Задачи транскриптомики ("РНК биоинформатика")

Ярослав Баранов
МНЛ«Компьютерные технологии», Университет ИТМО

Overview

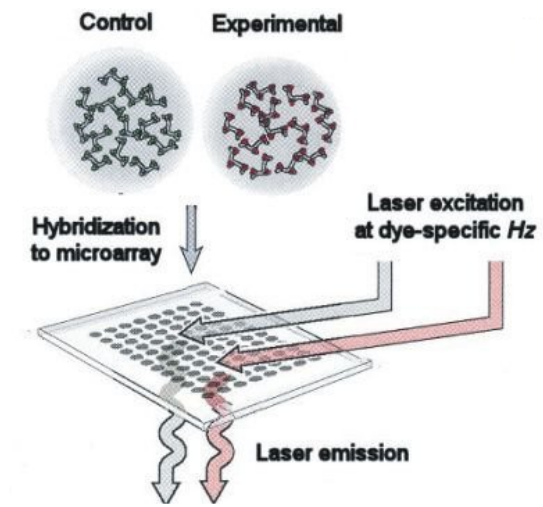
- Microarray
- RNA-Seq:

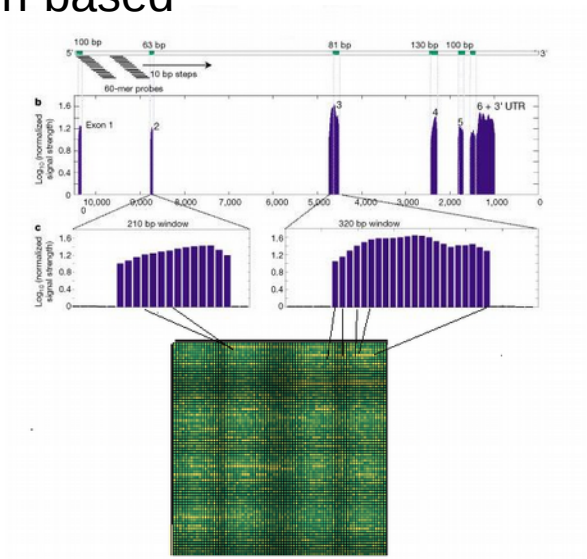
  Transcriptome assembly

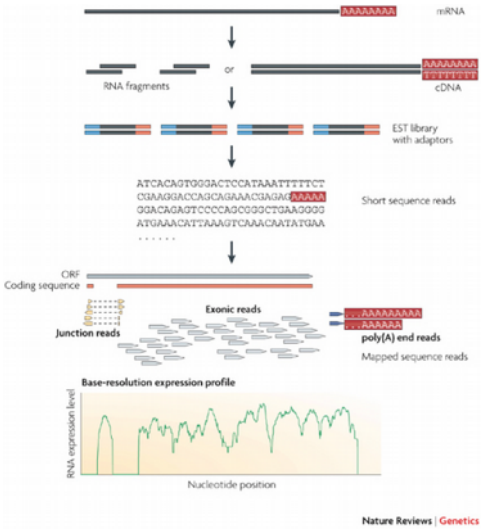
  Quantification of transcripts (diff. gene expression)
- Single-cell RNA-Seq

The evolution of transcriptomics

Hybridization-based







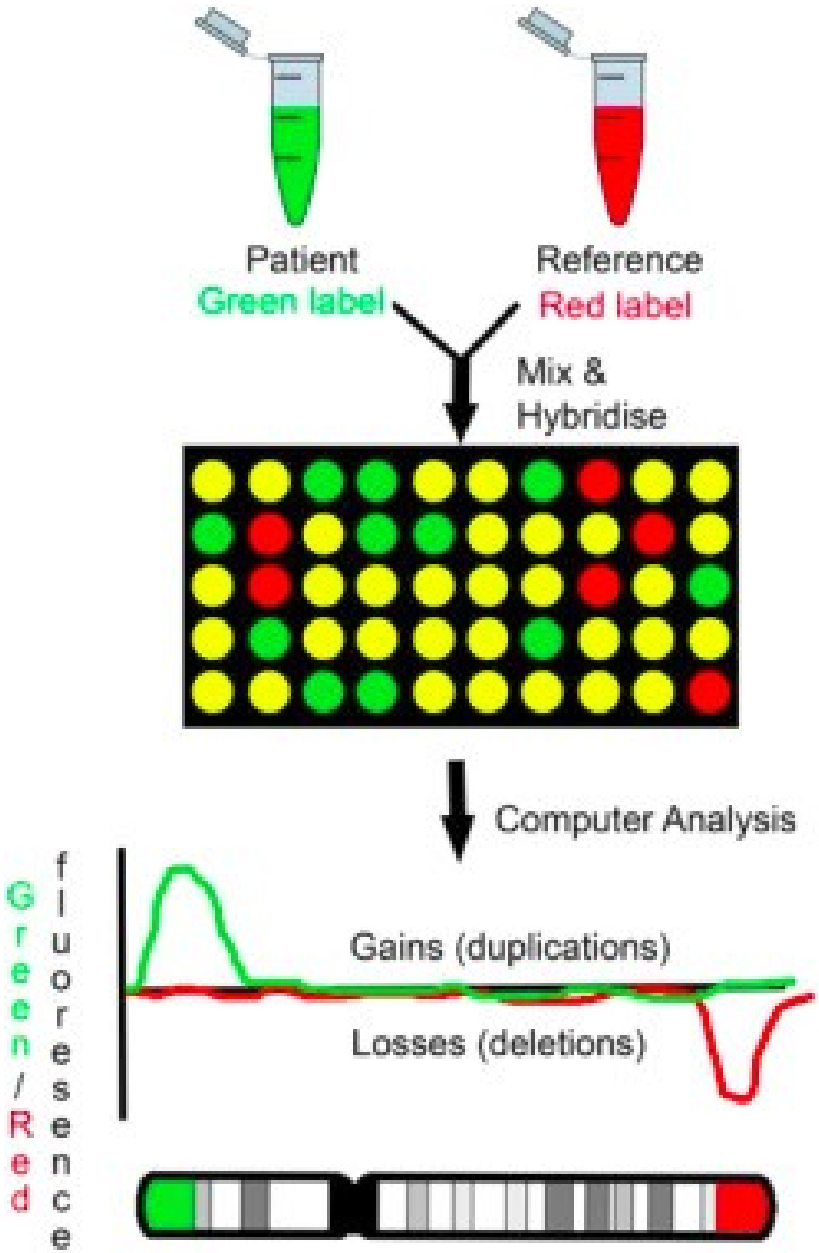
1995 P. Brown, et. al.
Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

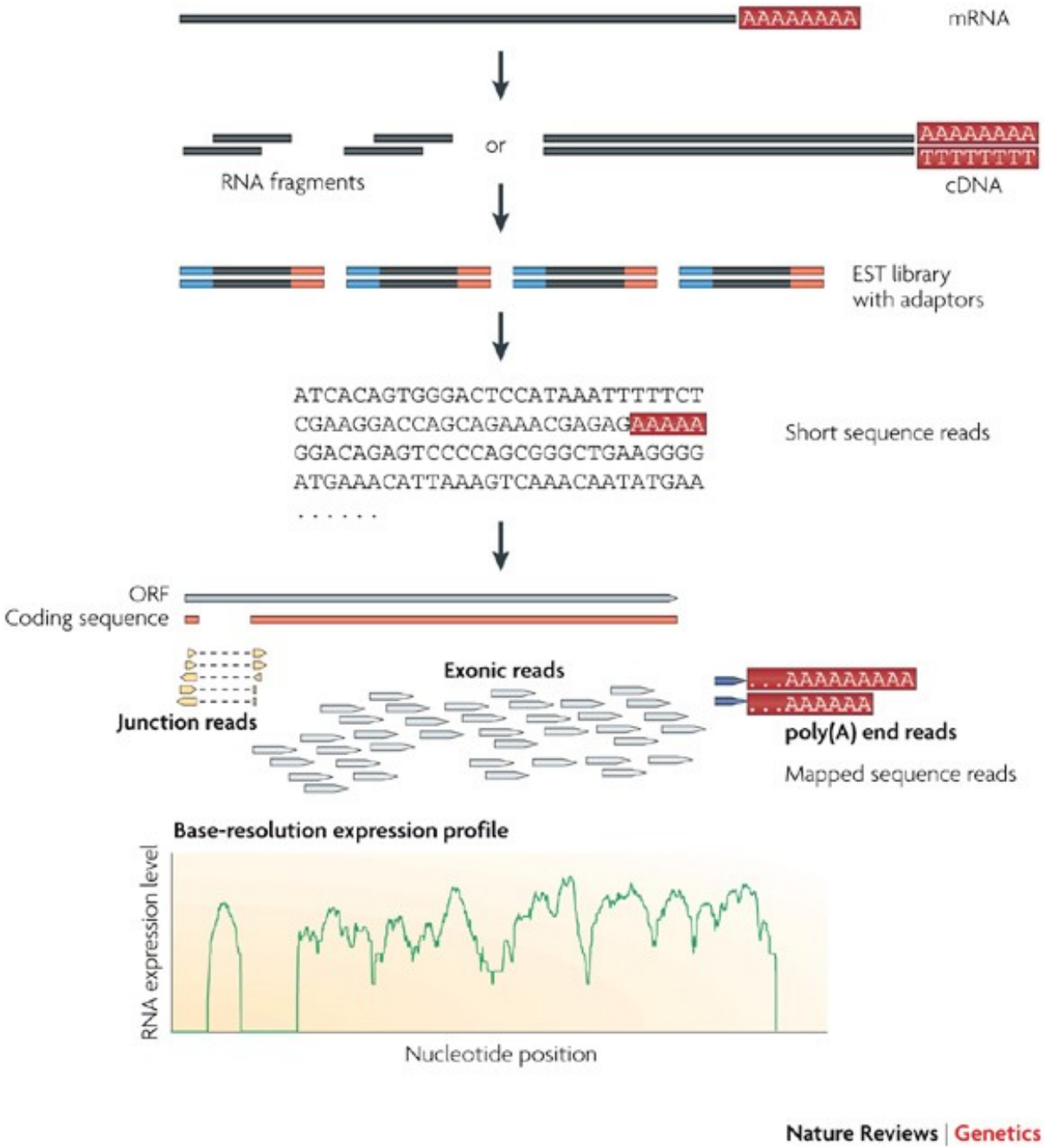
RNA-seq is still a technology under active development

Microarray (микрочип)



Spot	Patient	Control	Green : Red
	2 copies	2 copies	1.0 : 1.0
	3 copies	2 copies	1.5 : 1.0
	1 copy	2 copies	0.5 : 1.0

How RNA-seq works



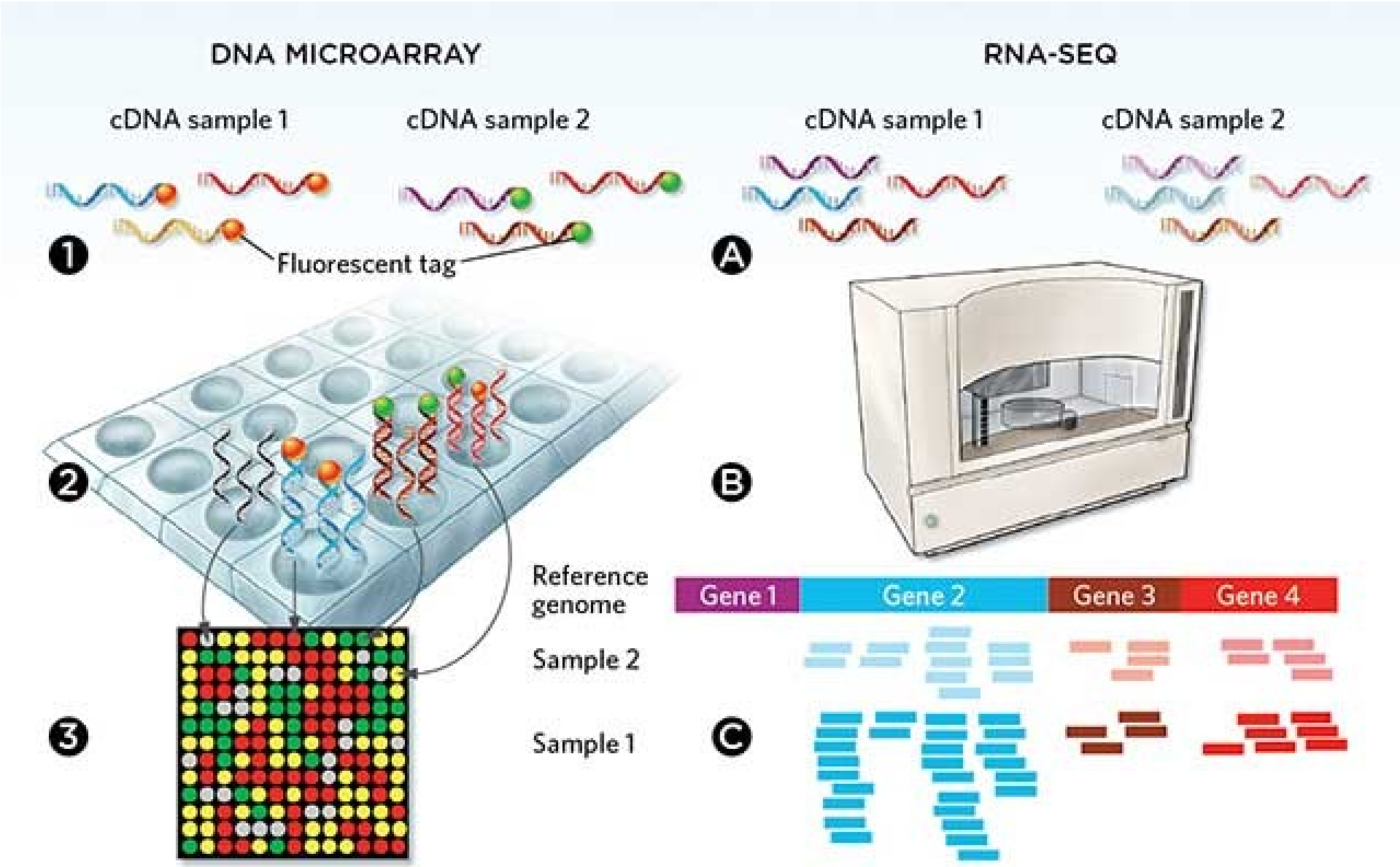
Sample preparation

Next generation sequencing (NGS)

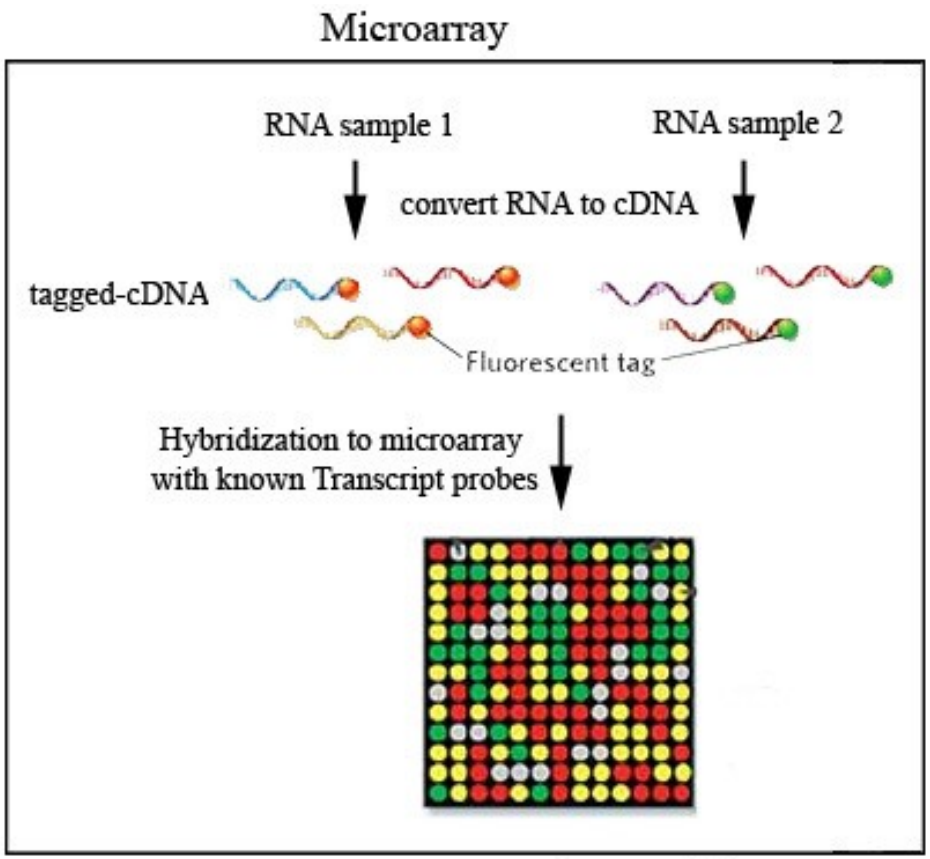
Data analysis:
- ✓Mapping reads
- ✓Visualization (Gbrowser)
- ✓De novo assembly
- ✓Quantification

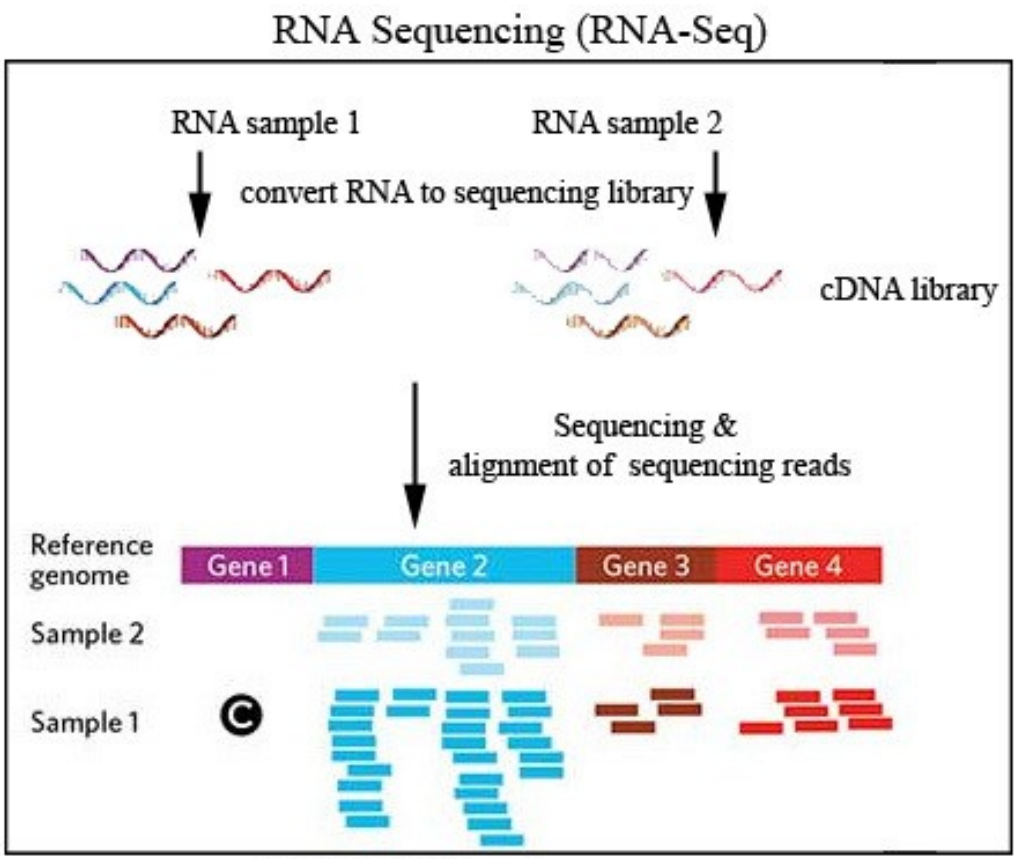
Figure from Wang et. al, **RNA-Seq: a revolutionary tool for transcriptomics,** Nat. Rev. Genetics 10, 57-63, 2009).

Microarray vs. RNA-Seq



Microarray vs. RNA-Seq





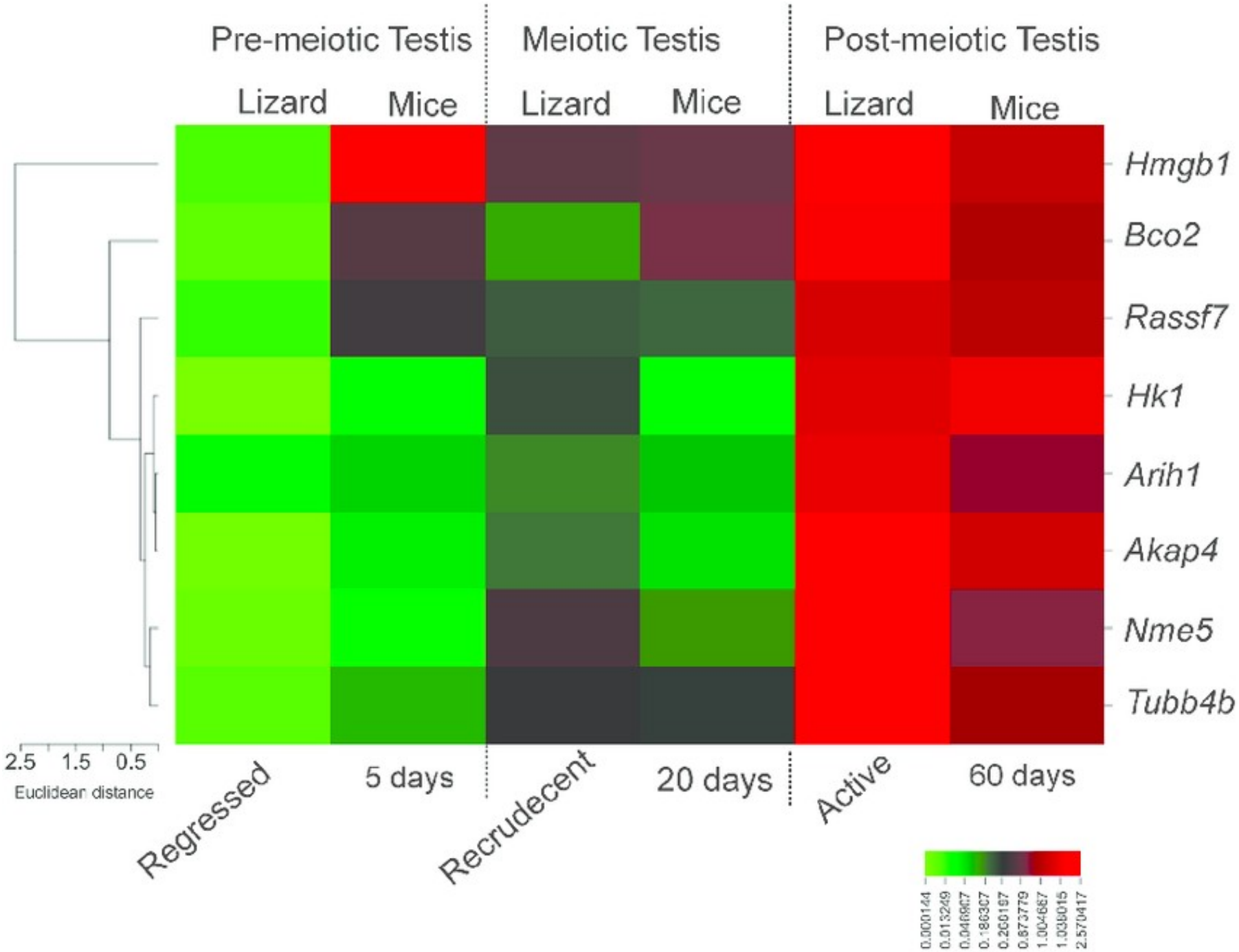
relative intensity
=
expression levels

Low sensitivity
Low dynamic range
known transcript only
No alternative splicing information
lower cost

High sensitivity
High dynamic range
Novel transcripts sequences identified
structural variation & alternative splicing revealed
unlimited sample comparisons

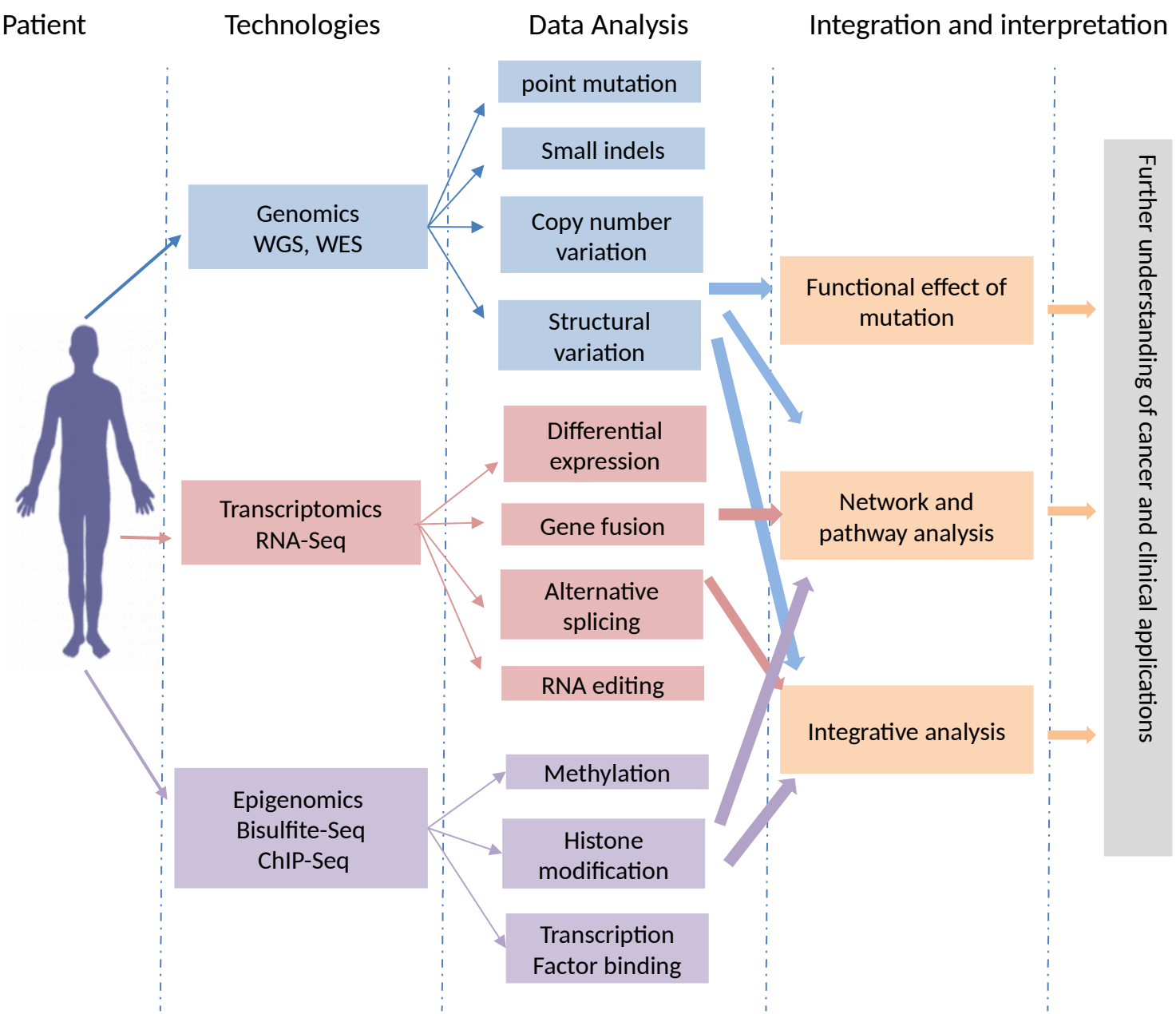
Sequencing Reads
=
expression levels

Gene expression heatmap



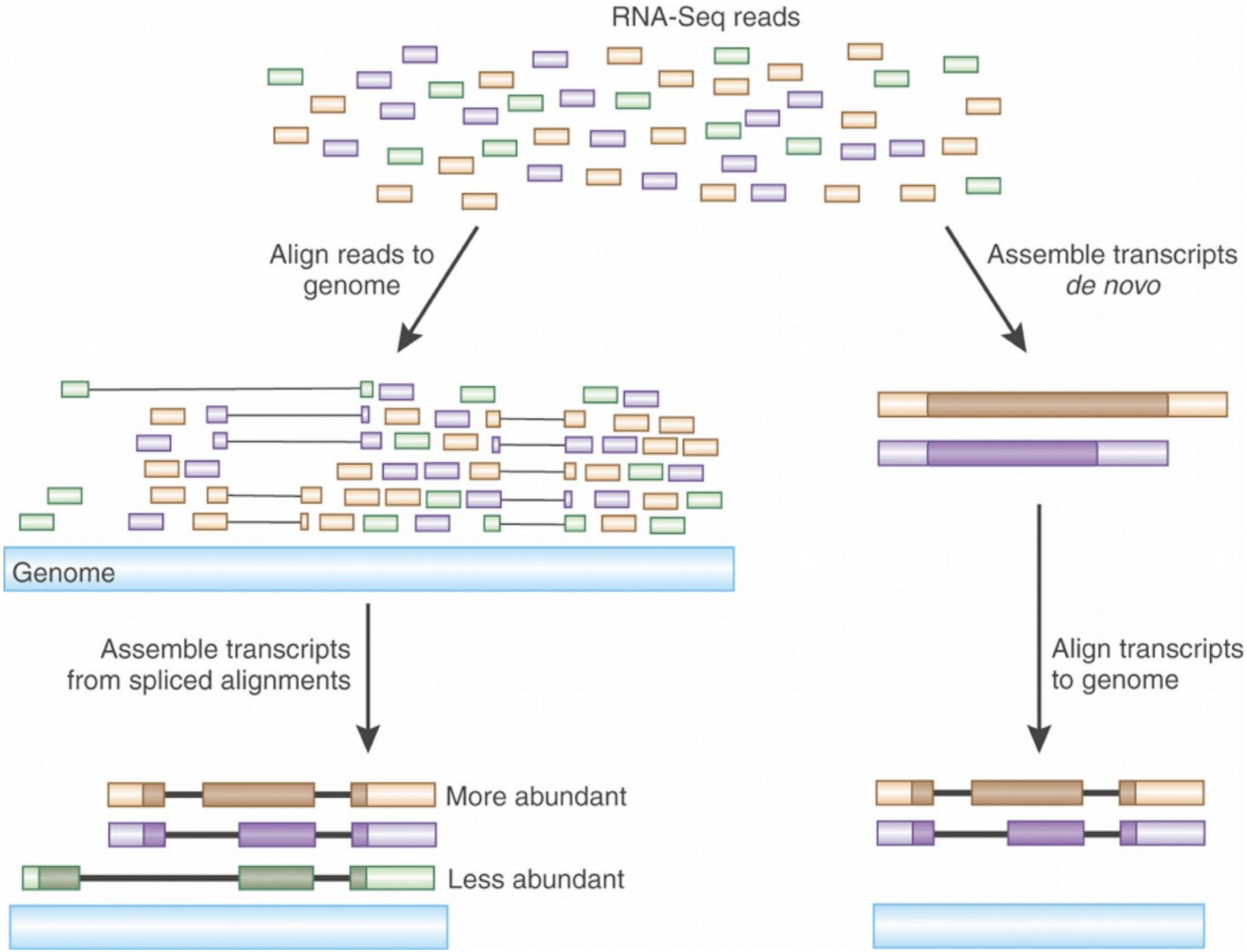
Next generation sequencing (NGS) techniques

	454 Sequencing	Illumina/Solexa	ABI SOLiD
Sequencing Chemistry	Pyrosequencing	Polymerase-based sequence-by-synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired end (PED) separation	**3 kb**	200-500 bp	**3 kb**
Mb per run	100 Mb	1300 Mb	3000 Mb
Time per PED run	<0.5 day	4 days	5 days
Read length (update)	250-400 bp	35, 75 and 100 bp	35 and 50 bp
Cost per run	$ 8,438 USD	$ 8,950 USD	$ 17,447 USD
Cost per Mb	$ 84.39 USD	**$ 5.97 USD**	**$ 5.81 USD**



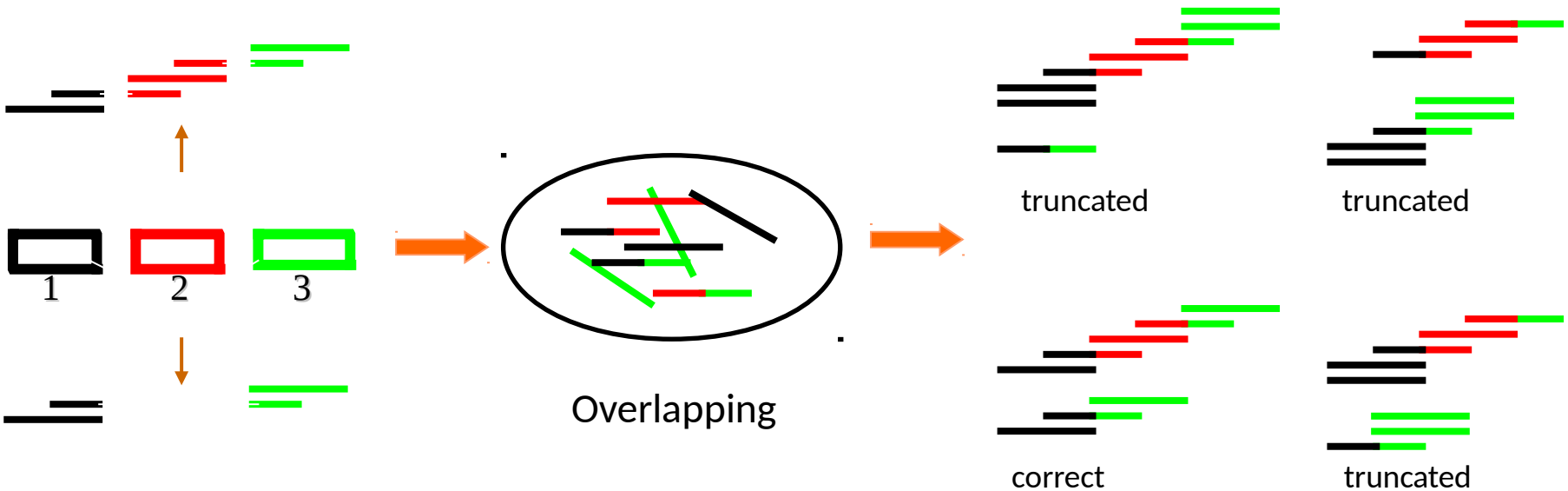
Shyr D, Liu Q. Biol Proced Online. (2013)15,4

Transcriptome assembly



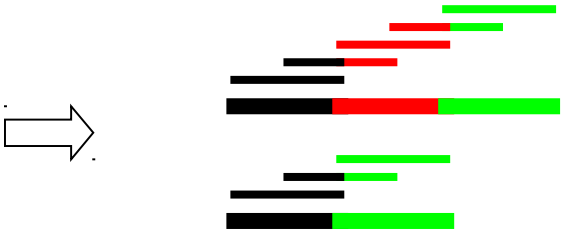
How to assemble multiple alternative spliced transcripts?

In the presence of AS, conventional assembly may be erroneous, ambiguous, or truncated.



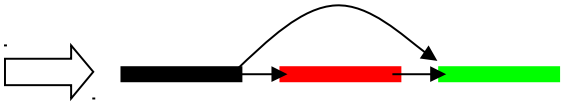
Splice graph approach

Replace the problem of finding a list of consensus sequences



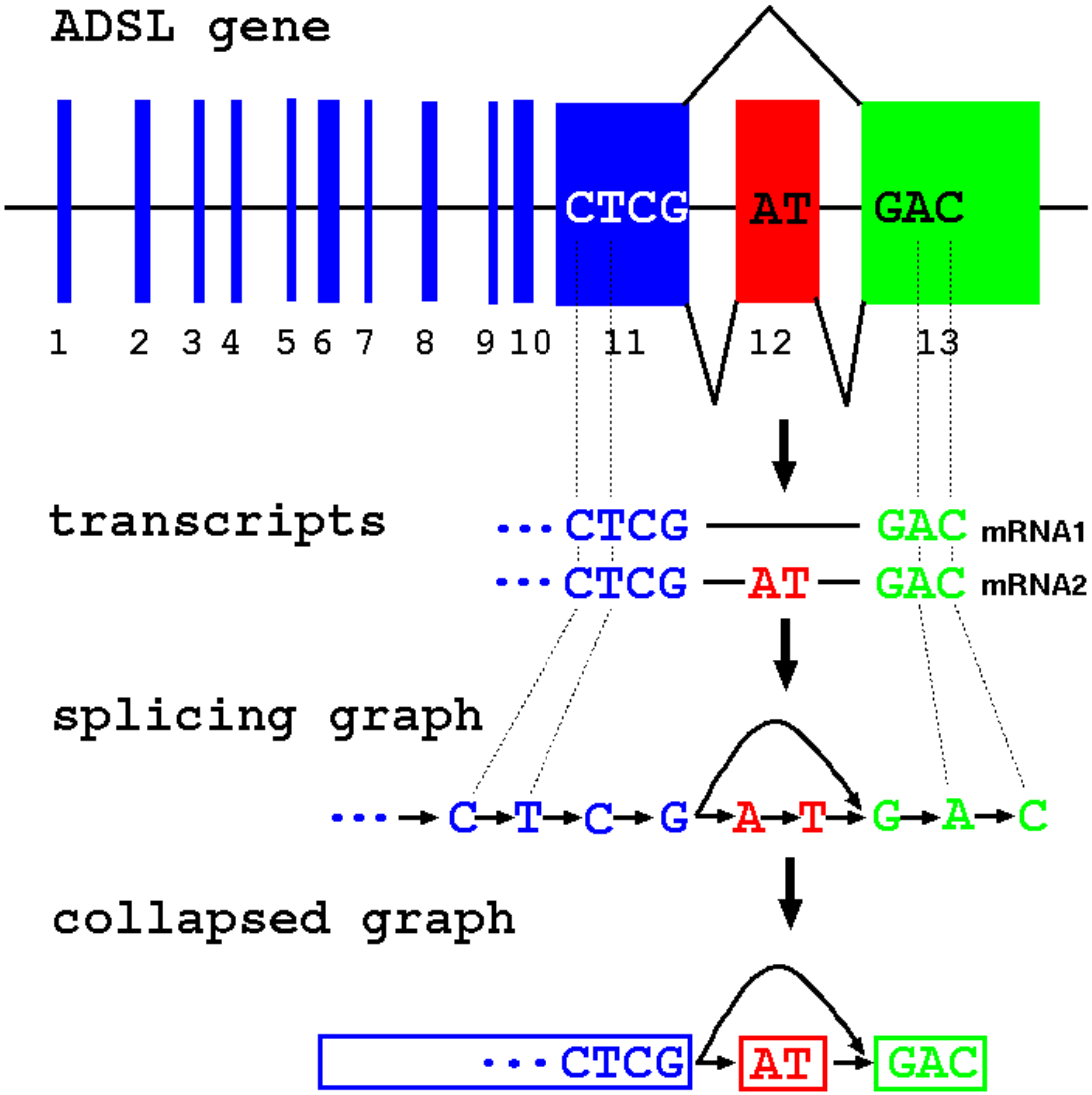
with ***Graph Reconstruction Problem***:

Given an set of expressed sequence, find a minimal graph (*splicing graph*) representing **all** transcripts as paths.



Heber, et. al. ISMB 2002

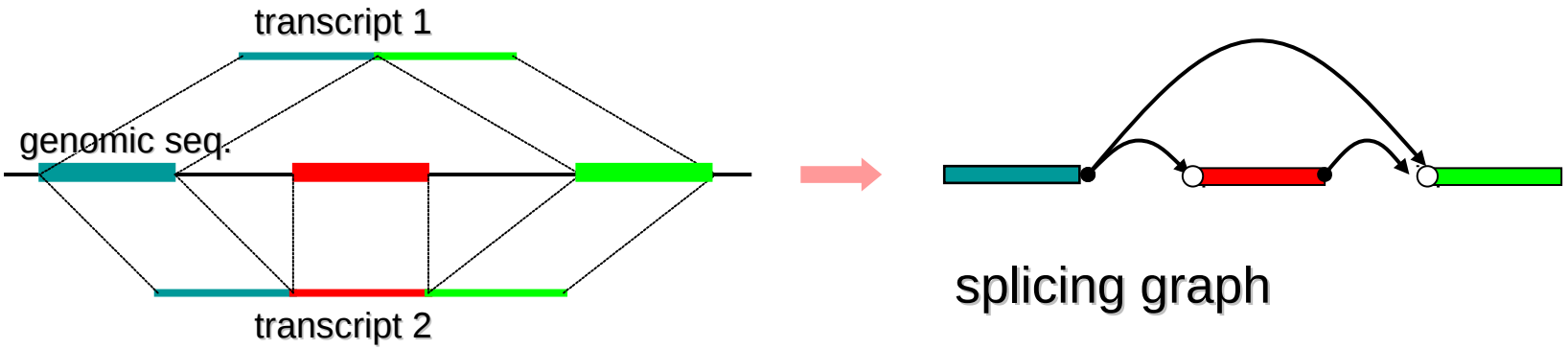
Splicing graphs



Splicing graph construction

If a reference genome is used:

- Map reads to the reference genome (short read aligner)
- Check alignment (splice sites, quality)
- Connect consecutive positions
- Build splicing graph



Splicing graph construction

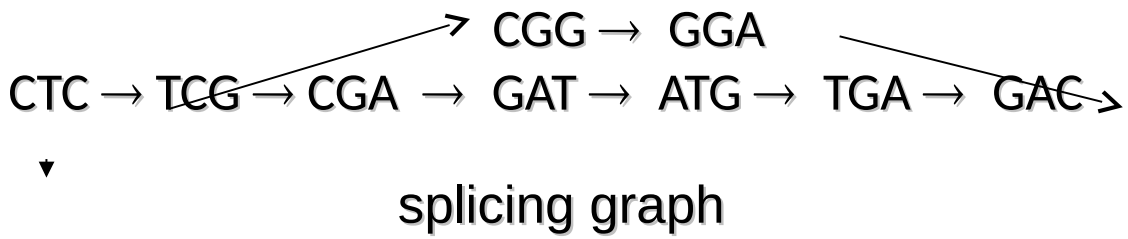
If a reference genome sequence is not used:

- Break sequences into *k*-mers (20-mers).
- Build graph using *k*-mers as vertices, connect them iff they occur consecutively in a sequence [Pevzner et al., 2001].

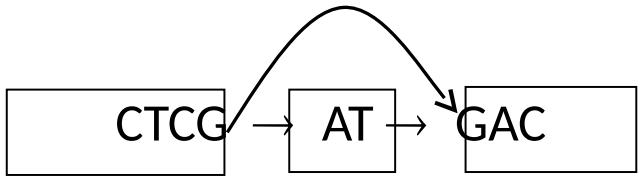
Example (3-mers):

Sequences: CTCGATGAC, CTCGGAC

Vertices: {CTC, TCG, CGA, GAT, ATG, TGA, GAC, CGG, GGA}

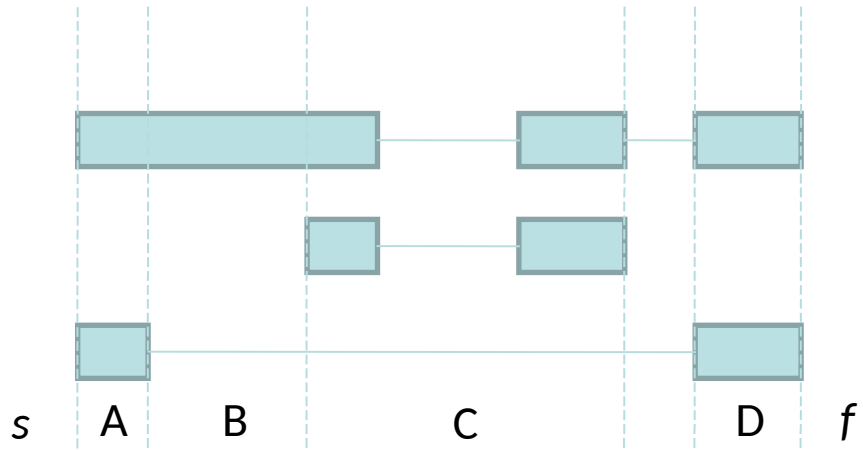


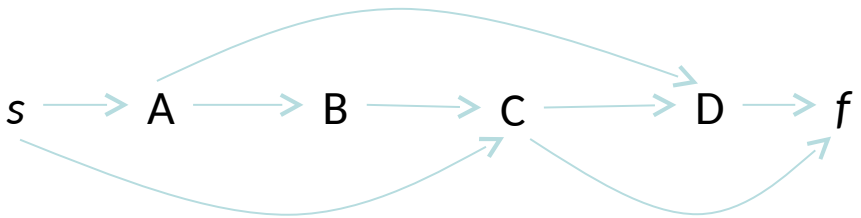
splicing graph



collapsed splicing graph

Splicing graph and splicing variants





An edge in the splicing graph, called a *block*, represents a maximal sequence of adjacent exons or exon fragments that always appear together in a given set of splicing variants. Therefore, variants can be represented by sequence of blocks, e.g. {ABCD, C, AD}.

Vertices *s* and *f* are included into graph, and are linked to the 5’ and 3’ of each variant, respectively. Each splicing variant corresponds to a directed path that goes from *s* to *t*. But note that some paths in the splicing graph do not correspond to real variants, e.g. {ABC, CD}.

Larcroix, et. al. WABI, 2009

Splicing graph construction

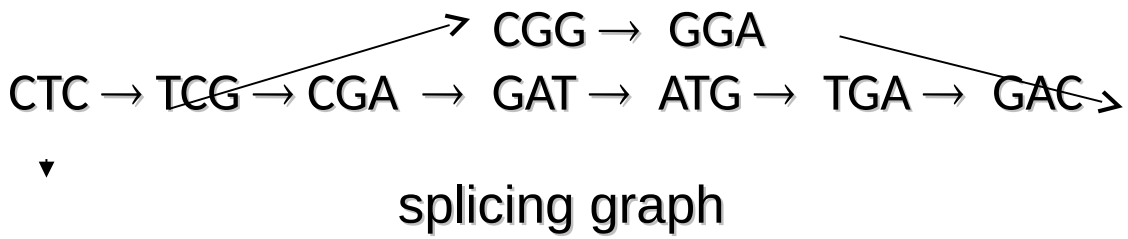
If a reference genome sequence is not used:

- Break sequences into *k*-mers (20-mers).
- Build graph using *k*-mers as vertices, connect them iff they occur consecutively in a sequence [Pevzner et al., 2001].

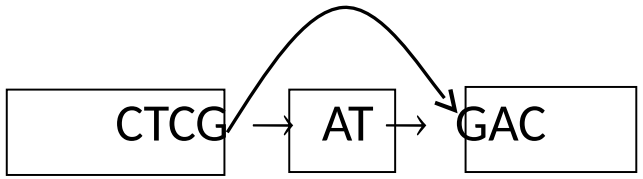
Example (3-mers):

Sequences: CTCGATGAC, CTCGGAC

Vertices: {CTC, TCG, CGA, GAT, ATG, TGA, GAC, CGG, GGA}



splicing graph



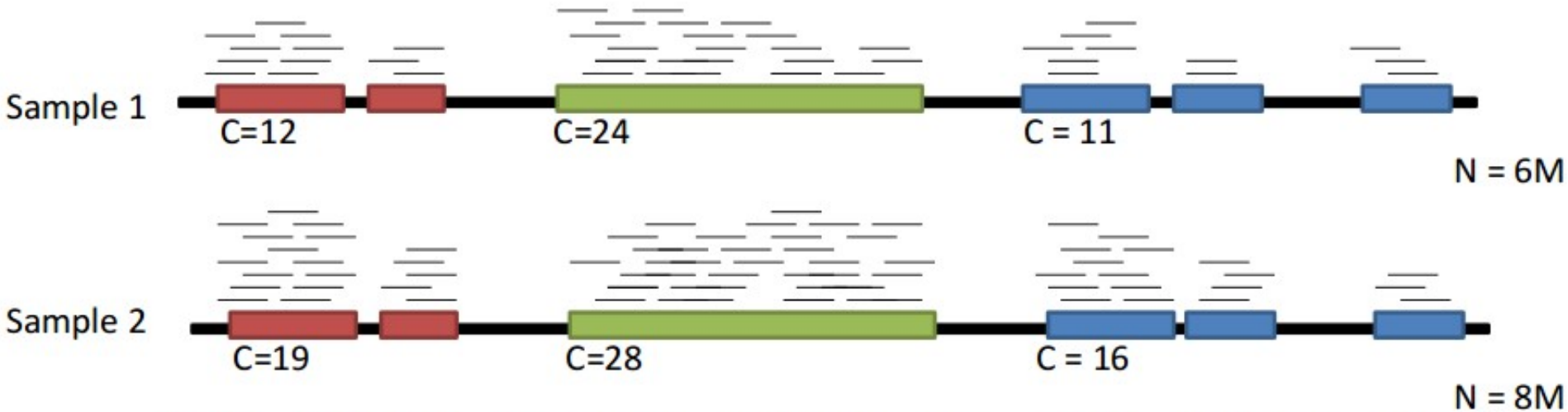
collapsed splicing graph

Transcript quantification

RPKM Example

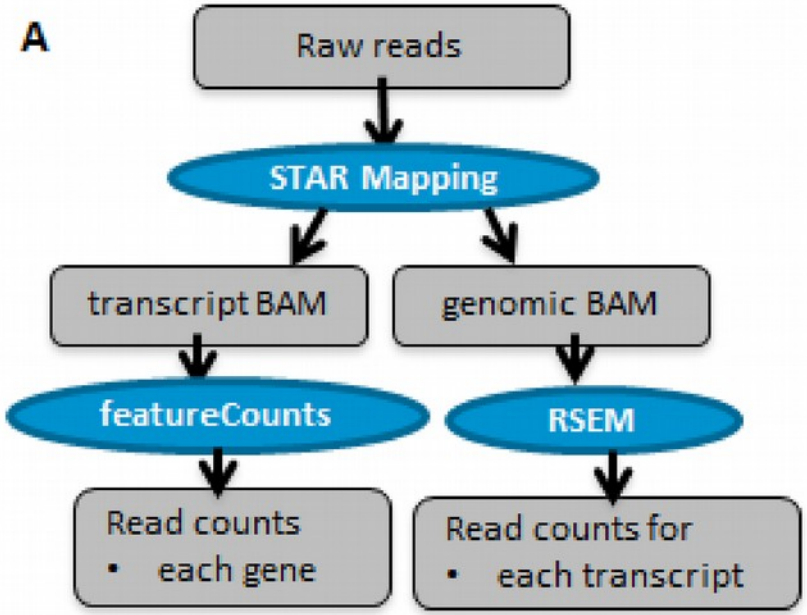
Gene A 600 bases Gene B 1100 bases Gene C 1400 bases

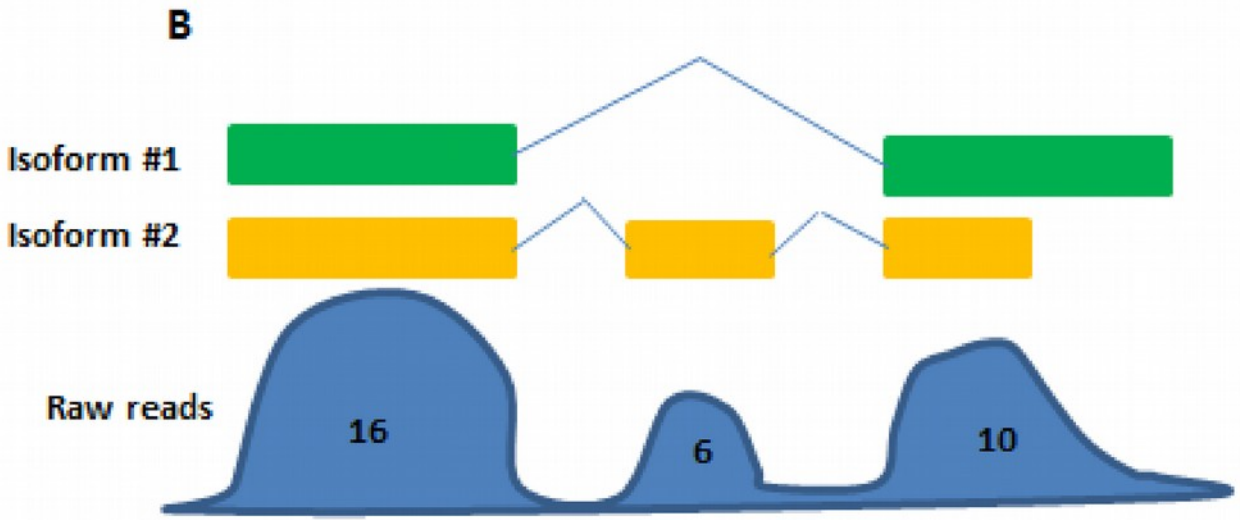
RPKM = 12/(0.6*6) = 3.33 RPKM = 24/(1.1*6) =3.64 RPKM = 11/(1.4*6) = 1.31

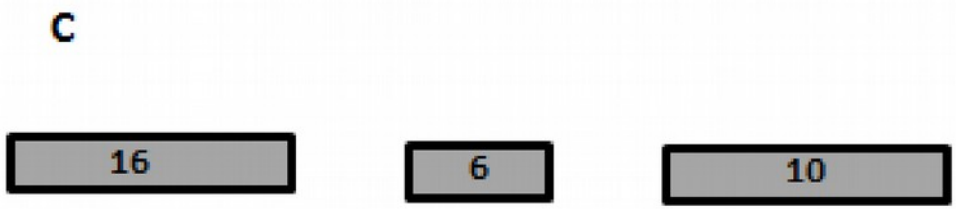


RPKM = 19/(0.6*8) = 3.96 RPKM = 28/(1.1*8) =1.94 RPKM = 16/(1.4*8) = 1.43

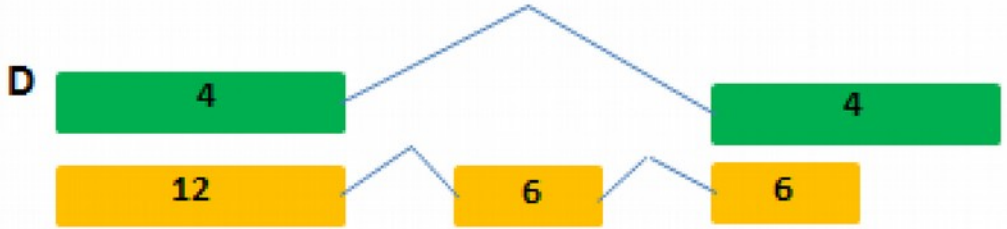
Transcript quantification





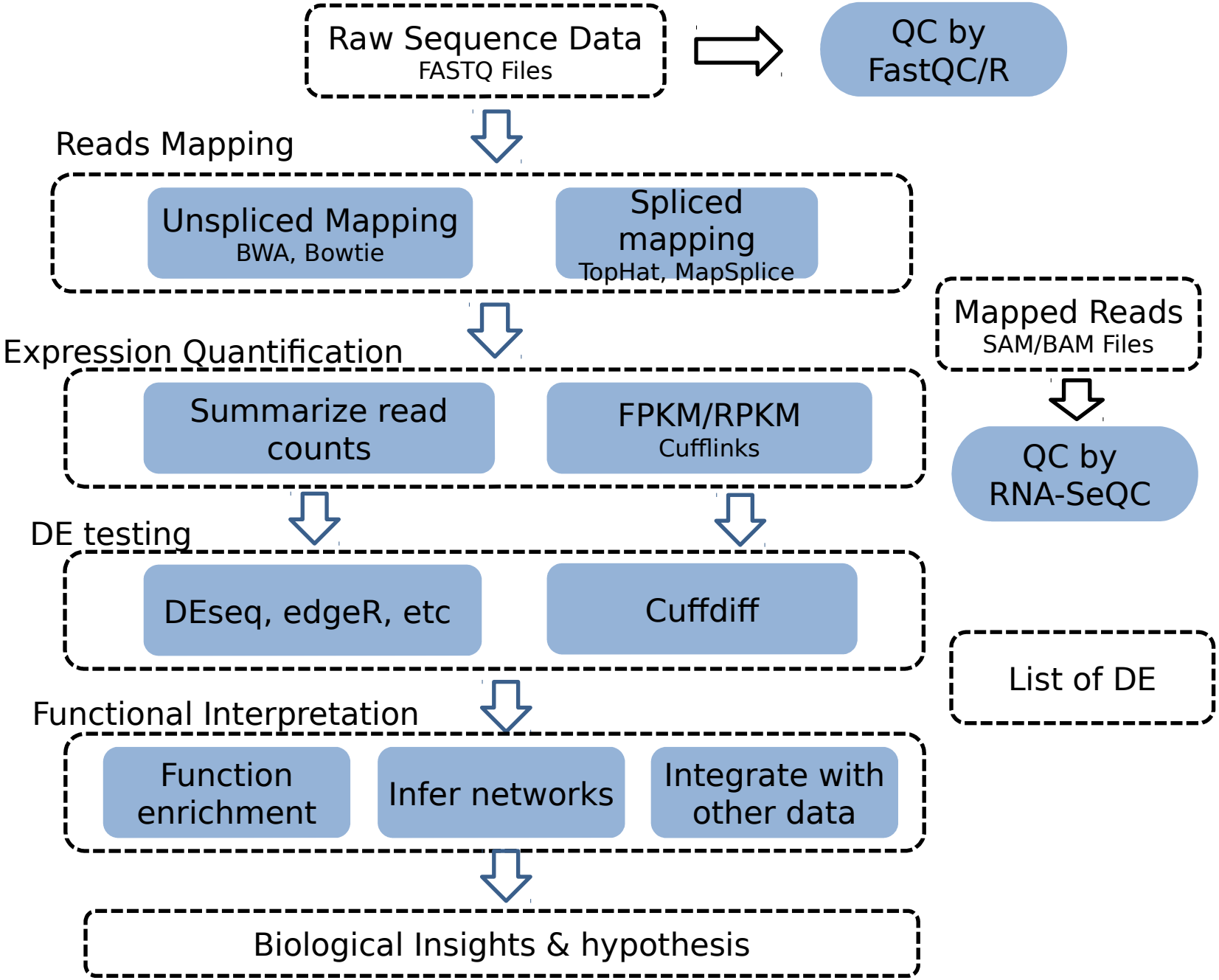


Total gene length after exon flattening: **5kb**
Total reads: **32**
RPKM for gene: **6.4 (=32/5)**



Relative isoform abundance (#1/#2): **25% / 75%**
RPKM for isoform #1 and #2: **2 and 6**
RPKM for gene (=sum of isoforms): **8 (=2+6)**

From reads to differential expression



FASTQ files

Line1: Sequence identifier
Line2: Raw sequence
Line3: meaningless
Line4: quality values for the sequence

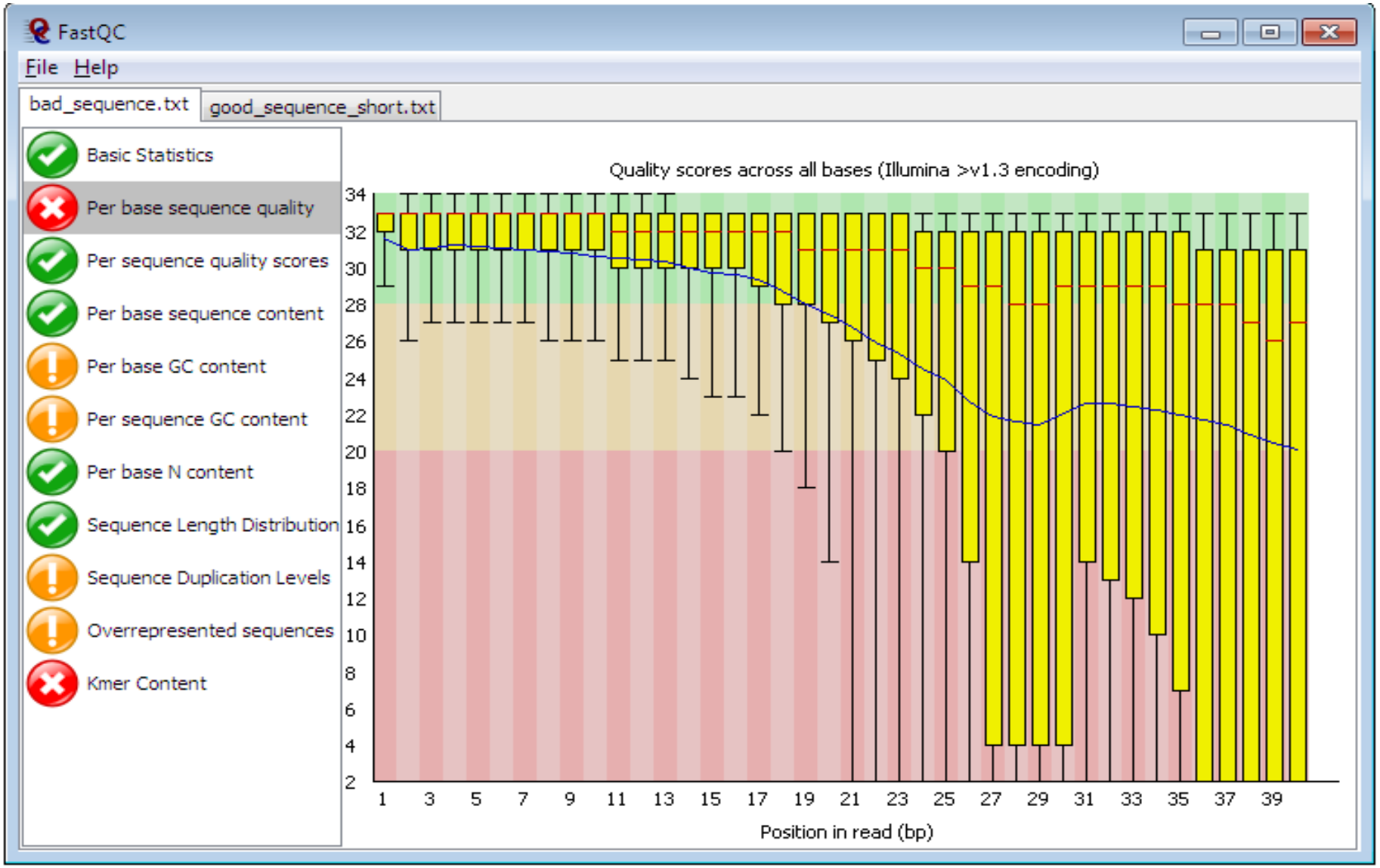
```
@HWI-ST508:210:C0EDTACXX:1:1101:1872:1227 1:N:0:
AATTGTGAAAACCCAAAAGGTGGAGCAGCCATTNTTATACATTGCAGAAGGGNGANNNANCNTTATGAAATTTAGCACCTGCCTTCCTGAATGATAAATGG
+
@CCFFEFFHHHHHJJJJIJJCGHEIIIJIJJJJ#1BFHIJJJJJJIJJIJJI#-;###-#-#-5?BFFFFEEEEECCDDDDDDDDDCCDDDDDDCCEED
@HWI-ST508:210:C0EDTACXX:1:1101:1895:1233 1:N:0:
TGACATAAGCTTGCATTTGAAAAGCACCTCCGAAAGCTTCCCAGCCTCAAAGNCANNATCGNCTTCTGATGCAGTTAGGCACCACAAGAGCTTCCCCACAA
+
CCCFFFFFHHHHGJJJJJIIJJIIIHJJIJJJJJJJJJJJJJJJJJJJGHJJ#.;##--;C#-5CEFFFFFEEECCEEDDDDDDDDDBDDDDDDDDDDDDB
@HWI-ST508:210:C0EDTACXX:1:1101:1761:1235 1:N:0:
GCTCTACTAAAAATATAAAAATTGGCCAGGCGCAGTGACACATGCCTGTAGTCCCNGCTATTCGGGAGGCTGACACACAAGAATCAATCACTTGAACCCAG
+
CCCFFFFFHGHHHJJJJJJJIJJJJJJJJIEIIIJFHGIIIIJJIJJJJHIJJIJ#-;FGGIJIJHHFFDDEEDDCCDDDDCCDDDDDDCDDDDDDDDDDD
@HWI-ST508:210:C0EDTACXX:1:1101:1971:1236 1:N:0:
CAGGATGAAAGAGGTCTGGCCAGGTGCTGGGTGCAGTGGCTCACACCTGTAATCCCAGCACTTTGGGAGGCCGAGGTGGGCGGATCACGAAGTCAGGAGTT
+
CCCFFFFFHGHHGJHIJIIJJJJI3CFGIJJ9DFHJDEHGIJIJJJJJIIJJJGGIJJJJJJJJFIJHFFFFDDDB/?BB@BD<39?CD@B8+:@CDCB##
@HWI-ST508:210:C0EDTACXX:1:1101:1830:1239 1:N:0:
TATTGATTCTTTTTTTTTTTTTTTTTTTTTTTTTTTCCTGGGGTCCTGCTTTGGGGGTCTCGGGGTCCCAAATTGCTGGTTTTACACCTCCCCCCCCCCG
+
?@@DD?DEFBBFDHEHIGEDA@FH>C??BBBCB6B#################################################################
@HWI-ST508:210:C0EDTACXX:1:1101:1999:1240 1:N:0:
AAAGAGTGAGAGAAGCAAGGCTTGTGTGAAGAGAGCAAAACTTAGAATCAACATTGGTTGAGCATCTCCTATGAGCTAATATTAATTAGCACTTTACATGC
+
@@@DDA2?FHBHHEGEHIHGIGGHBFCGIEHGAEGGIIEGIIIIGHIGEHEGHIGIGBFHEHIEAHGHHFHEH;B@DEBDCDEEBCDDCCCCC@@CCCDCC
@HWI-ST508:210:C0EDTACXX:1:1101:1806:1245 1:N:0:
ACATGCTAATATATGTACTGATATGGAACAATCTTTAAGATGTATTATTACATGGAATAAACCAACCAGACCACAAAACAGATGTTTTTGCTTTTGCTAAA
+
CCCFFFFFHHHHGJJHJJJJIJIJJJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJJJJJJJJJGIJJJJJIJJJHJIJGIHHHGFFFEFFEDDDDDDDDDDEDD
```

Sequencing QC

Information we need to check

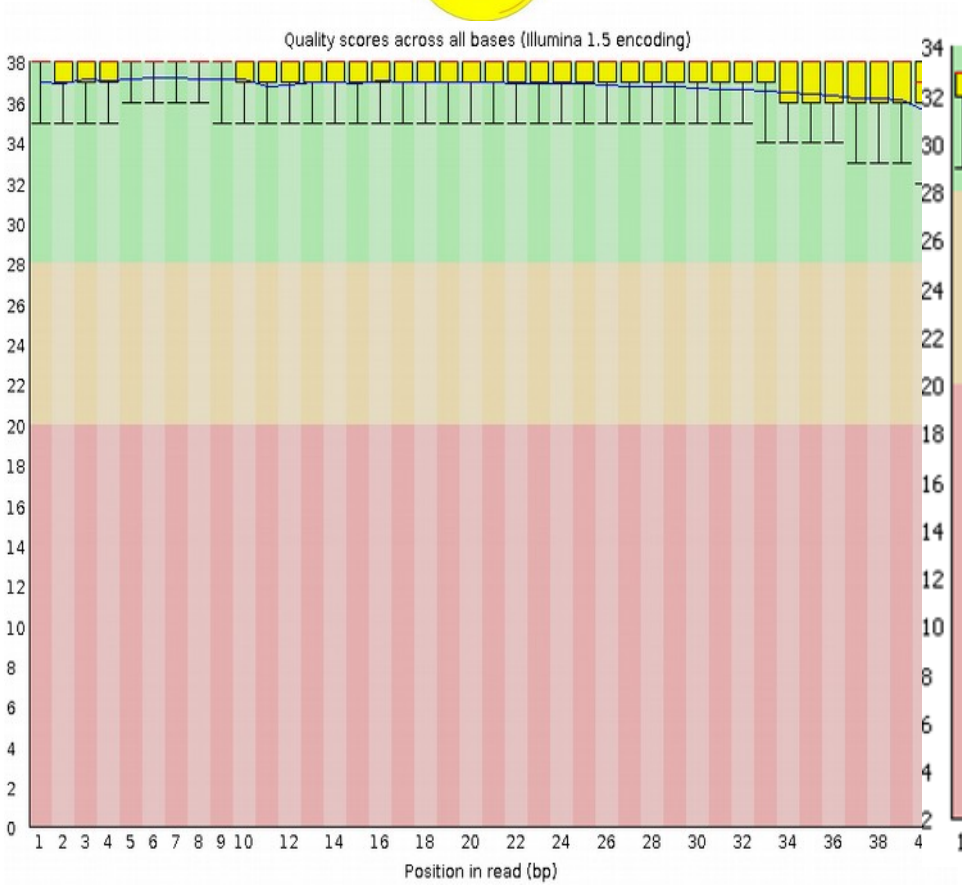
- Basic information(total reads, sequence length, etc.)
- Per base sequence quality
- Overrepresented sequences
- GC content
- Duplication level
- Etc.

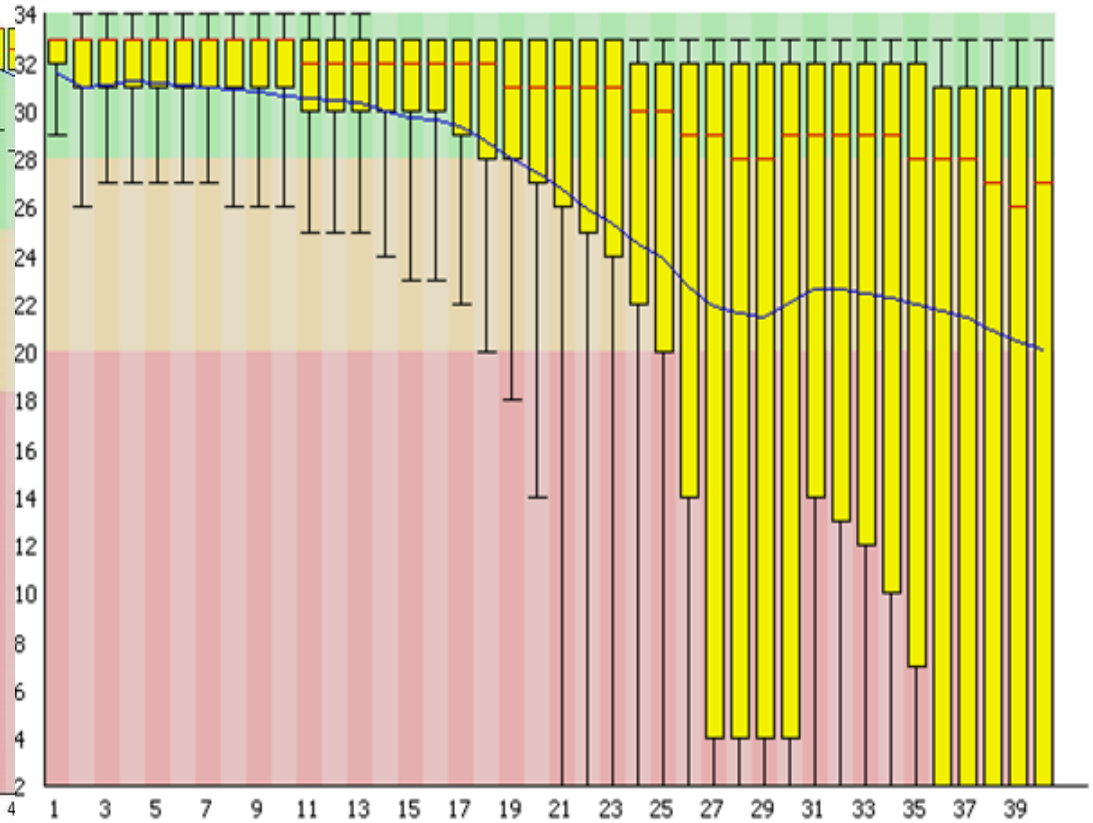
FastQC



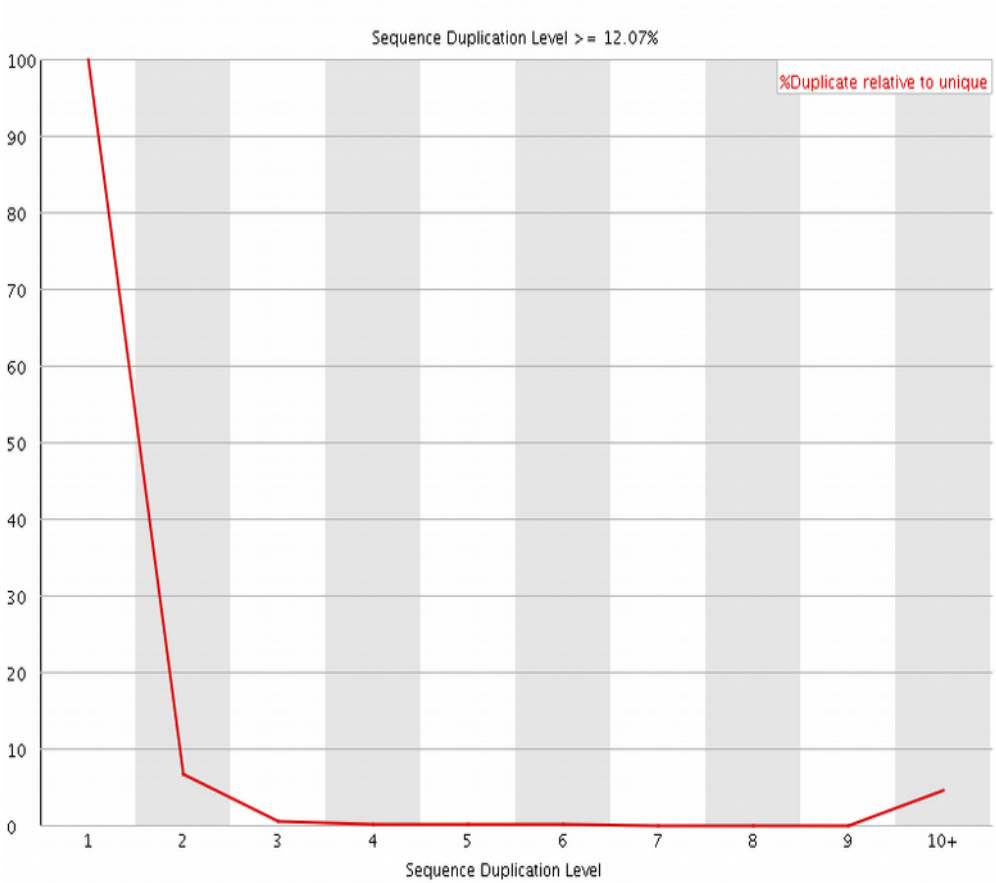
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

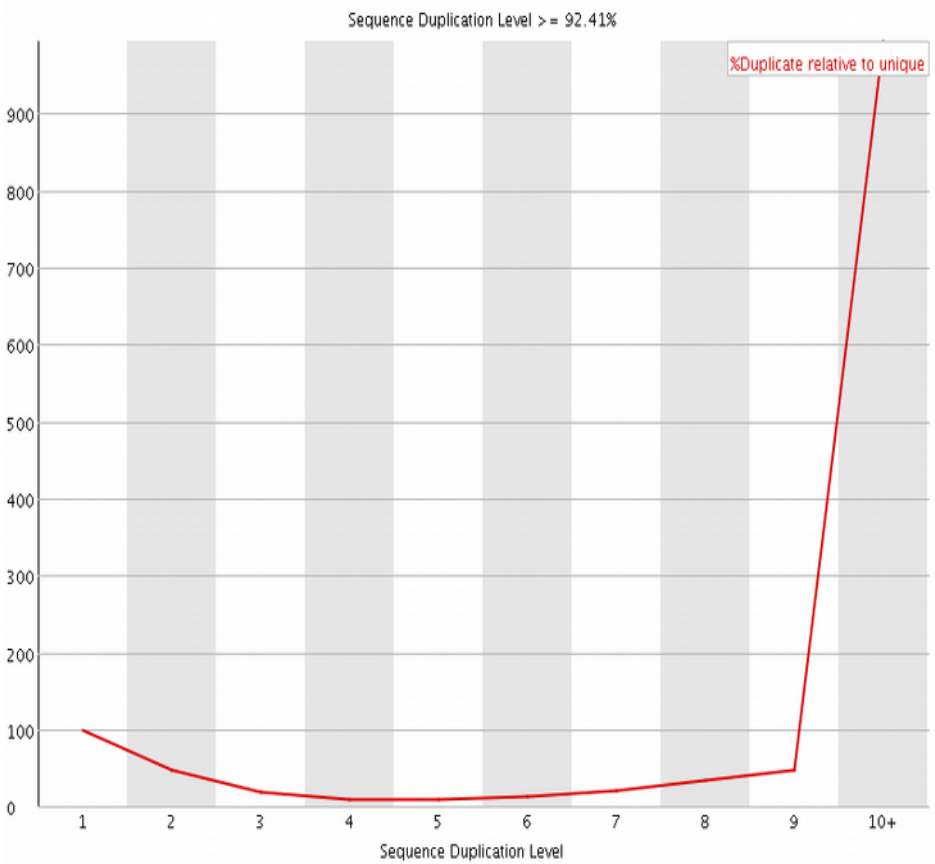
Per base sequence quality





Duplication level





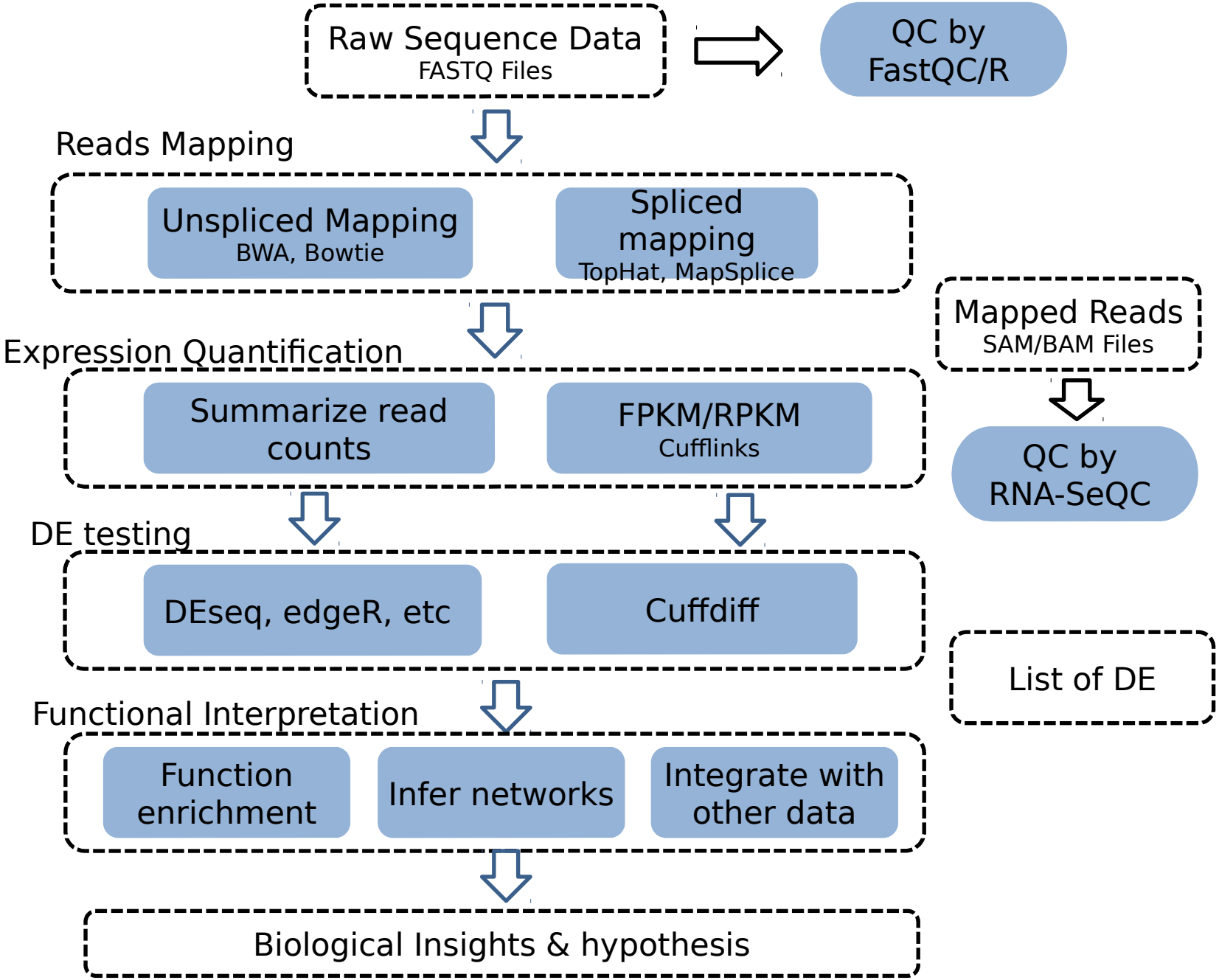
Overrepresented Sequences

Overrepresented sequences

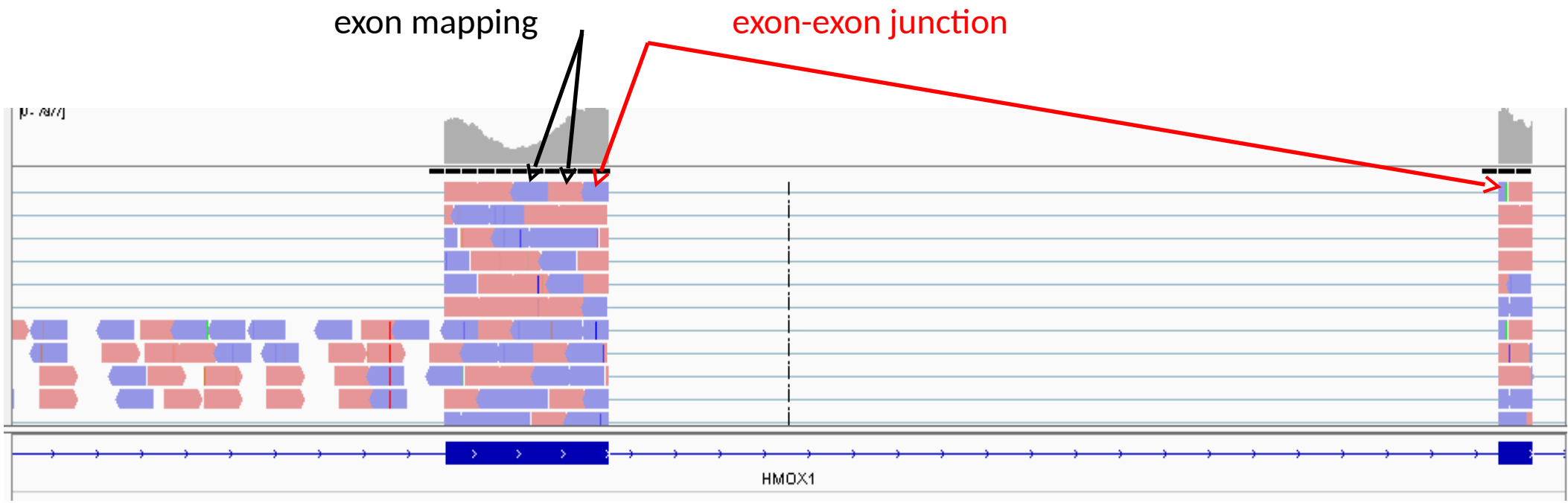
Sequence	Count	Percentage	Possible Source
GTGTCAGTCACTTCCAGCGGTCGTATGCCGTCTTCT	2667259	7.236020826756234	No Hit
TATCCCCGCCTGTCACGCGGGACGTGTCAGTCACTT	503193	1.907695950497944	No Hit
CTCGCTCCTCTCCTACTTGGATAACTCGTGTCAGTC	352107	0.9552329133566171	No Hit
TGTCAGTCACTTCCAGCGGTCGTATGCCGTCTTCTG	351690	0.9541016318857297	No Hit
CTCCTCTCCTACTTGGATAACTCGTGTCAGTCACTT	247800	0.6722579100380558	No Hit
CATCATATGGTGACCTCCCGCGTGTCAGTCACTTCC	192614	0.5225435233416872	No Hit
CATCAATATGGTGACCTCCCGCGTGTCAGTCACTTC	192513	0.5222695199158848	No Hit
CATCAATATGGTGACCTCCCGGAACGTGTCAGTCAC	191604	0.5198034890836628	No Hit
CATCAATATGGTGACCTCCCGCGTGTCAGTCACTTCC	163498	0.4435545753648186	No Hit
CATCATATGGTGACCTCCCGCGTGTCAGTCACTTCCA	158547	0.43012298169008734	No Hit
TATCCCCGCCTCACGCGGGACGTGTCAGTCACTTCC	131347	0.3563319600878471	No Hit
AAAACGTGTCAGTCACTTCCAGCGGTCGTATGCCGT	127345	0.34547491345357634	No Hit
CATGAGACTCTTAATCTCACGTGTCAGTCACTTCCA	109695	0.29759213656829914	No Hit

Adapter

From reads to differential expression



Read mapping



Unlike DNA-Seq, when mapping RNA-Seq reads back to reference genome, we need to pay attention to **exon-exon junction reads**

SAM/BAM format

Two section: header section, alignment section

Each alignment line has 11 mandatory fields. These fields always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The following table gives an overview of the mandatory fields in the SAM format:

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	
11	QUAL	String	`[!-~]+`	

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

http://samtools.sourceforge.net/SAM1.pdf

One example: SAM file

Read ID Flag pos MQ

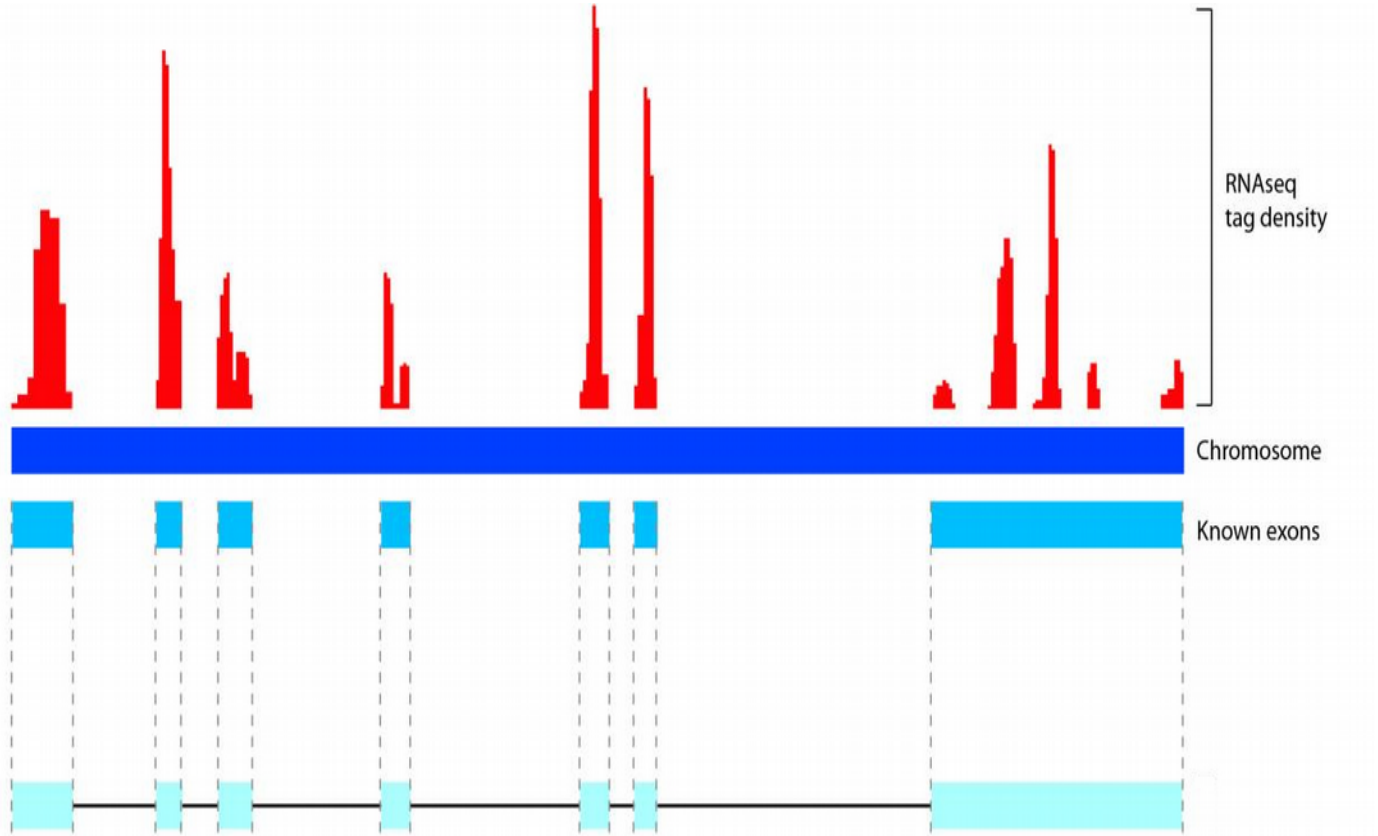
```
HWI-ST508_0109:6:1106:19590:4489#ATCACG 83      chr1    16230   255     81M296N19M      =       16179   -447    T
CAGTTGCACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTAGACTGGGAGATACAGCAGTGAAGCTGAAGGAGACGCGCTGCT     ########
#@D.BDGFGGGGEGGGDBEE@EFF?FECBADEEBEEECE@DC?DCB@EEE@EBEEE?B<=?FFEFFFF?FFD8FFEDGFDFFGGGGGDGBG     NM:i:2  XS:A:-  N
H:i:1
HWI-ST508_0109:7:1106:5833:71661#ATCACG 83      chr1    16234   255     77M296N23M      =       16184   -446    T
TGCCCACGCGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGACACAGCAGTGAAGCTGAAGGAGACGCGCTGCTGCTG     ########
#C?B?C8BFDEBEEEE4<9>7AECDE?7?>>3:?2?>9:AB5=9+<8D)DDD>DDC@@3=;?;=DD?DFDEFFFFFE<BDF<9:>24+83:     NM:i:2  XS:A:-  N
H:i:1
HWI-ST508_0109:8:2103:19403:137111#ATCACG       83      chr1    16234   255     100M    =       16155   -179    T
TGCACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGACACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTG     ########
#A:AABFGB;GGGGGGEDBACCCDE5>?<@>DE<?D?FCBFEEBDBFDFFFC>@>CDDADD>FDFFCECEEDGGFGEGGEGGGGGGGEGGF     NM:i:0  NH:i:1
HWI-ST508_0109:7:1204:3497:194785#ATCACG        163     chr1    16237   255     100M    =       16357   220     C
ACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTAGACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAG     DD@D=DEEE
E@GGEEGGFDF<GD@CEEEEEG=FFGFBFBFHHGHDEGGF@EEEBD>>=B:DF=@FEGDGBD/DDD@DD=CBFFGFDC@/>BCDC######     NM:i:2  NH:i:1
HWI-ST508_0109:6:1104:12243:43788#ATCACG        355     chr1    16241   3       100M    =       16337   196     C
ACGAGCCAGCAGAGGCGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAGTTTG     HHHHFHHHH
HCHHHHHHHHGHGHEHFHCHHHHHHHHHHHHHHFEHHHEHHHHHAFE?FCFFFFHEHDFFEEFEEGEGFGHHH?GDCFGGHHHF?FCGGC     NM:i:2  NH:i:2  C
C:Z:chr15 CP:i:102514823        HI:i:0
```

83= 1+2+16+64

read paired; read mapped in proper pair; read reverse strand; first in pair

Expression quantification

- Count data
  - Summarized mapped reads to CDS, gene or exon level



Count-based methods (R packages)

1. **DESeq** -- based on negative binomial distribution
2. **edgeR** -- use an overdispersed Poisson model
3. **baySeq** -- use an empirical Bayes approach
4. **TSPM** -- use a two-stage poisson model

Anders and Huber *Genome Biology* 2010, 11:R106
http://genomebiology.com/2010/11/10/R106

Genome Biology

METHOD Open Access

Differential expression analysis for sequence count data

Simon Anders*, Wolfgang Huber

Hardcastle and Kelly *BMC Bioinformatics* 2010, 11:422
http://www.biomedcentral.com/1471-2105/11/422

BMC Bioinformatics

RESEARCH ARTICLE Open Access

baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data

Thomas J Hardcastle*, Krystyna A Kelly

BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 1 2010, pages 139–140
doi:10.1093/bioinformatics/btp616

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]
[1]Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and [2]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Received on March 29, 2009; revised on October 19, 2009; accepted on October 23, 2009
Advance Access publication November 11, 2009

Statistical Applications in Genetics and Molecular Biology

lume 10, *Issue* 1 2011 *Article* 26

A Two-Stage Poisson Model for Testing RNA-Seq Data

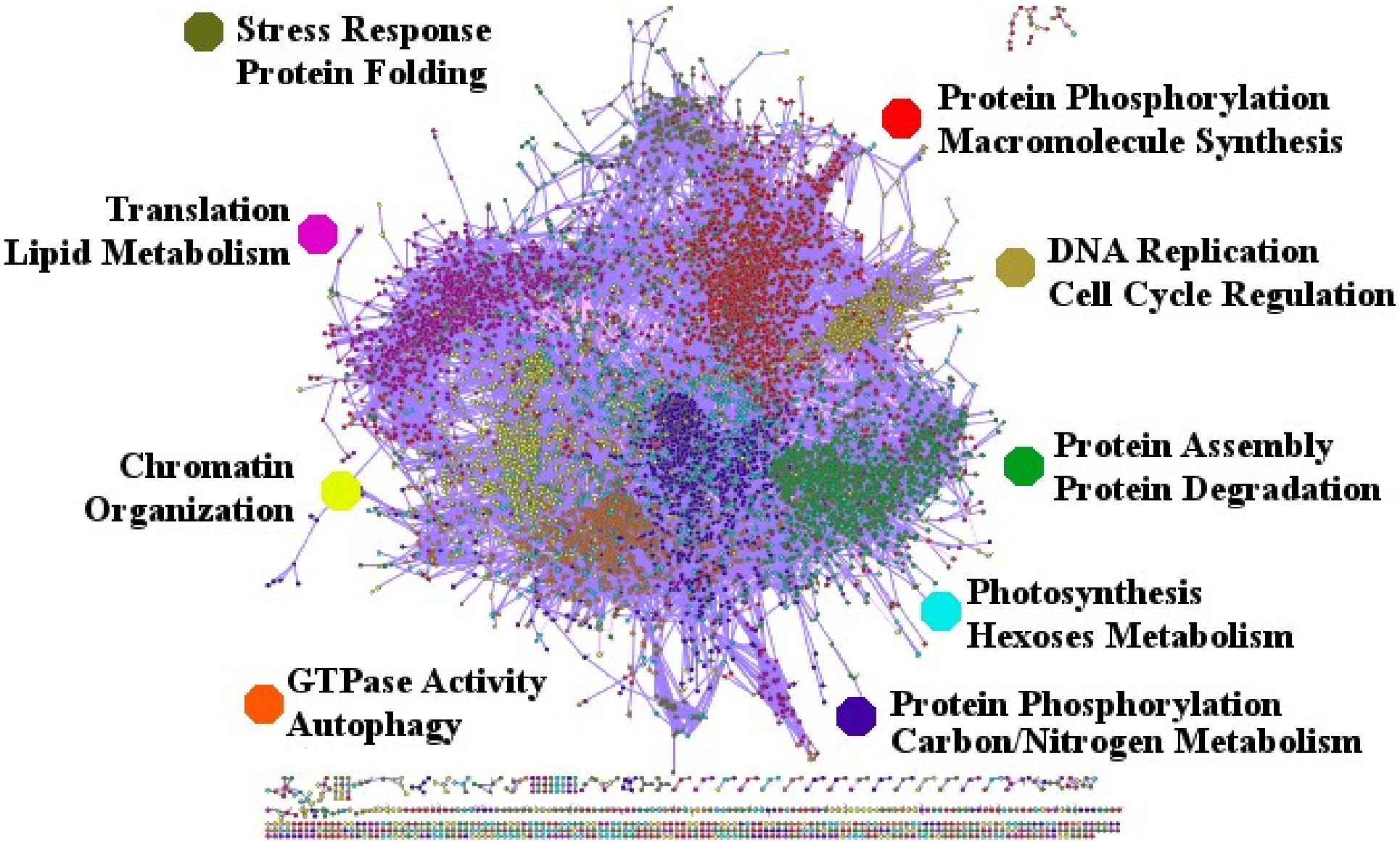
aul L. Auer, *Fred Hutchinson Cancer Research Center*
Rebecca W. Doerge, *Purdue University*

RPKM/FPKM-based methods

- Cufflinks & Cuffdiff
- Other differential analysis methods for microarray data
  - t-test, limma etc.

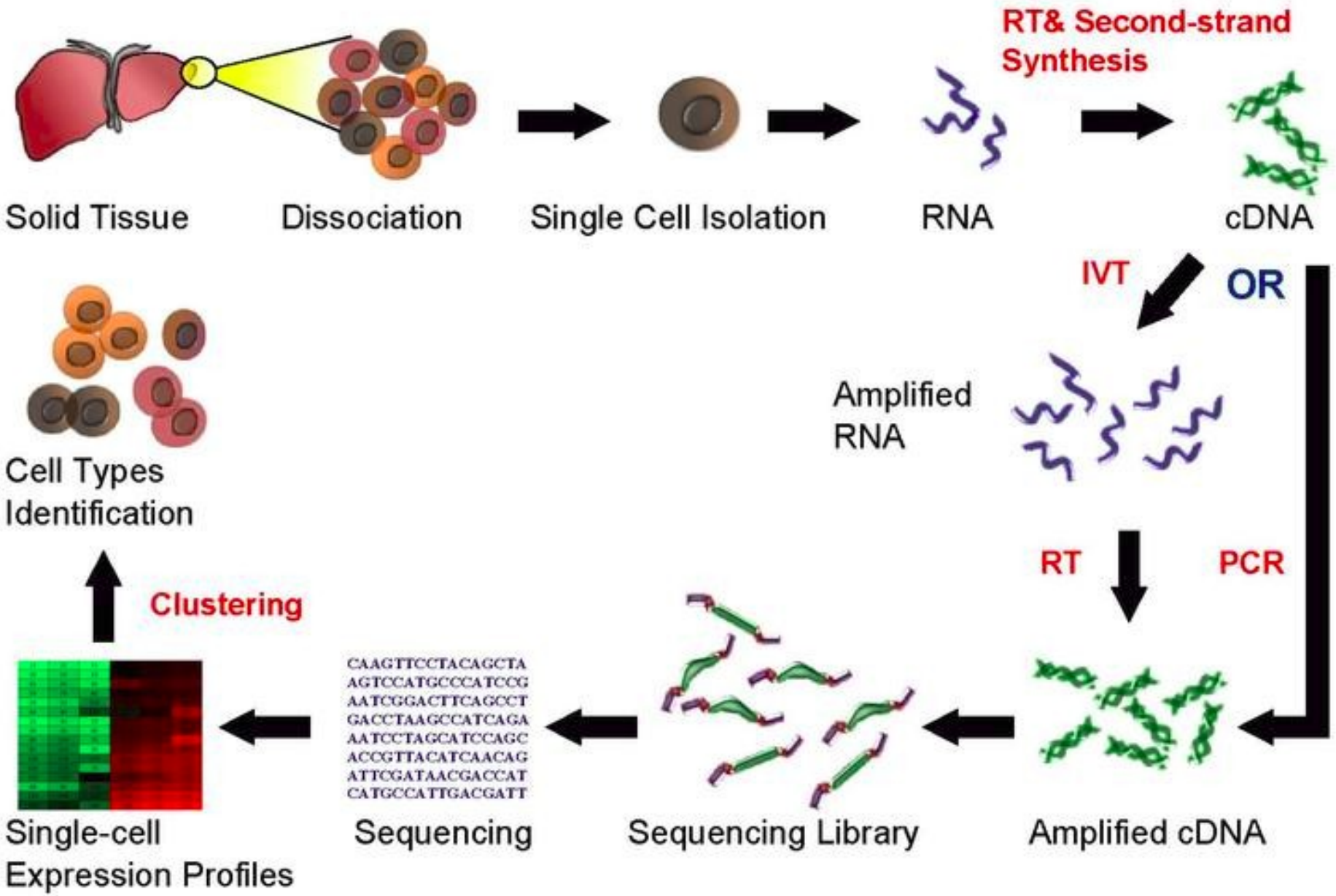
Chlamydomonas reinhardtii Gene Co-expression Network

link

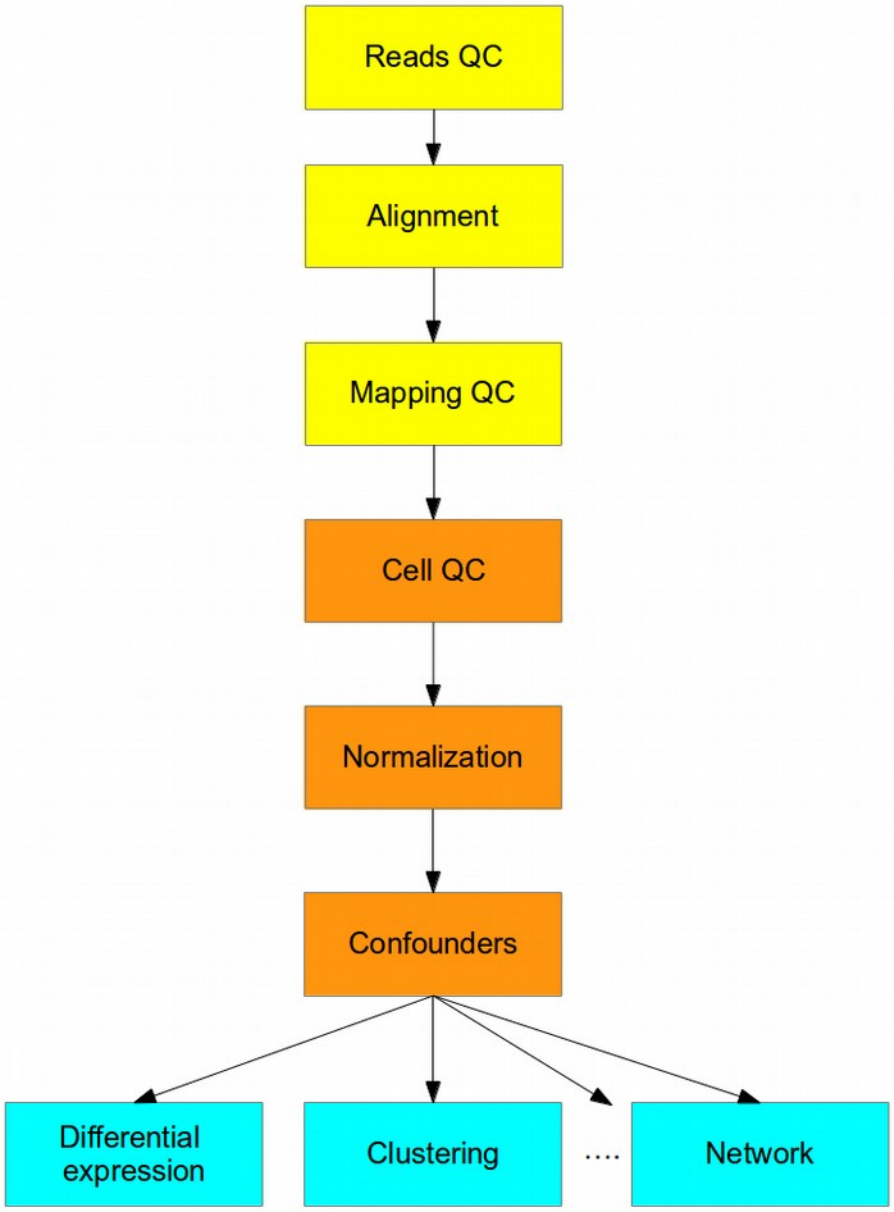


Single-cell RNA-Seq

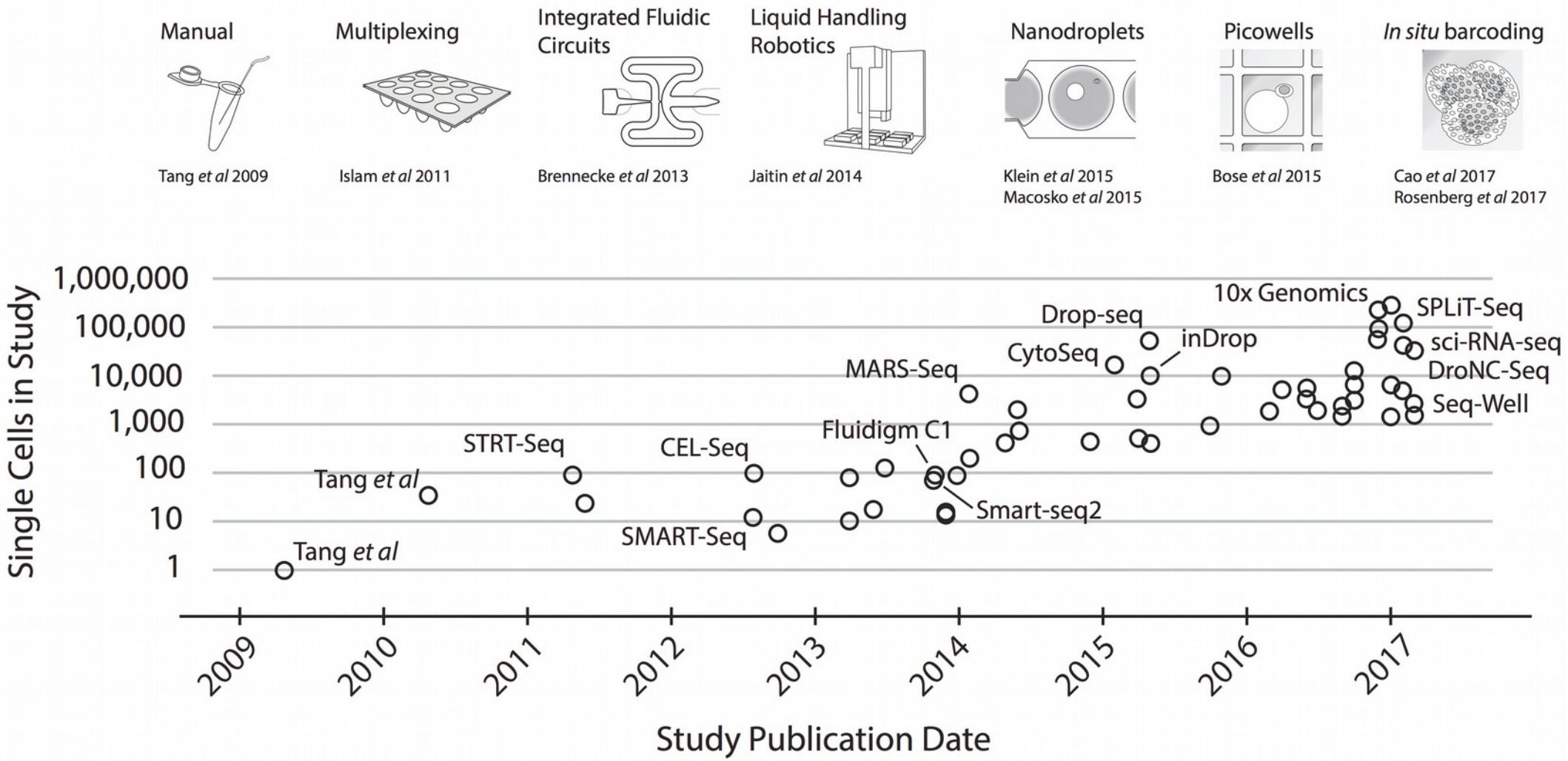
Single Cell RNA Sequencing Workflow



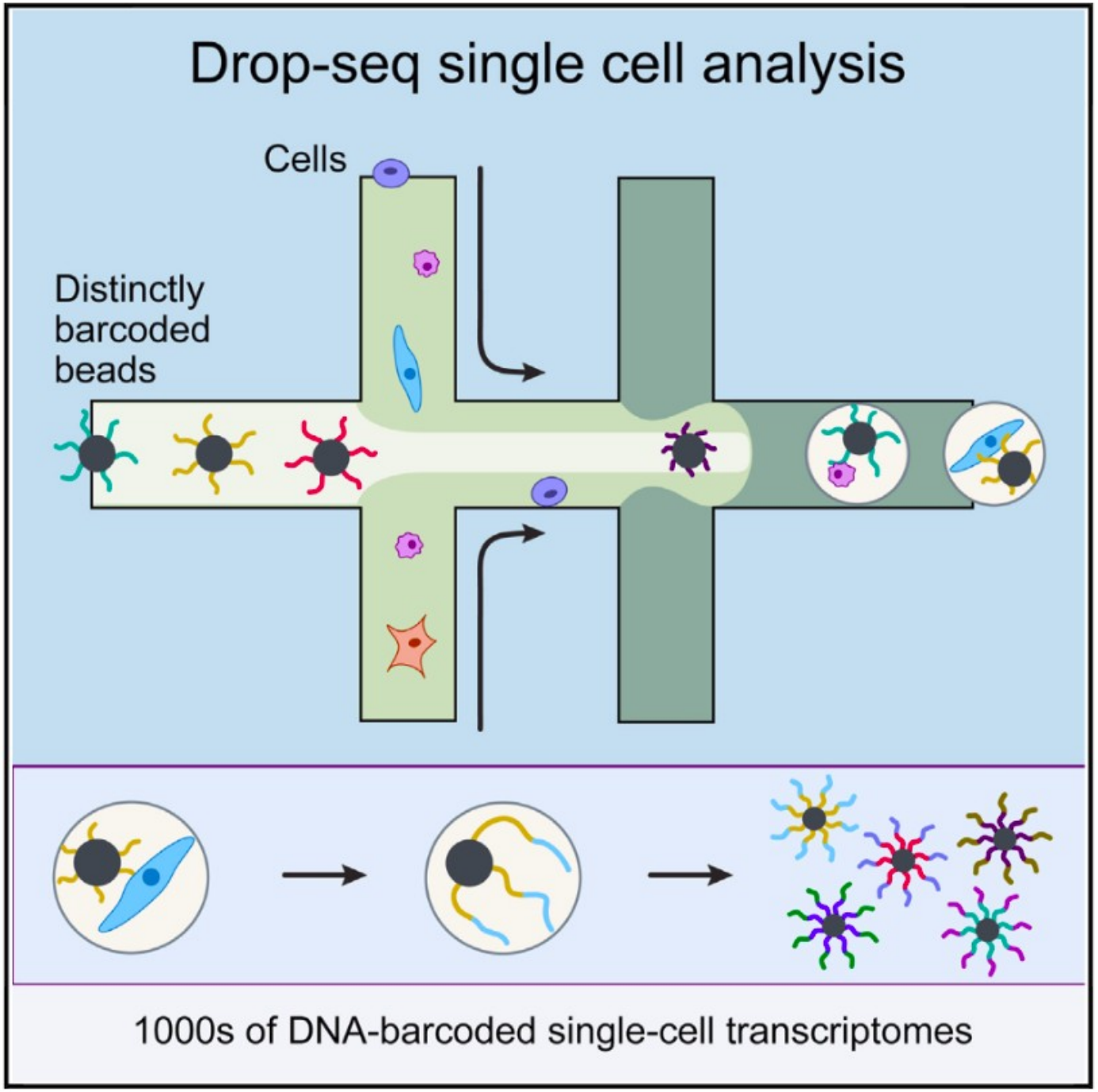
scRNA-seq analysis



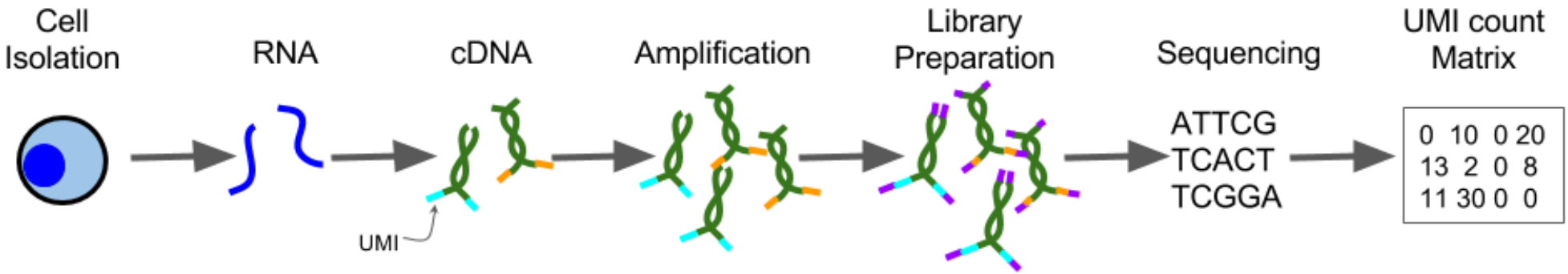
Moore's law in single cell transcriptomics

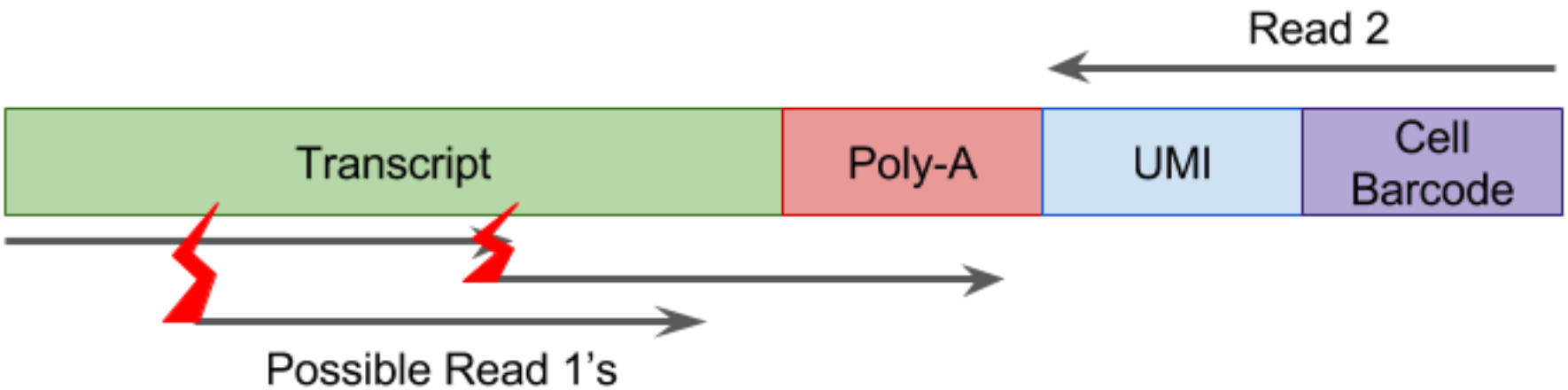


Schematic overview of the drop-seq method

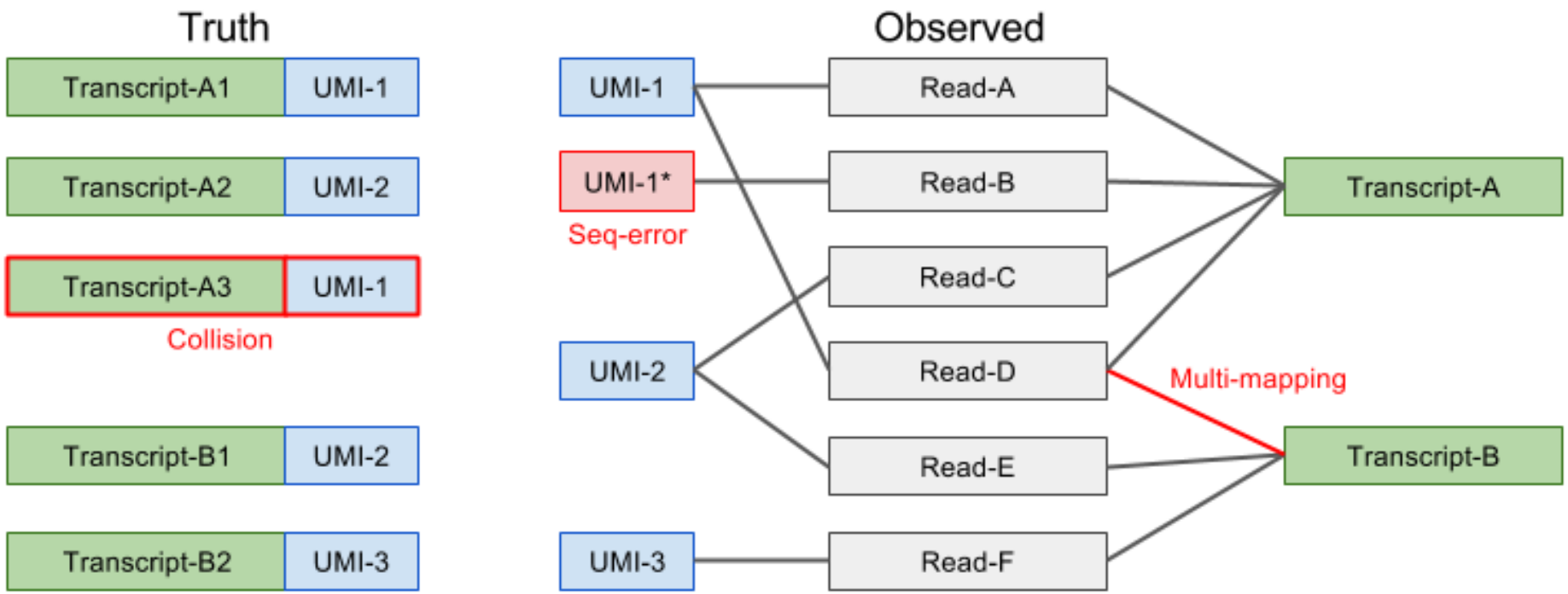


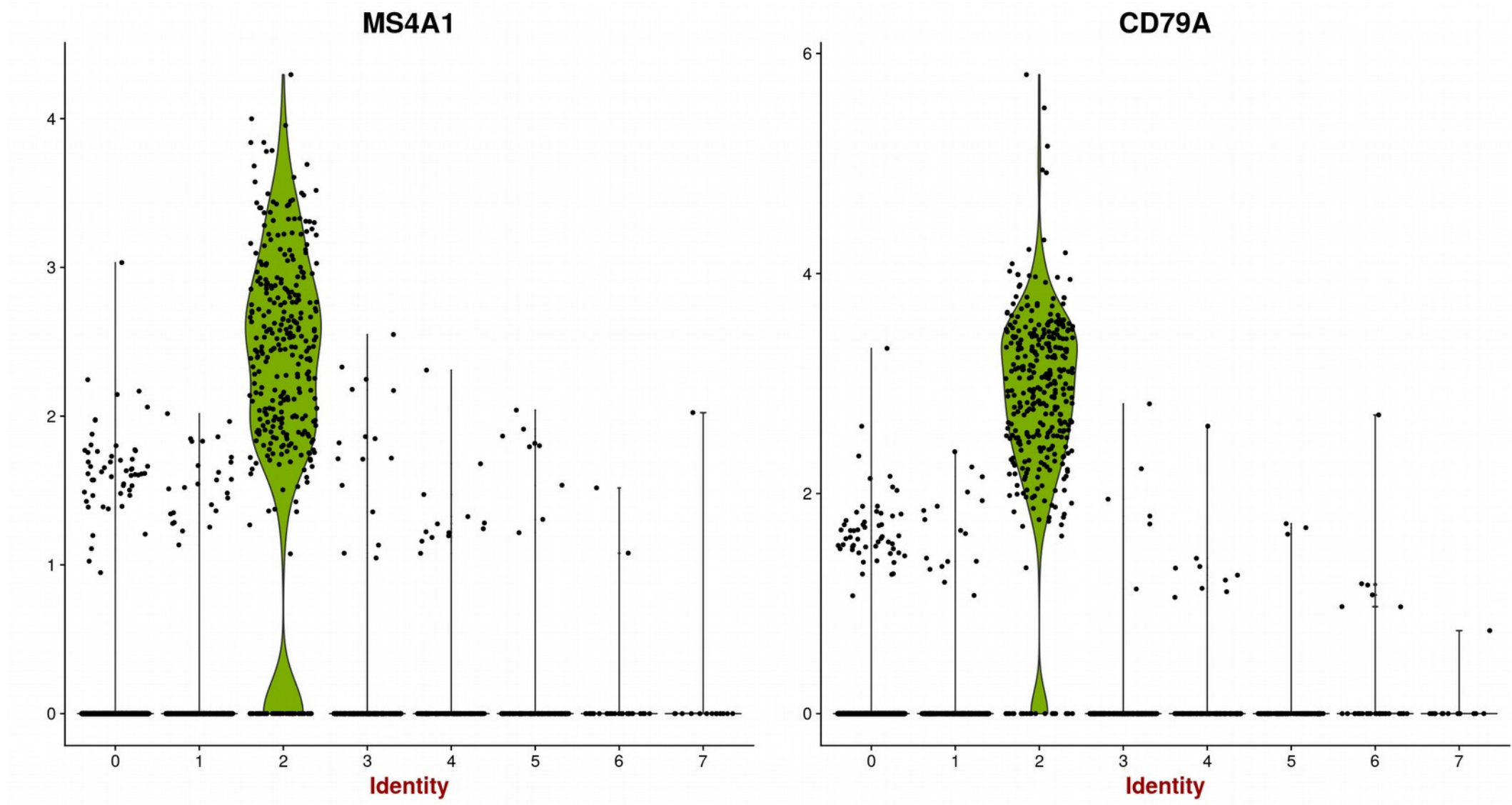
UMI sequencing protocol



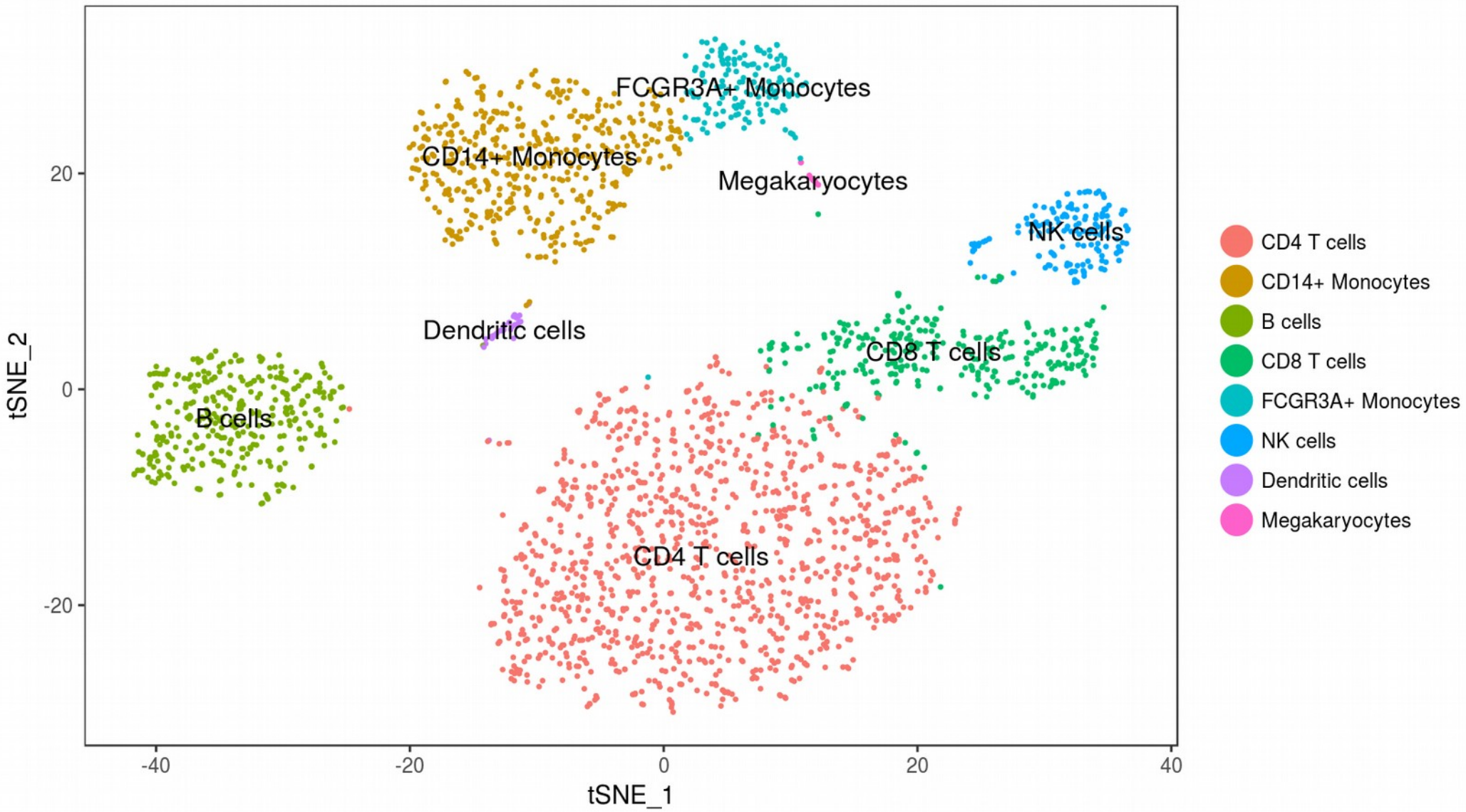


Potential Errors in UMIs

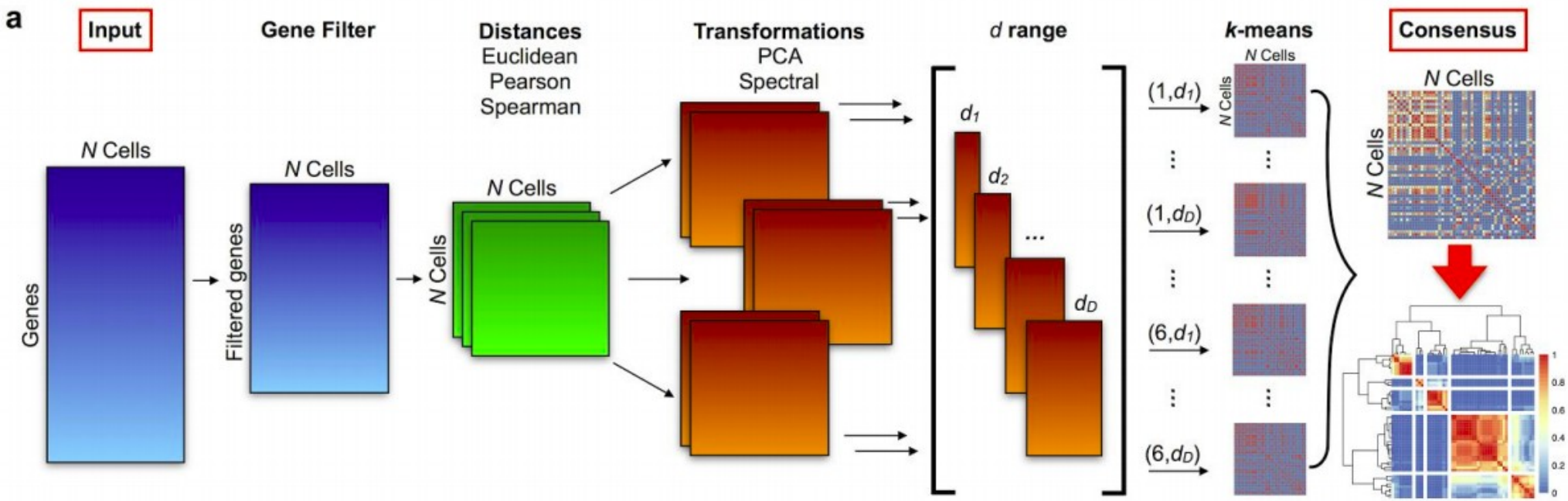



MS4A1
CD79A
Identity
Identity
0
1
2
3
4
5
6
7

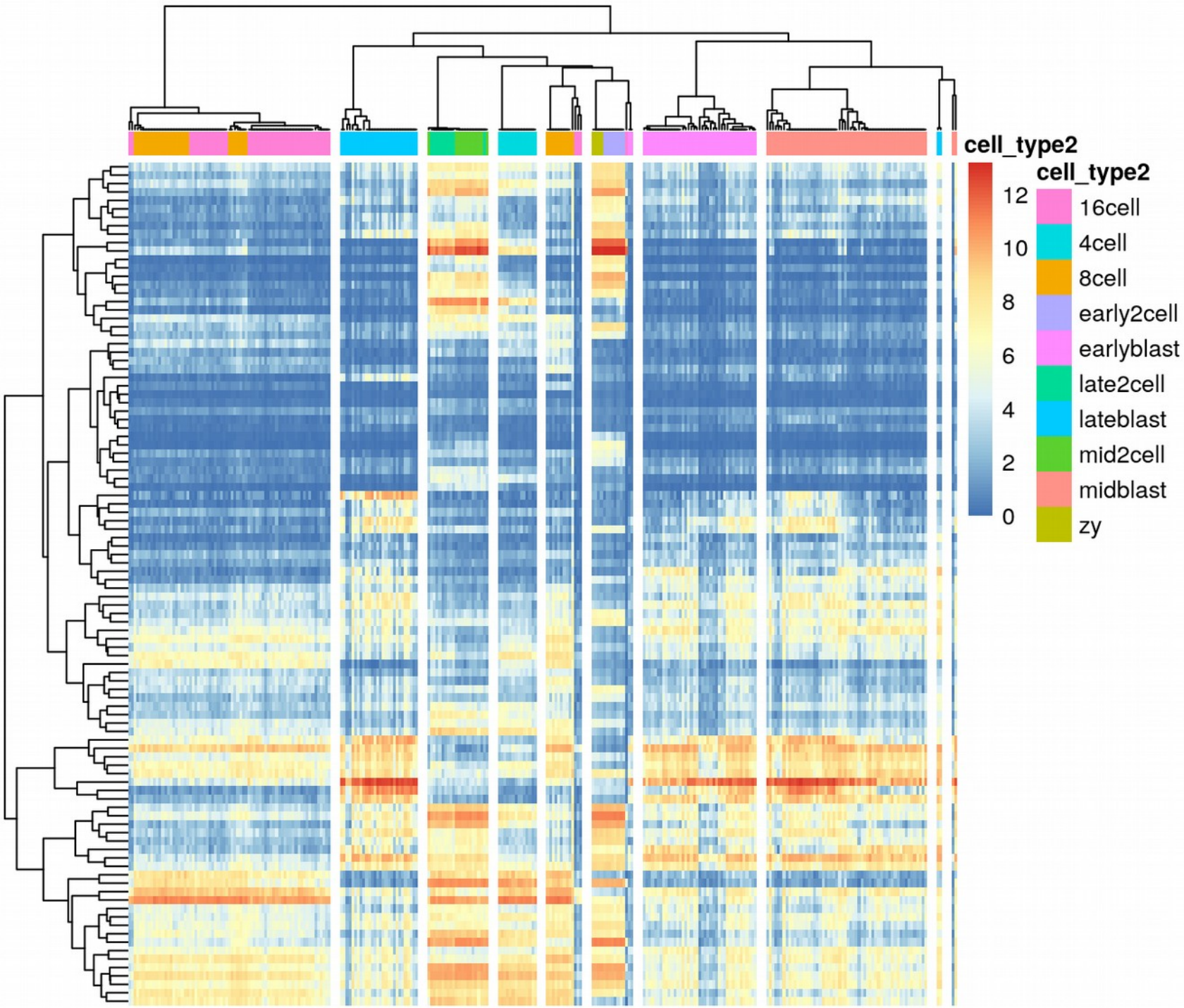
tSNE plots


FCGR3A+ Monocytes
CD14+ Monocytes
Megakaryocytes
NK cells
Dendritic cells
CD8 T cells
B cells
CD4 T cells
20
0
-20
-40
-20
0
20
40
tSNE_2
tSNE_1
CD4 T cells
CD14+ Monocytes
B cells
CD8 T cells
FCGR3A+ Monocytes
NK cells
Dendritic cells
Megakaryocytes

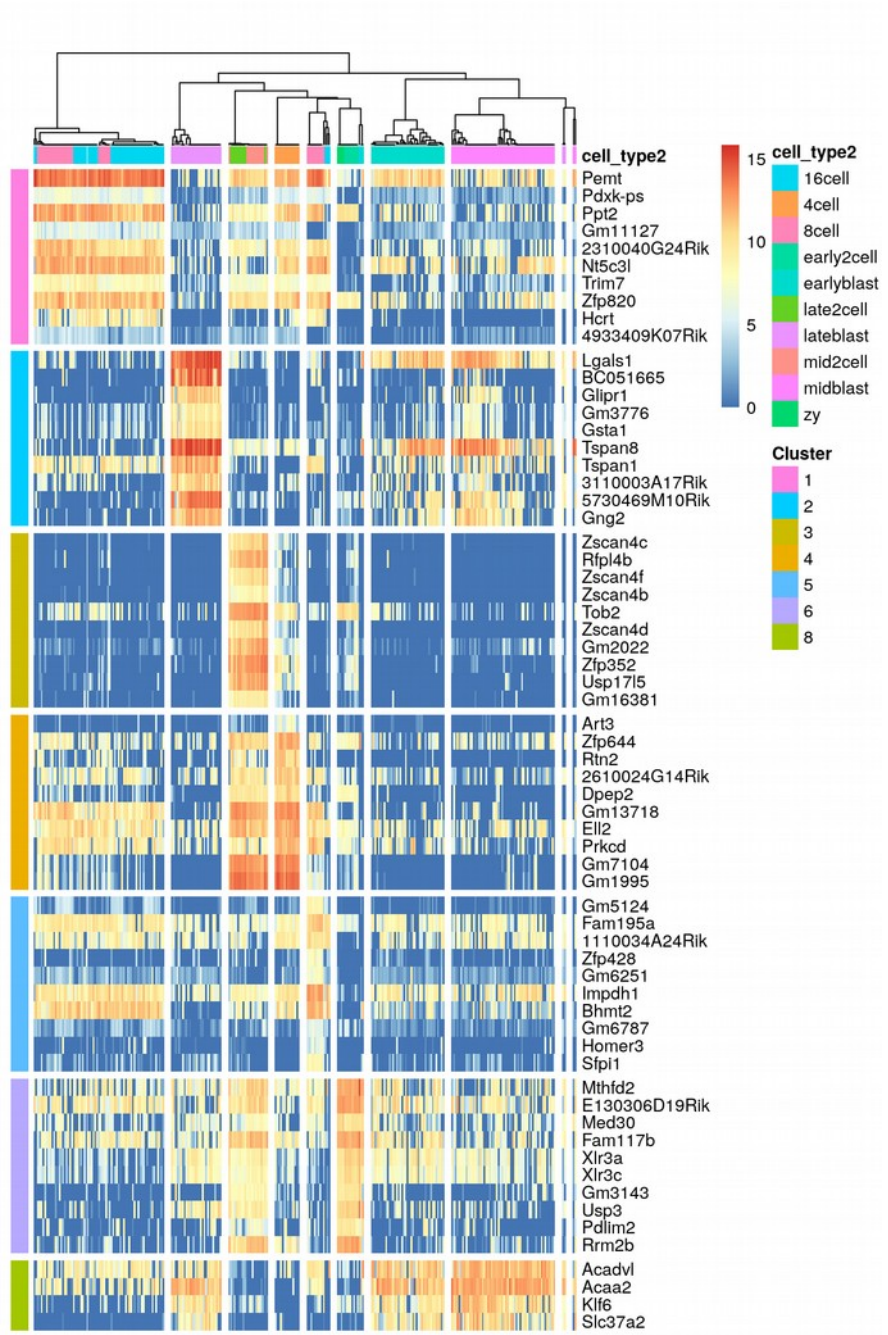
SC3 pipeline



Heatmap of the expression matrix



Identified marker genes



Спасибо за внимание!