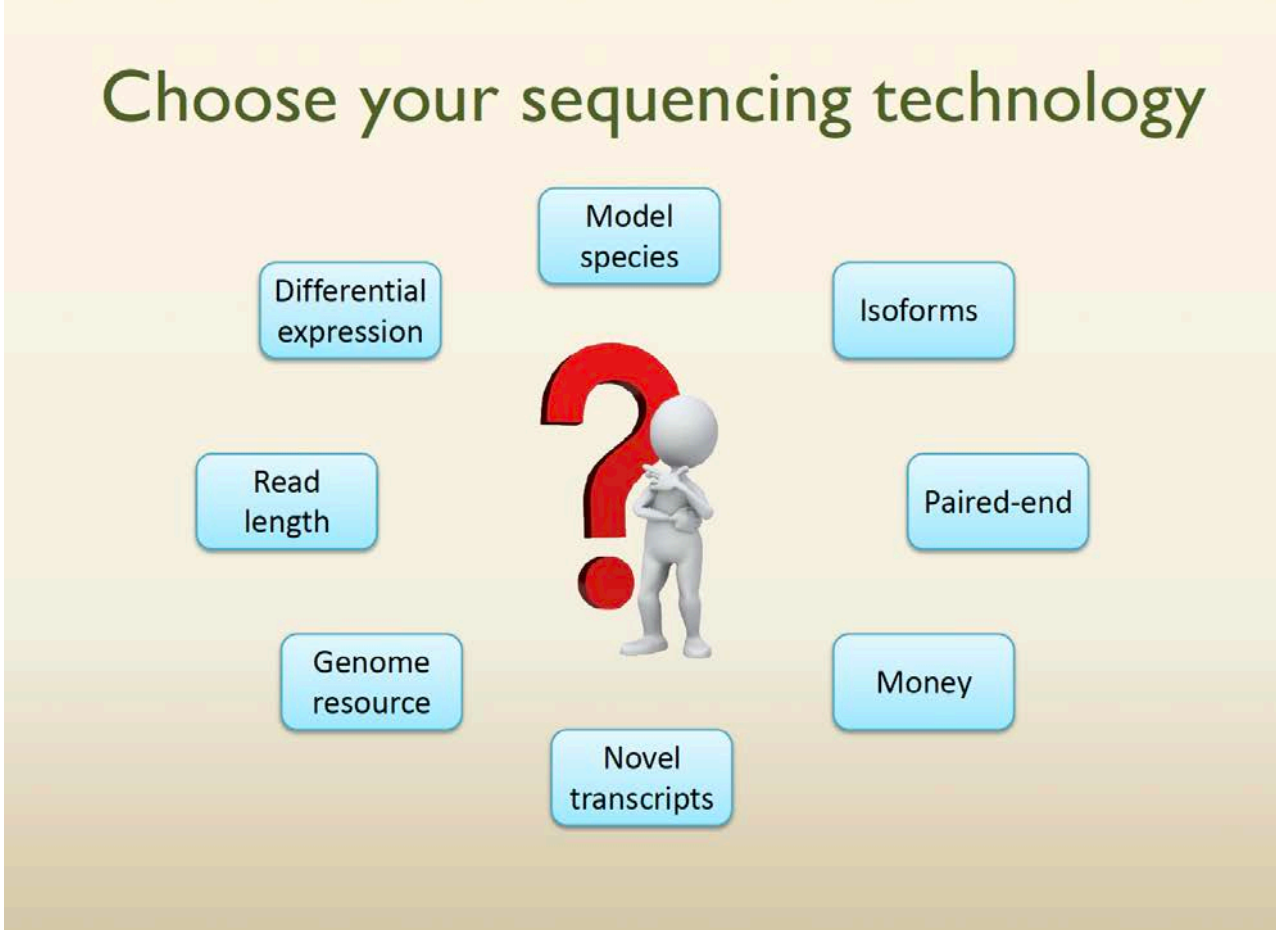
Meta-transcriptomics



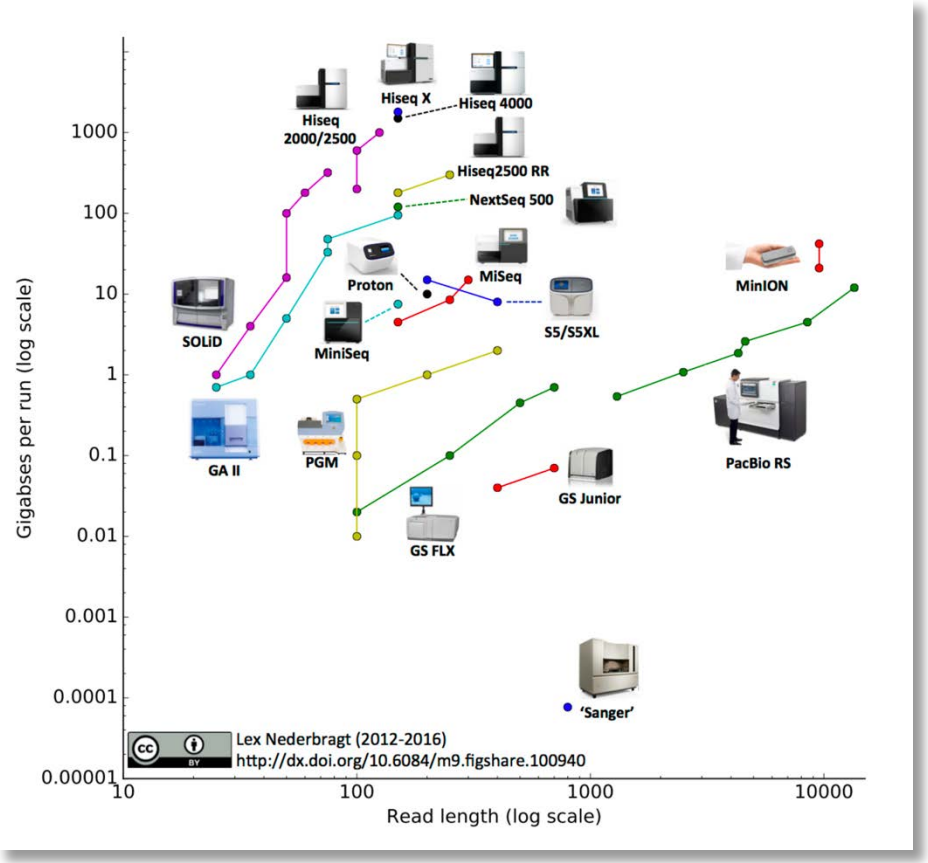
“Next generation” transcriptomics

- Transcriptome and expression in one go
- No need for prior gene/genome sequence information
- High throughput
- Can be outsourced
- Can be costly




Choose your sequencing technology
Model species
Differential expression
Isoforms
Read length
Paired-end
Genome resource
Money
Novel transcripts

Next generation sequencing (NGS/HTS)



https://flxlexblog.wordpress.com/

Illumina

- Short read paired-end technology
- 2 flowcells – 8 lanes each
- ~150 bp PE reads
- Reasonable reconstruction of isoforms
- Reasonable detection of novel transcripts
- Expression analysis
- Makes decent reference transcriptomes

Illumina HiSeq 4000

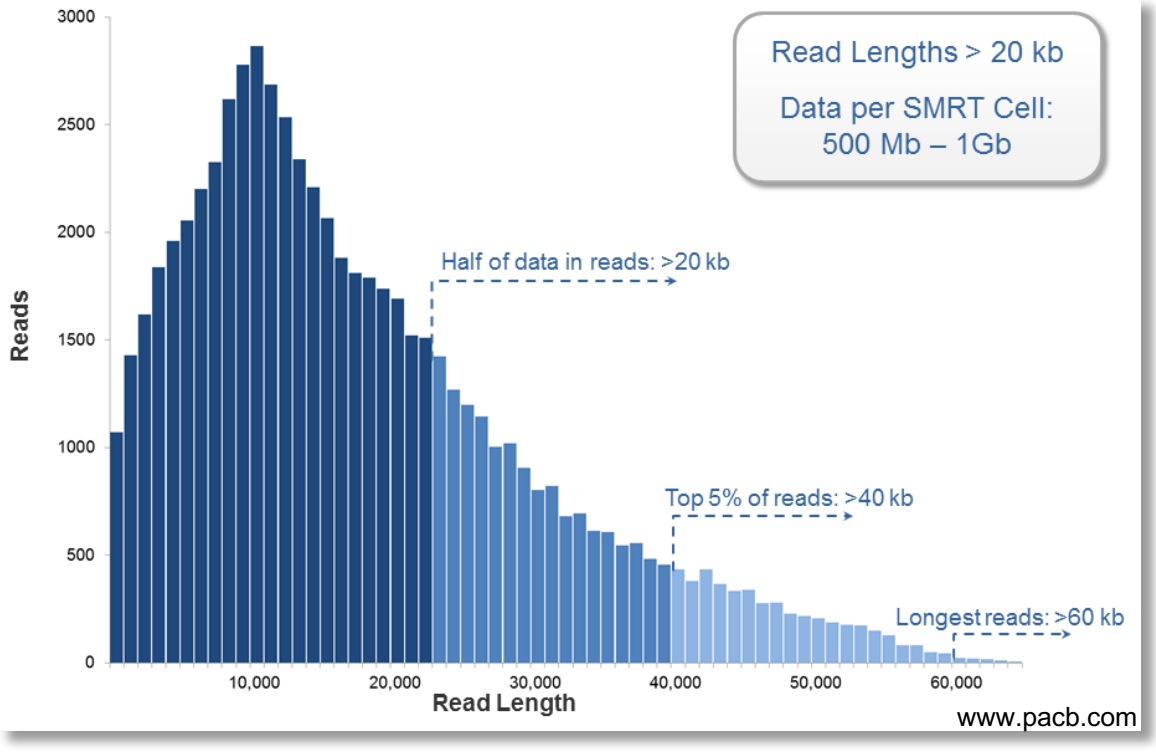
PacBio - single-molecule sequencing

- Long read sequencing technology
- 16 SMRT Cells
- Sequences entire RNAs up to 10 kb
- Reconstruction of isoforms
- Detection of novel transcripts
- Expression analysis
- Great for reference transcriptomes

PacBio sequencing well

Prof. Kjetill Jakobsen in fron t of the NSC' PacBio

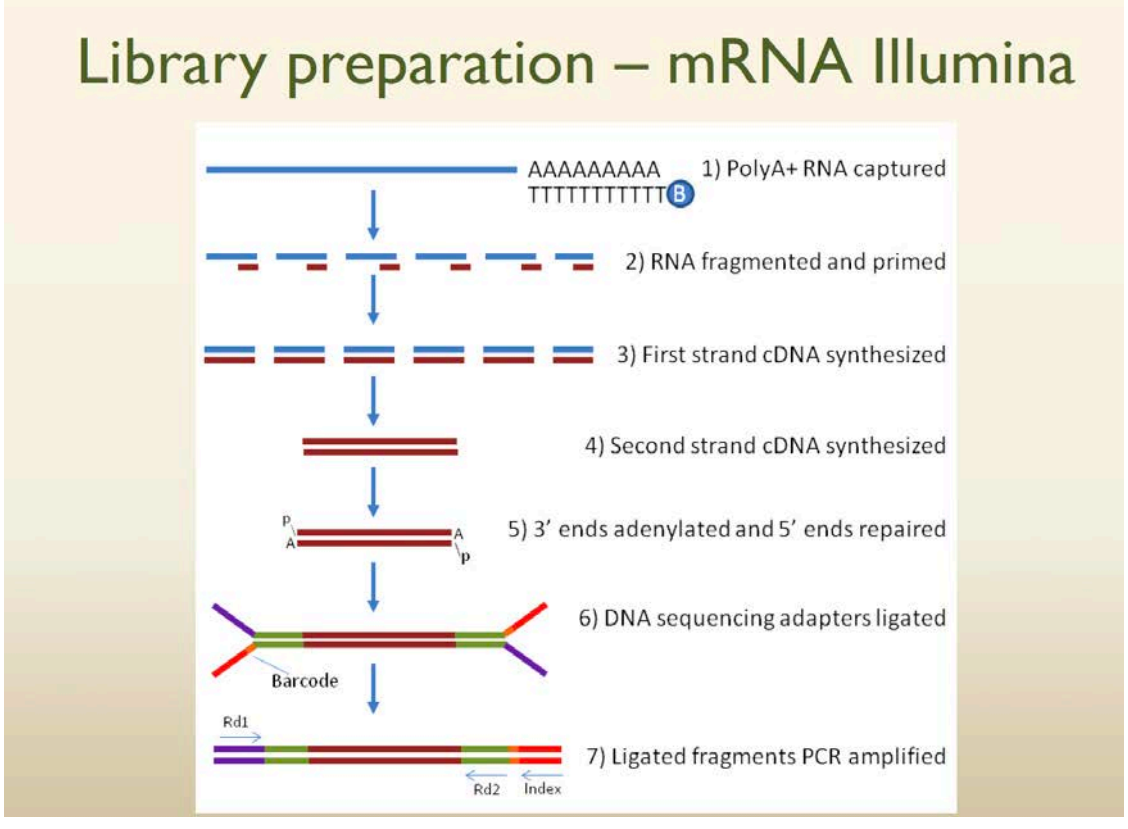
PacBio - single-molecule sequencing



Briefly about sample preparation

- Depending on focus you may perform:
  - rRNA depletion
  - mRNA selection
  - Abundant transcript removal
  - smallRNA conservation
  - Skip library amplification
  - Strand specific library preparation

Briefly about sample preparation



Illumina paired-end sequencing

- You get two reads from the same RNA fragment
- Can be long gap in between
- Easier to assemble.

Detect novel splice variants.

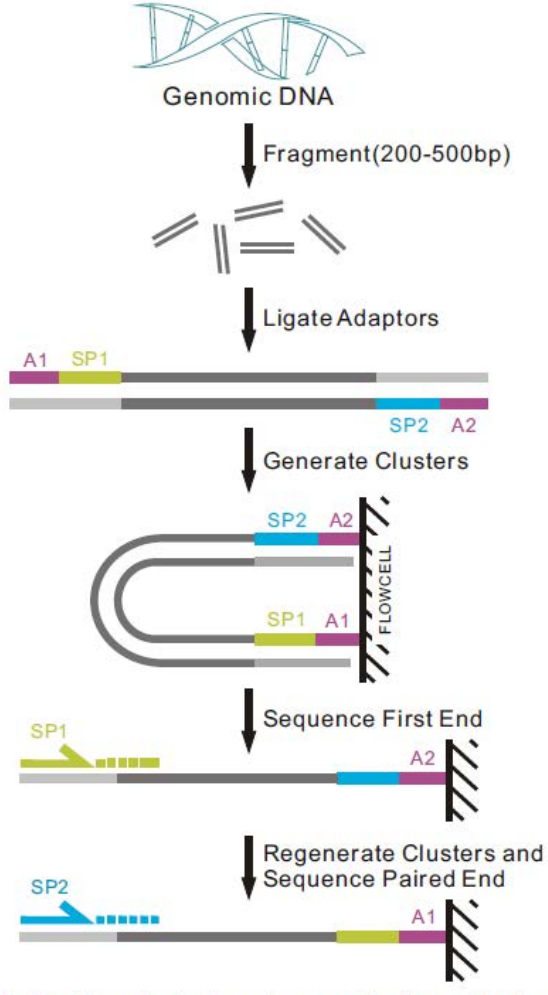
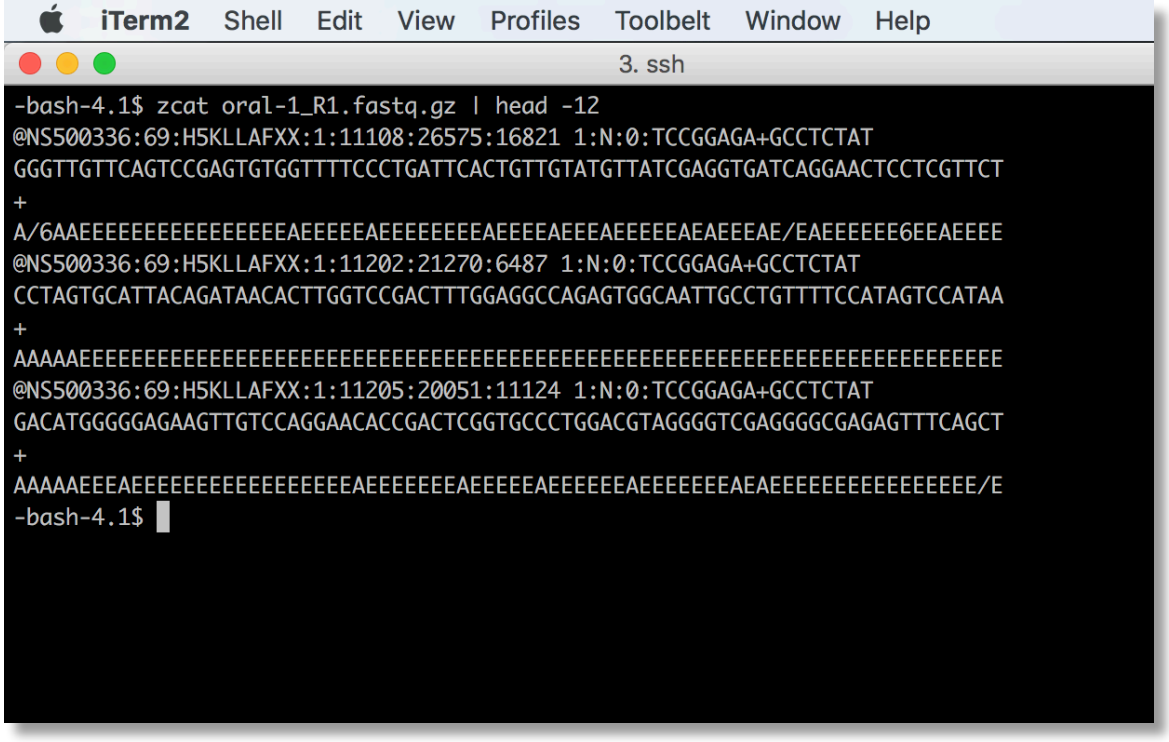


Figure 1-2-1 Pipeline of paired-end sequencing (www.illumina.com)

Fastq format (.fastq)

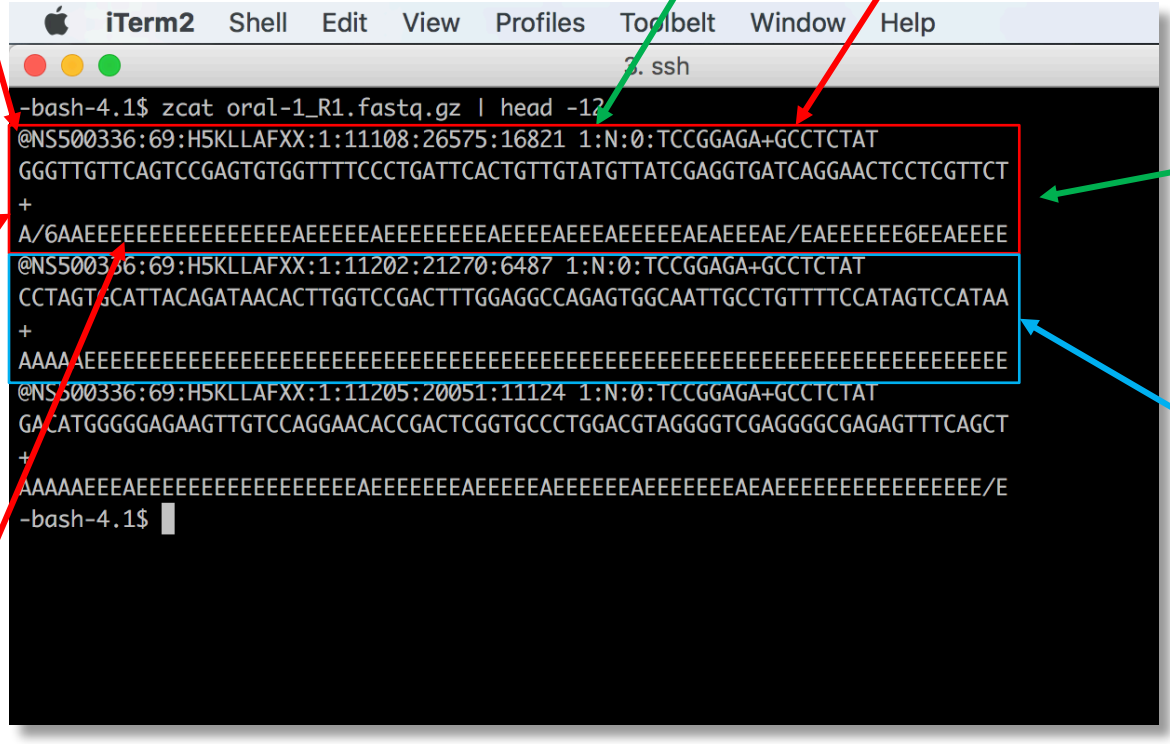


Fastq format (.fastq)

When paired-end
Pair mate has ‘2’ here

One read (each read cover 4 lines)

Each read starts with ‘@’



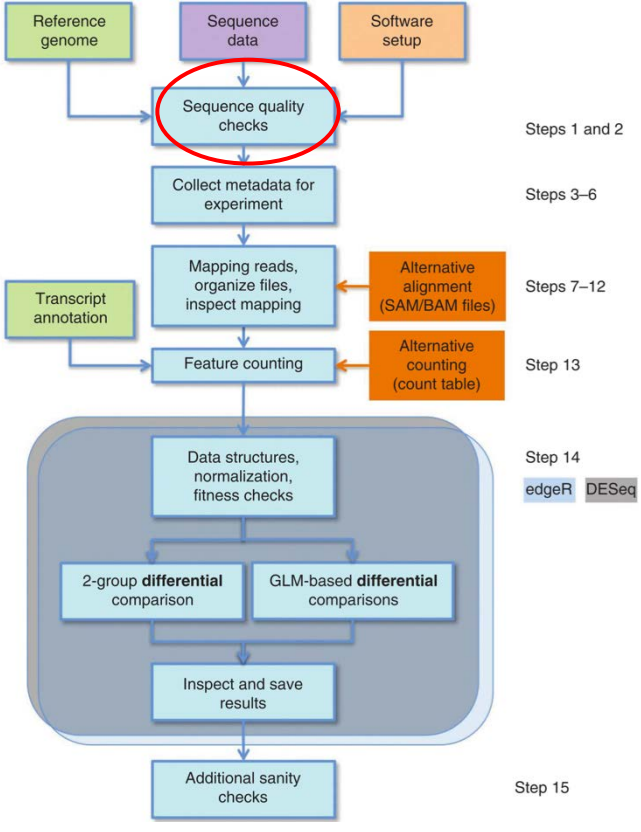
Separator

Sequence

Another read

Sequence quality scores

A general RNA-seq pipeline



PROTOCOL

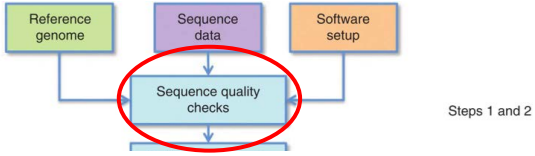
Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]

[1]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [2]Department of Statistics, University of Oxford, Oxford, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Bioinformatics Division, Walter and Eliza Hall Institute, Parkville, Victoria, Australia. [5]Department of Medical Biology, University of Melbourne, Melbourne, Victoria, Australia. [6]Functional Genomics Center UNI ETH, Zurich, Switzerland. [7]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, Australia. [8]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [9]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. Correspondence should be addressed to M.D.R. (mark.robinson@imls.uzh.ch) or W.H. (whuber@embl.de).

Published online 22 August 2013; doi:10.1038/nprot.2013.099

RNA sequencing (RNA-seq) has been rapidly adopted for the profiling of transcriptomes in many areas of biology, including studies into gene regulation, development and disease. Of particular interest is the discovery of differentially expressed genes across different conditions (e.g., tissues, perturbations) while optionally adjusting for other systematic factors that affect the data-collection process. There are a number of subtle yet crucial aspects of these analyses, such as read counting, appropriate treatment of biological variability, quality control checks and appropriate setup of statistical modeling. Several variations have been presented in the literature, and there is a need for guidance on current best practices. This protocol presents a state-of-the-art computational and statistical RNA-seq differential expression analysis workflow largely based on the free open-source R language and Bioconductor software and, in particular, on two widely used tools, DESeq and edgeR. Hands-on time for typical small experiments (e.g., 4–10 samples) can be <1 h, with computation time <1 d using a standard desktop PC.

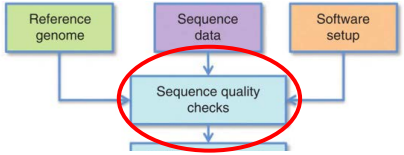


Steps 1 and 2

Sequence quality

Raw read trimming

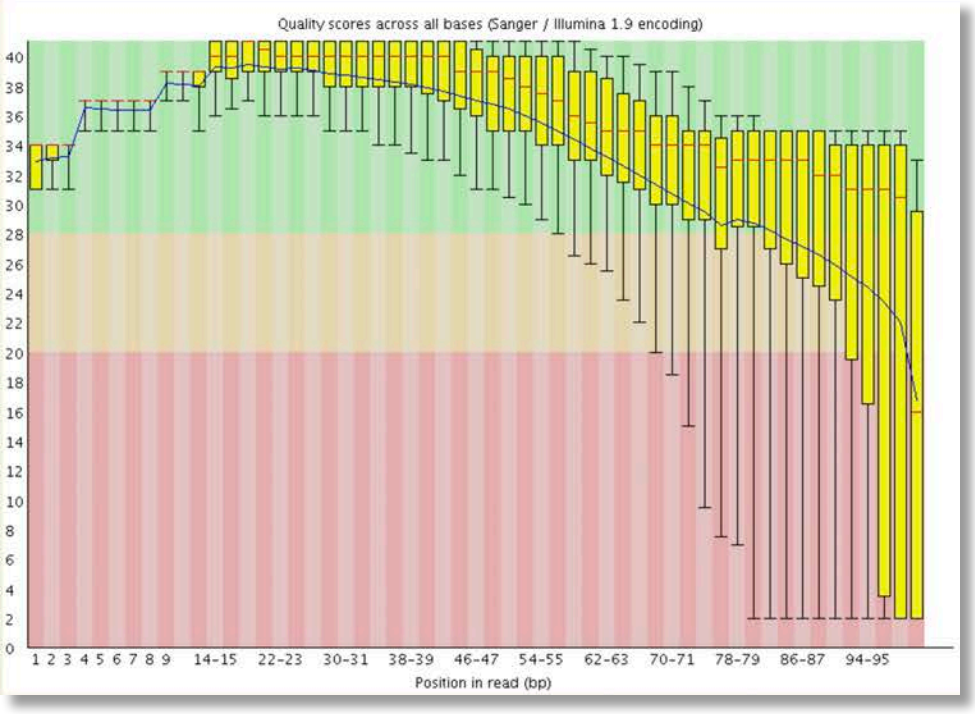
- Adapter trim: based on sequence similarity
  - - adapter/sequencing primer removal
- Hard trim: set number of bases
  - Certain primers (tags)
  - Known bias
- Soft trim: set quality threshold
  - Quality trimming



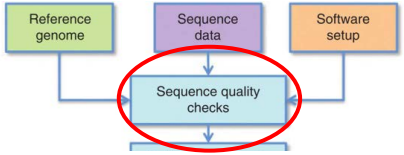
Steps 1 and 2

Sequence quality

FastQC



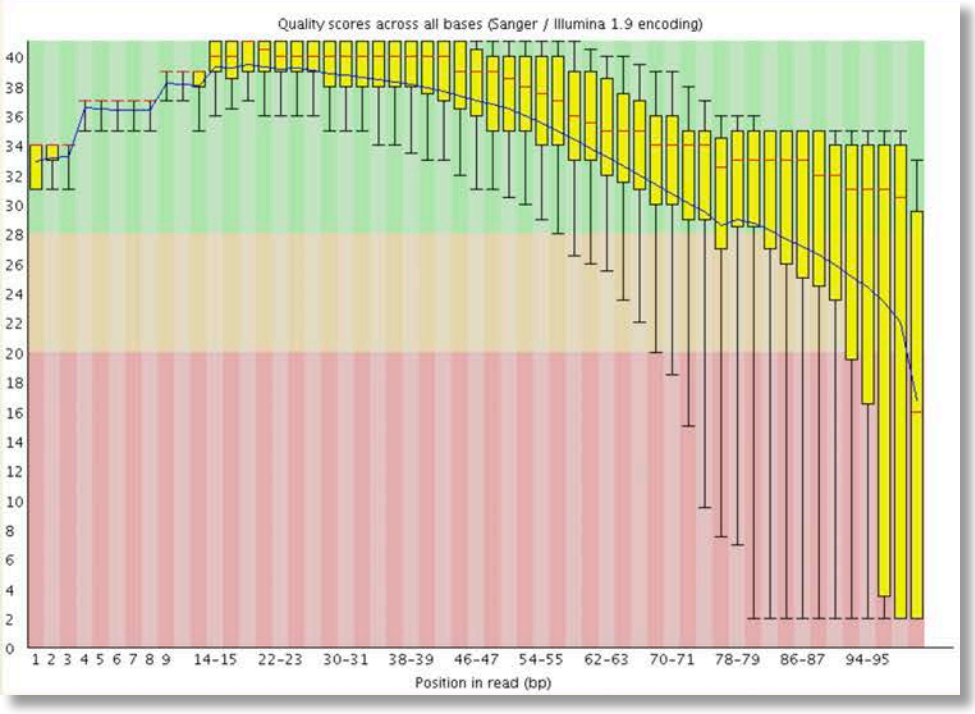
Phred Quality Score 30 = 99.9% accuray (20 = 99%)

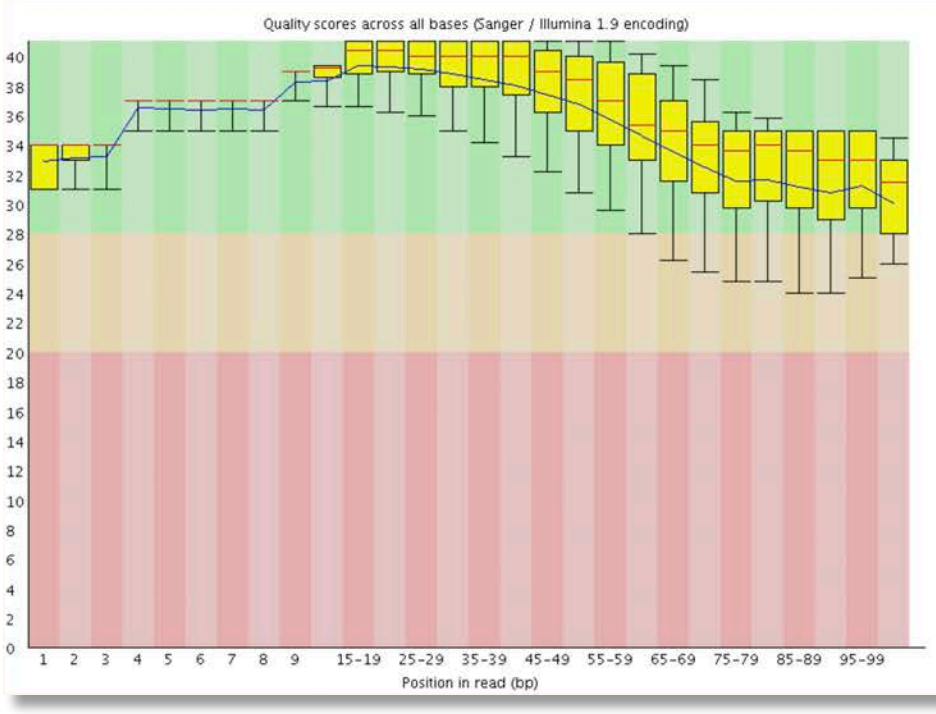


Steps 1 and 2

Sequence quality

FastQC

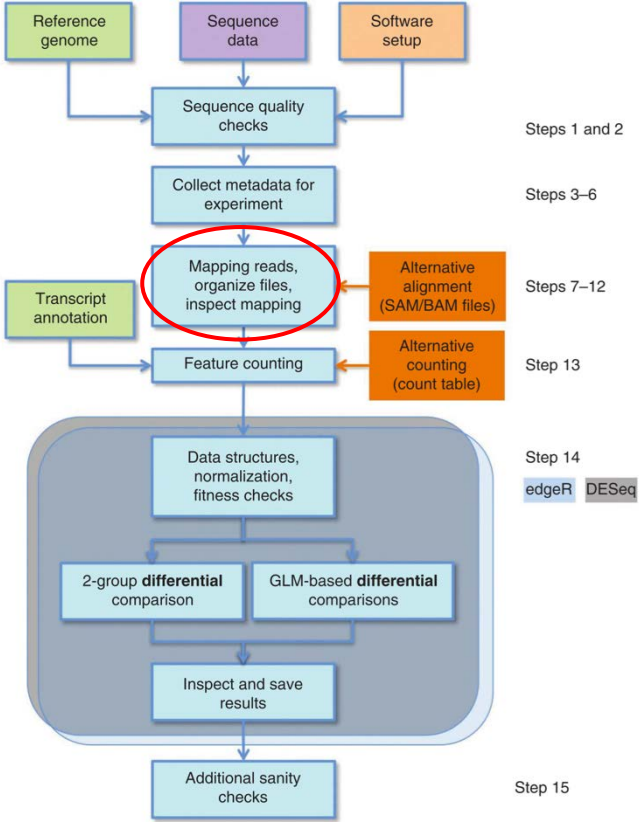


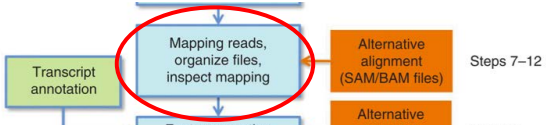


Phred Quality Score 30 = 99.9% accuray (20 = 99%)

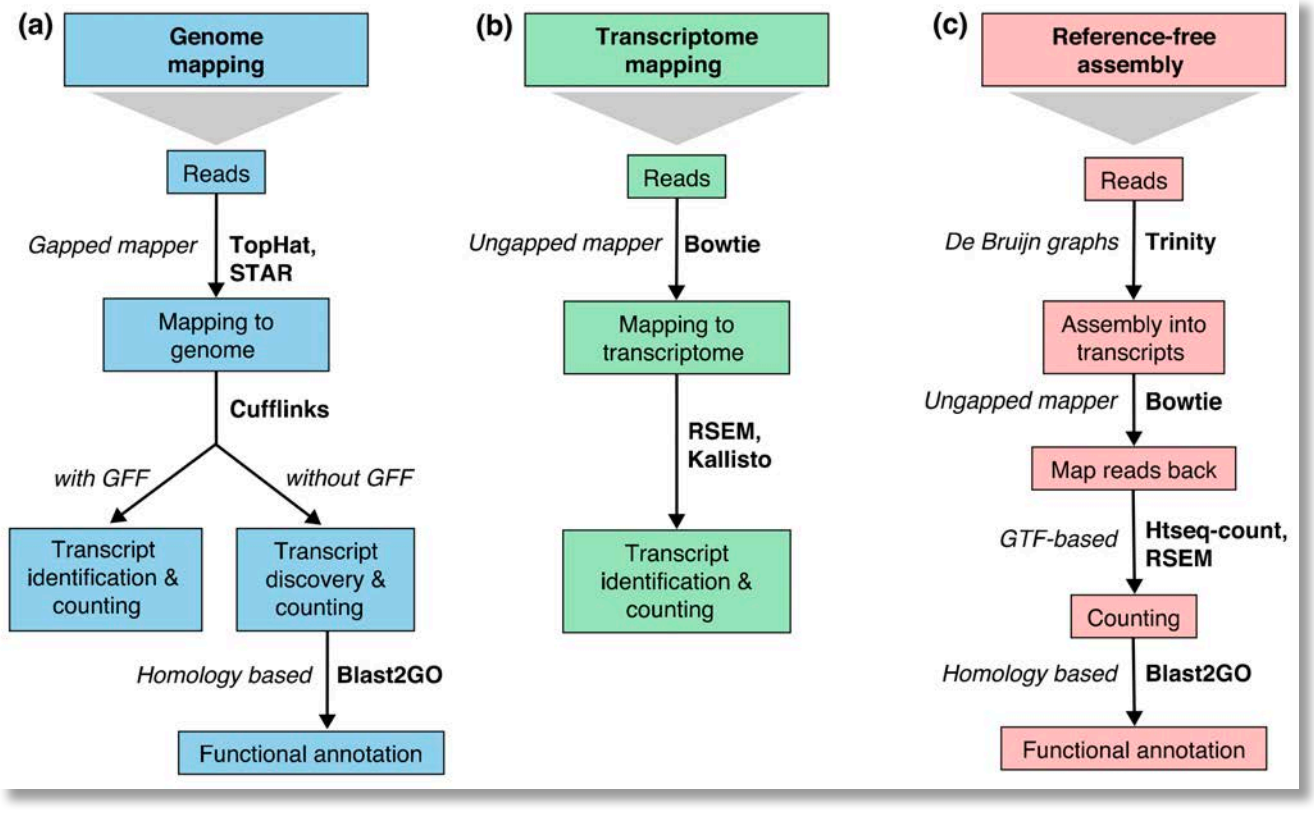
Exercise 1 – Quality assessment

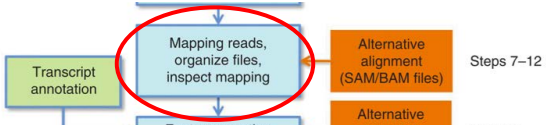
Mapping / *de novo* assembly





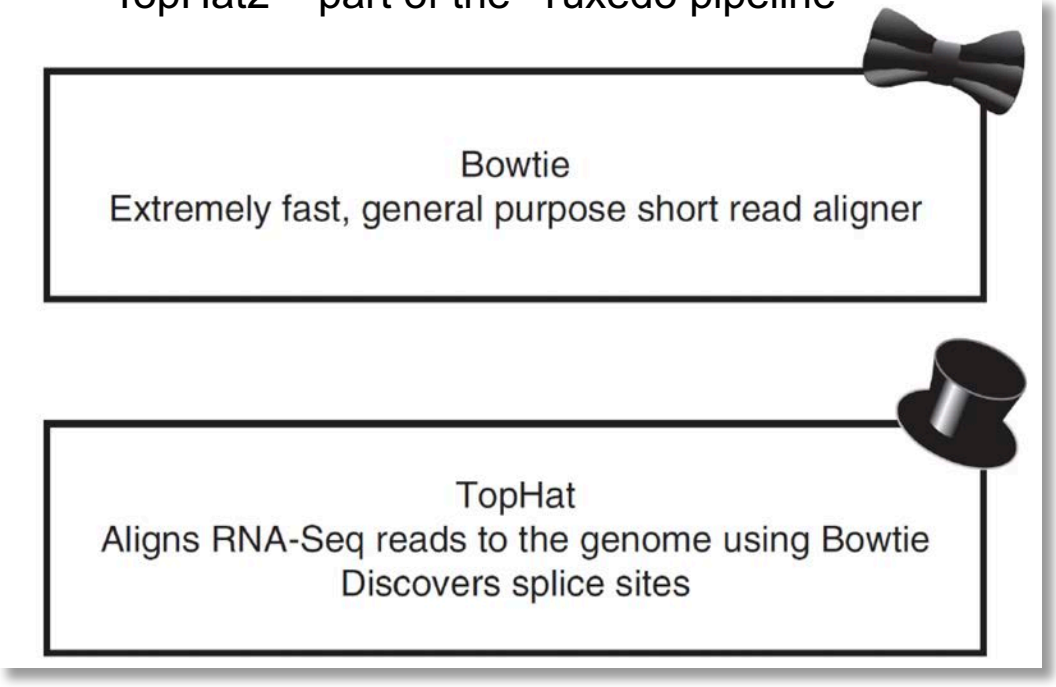
Mapping / *de novo* assembly





Reference genome

TopHat2 – part of the “Tuxedo pipeline”



PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [3]Department of Computer Science, University of California, Berkeley, California, USA. [4]Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [6]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [7]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [8]Department of Mathematics, University of California, Berkeley, California, USA. [9]Department of Molecular and Cell Biology, University of California, Berkeley, California, USA. Correspondence should be addressed to C.T. (cole@broadinstitute.org).

Published online 1 March 2012; corrected after print 7 August 2014; doi:10.1038/nprot.2012.016

HISAT – replaces TopHat2?

PROTOCOL

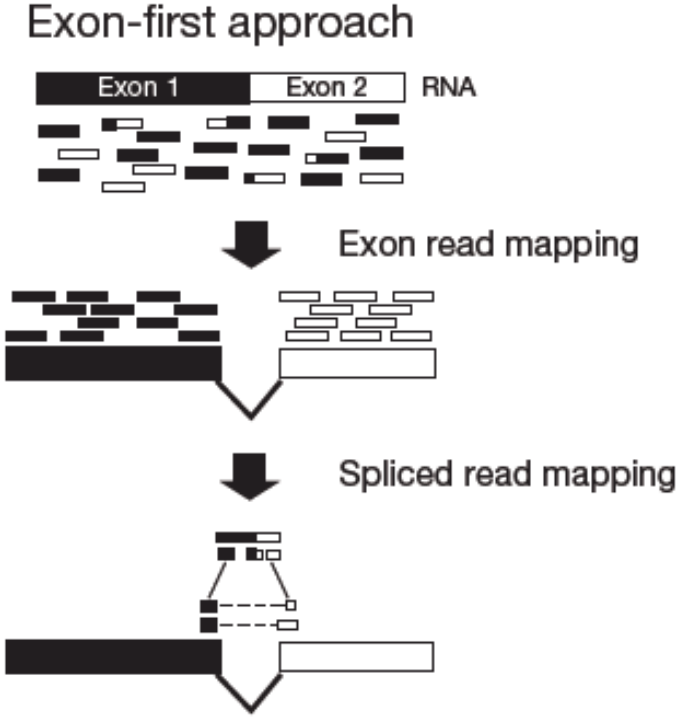
Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea[1,2], Daehwan Kim[1], Geo M Pertea[1], Jeffrey T Leek[3] & Steven L Salzberg[1–4]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [2]Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA. [3]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. Correspondence should be addressed to S.L.S. (salzberg@jhu.edu).

Published online 11 August 2016; doi:10.1038/nprot.2016.095

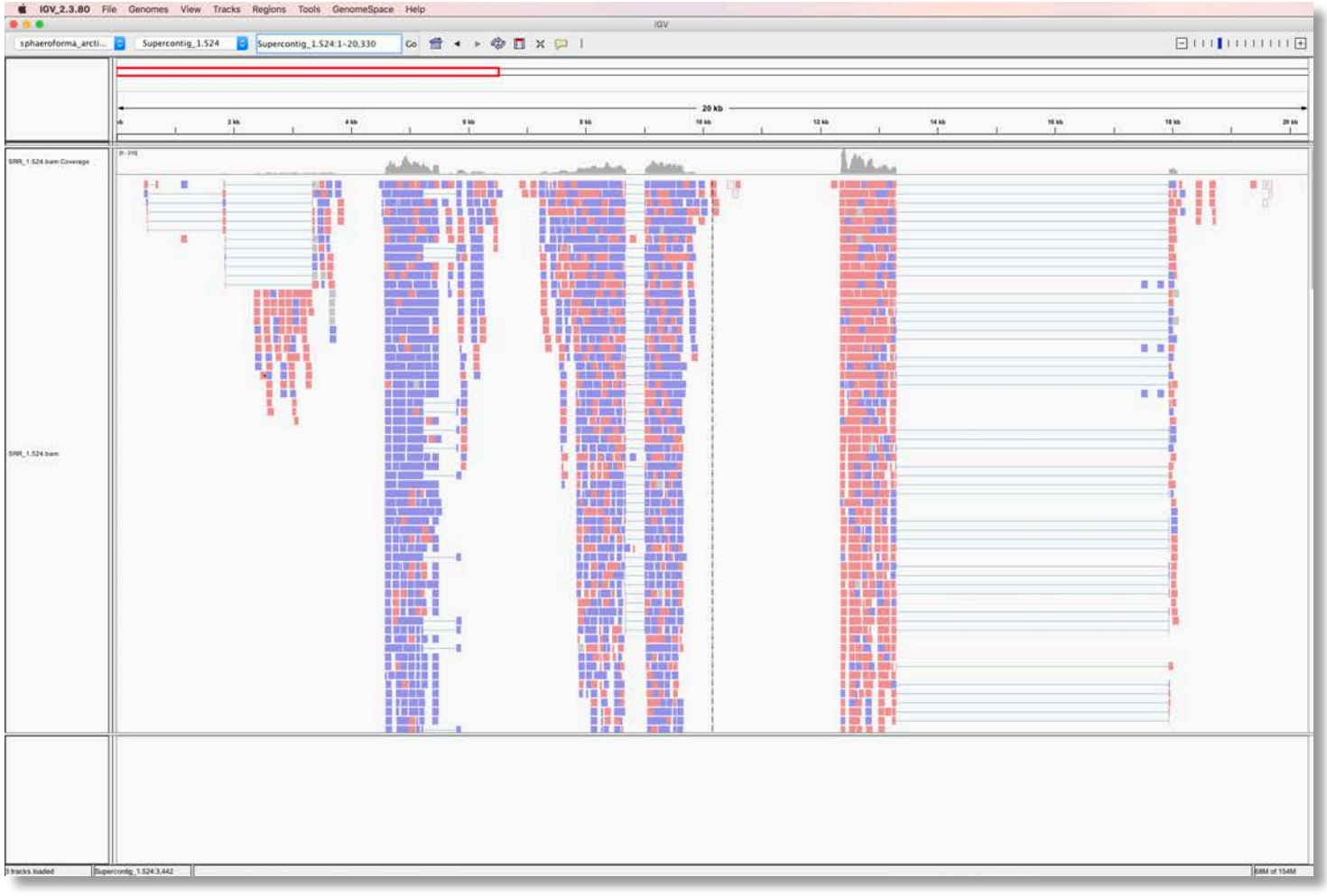
TopHat2



1. aligns reads to the transcriptome using bowtie

2. Unmapped reads are mapped to the genome (discovery of novel transcripts)

3. Remaining reads are split up and mapped to find splice junctions


IGV_2.3.80
File
Genomes
View
Tracks
Regions
Tools
GenomeSpace
Help
IGV
sphaeroforma_arcti...
Supercontig_1.524
Supercontig_1.524:1-20,330
Go
20 kb
SRR_1.524.bam Coverage
SRR_1.524.bam
Supercontig_1.524:3,442

Exercise 2 – Mapping