Module 2 : RNAseq bioinformatic
from sequence to expression level
smallRNA & isoform (eukaryote)

Claire Toffano-Nioche

I2BC - Institut de Biologie Intégrative de la Cellule, Orsay

RNAseq analyses

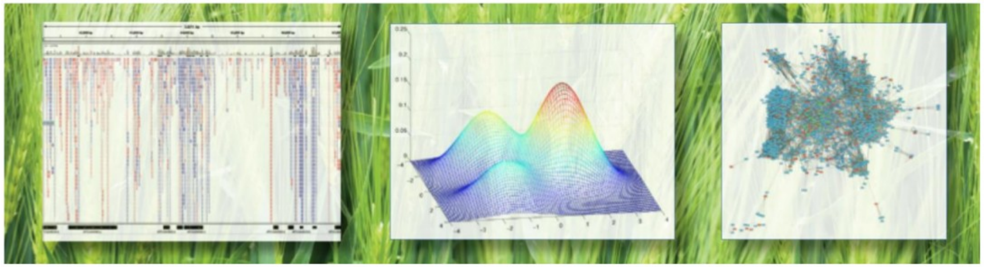
RNAseq: From sequence data (reads) to expression level (count)

Classical analyses of RNA-Seq:

- Check quality, Trimming
- Mapping / counts
- Assembly

Others usages of RNA-Seq:

- smallRNA study
- expression at isoform level



RNAseq analyses

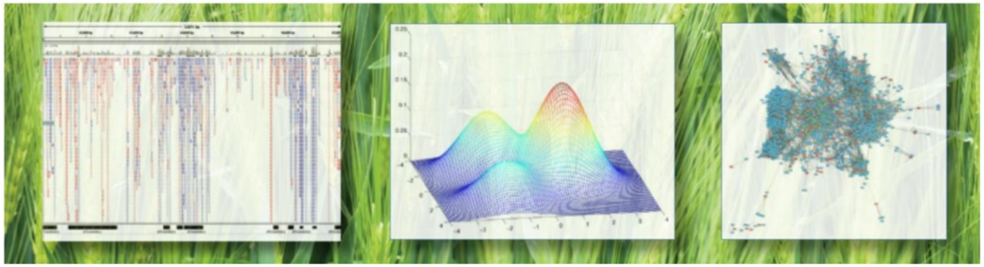
RNAseq: From sequence data (reads) to expression level (count)

Classical analyses of RNA-Seq:

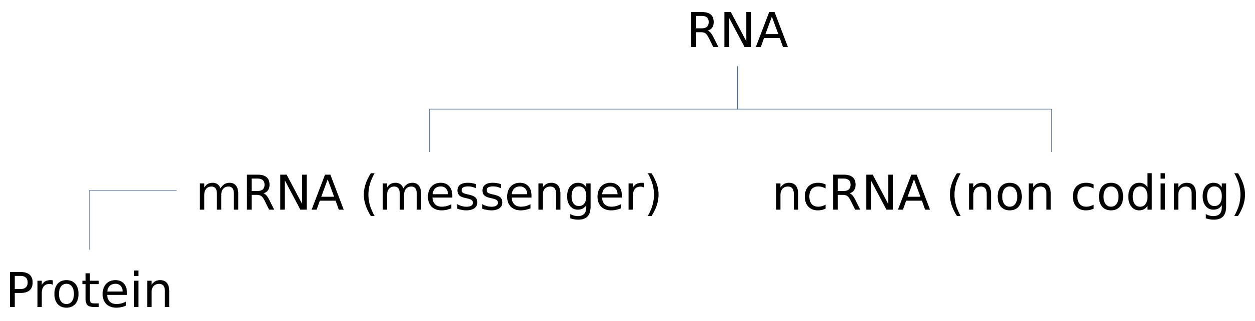
- Check quality, Trimming
- Mapping / counts
- Assembly

Others usages of RNA-Seq:

- **smallRNA study**
- expression at isoform level



RNA world



RNA

mRNA (messenger)

ncRNA (non coding)

Protein

Not predicted by gene prediction tools

- No specific signal (start, stop, splicing sites...)
- Multiple location (intergenic, intronic, coding, antisense)
- Variable size
- No strong sequence conservation in general

Function related to structure : ncRNA of the same family have a conserved structure

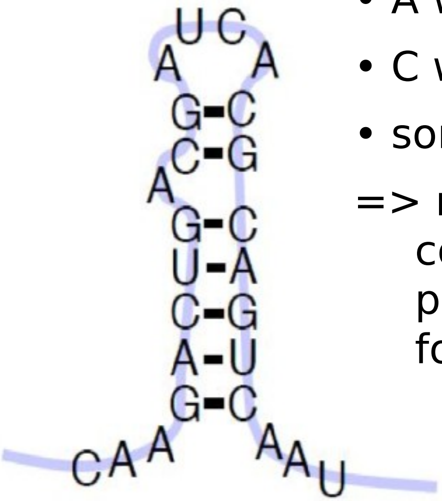
RNA folding

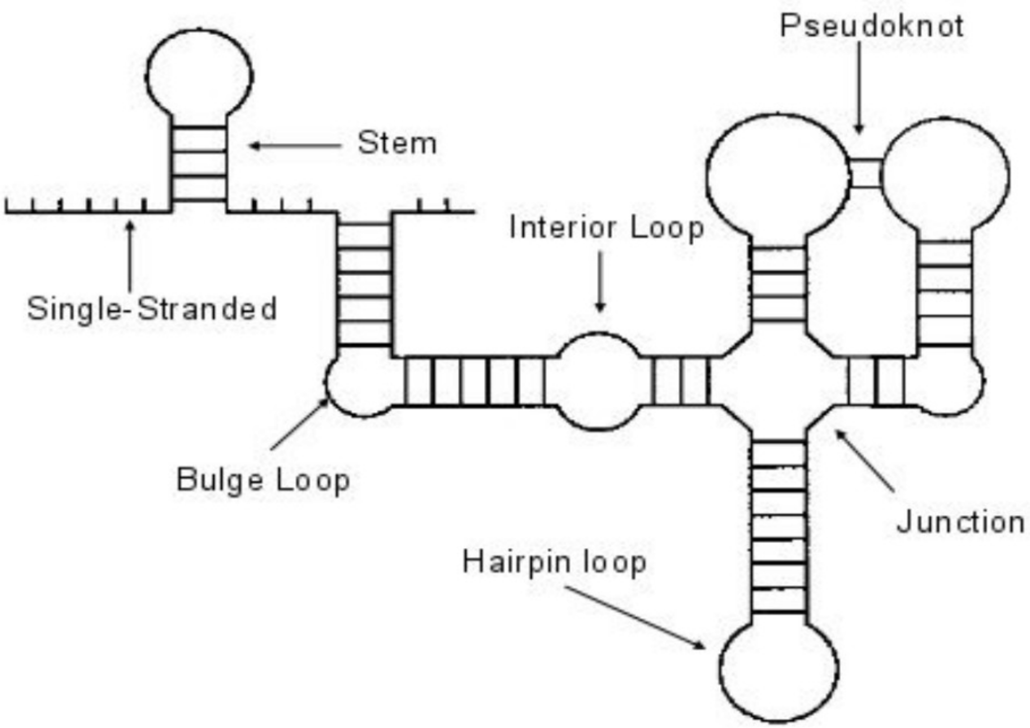
Folding = Secondary structure

RNA folds on itself by base pairing :

- A with U
- C with G
- sometimes G with U

=> more combinatorial possibilities of folding than DNA

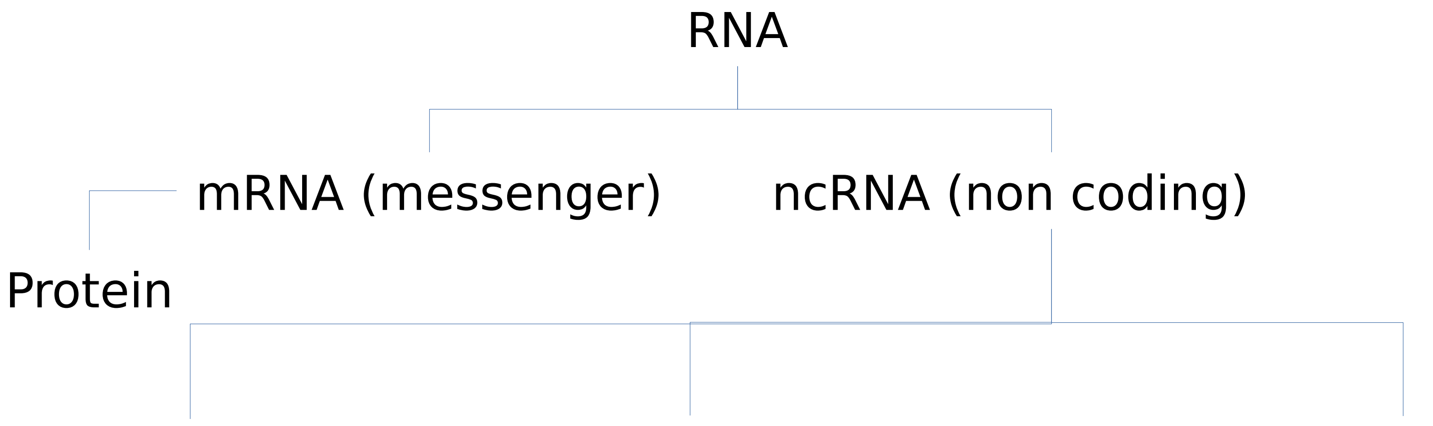




RNA folds:

- on itself
- with another: RNA duplex

RNA world



RNA

mRNA (messenger)

ncRNA (non coding)

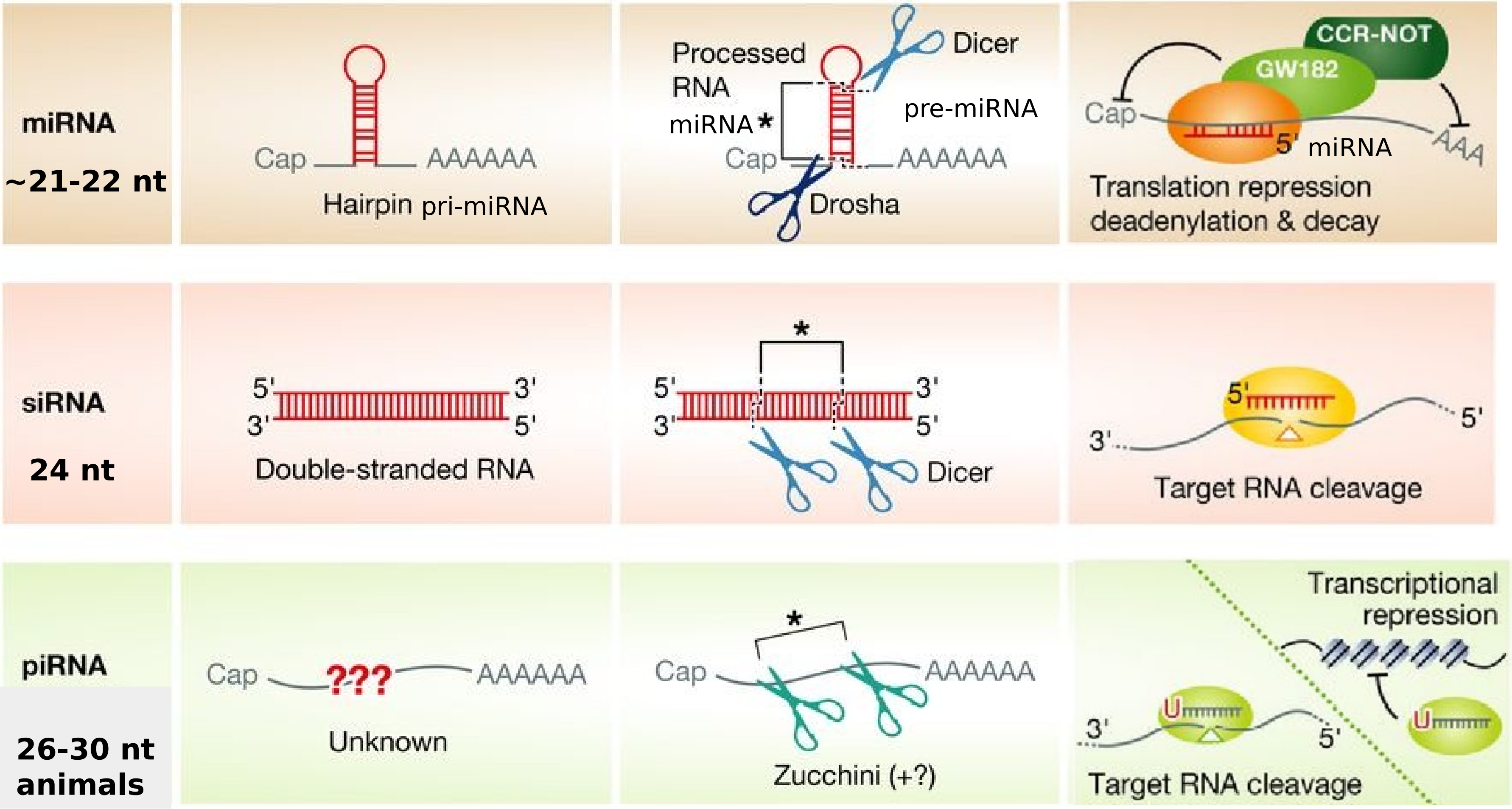
Protein

Ribozyme RNA (catalytic)
- Hairpin
- Hammerhead
- ...

Well-knowed RNA (housekeeping)
- tRNA
- rRNA
- RNaseP RNA
- telomerase RNA (replication)
- snRNA (maturation-splicing)
- snoRNA, gRNA (modification, editing)
- ...

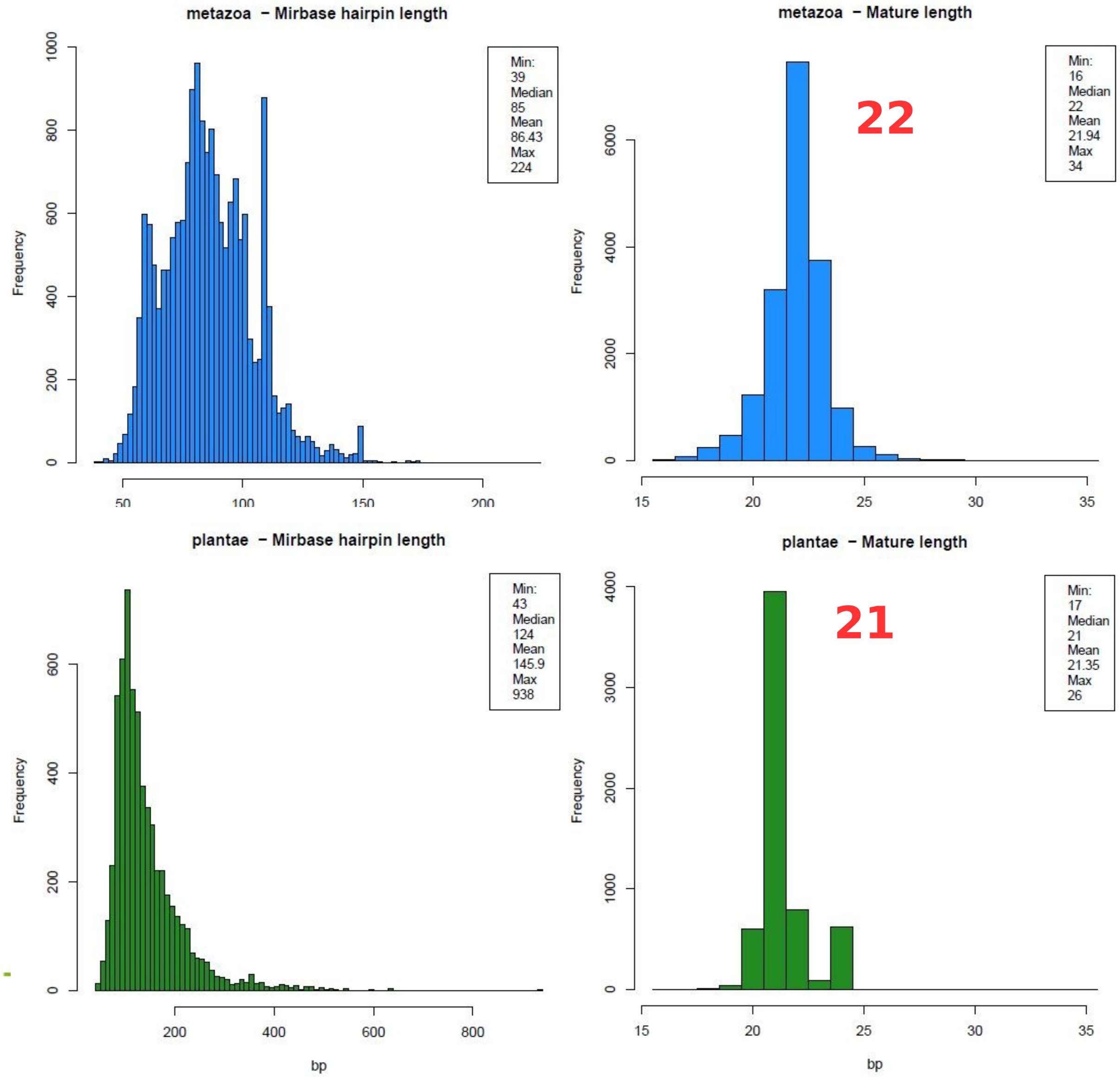
Regulatory RNA (env. adaptation)
- lncRNA (eukaryotes)
- smallRNA (adaptative response in bacteria)
- miRNA (development...)
- siRNA (defense)
- piRNA (epigenetic and post-transcriptional gene silencing in spermatogenesis)
- ...

Eukaryotic regRNA



Hirose T, Mishima Y, Tomari Y. (2014): Elements and machinery of non-coding RNAs: toward their taxonomy. EMBO Rep. 2014 May;15(5):489-507

Size maters



Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008 Jan;36(Database issue):D154-8

smallRNA & NGS

Classical RNAseq not suited for smallRNA (protocol and size)

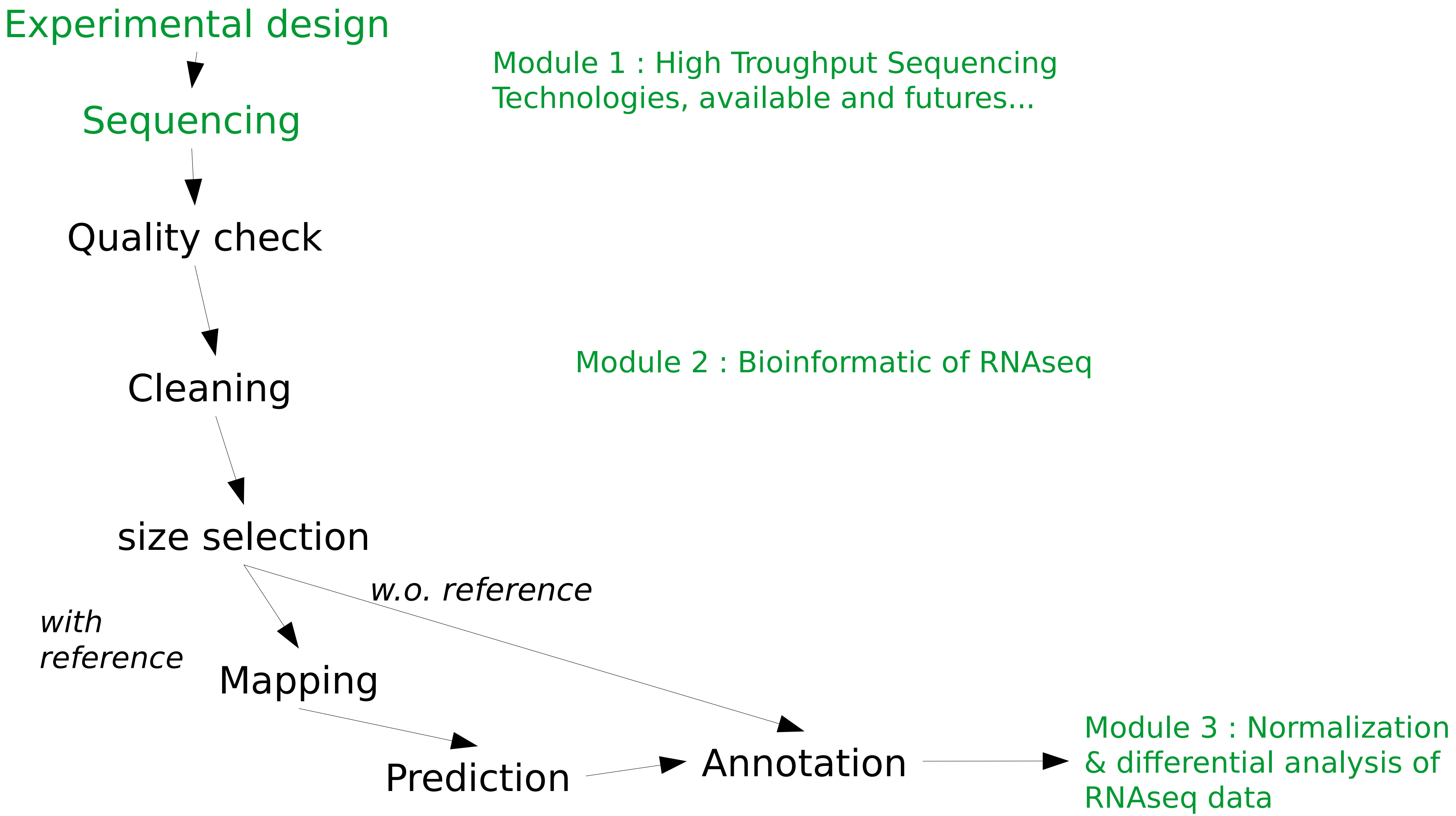
smallRNAs lack a common sequence (e.g. a poly(A) tail) that can be used for selective enrichment or as a universal primer-binding site for reverse transcription

Strategies to enrich RNA sample in smallRNA (*):

- smallRNA cloning, deep sequencing : significant biases are introduced during small RNA cDNA library preparation (often more than 3 orders of magnitude for individual miRNAs ; one major source: RNA ligation)
- RNA Immuno-Precipitation (RIP-seq)
- Total RNA extraction + size selection

(*) Module 1 : High Troughput Sequencing Technologies, available and futures...

smallRNAseq pipeline



Experimental design

Module 1 : High Troughput Sequencing Technologies, available and futures...

Sequencing

Quality check

Module 2 : Bioinformatic of RNAseq

Cleaning

size selection

w.o. reference

with reference

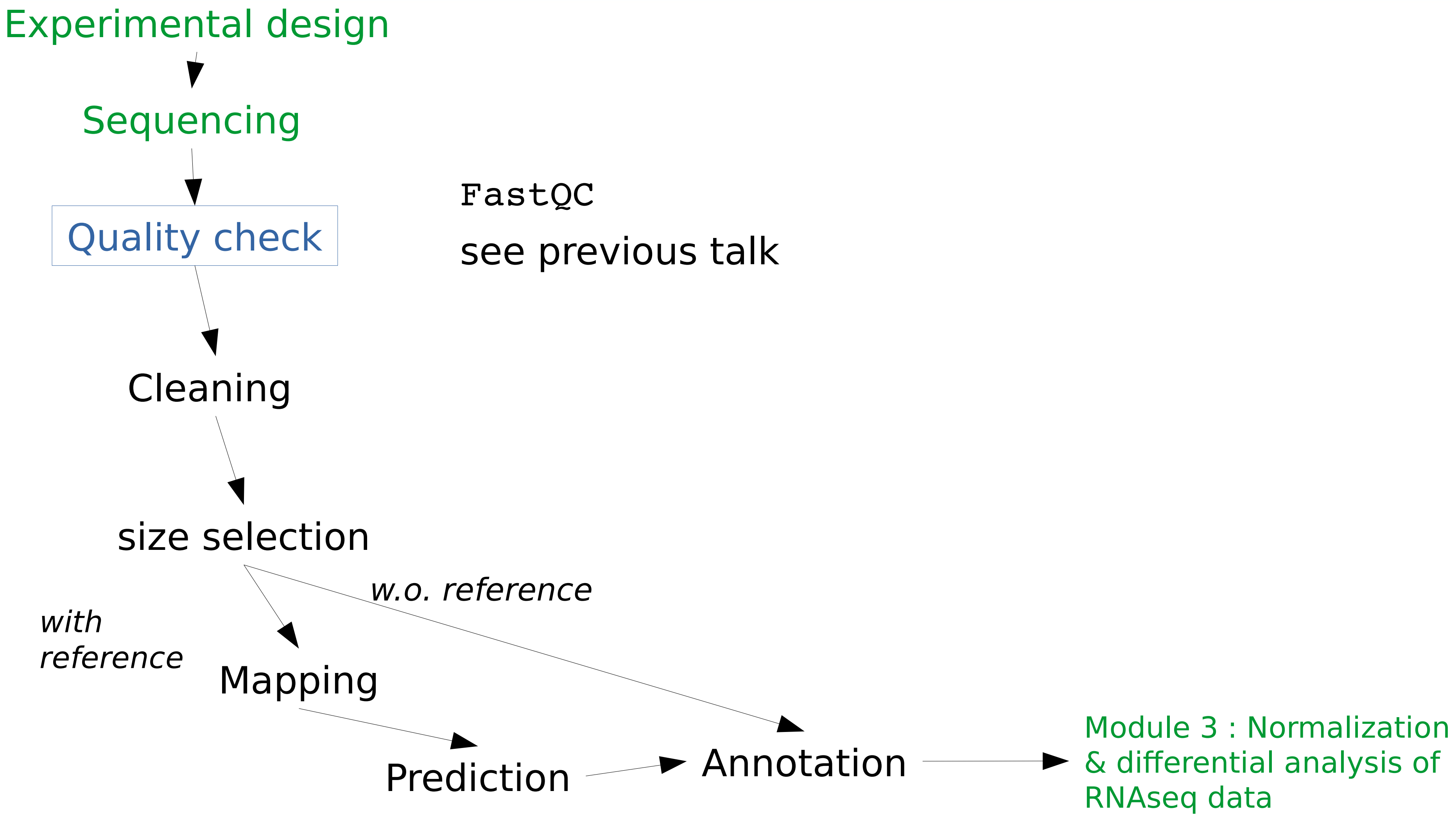
Mapping

Prediction

Annotation

Module 3 : Normalization & differential analysis of RNAseq data

smallRNAseq pipeline



Experimental design

Sequencing

Quality check

FastQC

see previous talk

Cleaning

size selection

with reference

w.o. reference

Mapping

Prediction

Annotation

Module 3 : Normalization & differential analysis of RNAseq data

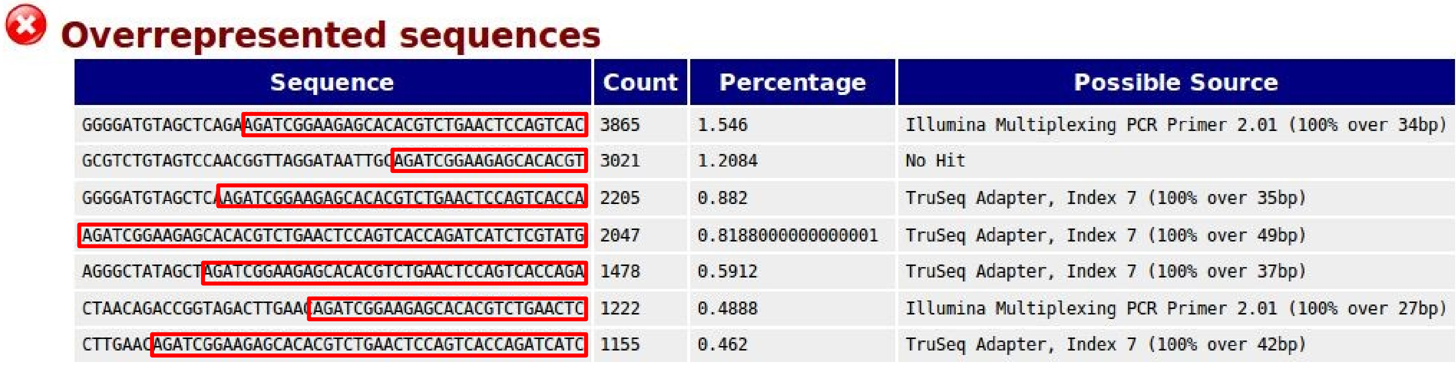
smallRNAseq pipeline

Experimental design

Sequencing

Quality check

Read size > regRNA size
PCRprimer sequences & adapter are included in read



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGGGATGTAGCTCAGAAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	3865	1.546	Illumina Multiplexing PCR Primer 2.01 (100% over 34bp)
GCGTCTGTAGTCCAACGGTTAGGATAATTGCAGATCGGAAGAGCACACGT	3021	1.2084	No Hit
GGGGATGTAGCTCAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCA	2205	0.882	TruSeq Adapter, Index 7 (100% over 35bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATG	2047	0.8188000000000001	TruSeq Adapter, Index 7 (100% over 49bp)
AGGGCTATAGCTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGA	1478	0.5912	TruSeq Adapter, Index 7 (100% over 37bp)
CTAACAGACCGGTAGACTTGAACAGATCGGAAGAGCACACGTCTGAACTC	1222	0.4888	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
CTTGAACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATC	1155	0.462	TruSeq Adapter, Index 7 (100% over 42bp)

Cleaning

=> trimming step : remove adapter in 3’, no quality trimming, remove rRNA, tRNA

`trimgalore, cutadapt, trimmomatic, …`

size selection

w.o. reference

with reference

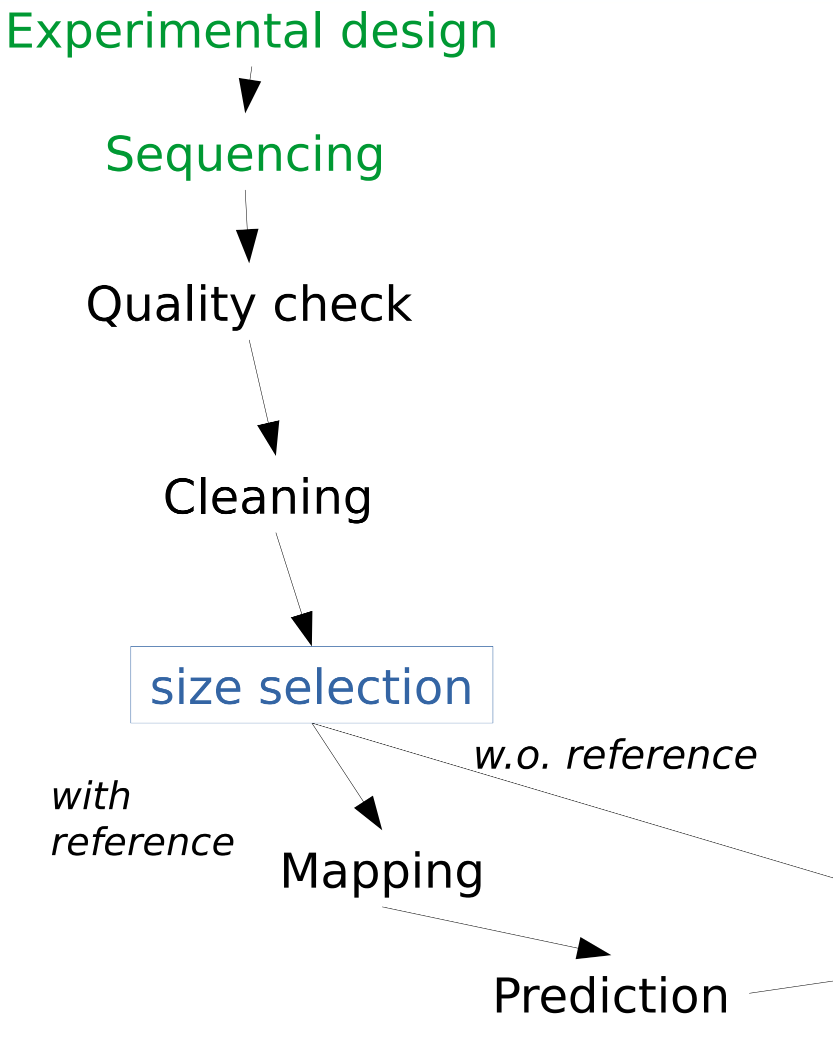
Mapping

Prediction

Annotation

Module 3 : Normalization & differential analysis of RNAseq data

smallRNAseq pipeline



Experimental design

Sequencing

Quality check

Cleaning

size selection

with reference

w.o. reference

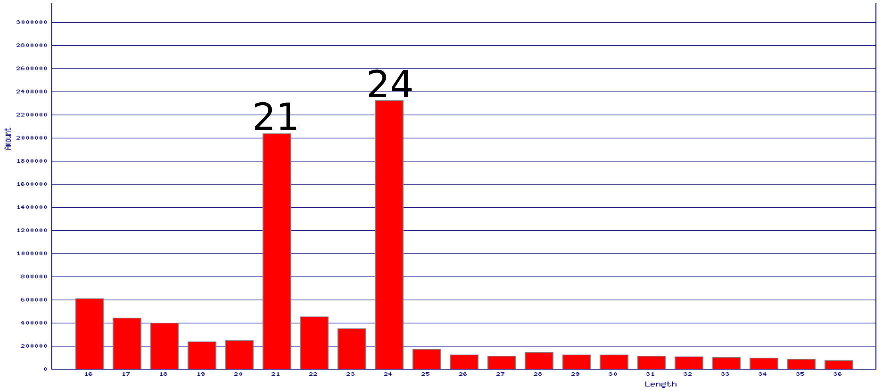
Mapping

Prediction

Annotation

Module 3 : Normalization & differential analysis of RNAseq data

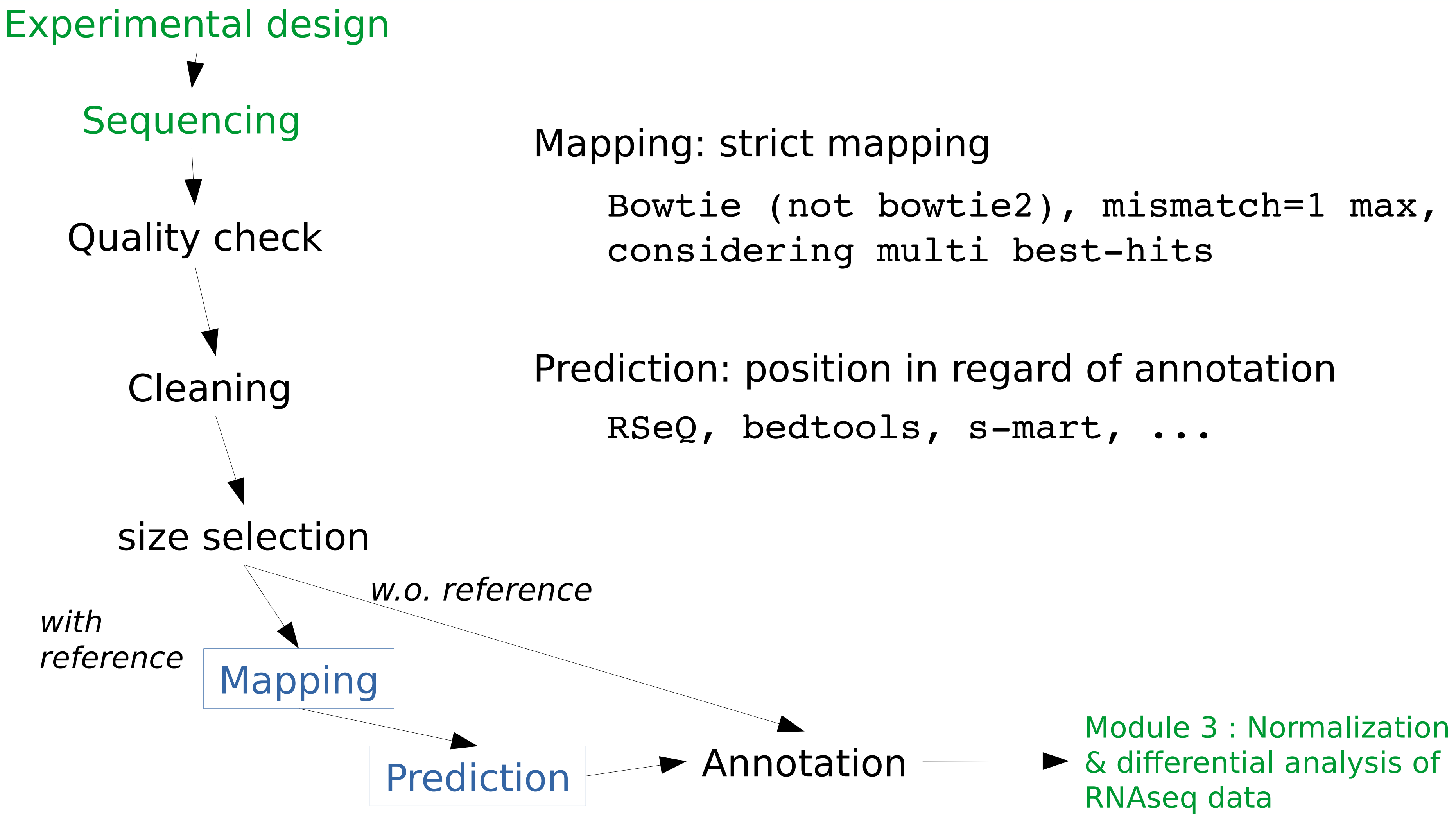
Check the distribution of length reads
Plant: miRNA=21nt, siRNA=24nt



detect redondundy

`fasta_clipping_histogram`,
duplicate `fastx_collapser`
from fastx-toolkit

smallRNAseq pipeline



Experimental design

Sequencing

Quality check

Cleaning

size selection

with reference

w.o. reference

Mapping

Prediction

Annotation

Module 3 : Normalization & differential analysis of RNAseq data

Mapping: strict mapping

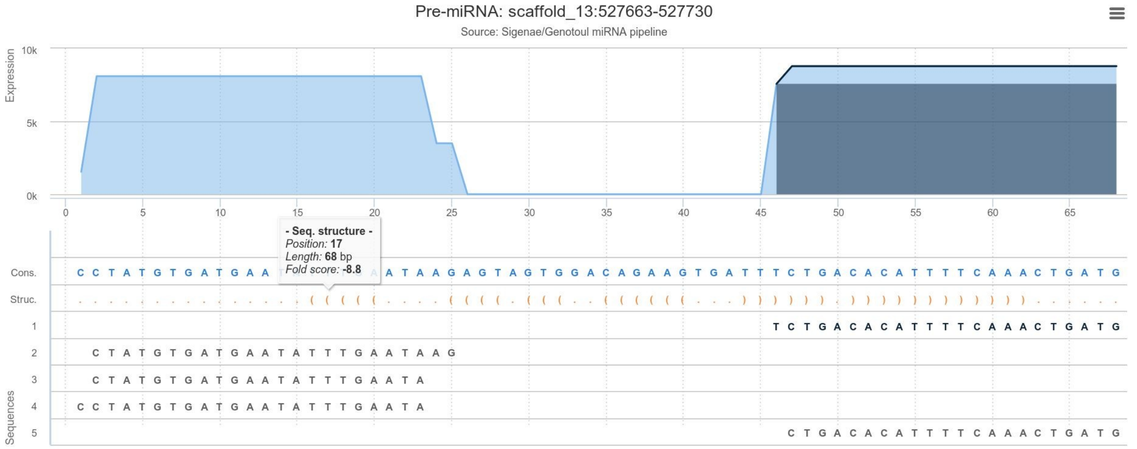
```
Bowtie (not bowtie2), mismatch=1 max,
considering multi best-hits
```

Prediction: position in regard of annotation

```
RSeQ, bedtools, s-mart, ...
```

http://bowtie-bio.sourceforge.net/index.shtml, http://rseqc.sourceforge.net/,
http://bedtools.readthedocs.io/en/latest/, https://urgi.versailles.inra.fr/Tools/S-Mart

smallRNAseq pipeline



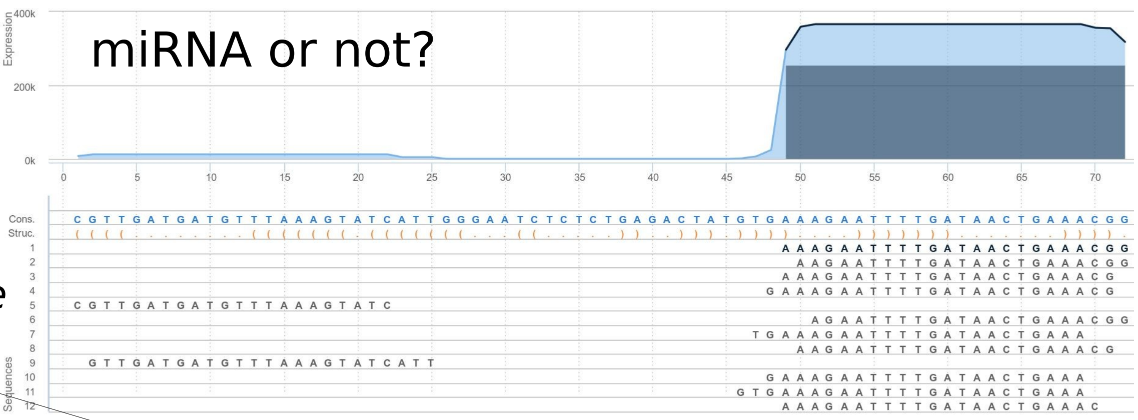
Pre-miRNA: scaffold_13:527663-527730

Source: Sigenae/Genotoul miRNA pipeline

- Seq. structure -
Position: **17**
Length: **68** bp
Fold score: **-8.8**

```
Cons.  CCTATGTGATGAA[illegible]ATAAGAGTAGTGGACAGAAGTGATTTCTGACACATTTTCAAACTGATG
Struc. ...............(((((....((((.(((..((((((...)))))).))))))))))))......
1                                                   TCTGACACATTTTCAAACTGATG
2       CTATGTGATGAATATTTGAATAAG
3       CTATGTGATGAATATTTGAATA
4      CCTATGTGATGAATATTTGAATA
5                                                    CTGACACATTTTCAAACTGATG
```

miRNA or not?



```
Cons.  CGTTGATGATGTTTAAAGTATCATTGGGAATCTCTCTGAGACTATGTGAAAGAATTTTGATAACTGAAACGG
Struc. ((((........((((((.(((((((...((.....))..))).)))).....))))))........)))).
1                                                      AAAGAATTTTGATAACTGAAACGG
2                                                       AAGAATTTTGATAACTGAAACGG
3                                                      AAAGAATTTTGATAACTGAAACG
4                                                     GAAAGAATTTTGATAACTGAAACG
5      CGTTGATGATGTTTAAAGTATC
6                                                        AGAATTTTGATAACTGAAACGG
7                                                     TGAAAGAATTTTGATAACTGAAA
8                                                       AAGAATTTTGATAACTGAAACG
9       GTTGATGATGTTTAAAGTATCATT
10                                                    GAAAGAATTTTGATAACTGAAA
11                                                  GTGAAAGAATTTTGATAACTGAAA
12                                                     AAAGAATTTTGATAACTGAAAC
```

Experimental design

Sequencing

Quality check

Cleaning

size selection

with reference

Mapping

Prediction

w.o. reference

Annotation

Module 3 : Normalization & differential analysis of RNAseq data

smallRNAseq pipeline

Experimental design

Sequencing

Quality check

Cleaning

size selection

with reference

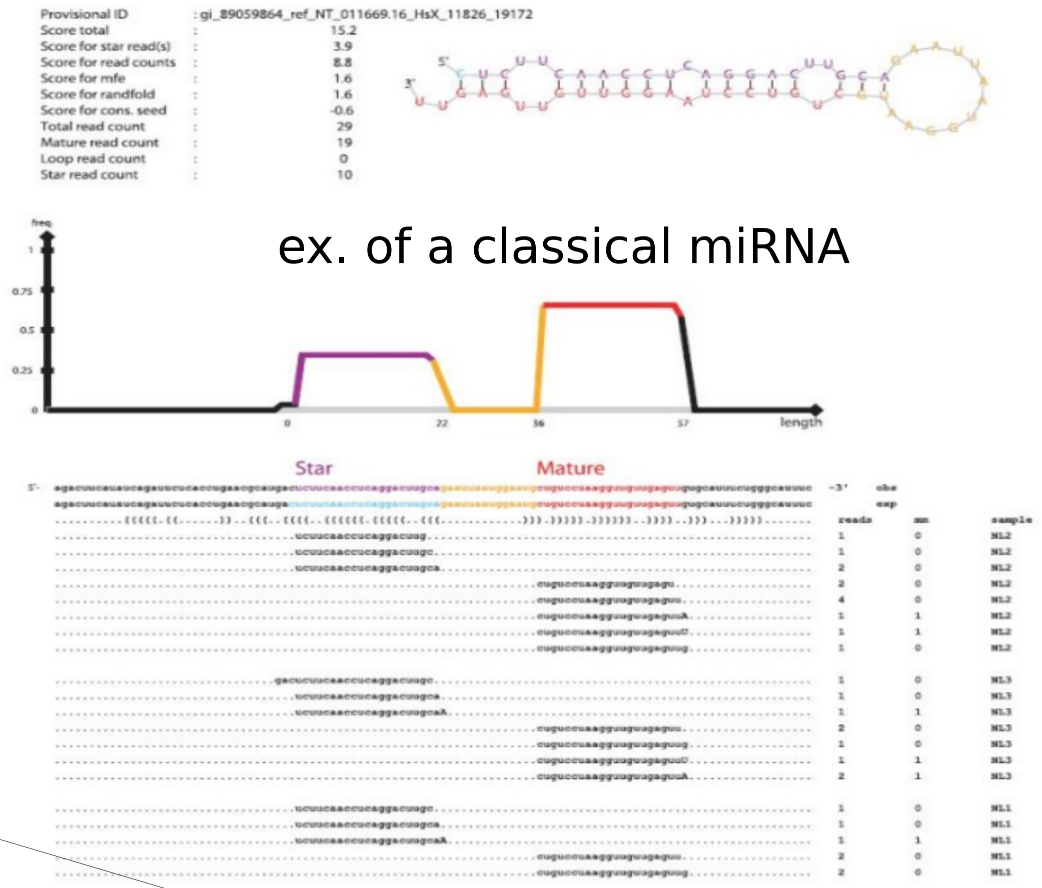
w.o. reference

Mapping

Prediction

Annotation

Module 3 : Normalization & differential analysis of RNAseq data



Provisional ID	:	gi_89059864_ref_NT_011669.16_HsX_11826_19172
Score total	:	15.2
Score for star read(s)	:	3.9
Score for read counts	:	8.8
Score for mfe	:	1.6
Score for randfold	:	1.6
Score for cons. seed	:	-0.6
Total read count	:	29
Mature read count	:	19
Loop read count	:	0
Star read count	:	10

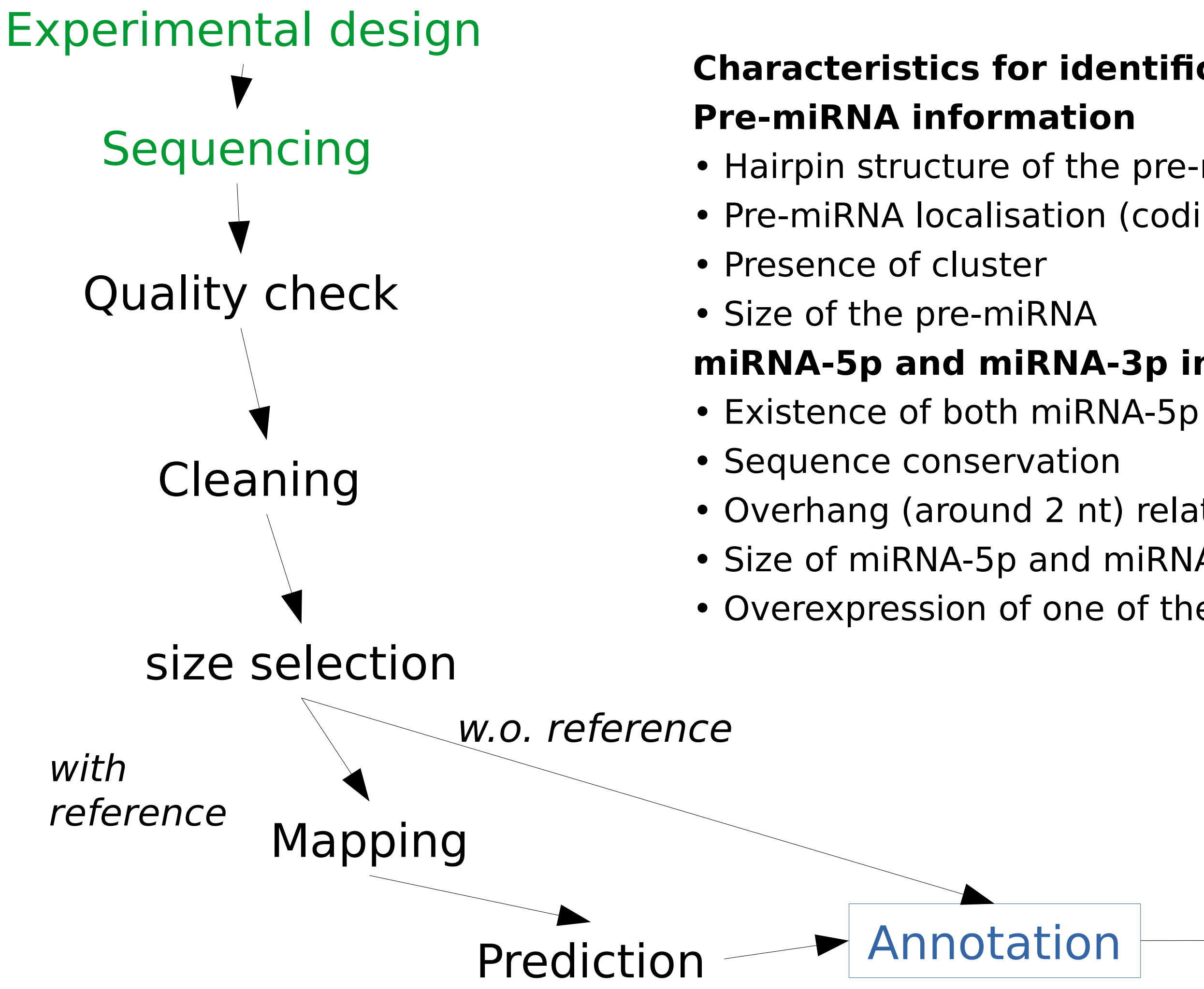
ex. of a classical miRNA

```
5'- agacuucauaucagauucucaccugaacgcaugacucuucaaccucaggacuugcagaauuaauggaaugcuguccuaagguuguugaguugugcauuucugggcauuuc -3'  obs
    agacuucauaucagauucucaccugaacgcaugacucuucaaccucaggacuugcagaauuaauggaaugcuguccuaagguuguugaguugugcauuucugggcauuuc      exp
    ..........(((((.((......))..(((..((((..((((((.(((((..(((.............))).))))).))))))..))))..)))...)))))......      reads   mm   sample
    ...................................ucuucaaccucaggacuug........................................................      1       0    NL2
    ...................................ucuucaaccucaggacuugc.......................................................      1       0    NL2
    ...................................ucuucaaccucaggacuugca......................................................      2       0    NL2
    ......................................................................cuguccuaagguuguugagu....................      2       0    NL2
    ......................................................................cuguccuaagguuguugaguu...................      4       0    NL2
    ......................................................................cuguccuaagguuguugaguuA..................      1       1    NL2
    ......................................................................cuguccuaagguuguugaguuU..................      1       1    NL2
    ......................................................................cuguccuaagguuguugaguug..................      1       0    NL2

    ................................gacucuucaaccucaggacuugc.......................................................      1       0    NL3
    ...................................ucuucaaccucaggacuugca......................................................      1       0    NL3
    ...................................ucuucaaccucaggacuugcaA.....................................................      1       1    NL3
    ......................................................................cuguccuaagguuguugaguu...................      2       0    NL3
    ......................................................................cuguccuaagguuguugaguug..................      1       0    NL3
    ......................................................................cuguccuaagguuguugaguuU..................      1       1    NL3
    ......................................................................cuguccuaagguuguugaguuA..................      2       1    NL3

    ...................................ucuucaaccucaggacuugc.......................................................      1       0    NL1
    ...................................ucuucaaccucaggacuugca......................................................      1       0    NL1
    ...................................ucuucaaccucaggacuugcaA.....................................................      1       1    NL1
    ......................................................................cuguccuaagguuguugaguu...................      2       0    NL1
    ......................................................................cuguccuaagguuguugaguug..................      2       0    NL1
```

smallRNAseq pipeline



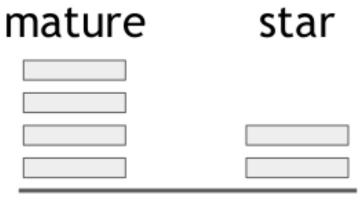
Characteristics for identification

Pre-miRNA information

- Hairpin structure of the pre-miRNA
- Pre-miRNA localisation (coding/non coding TU intronic/exonic)
- Presence of cluster
- Size of the pre-miRNA

miRNA-5p and miRNA-3p information

- Existence of both miRNA-5p and miRNA-3p
- Sequence conservation
- Overhang (around 2 nt) related to Drosha and Dicer cuts
- Size of miRNA-5p and miRNA-3p
- Overexpression of one of the miRNA-5p and miRNA-3p



Module 3 : Normalization & differential analysis of RNAseq data

smallRNAseq pipeline

Experimental design

Sequencing

Quality check

Cleaning

size selection

with reference

Mapping

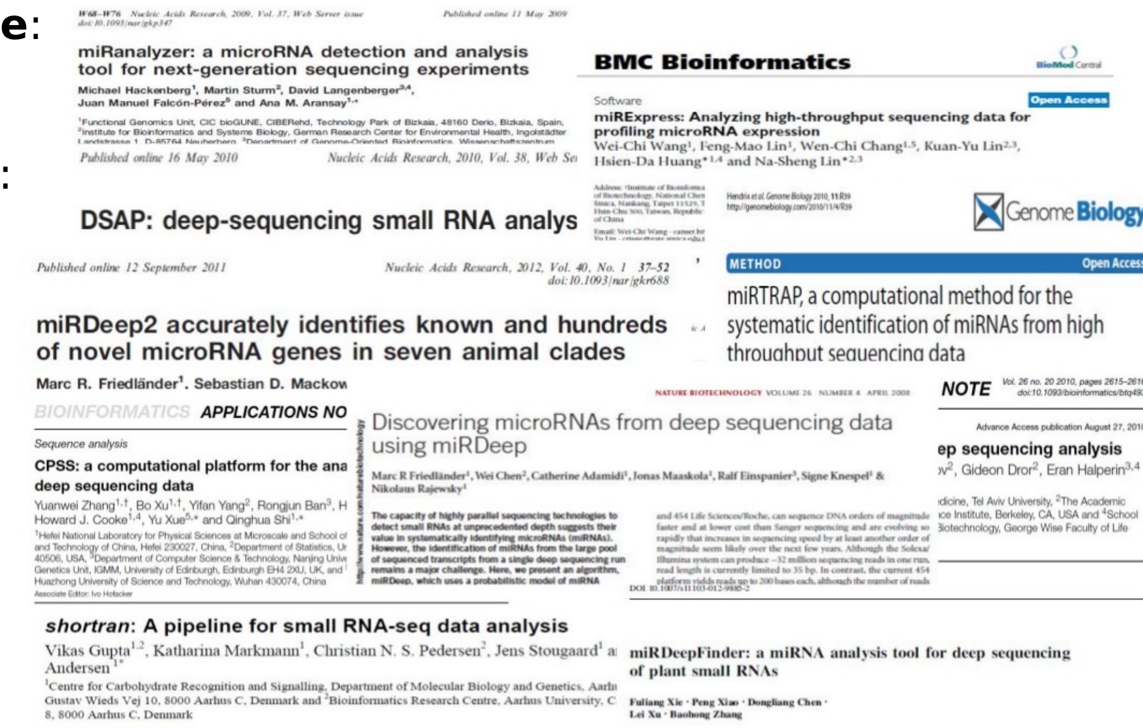
Prediction

w.o. reference

Annotation

Module 3 : Normalization & differential analysis of RNAseq data

Software:



W68–W76 Nucleic Acids Research, 2009, Vol. 37, Web Server issue Published online 11 May 2009
doi:10.1093/nar/gkp347

miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments

Michael Hackenberg[1], Martin Sturm[2], David Langenberger[3,4], Juan Manuel Falcón-Pérez[5] and Ana M. Aransay[1,*]

[1]Functional Genomics Unit, CIC bioGUNE, CIBERehd, Technology Park of Bizkaia, 48160 Derio, Bizkaia, Spain, [2]Institute for Bioinformatics and Systems Biology, German Research Center for Environmental Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, [3]Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum

Published online 16 May 2010 Nucleic Acids Research, 2010, Vol. 38, Web Se

DSAP: deep-sequencing small RNA analys

BMC Bioinformatics

BioMed Central

Software

Open Access

miRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression

Wei-Chi Wang[1], Feng-Mao Lin[1], Wen-Chi Chang[1,5], Kuan-Yu Lin[2,3], Hsien-Da Huang*[1,4] and Na-Sheng Lin*[2,3]

Address: [1]Institute of Bioinforma of Biotechnology, National Chen Sinica, Nankang, Taipei 11529, T Hsin-Chu 300, Taiwan, Republic of China

Email: Wei-Chi Wang - canser.bi Yu Lin - [illegible]

Hendrix et al. Genome Biology 2010, 11:R39
http://genomebiology.com/2010/11/4/R39

Genome Biology

METHOD Open Access

miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data

Published online 12 September 2011 Nucleic Acids Research, 2012, Vol. 40, No. 1 37–52
doi:10.1093/nar/gkr688

miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades

Marc R. Friedländer[1]. Sebastian D. Mackow

BIOINFORMATICS APPLICATIONS NO

Sequence analysis

CPSS: a computational platform for the ana deep sequencing data

Yuanwei Zhang[1,†], Bo Xu[1,†], Yifan Yang[2], Rongjun Ban[3], H Howard J. Cooke[1,4], Yu Xue[5,*] and Qinghua Shi[1,*]

[1]Hefei National Laboratory for Physical Sciences at Microscale and School of and Technology of China, Hefei 230027, China, [2]Department of Statistics, Ur 40506, USA, [3]Department of Computer Science & Technology, Nanjing Unive Genetics Unit, IGMM, University of Edinburgh, Edinburgh EH4 2XU, UK, and Huazhong University of Science and Technology, Wuhan 430074, China

Associate Editor: Ivo Hofacker

NATURE BIOTECHNOLOGY VOLUME 26 NUMBER 4 APRIL 2008

http://www.nature.com/naturebiotechnology

Discovering microRNAs from deep sequencing data using miRDeep

Marc R Friedländer[1], Wei Chen[2], Catherine Adamidi[1], Jonas Maaskola[1], Ralf Einspanier[3], Signe Knespel[1] & Nikolaus Rajewsky[1]

The capacity of highly parallel sequencing technologies to detect small RNAs at unprecedented depth suggests their value in systematically identifying microRNAs (miRNAs). However, the identification of miRNAs from the large pool of sequenced transcripts from a single deep sequencing run remains a major challenge. Here, we present an algorithm, miRDeep, which uses a probabilistic model of miRNA

and 454 Life Sciences/Roche, can sequence DNA orders of magnitude faster and at lower cost than Sanger sequencing and are evolving so rapidly that increases in sequencing speed by at least another order of magnitude seem likely over the next few years. Although the Solexa/Illumina system can produce ~32 million sequencing reads in one run, read length is currently limited to 35 bp. In contrast, the current 454 platform yields reads up to 200 bases each, although the number of reads

DOI 10.1007/s11103-012-9885-2

NOTE Vol. 26 no. 20 2010, pages 2615–261
doi:10.1093/bioinformatics/btq49

Advance Access publication August 27, 201

ep sequencing analysis

w[2], Gideon Dror[2], Eran Halperin[3,4]

dicine, Tel Aviv University, [2]The Academic ce Institute, Berkeley, CA, USA and [4]School Biotechnology, George Wise Faculty of Life

shortran: A pipeline for small RNA-seq data analysis

Vikas Gupta[1,2], Katharina Markmann[1], Christian N. S. Pedersen[2], Jens Stougaard[1] a Andersen[1*]

[1]Centre for Carbohydrate Recognition and Signalling, Department of Molecular Biology and Genetics, Aarh Gustav Wieds Vej 10, 8000 Aarhus C, Denmark and [2]Bioinformatics Research Centre, Aarhus University, C 8, 8000 Aarhus C, Denmark

miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs

Fuliang Xie · Peng Xiao · Dongliang Chen · Lei Xu · Baohong Zhang

Database:
- Rfam
- miRbase
- Silva
- GtRNAdb
- piRNA databank

...

Hoeppner MP, Barquist LE, Gardner PP.: An introduction to RNA databases. Methods Mol Biol. 2014;1097:107-23

smallRNAseq conclusion

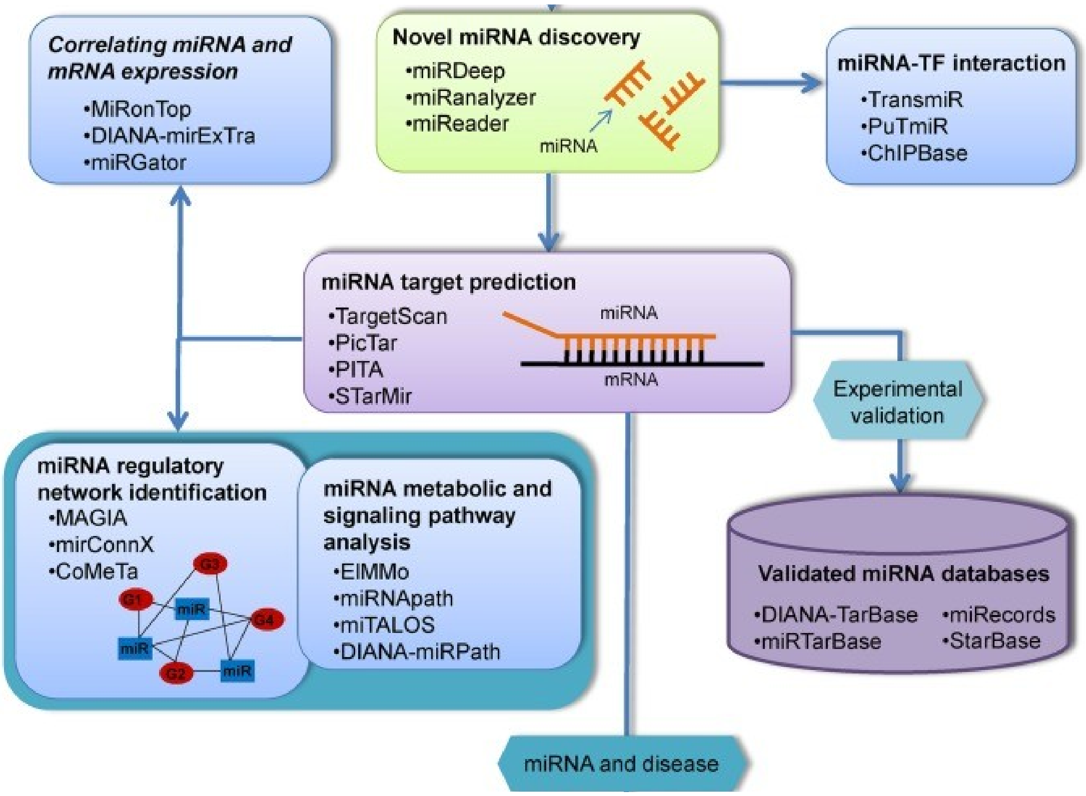
Presented example: miRNA

Experimental design

Sequence characteristics to select the expected smallRNA

Annotation: first step of the biological analysis

RNAseq or high-throughput qPCR?



Correlating miRNA and mRNA expression
- MiRonTop
- DIANA-mirExTra
- miRGator

Novel miRNA discovery
- miRDeep
- miRanalyzer
- miReader

miRNA-TF interaction
- TransmiR
- PuTmiR
- ChIPBase

miRNA target prediction
- TargetScan
- PicTar
- PITA
- STarMir

miRNA regulatory network identification
- MAGIA
- mirConnX
- CoMeTa

miRNA metabolic and signaling pathway analysis
- ElMMo
- miRNApath
- miTALOS
- DIANA-miRPath

Validated miRNA databases
- DIANA-TarBase
- miRTarBase
- miRecords
- StarBase

Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD.: Bioinformatic tools for microRNA dissection. Nucleic Acids Res. 2016 Jan 8;44(1):24-44.

RNAseq analyses

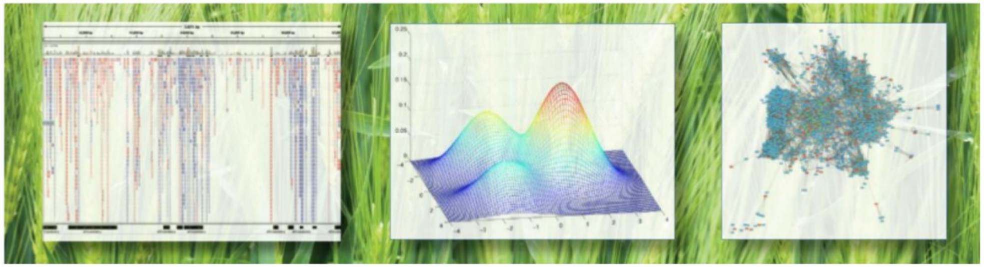
RNAseq: From sequence data (reads) to expression level (count)

Classical analyses of RNA-Seq:

- Check quality, Trimming
- Mapping / counts
- Assembly

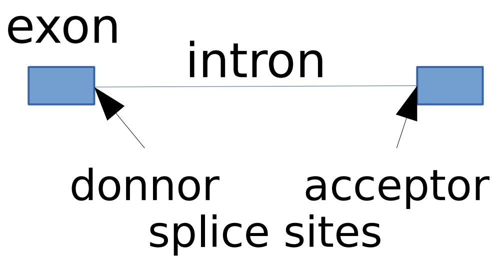
Others usages of RNA-Seq:

- smallRNA study
- **expression at isoform level**



Isoform

Eukaryotic gene includes introns:

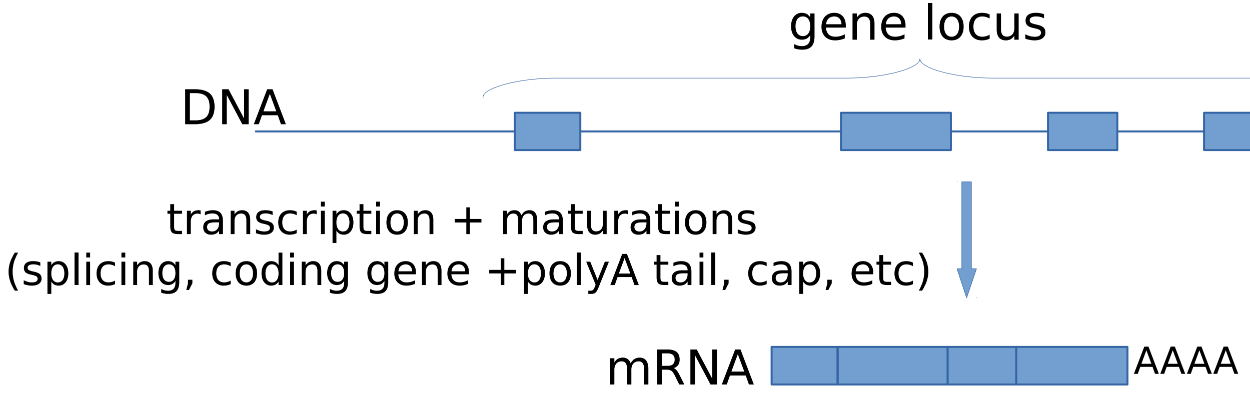


exon

intron

donnor acceptor

splice sites



gene locus

DNA

transcription + maturations
(splicing, coding gene +polyA tail, cap, etc)

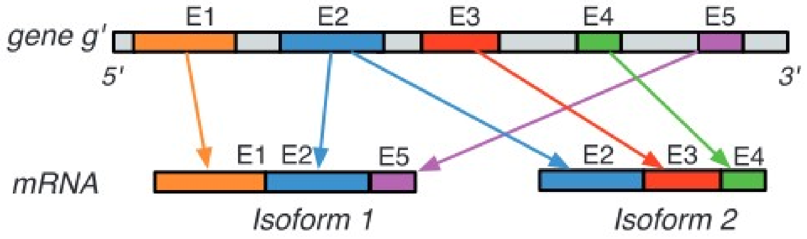
mRNA

AAAA

One gene locus may rise diverse transcripts with different usages of exons

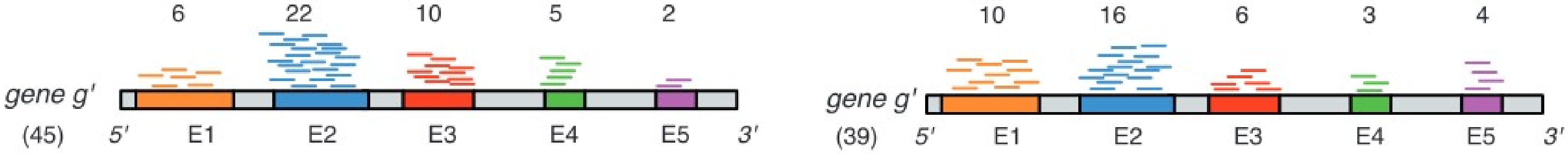
Alternative Splicing Event => isoforms

Gene level



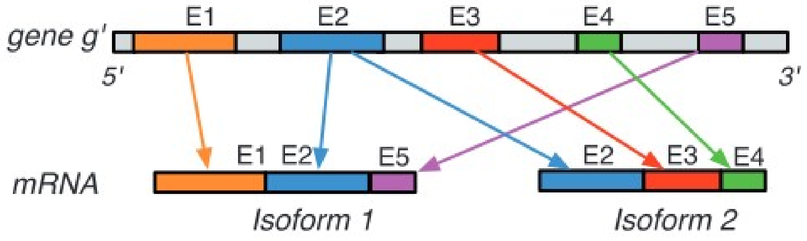
Condition 1

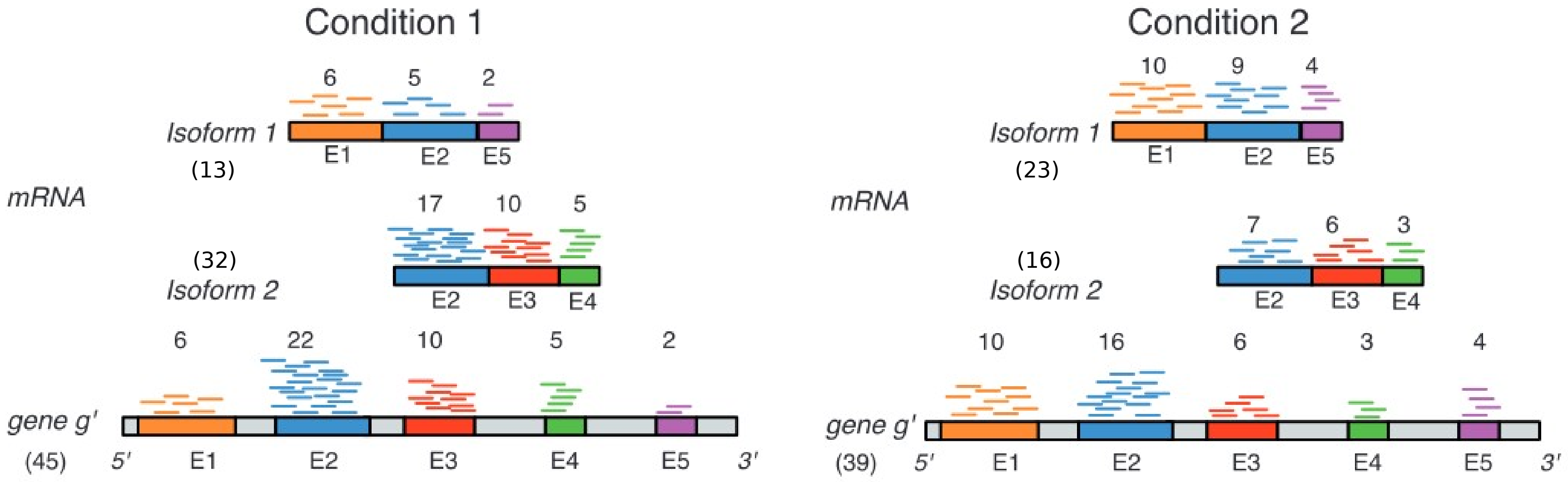
Condition 2



Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C.: EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.Bioinformatics. 2013 Apr 15;29(8):1035-43

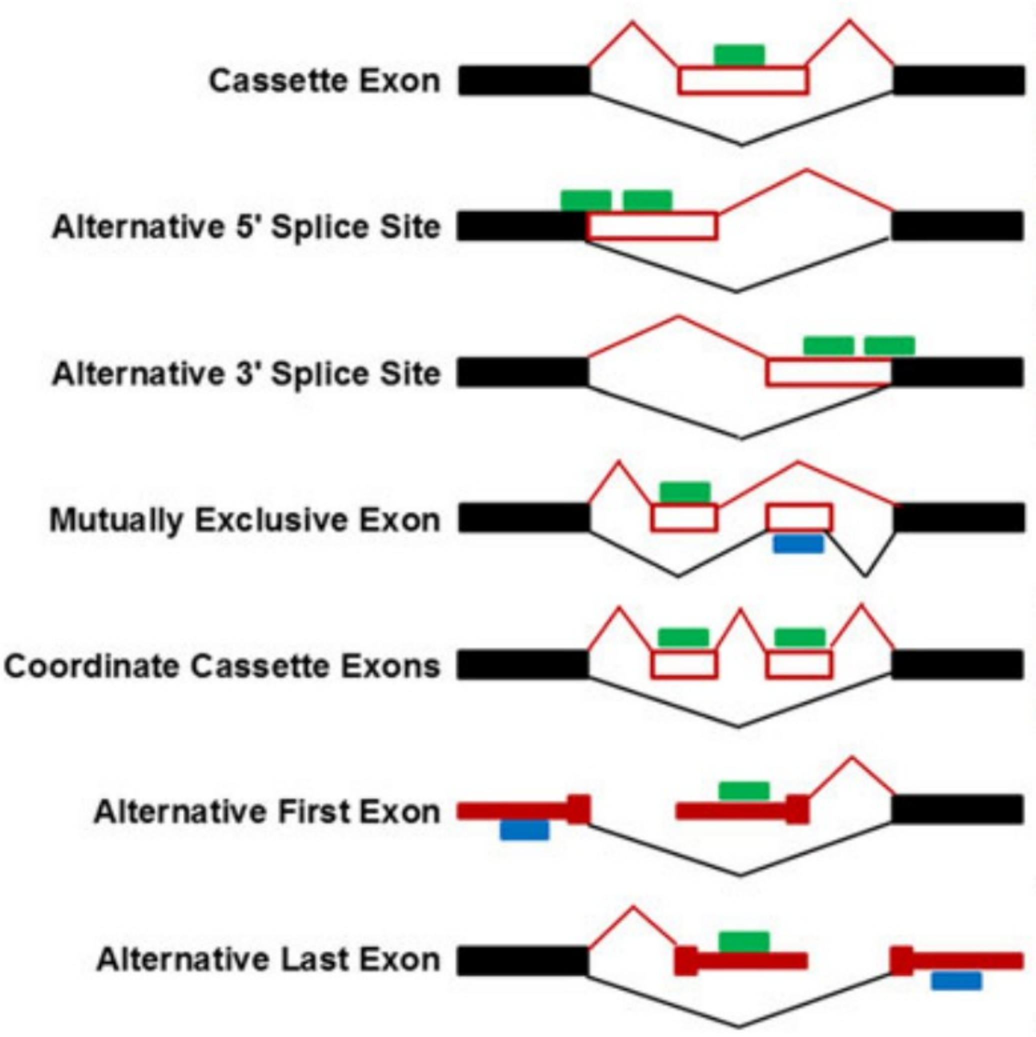
Transcript level





Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C.: EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.Bioinformatics. 2013 Apr 15;29(8):1035-43

ASE-Alternative Splicing Events

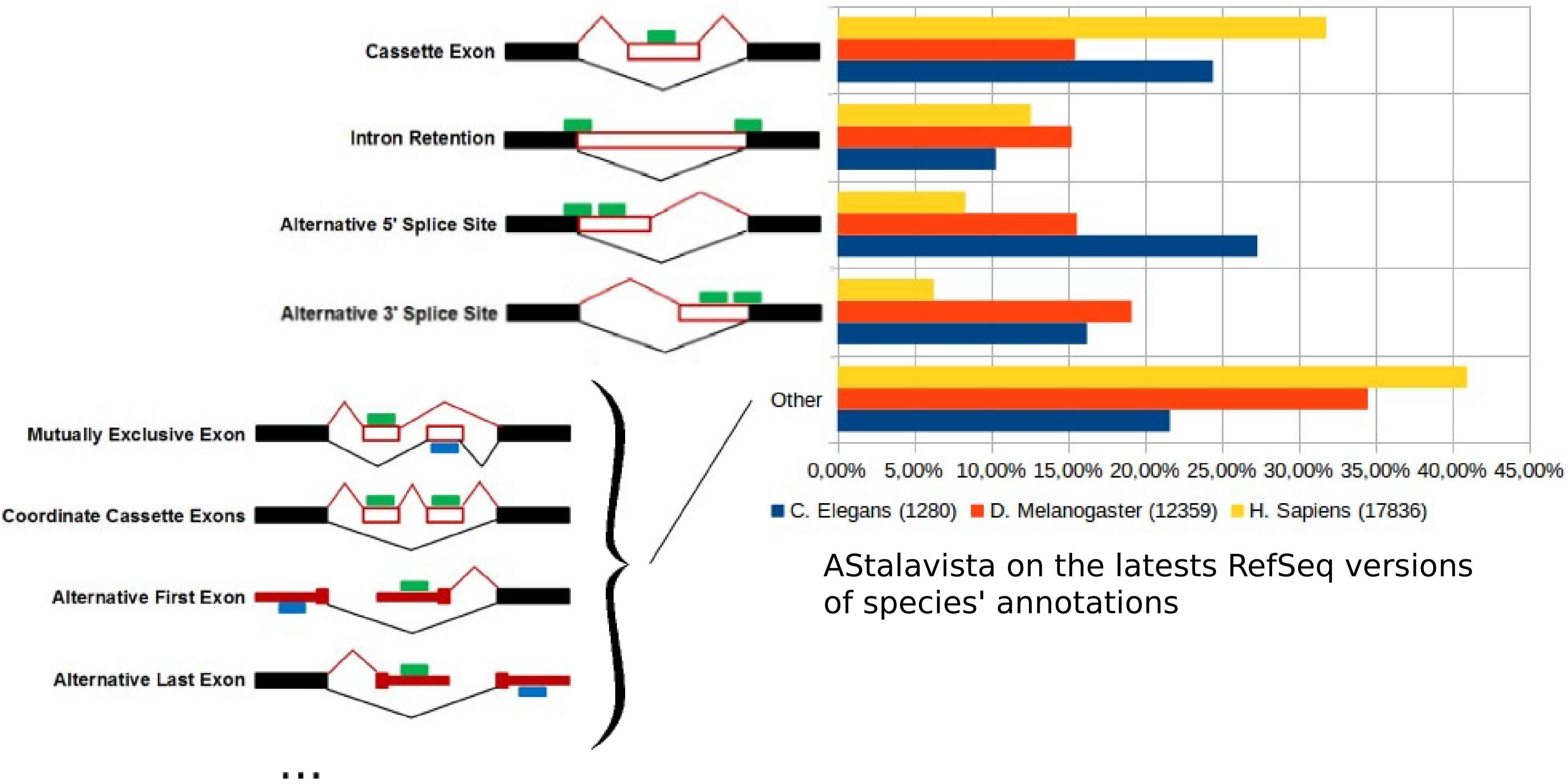


━ 2 isoforms

🟩 🟦 isoforms discriminating reads

Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L. (2014) Genome-wide
analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. BMC Genomics, 15:431

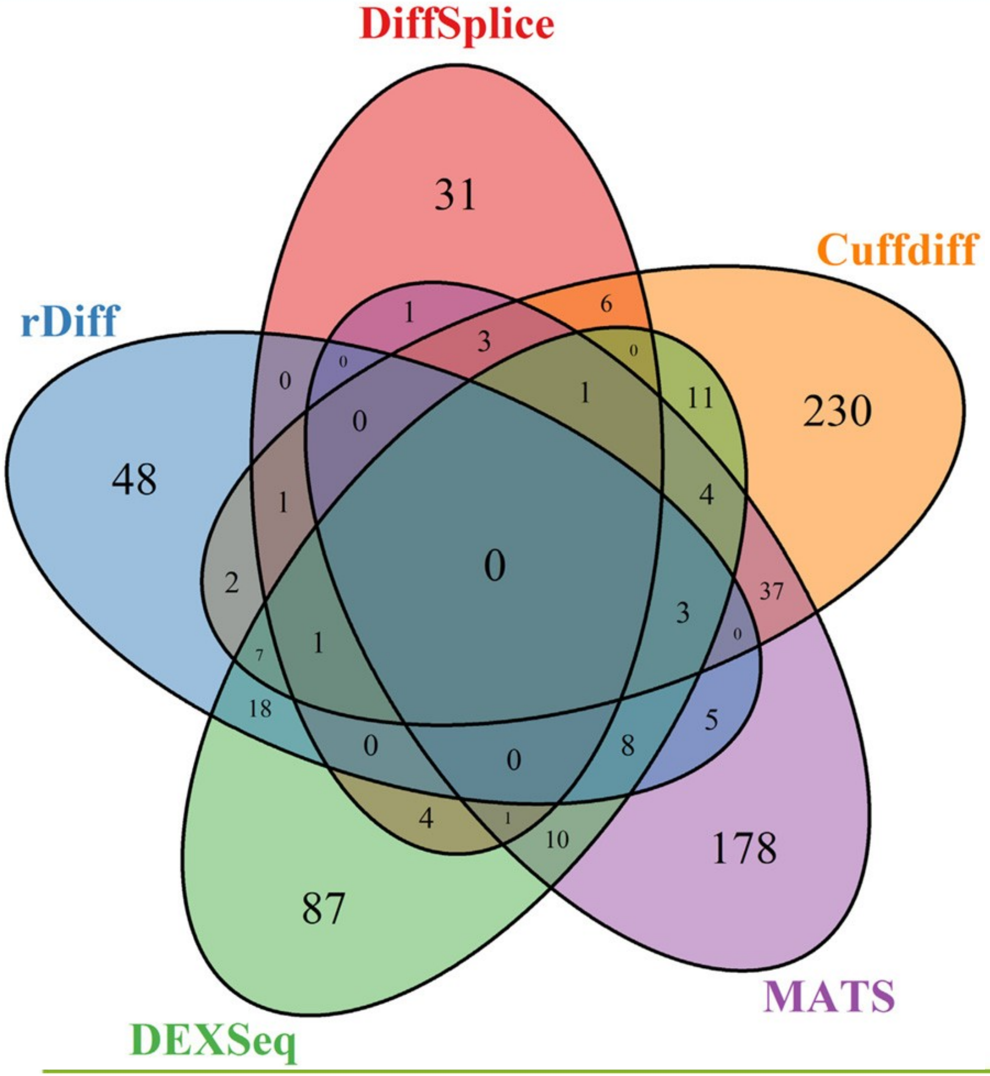
Organism specificity



AStalavista on the latests RefSeq versions of species' annotations

Foissac S, Sammeth M (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acids Research 35:W297-299

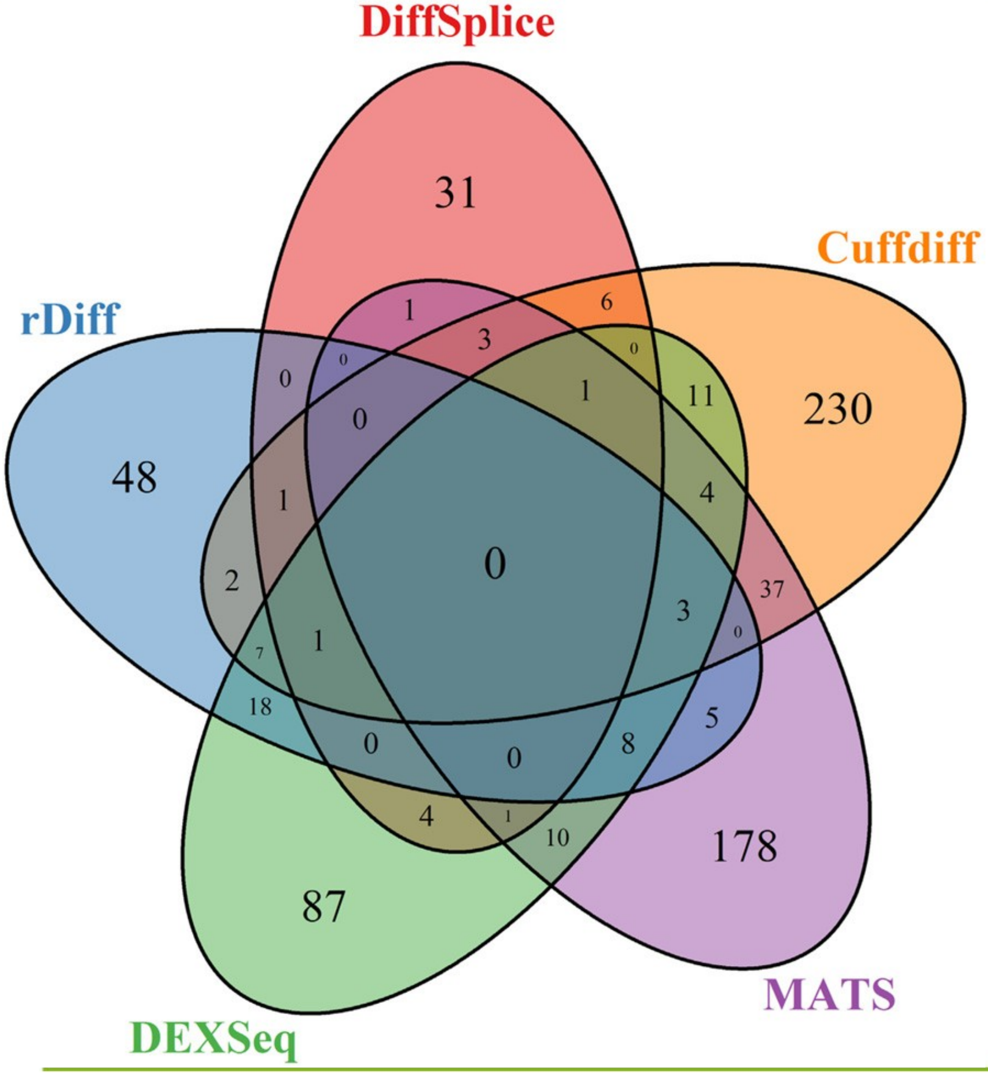
Comparing DE tools



- 5 common tools for DE
- RNAseq data:
  simulated data, *A.thaliana*
- number of common DE genes

Liu R, Loraine AE, Dickerson JA.Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems *BMC Bioinformatics*. 2014, Dec 16;15:364

Comparing DE tools



- 5 common tools
- RNAseq data:
  simulated data, *A.thaliana*
- number of common DE genes

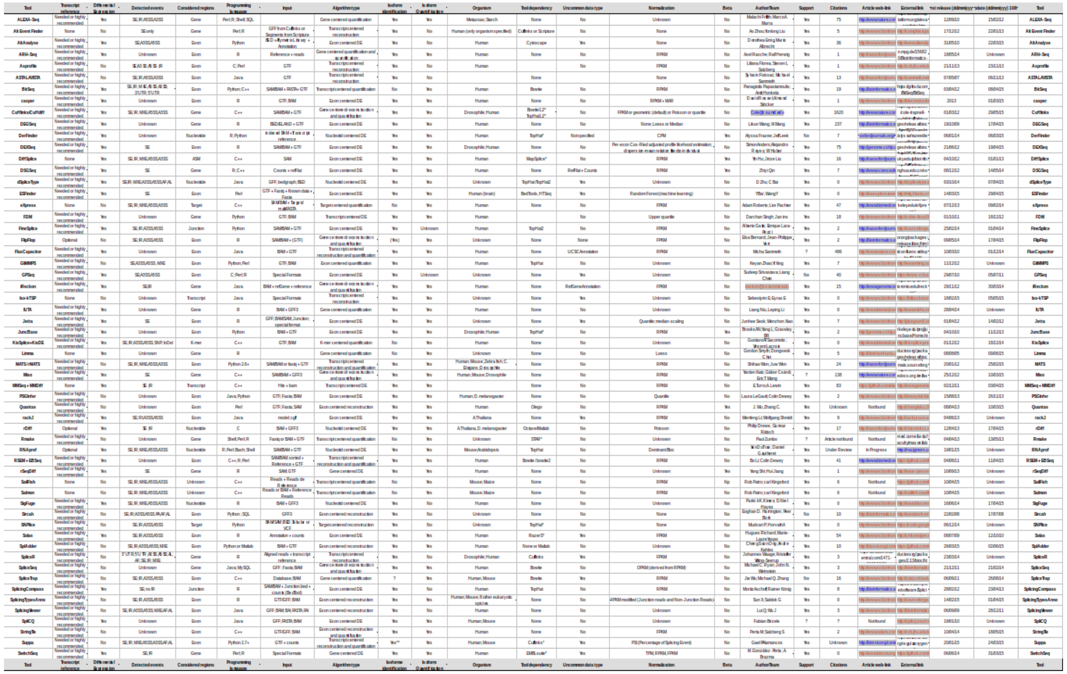
Hypotheses :

- level of analysis:
  transcript ≠ exon ≠ region
- organism specificity
- different methods/algorithms:
  mapping, counting, and DE

Liu R, Loraine AE, Dickerson JA.Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems *BMC Bioinformatics*. 2014, Dec 16;15:364

Transcript level tools

Tools for DE : >> 100, Tools for ASE : ~ 60, (2016)



Tool	Transcript reference	Differential Expression	Detected events	Considered regions	Programming language	Input	Algorithm type	Isoform identification	Isoform Quantification	Organism	Tool dependency	Uncommon data type	Normalization	Beta	Author/Team	Support	Citations	Article web-link	External link	First release (dd/mm/yyyy)	Update date (dd/mm/yyyy) 10/0	Tool
ALEXA-Seq	Needed or highly recommended	Yes	SE; IR; A5SS; A3SS	Gene	Perl; R; Shell; SQL		Gene centered quantification	Yes	Yes	Metazoas; Starch	None	No	Unknown	No	Malachi Griffith; Marco A. Marra	Yes	75	http://www.nature.com	latform.org/alexa	12/09/10	15/02/12	ALEXA-Seq
Alt Event Finder	None	No	SE only	Gene	Perl; R	GFF from Cufflinks or Segments from Scripture	Transcript centered reconstruction	Yes	No	Human (only organism specified)	Cufflinks or Scripture	No	None	No	Ao Zhou; Yunlong Liu	Yes	5	http://www.ncbi.nlm.ni	http://compbio.iupu	17/12/12	22/01/13	Alt Event Finder
AltAnalyse	Needed or highly recommended	Yes	SE; A3SS; A5SS	Exon	Python	BED + Affymetrix Library + Annotation	Exon centered DE	Yes	No	Human	Cytoscape	Yes	None	No	Dorothea Emig; Mario Albrecht	Yes	36	http://www.ncbi.nlm.ni	http://www.altanaly	31/05/10	22/03/15	AltAnalyse
ARH-Seq	Needed or highly recommended	Yes	Unknown	Exon	R	Reference + reads	Gene centered quantification and quantification	Yes	Yes	Human	None	No	RPKM	No	Axel Rasche; Ralf Herwig	Yes	1	http://nar.oxfordjourn	n.mpg.de/15682_0/Bioinformatics-	18/05/14	Unknown	ARH-Seq
Asprofile	Needed or highly recommended	No	SE; A3SS; A5SS; IR	Exon	C; Perl	GTF	Transcript centered reconstruction	Yes	No	Human		No	FPKM	No	Liliana Florea; Steven L. Salzberg	Yes	1	http://www.ncbi.nlm.ni	http://ccb.jhu.edu/s	21/11/13	23/11/13	Asprofile
ASTALAVISTA	Needed or highly recommended	No	SE; IR; A5SS; A3SS	Exon	Java	GTF	Transcript centered reconstruction	Yes	No		None		None	No	Sylvain Foissac; Michael Sammeth	Yes	13	http://nar.oxfordjourn	http://sammeth.net	07/05/07	06/11/13	ASTALAVISTA
BitSeq	Needed or highly recommended	Yes	SE; IR; MXE; A5SS; A3SS; 3'UTR; 5'UTR	Exon	Python; C++	SAM/BAM + FASTA + GTF	Transcript centered quantification	No	Yes	Human	Bowtie	No	RPKM	No	Panagiotis Papastamoulis; Antti Honkela	Yes	19	http://bioinformatics.o	https://github.com/BitSeq/BitSeq	03/04/12	08/04/15	BitSeq
casper	Needed or highly recommended	Yes	Unknown	Exon	R	GTF; BAM	Exon centered DE	Yes	Yes	Human	None	No	RPKM + MAR	No	David Rossell; Almond Stöcker	Yes	1	http://www.ncbi.nlm.ni	http://bioconductor	2013	01/03/15	casper
Cufflinks/Cuffdiff	Needed or highly recommended	Yes	SE; IR; MXE; A5SS; A3SS	Gene	C++	SAM/BAM + GTF	Gene centered reconstruction and quantification	Yes	Yes	Drosophile; Human	Bowtie1;2* TopHat1;2*	No	FPKM or geometric (default) or Poisson or quartile	No	Cole@cs.umd.edu	Yes	1620	http://www.nature.com	/cole-trapnell-	01/03/12	29/05/15	Cufflinks
DEGSeq	Needed or highly recommended	Yes	Unknown	Gene	R	BED/ELAND + GTF	Exon centered DE	Yes	Yes	Human	None	No	None; Loess or Median	No	Likun Wang; Xi Wang	Yes	237	http://bioinformatics.o	ges/release/bioc	19/10/09	17/04/15	DEGSeq
DerFinder	Needed or highly recommended	Yes	Unknown	Nucleotide	R; Python	Indexed BAM + Transcript reference	Nucleotid centered DE	Yes	Yes	Human	TopHat*	Not specified	CPM	Yes	Alyssa Frazee; Jeff Leek	No	7	oxfordjournals.org/	alyssafrazee/de	06/01/14	06/03/15	DerFinder
DEXSeq	Needed or highly recommended	Yes	SE	Exon	R	SAM/BAM + GTF	Exon centered DE	Yes	Yes	Drosophile; Human	None	No	Per-exon Cox-Ried adjusted profile likehood estimation; dispersion-mean relation fit to individual	No	Simon Anders; Alejandro Reyes; W Huber	Yes	75	http://genome.cshlp.o	ges/release/bioc	21/06/12	19/04/15	DEXSeq
DiffSplice	None	Yes	SE; IR; MXE; A5SS; A3SS	ASM	C++	SAM	Exon centered DE	Yes	Yes	Human	MapSplice*	No	RPKM	Yes	Yin Hu; Jinze Liu	Yes	16	http://nar.oxfordjourn	uky.edu/bioinfo	04/10/12	01/01/13	DiffSplice
DSGSeq	Needed or highly recommended	Yes	SE	Gene	R; C++	Counts + refFlat	Exon centered DE	Yes	Yes	Human	None	RefFlat + Counts	RPKM	Yes	Zhiyi Qin	Yes	7	http://www.sciencedi	nghua.edu.cn/bio	08/12/12	14/05/14	DSGSeq
dSpliceType	Needed or highly recommended	Yes	SE; IR; MXE; A3SS; A5SS; AF; AL	Nucleotide	Java	GFF; bedgraph; BED	Nucleotid centered DE	Yes	Yes	Unknown	TopHat/TopHat2	Yes	Unknown	No	D Zhu; C Bai	Yes	0	http://www.ncbi.nlm.ni	http://dsplicetype.e	03/10/14	07/04/15	dSpliceType
ESFinder	Needed or highly recommended	Yes	SE	Exon	Perl	GTF + Fastq + Known data + Fasta	Exon centered DE	Yes	Yes	Human (brain)	BedTools, HTSeq	Yes	Random Forest (machine learning)	No	YBai; WangY	Yes	0	http://ieeexplore.ieee	http://mlg.hit.edu.cn	14/03/15	29/04/15	ESFinder
eXpress	None	No	SE; IR; MXE; A5SS; A3SS	Target	C++	SAM/BAM + Targets/multiFASTA	Target centered quantification	No	Yes	Human	None	No	FPKM	No	Adam Roberts; Lior Pachter	Yes	47	http://www.biomedcen	keley.edu/eXpre	07/12/13	09/02/14	eXpress
FDM	Needed or highly recommended	Yes	Unknown	Gene	Python	GTF; BAM	Transcript centered DE	Yes	Yes	Human		No	Upper quartile	No	Darshan Singh; Janins	Yes	18	http://www.ncbi.nlm.ni	http://csbio-linux0	01/10/11	18/12/12	FDM
FineSplice	Needed or highly recommended	Yes	SE; IR; A3SS; A5SS	Junction	Python	SAM/BAM + GTF	Exon centered DE	Yes	Unknown	Human	TopHat2	No	FPKM	No	Alberto Gatto; Enrique Lara-Pezzi	Yes	2	http://nar.oxfordjourn	http://sourceforge.	25/02/14	01/04/14	FineSplice
FlipFlop	Optional	No	SE; IR; A3SS; A5SS	Exon	R	SAM/BAM + (GTF)	Gene centered reconstruction and quantification	(Yes)	Yes	Unknown	None	None	FPKM	No	Elsa Bernard; Jean-Philippe Vert	Yes	2	http://bioinformatics.o	or.org/packages/release/bioc/html/	09/05/14	17/04/15	FlipFlop
FluxCapacitor	Needed or highly recommended	Yes	Unknown	Exon	Java	BAM + GTF	Transcript centered reconstruction and quantification	Yes	Yes	Human	None	UCSC Annotation	RPKM	No	Micha Sammeth	?	486	http://www.nature.com	[illegible]	10/03/10	01/12/14	FluxCapacitor
GIMMPS	Needed or highly recommended	No	SE; A3SS; A5SS; MXE	Exon	Python; Perl	GTF; BAM	Exon centered quantification	Yes	Yes	Human	TopHat	No	Unknown	No	Keyan Zhao; Yi Xing	Yes	7	http://www.ncbi.nlm.ni	http://www.mimg.uc	11/12/12	Unknown	GIMMPS
GPSeq	Needed or highly recommended	Yes	SE; A3SS; A5SS	Exon	C; Perl; R	Special Formats	Exon centered DE	Yes	Unknown	Unknown	None	Yes	Unknown	No	Sudeep Srivastava; Liang Chen	No	40	http://www.ncbi.nlm.ni	https://www-rcf.us	29/07/10	05/07/11	GPSeq
iReckon	Needed or highly recommended	Yes	SE; IR	Gene	Java	BAM + refGene + reference	Gene centered reconstruction and quantification	Yes	Yes	Human	None	RefGeneAnnotation	FPKM	No	ireckon@cs.toronto.edu	Yes	15	http://www.genome.o	toronto.edu/reck	29/11/12	30/03/14	iReckon
Iso-kTSP	None	No	Unknown	Transcript	Java	Special Formats	Transcript centered reconstruction	Yes	Yes	Unknown	None	Yes	Unknown	No	Sebestyén E; Eyras E	Yes	0	http://www.ncbi.nlm.ni	https://bitbucket.or	18/02/15	05/05/15	Iso-kTSP
IUTA	Needed or highly recommended	Yes	Unknown	Gene	R	BAM + GFF3	Gene centered quantification	Yes	Yes	Human	None	No	Unknown	No	Liang Niu; Leping Li	Yes	0	http://www.biomedcen	http://www.niehs.ni	28/04/14	Unknown	IUTA
Jetta	Needed or highly recommended	Yes	SE	Exon	R	GFF; BAM/SAM; Junction; special format	Exon centered DE	Yes	Yes	Unknown	None	Yes	Quantile; median-scaling	No	Junhee Seok; Wenzhong Xiao	Yes	7	http://www.ncbi.nlm.ni	http://gluegrant1.st	01/04/12	14/02/12	Jetta
JuncBase	Needed or highly recommended	Yes	Unknown	Exon	Python	BAM + GTF	Exon centered DE	Yes	Yes	Drosophile; Human	TopHat*	No	RPKM	Yes	Brooks AN; Yang L; Graveley BR	Yes	2	http://genome.cshlp.o	rkeley.edu/projju ncbase/Home.ht	04/10/10	11/12/13	JuncBase
KisSplice+KisDE	Needed or highly recommended	Yes	SE; IR; A3SS; A5SS; SNP; InDel	K-mer	C++	GTF; BAM	K-mer centered quantification	No	Yes	Human	None	No	Unknown	No	Gustavo AT Sacomoto; Vincent Lacroix	Yes	0	http://www.biomedcen	http://kissplice.pra	01/12/12	18/12/14	KisSplice
Limma	None	Yes	Unknown	Gene	R		Gene centered quantification	No	Yes	Unknown	None	No	Loess	No	Gordon Smyth; Dongseok Choi	Yes	5	http://bioinf.wehi.edu.	ductor.org/packages/release/bioc	08/08/05	09/06/15	Limma
MATS/rMATS	Needed or highly recommended	No	SE; IR; MXE; A5SS; A3SS	Exon	Python 2.6+	SAM/BAM or fastq + GTF	Transcript centered reconstruction and quantification	Yes	Yes	Human; Mouse; Zebra fish; C. Elegans; Drosophile	None	No	RPKM	No	Shihao Shen; Juw Won	Yes	24	http://nar.oxfordjourn	mats.sourceforg	20/01/12	25/02/15	MATS
Miso	Needed or highly recommended	Yes	SE	Gene	C++	SAM/BAM + GFF3	Gene centered reconstruction and quantification	Yes	Yes	Human; Mouse; Drosophile	None	No	RPKM	No	Yarden Katz; Gábor Csárdi; Eric T Wang	?	138	http://www.nature.com	edocs.org/en/lat	25/12/12	10/03/15	Miso
MMSeq + MMDiff	None	Yes	SE; IR	Transcript	C++	Hits + bam	Transcript centered DE	Yes	Yes		None	No	RPKM	No	E.Turro; A.Lewin	Yes	83	https://github.com/eturro	http://www.genome	02/12/11	03/04/15	MMSeq + MMDiff
PSGInfer	Needed or highly recommended	No	Unknown	Exon	Java; Python	GTF; Fasta; BAM	Exon centered DE	Yes	Yes	Human; D. melanogaster	None	No	Quantile	No	Laura LeGault; Colin Dewey	Yes	2	http://www.ncbi.nlm.ni	http://deweylab.bio	15/08/13	26/11/13	PSGInfer
Quantas	Needed or highly recommended	No	Unknown	Exon	Perl	GTF; Fasta; SAM	Exon centered reconstruction	Yes	Yes	Human	Olego	No	RPKM	Yes	J. Wu; Zhang C.	Yes	Unknown	Not found	http://zhanglab.c2b	08/04/13	10/03/15	Quantas
rackJ	Needed or highly recommended	Yes	SE; RI; A3SS; A5SS	Exon	Java	model; cgf	Exon centered DE	Yes	Yes	A.Thaliana	None	Yes	RPKM	No	Wenfeng Li; Wolfgang Shmidt	Yes	9	http://www.ncbi.nlm.ni	http://rackj.source	04/06/13	Unknown	rackJ
rDiff	Optional	Yes	SE; IR	Nucleotide	C	BAM + GFF3	Nucleotid centered DE	Yes	Yes	A.Thaliana, D. melanogaster	Octave/Matlab	No	Poisson	No	Philipp Drewe; Gunnar Rätsch	Yes	17	http://nar.oxfordjourn	http://cbio.mskcc.o	12/04/13	17/04/15	rDiff
Rmake	Needed or highly recommended	No	Unknown	Gene	Shell; Perl; R	Fastq or BAM + GTF	Transcript centered quantification	No	Yes	Unknown	STAR*	No	Unknown	No	Paul Zumbo	?	Article not found	Not found	med.cornell.edu/faculty/mason/lab	04/04/13	13/05/13	Rmake
RNAprof	Optional	Yes	SE; IR; MXE; A5SS; A3SS	Nucleotide	R; Perl; Bash; Shell	SAM/BAM + GTF	Nucleotid centered DE	Yes	Yes	Mouse; Arabidopsis	TopHat	No	DominantBloc	No	Van Du Tran; Daniel Gautheret	Yes	Under Review	In Progress	http://rnaprof.u	19/01/15	Unknown	RNAprof
RSEM + EBSeq	Needed or highly recommended	Yes	Unknown	Exon	C++; R; Perl	SAM/BAM; sorted + Reference + GTF	Transcript centered reconstruction and quantification	Yes	Yes	Human	Bowtie / bowtie2	No	RPKM	No	Bo Li; Colin Dewey	Yes	41	http://www.biomedcen	https://github.com/	04/08/11	11/04/15	RSEM + EBSeq
rSeqDiff	Needed or highly recommended	Yes	SE	Gene	R	SAM; GTF	Gene centered DE	Yes	Yes	Human	None	No	Unknown	Yes	Yang Shi; Hui Jiang	Yes	1	http://www.ncbi.nlm.ni	http://www-person	10/09/13	Unknown	rSeqDiff
SailFish	None	No	SE; IR; MXE; A5SS; A3SS	Unknown	C++	Reads + Reads de Reference	Transcript centered quantification	No	Yes	Mouse; Maize	None	No	RPKM	Np	Rob Patro; carl Kingsford	Yes	6	Not found	https://github.com/	10/04/15	Unknown	SailFish
Salmon	None	No	SE; IR; MXE; A5SS; A3SS	Unknown	C++	Reads or BAM + Reference Reads	Transcript centered quantification	No	Yes	Mouse; Maize	None	No	RPKM	No	Rob Patro; carl Kingsford	Yes	6	Not found	http://sailfish.readt	10/04/15	Unknown	Salmon
SigFuge	Needed or highly recommended	Yes	SE; IR; MXE; A5SS; A3SS	Nucleotide	R	BAM + GFF3	Nucleotid centered DE	Yes	No	Human	None	No	Unknown	No	Patrick K. Kimes; D. Neil Hayes	Yes	0	http://www.ncbi.nlm.ni	http://www.biocond	16/06/14	17/04/15	SigFuge
Sircah	Needed or highly recommended	No	SE; IR; A3SS; A5SS; FA; AF; AL	Exon	Python; SQL	GFF3	Exon centered reconstruction	Yes	No	Unknown	None	No	None	No	Eoghan D. Harrington; Peer Bork	No	10	http://bioinformatics.o	http://www.bork.em	22/02/08	17/07/08	Sircah
SNPlice	Needed or highly recommended	No	SE; IR; A3SS; A5SS	Target	Python	BAM/SAM; BED; Tabular or VCF	Target centered reconstruction	Yes	No	Unknown	TopHat*	No	None	No	Mudvari P; Horvath A	Yes	0	http://www.ncbi.nlm.ni	https://code.googl	06/12/14	Unknown	SNPlice
Solas	Needed or highly recommended	Yes	SE; IR; A3SS; A5SS	Exon	R	Annotation + counts	Exon centered DE	Yes	Yes	Human	RazerS*	Yes	FPKM	No	Hugues Richard; Marie-Laure Yaspo	Yes	54	http://www.ncbi.nlm.ni	http://cmb.molgen.	08/07/09	12/10/10	Solas
SplAdder	Needed or highly recommended	No	SE; IR; A3SS; A5SS; MXE	Exon	Python or Matlab	BAM + GTF	Exon centered reconstruction	Yes	Yes	Human	None or Matlab	No	Unknown	No	Cheng Soon Ong; Andre Kahles	Yes	10	http://biorxiv.org/con	https://github.com/	26/03/15	02/06/15	SplAdder
SpliceR	Needed or highly recommended	No	3'UTR; 5'UTR; A3SS; A5SS; AL; AF; SE; IR; MXE	Gene	R	Aligned reads + transcript reference	Transcript centered reconstruction	Yes	No	Drosophile; Human	Cufflinks	Yes	FPKM	No	Johannes Waage; Kristoffer Vitting-Seerup	Yes	3	entral.com/1471-	ductor.org/packages/2.13/bioc/ht	23/03/14	Unknown	SpliceR
SpliceSeq	Needed or highly recommended	No	Unknown	Gene	Java; MySQL	GFF; Fasta; BAM	Gene centered reconstruction and quantification	Yes	Yes	Human	Bowtie	No	OPKM (derived from RPKM)	No	Michael C. Ryan; John N. Weinstein	Yes	3	http://www.ncbi.nlm.ni	http://bioinformatic	21/12/11	21/02/14	SpliceSeq
SpliceTrap	Needed or highly recommended	No	SE; IR; A3SS; A5SS	Exon	C++	Database; BAM	Gene centered quantification	?	Yes	Human, Mouse	Bowtie	Yes	RPKM	No	Jie Wu; Michael Q. Zhang	No	16	http://www.ncbi.nlm.ni	http://rulai.cshl.edu	06/09/11	26/08/14	SpliceTrap
SplicingCompass	Needed or highly recommended	Yes	SE; no IR	Junction	R	SAM/BAM + Junction.bed + counts (Bedtool)	Exon centered DE	Yes	No	Human	TopHat	No	RPKM	No	Moritz Aschoff; Rainer König	Yes	8	http://bioinformatics.o	e/software/Splici	28/02/12	23/04/13	SplicingCompass
SplicingTypesAnno	Needed or highly recommended	No	SE; IR; A3SS; A5SS	Exon	R	GTF/GFF; BAM	Exon centered reconstruction	Yes	Yes	Human; Mouse; 8 other eukaryotic spicies	None	No	RPKM modified (Junction reads and Non-Junction Reads)	No	Sun X; Sablok G	Yes	0	http://www.ncbi.nlm.ni	http://sourceforge.	14/02/15	01/04/15	SplicingTypesAnno
SplicingViewer	Needed or highly recommended	No	SE; IR; A3SS; A5SS; MXE; AF; AL	Exon	Java	GFF; BAM; BAI; FASTA; FAI	Exon centered reconstruction	Yes	No	Human	None	No	Unknown	No	Liu Q; Wu J	Yes	3	http://www.ncbi.nlm.ni	http://bioinformatic	06/09/09	28/12/11	SplicingViewer
SpliCQ	Needed or highly recommended	Yes	Unknown	Exon	Java	GFF; FASTA; BAM	Exon centered DE	Yes	Yes	Human; Mouse	None	No	Unknown	No	Fabian Birzele	?	7	Not found	http://splicq.sourc	19/01/10	Unknown	SpliCQ
StringTie	Needed or highly recommended	No	Unknown	Exon	C++	GTF/GFF; BAM	Exon centered reconstruction and quantification	Yes	Yes	Human	None	No	FPKM	No	Perta M; Salzberg S	Yes	2	http://www.nature.com	http://ccb.jhu.edu/s	10/04/14	18/05/15	StringTie
Suppa	Needed or highly recommended	No	SE; IR; MXE; A5SS; A3SS; AF; AL	Exon	Python 2.7+	GTF + counts	Transcript centered reconstruction and quantification	Yes***	Yes	Human; Mouse	Cufflinks*	Yes	PSI (Percentage of Splicing Event)	No	Gael P Alamancos	Yes	Unknown	http://biorxiv.org/cont	rg/regulatorygen	20/01/15	24/03/15	Suppa
SwitchSeq	Needed or highly recommended	Yes	SE; IR	Gene	Perl; R	Special Formats	Gene centered DE	Yes	Yes	Human	EMBL suite*	Yes	TPM, RPKM, FPKM	No	M.González-Porta; A Brazma	Yes	0	http://www.biorxiv.org	https://github.com/	06/06/14	31/03/15	SwitchSeq
Tool	Transcript reference	Differential Expression	Detected events	Considered regions	Programming language	Input	Algorithm type	Isoform identification	Isoform Quantification	Organism	Tool dependency	Uncommon data type	Normalization	Beta	Author/Team	Support	Citations	Article web-link	External link			Tool

ALEXA-Seq, Alt Event Finder, AltAnalyse, ARH-Seq, Asprofile, ASTALAVISTA, BitSeq, casper, Cufflinks/Cuffdiff, DEGSeq, DerFinder, DEXSeq, DiffSplice, DSGSeq, dSpliceType, ESFinder, eXpress, FDM, FineSplice, FlipFlop, FluxCapacitor, GliMMPS, GPSeq, iReckon, Iso-kTSP, IUTA, Jetta, JuncBase, KisSplice/KisDE, Limma, MATS/rMATS, Miso, MMSeq/MMDiff, PSGInfer, Quantas, rackJ, rDiff, Rmake, RNAprof, RSEM/EBSeq, rSeqDiff, SailFish, Salmon, SigFuge, Sircah, SNPlice, Solas, SplAdder, SpliceR, SpliceSeq, SpliceTrap, SplicingCompass, SplicingTypesAnno, SplicingViewer, SpliCQ, StringTie, Suppa, SwitchSeq

How to choose one ? Benchmarking them !

Benchmarking

Softwares?

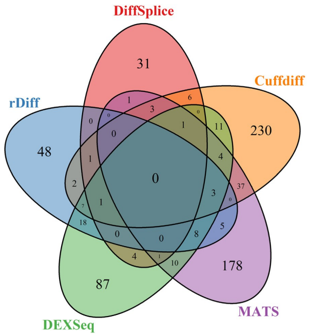
- the number of tested methods is limited

Data?

- RNAseq => how to know the truth?
  - qPCR studies? Only small set of genes & cross-hybridization between isoforms
  - RNA spikes? exogene sequences in controled quantity
- Simulated data => how to be as close as possible to the variability of the real data?

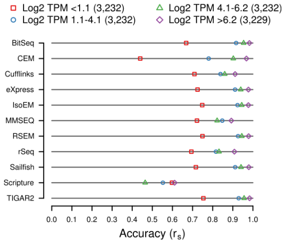
Task?

Some benchmarking tasks



DE tools

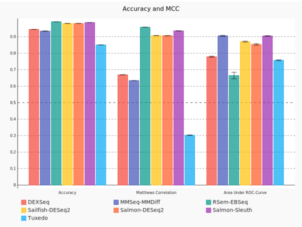
How many common DE genes?



Isoform quantification

Expression rate, Exon number/transcript

Number of isoform/gene, ASE type



ASE detection

Quantification: expression rate

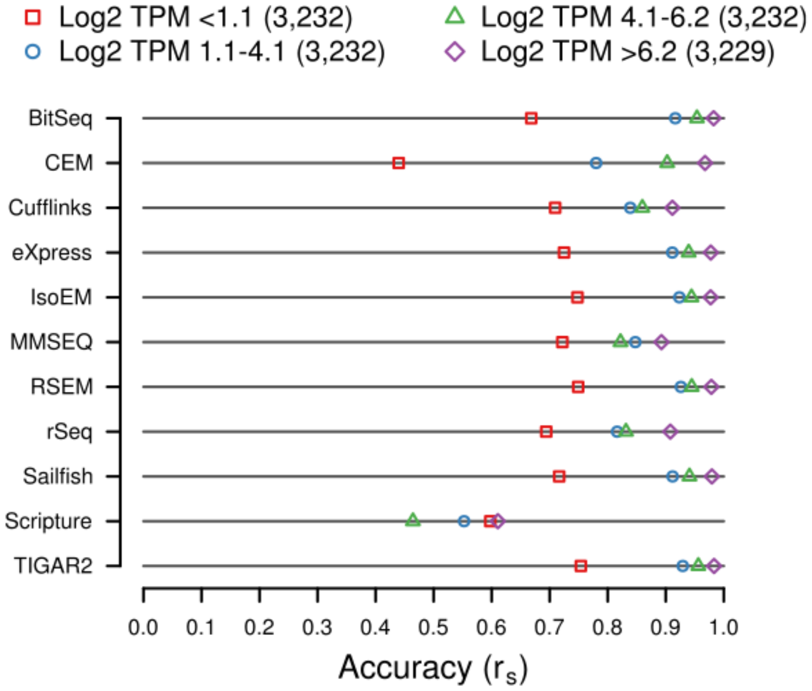
Simulation:
Flux simulator
Human data set
RNAseq single-end
sequencing depth: 30 million reads

restricted on expressed transcripts
(10% of human transcripts)

Spearman correlation coefficient (r_s) between the estimates and the known input levels

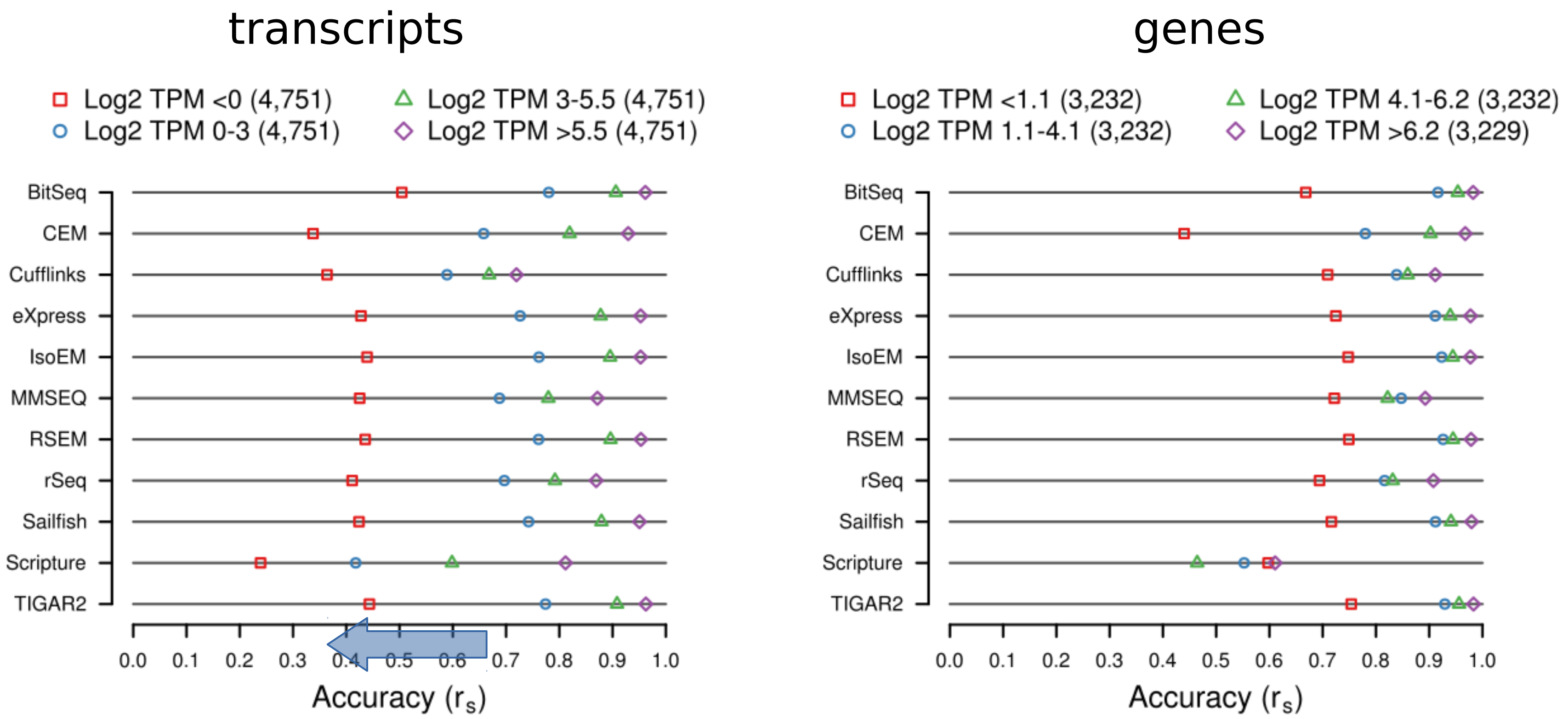
4 bins of expression levels
(Log2 TPM)

genes



Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 2015 Jul 23;16:150

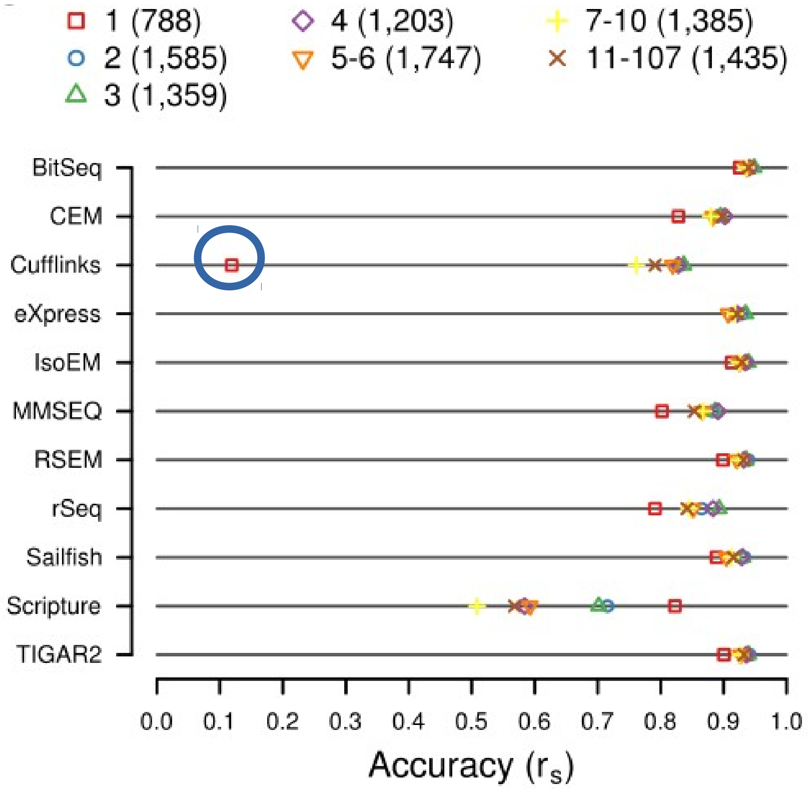
Quantification: expression rate



Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 2015 Jul 23;16:150

Quantification: exon number

#exons / transcript



Median expression levels:
0< Log2 TPM < 5.5

Cufflinks uses read-overlapping junction

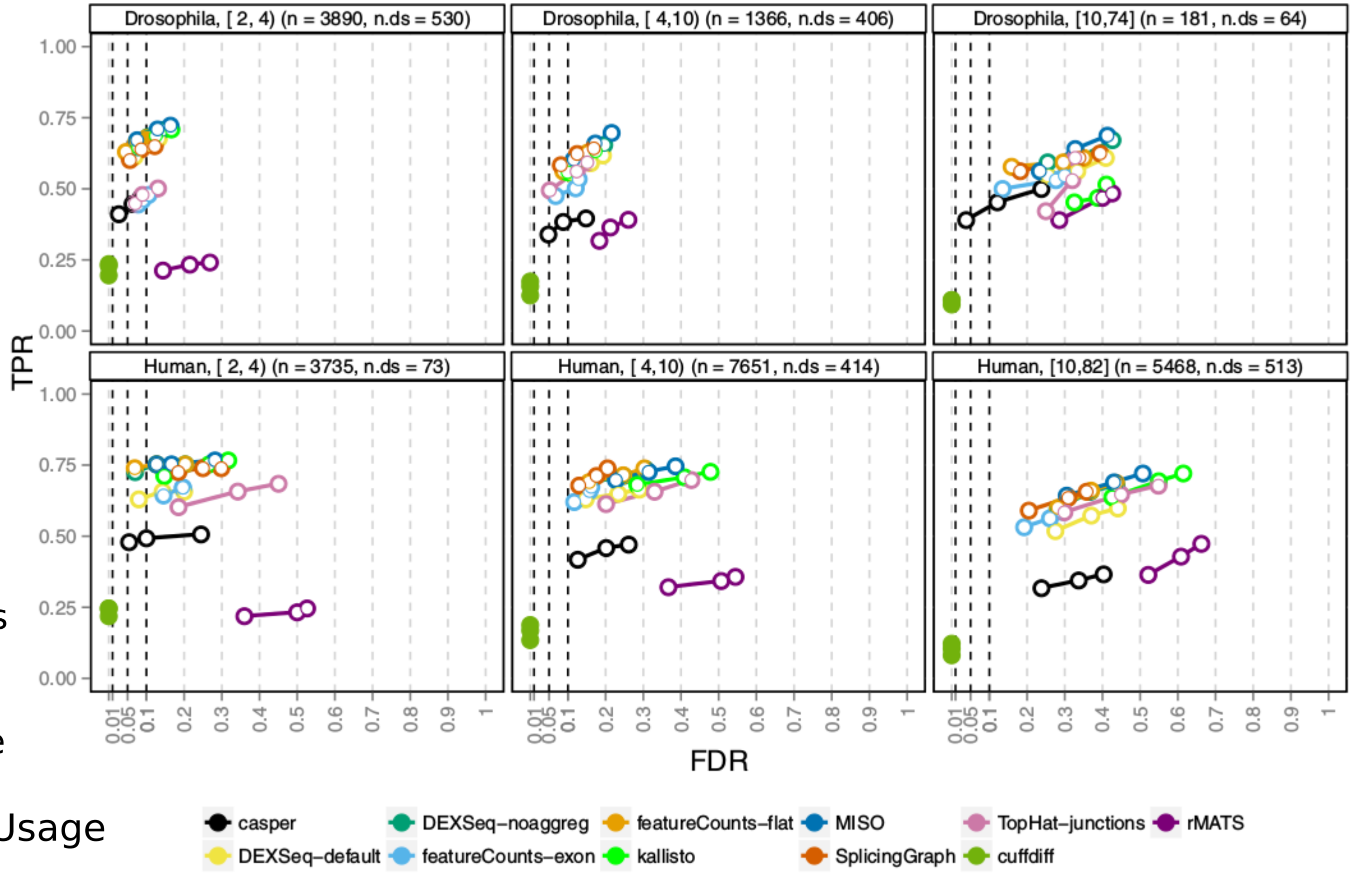
Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 2015 Jul 23;16:150

Quantification: isoform number

Simulation:
RSEM model from real data set + DTU for 1000 genes (switch of the relative abundances for the 2 most abundant isoforms between the conditions)

[i,j) i to j isoforms
n: gene number
n.ds: with DTU

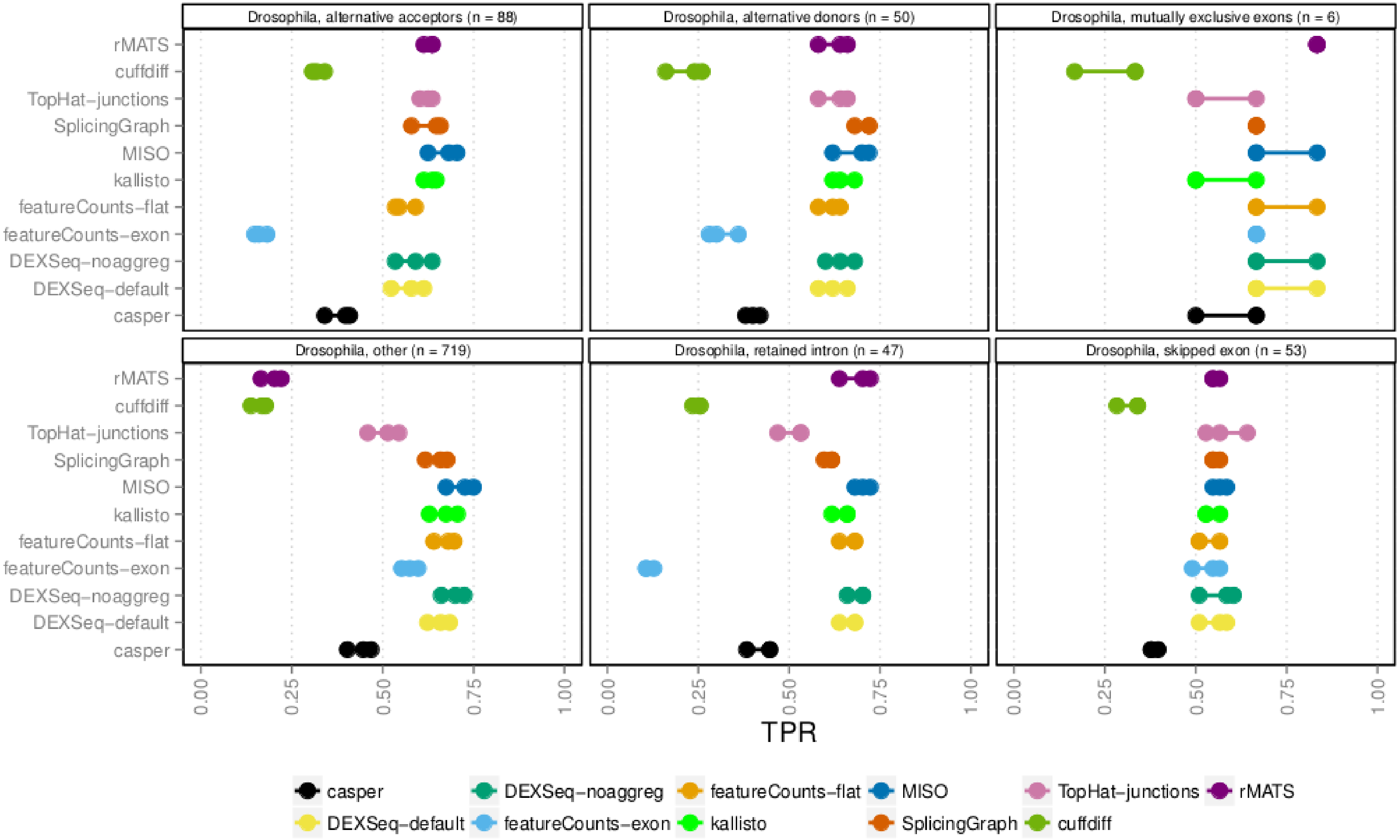
3 circles, usual FDR thresholds (0.01, 0.05, 0.1): ideally, each circle should fall to the left of the corresponding vertical line



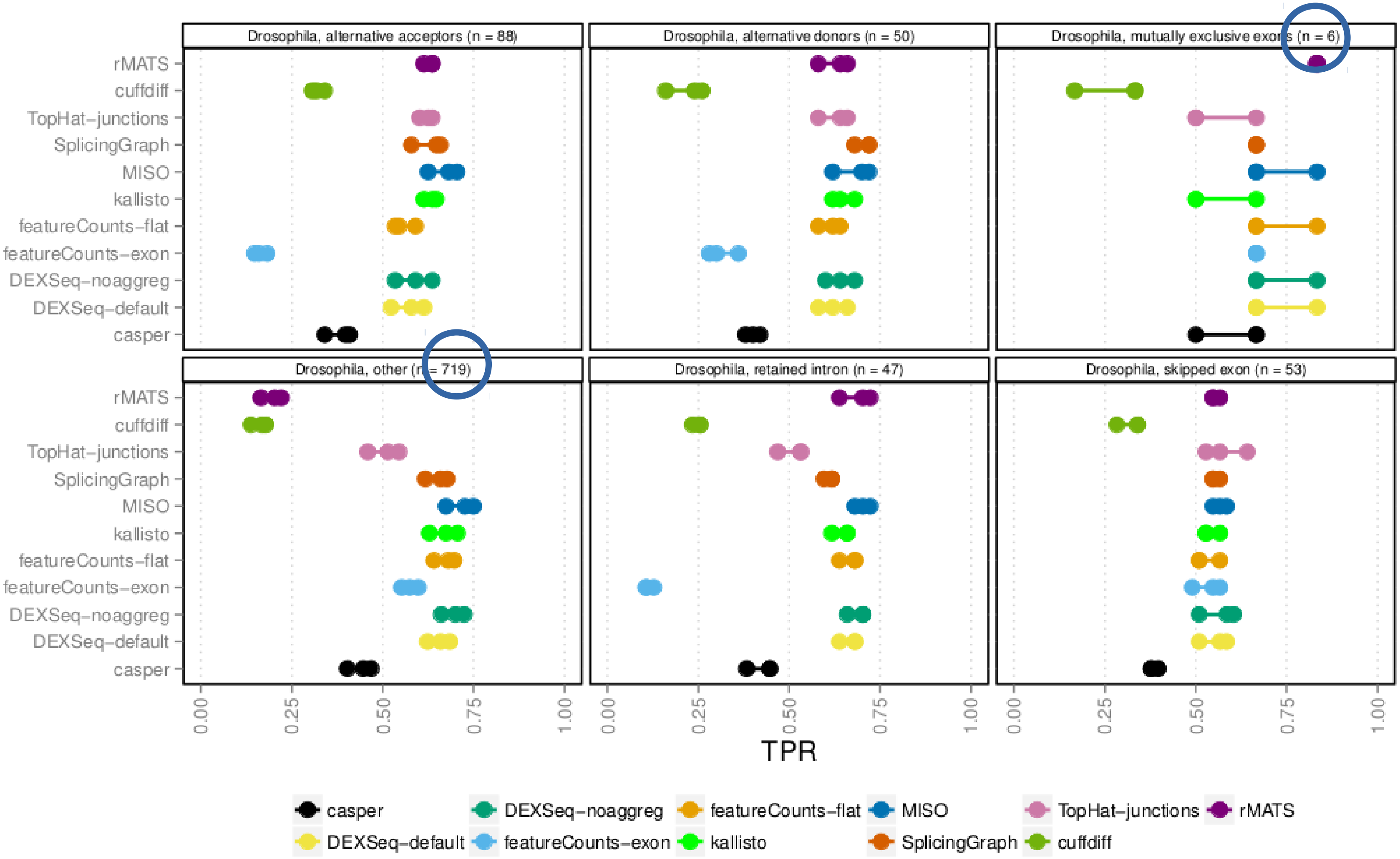
DTU: Differential Transcript Usage
FDR: False Discovery Rate
TPR: True Positive Rate

Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol*. 2016 Jan 26;17(1):12

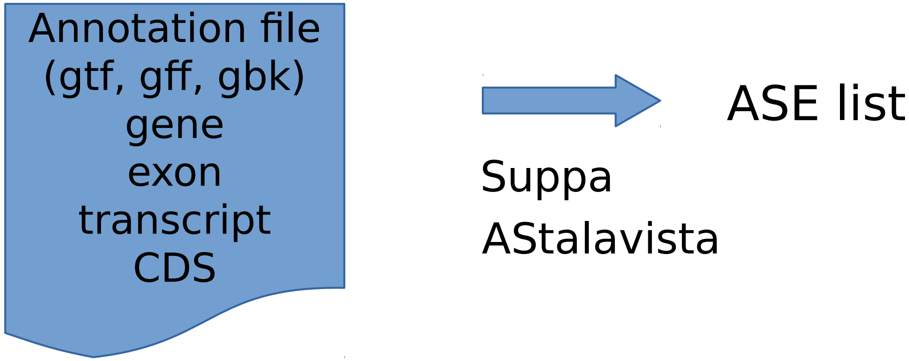
Quantification: ASE ?



Quantification: ASE ?



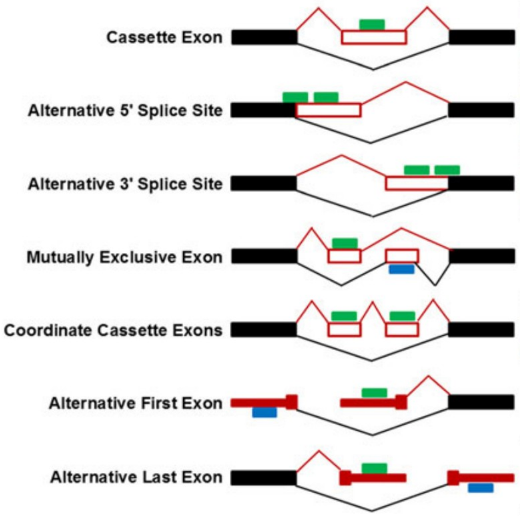
ASE analysis



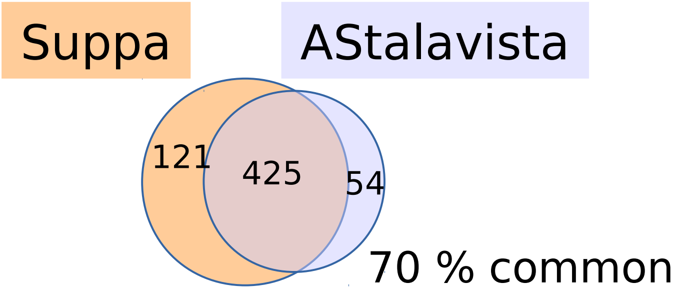
Annotation file
(gtf, gff, gbk)
gene
exon
transcript
CDS

Suppa
AStalavista

ASE list



chr22 hum.	Suppa	AStalavista
with ASE	539	479
without ASE	733	793
total	1272	1272

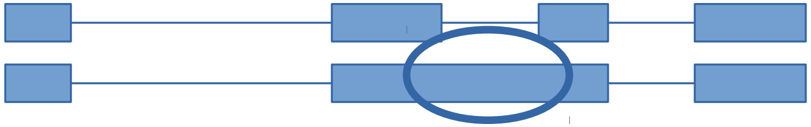


70 % common

Alamancos GP, Pagès A, Trincado JL, Bellora N, Eyras E. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*. 2015 Sep;21(9):1521-31
Foissac S, Sammeth M. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res*. 2007 Jul;35(Web Server issue):W297-9

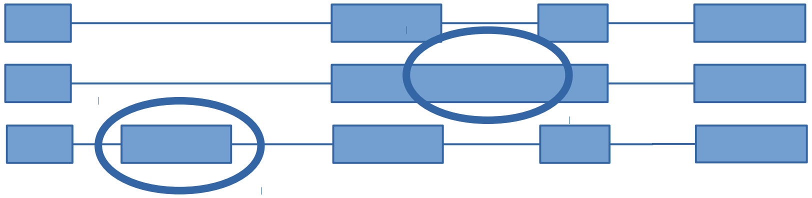
ASE analysis need reference

ASE analysis need reference



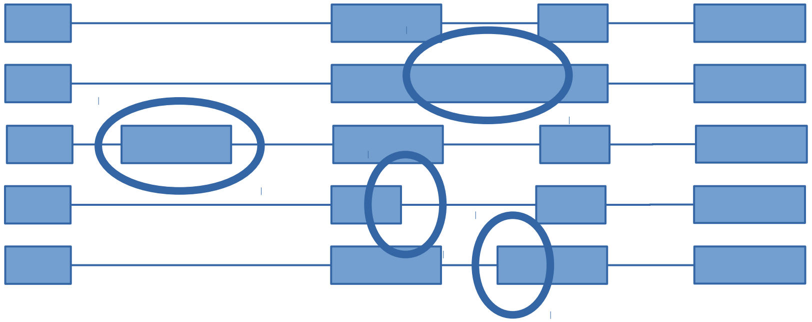
Retained Intron

ASE analysis need reference



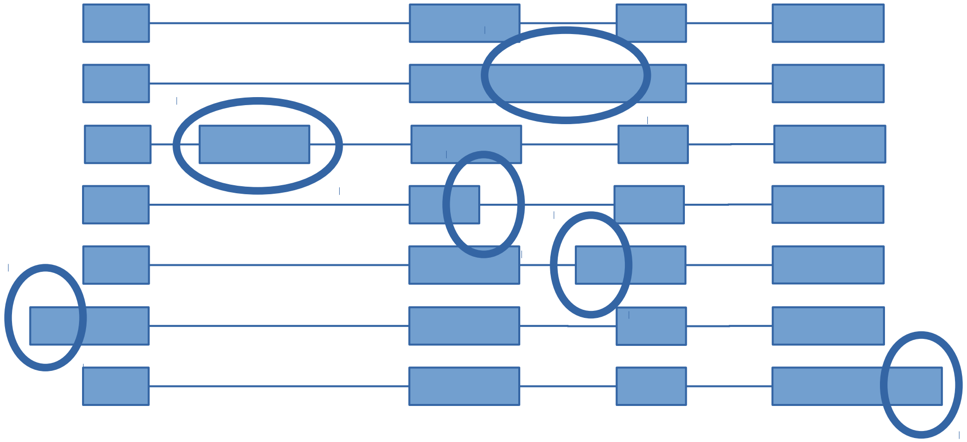
Retained Intron
Skipped Exon

ASE analysis need reference



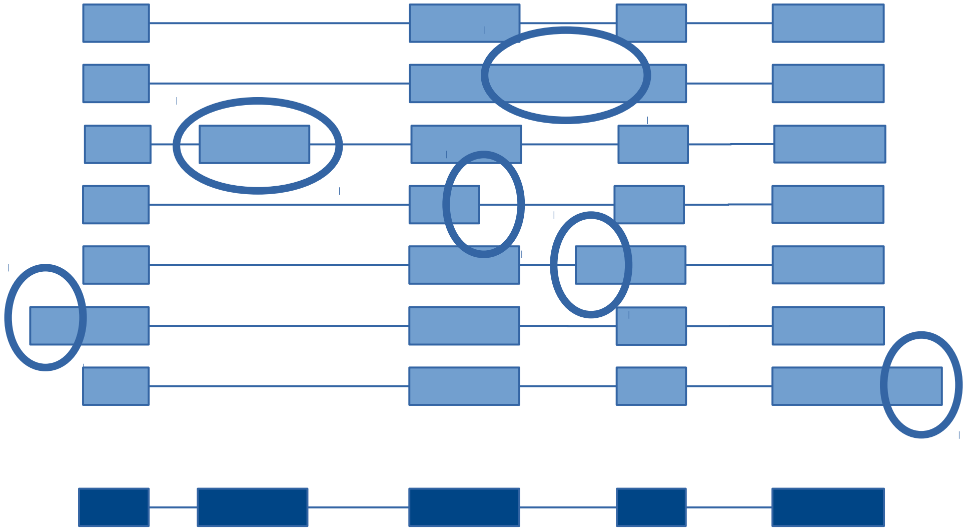
Retained Intron
Skipped Exon
Alternative Donnor/Acceptor
Splicing Site

ASE analysis need reference



Retained Intron
Skipped Exon
Alternative Donnor/Acceptor
Splicing Site
Alternative First/Last Exon

ASE analysis need reference



Retained Intron
Skipped Exon
Alternative Donnor/Acceptor
Splicing Site
Alternative First/Last Exon

Define a Reference Transcript :
1 - **largest set** of **non-overlapping** exons
2 - that appends the **most frequently** among isoforms
3 - that covers the **widest area** over the gene region

=> may be a non “real” transcript
=> specific for each project

Benchmark: ASE detection

Evaluate tools in their capacity to detect ASE from RNAseq data (neither the « right » rate of transcript expression, nor discovery of new expressed loci)

Simulated data: controled expression rate of each isoform & the presence of each type of ASE :

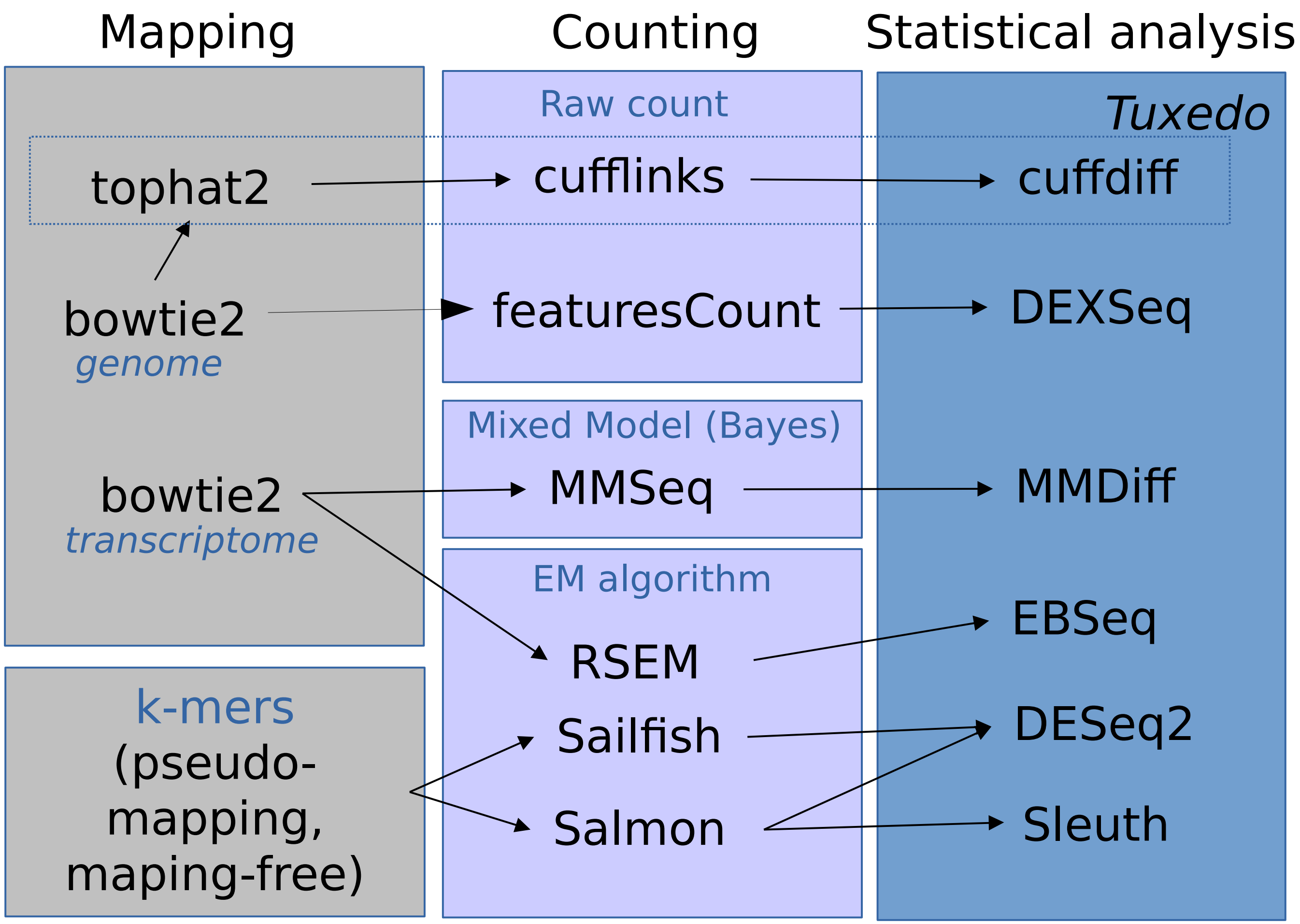
- human chromosome 22, 744 genes
- 2 conditions x 3 replicates
- reads : pairend-end, 2 x 100 bp
- expression : 100 reads / transcript (no DE variation)
- For each type of ASE :
  - 10% of the transcript in one condition
  - Only reference transcripts in the other condition
  - 1 ASE/transcript/gene

Condition 1
100% references

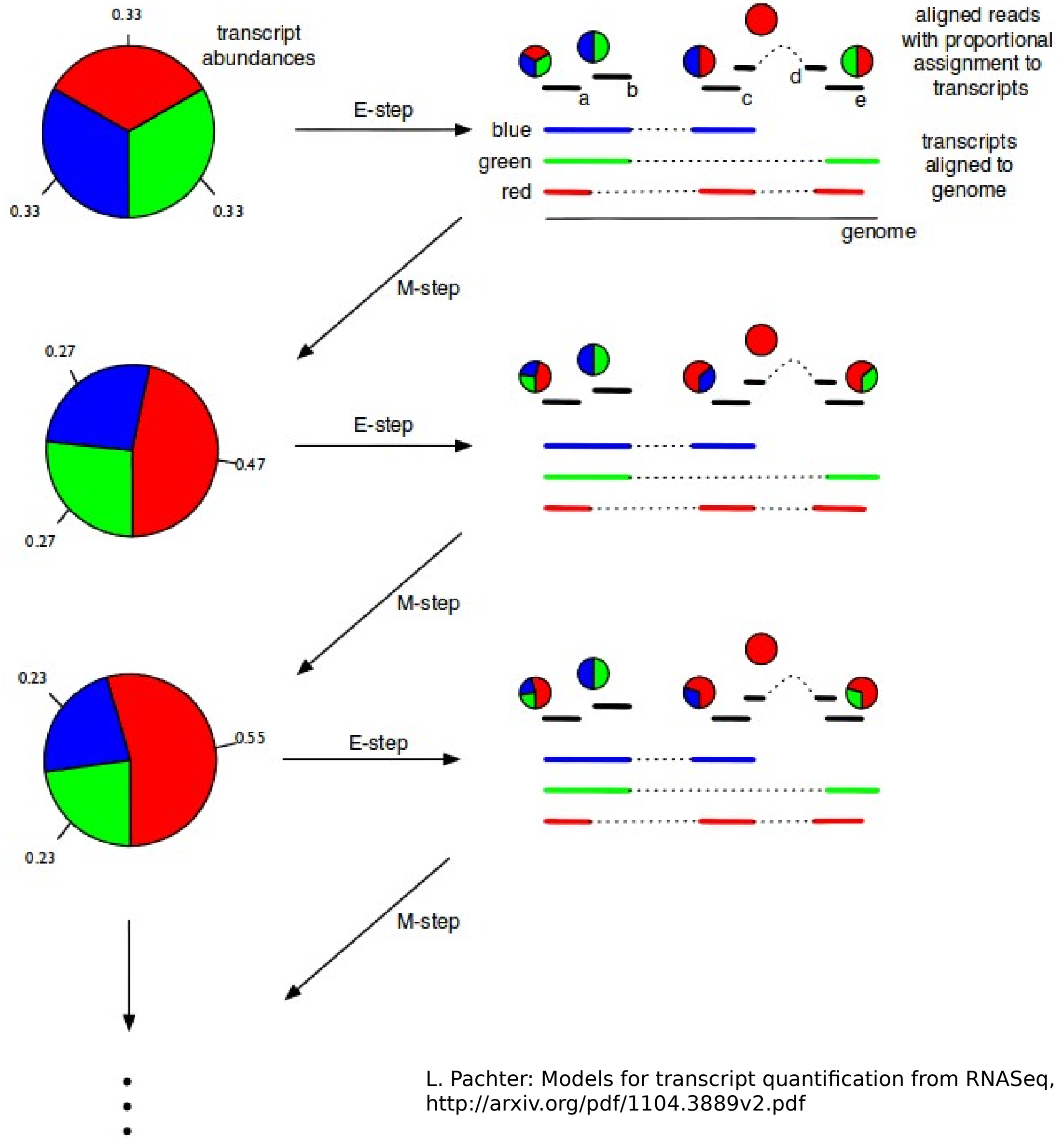
Condition 2
90% references
10% with 1 ASE

Identification of a DE transcript => the method detects the ASE type

Methods & tools benchmark



Count estimation with EM



L. Pachter: Models for transcript quantification from RNASeq, http://arxiv.org/pdf/1104.3889v2.pdf

Based-on the Expected-Maximisation algorithm

First 3 cycles of EM algorithm :

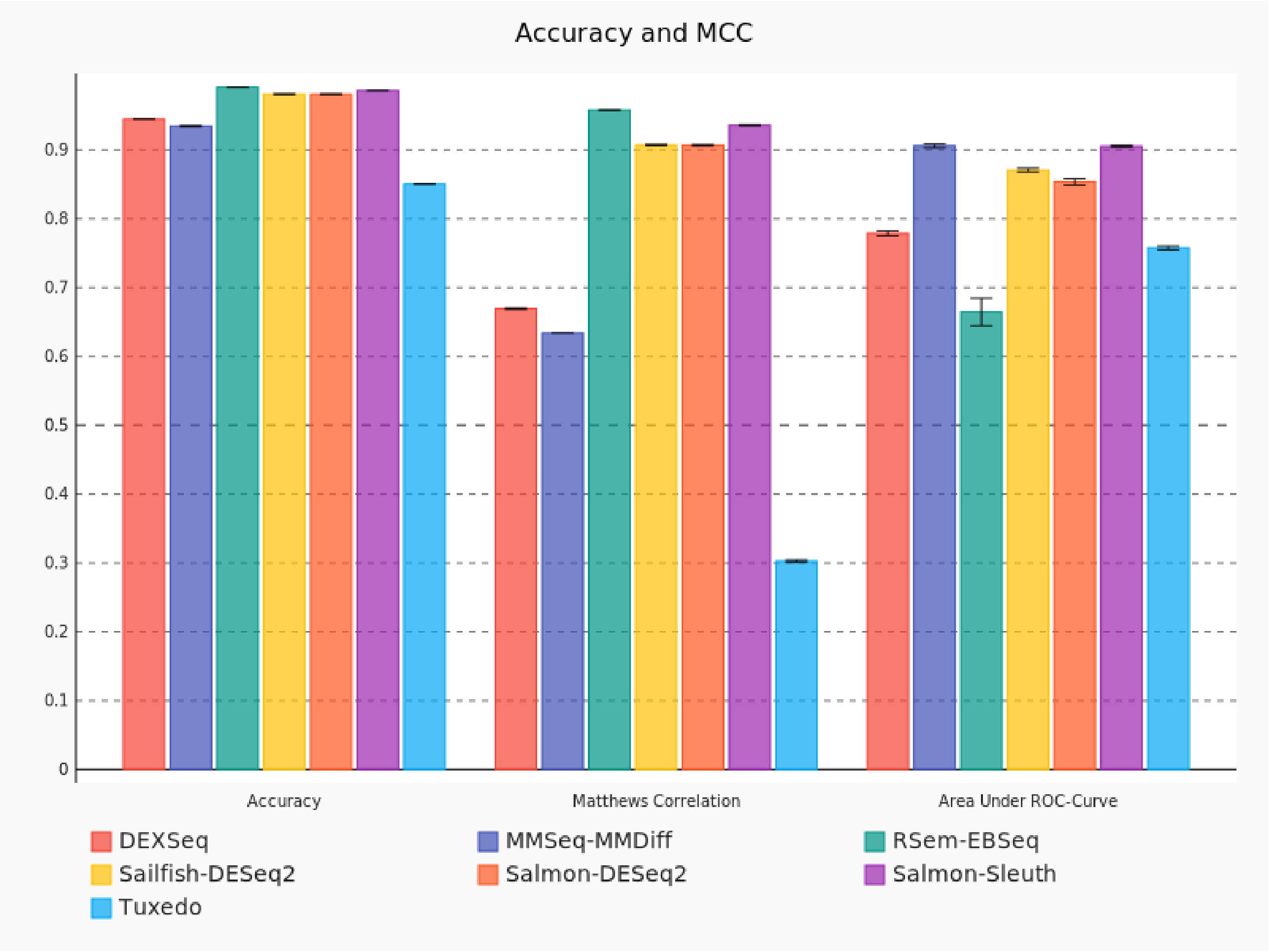
ex. : abundance of **red** transcipt estimated after the 1srt M-step: (1/3 read a + 1/2 read b + 1 read d + 1/2 read e)/(total read number)

or (0.33+0.5+1+0.5)/5 =0.47

- proved to converge

- stop criterion implementation: when all probabilities that a fragment is derived from a transcript ≥ 10^{7} have a relative change of ≤ than 10^{3}

Results, alternative donnor site



Accuracy: are tool predictions correct?

RSEM/EBSeq, Sailfish, Salmon

MCC: if >0.5 then using tool is better than random

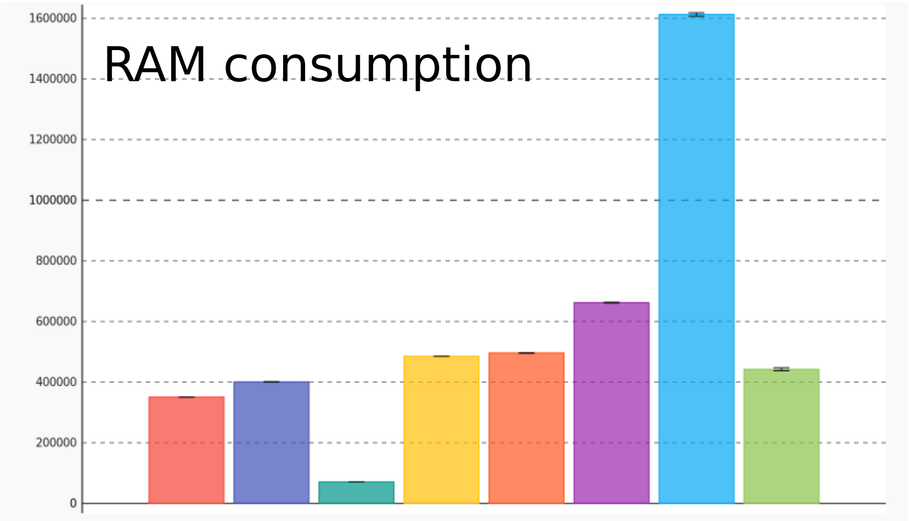
RSEM/EBSeq, Salmon-Sleuth
Tuxedo

AUC of ROC curve: confidence in the results of the tool

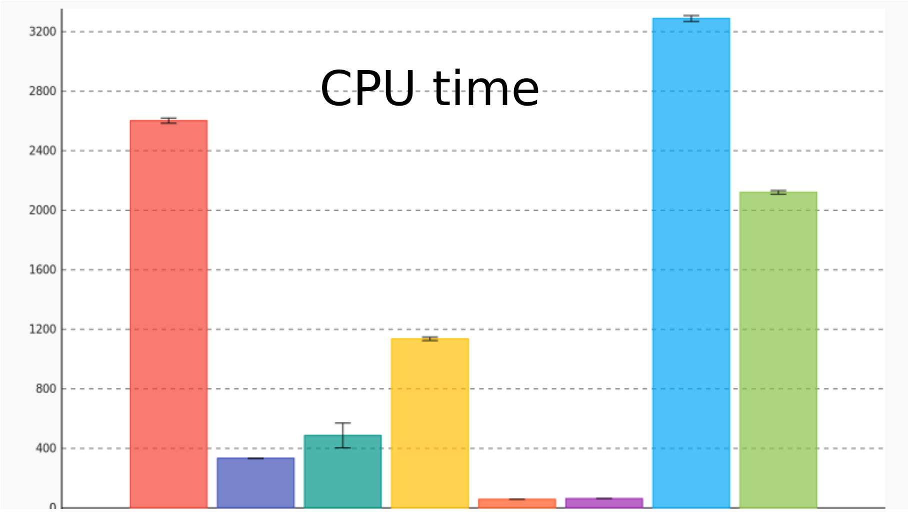
Salmon-Sleuth

Performaces (ADss)

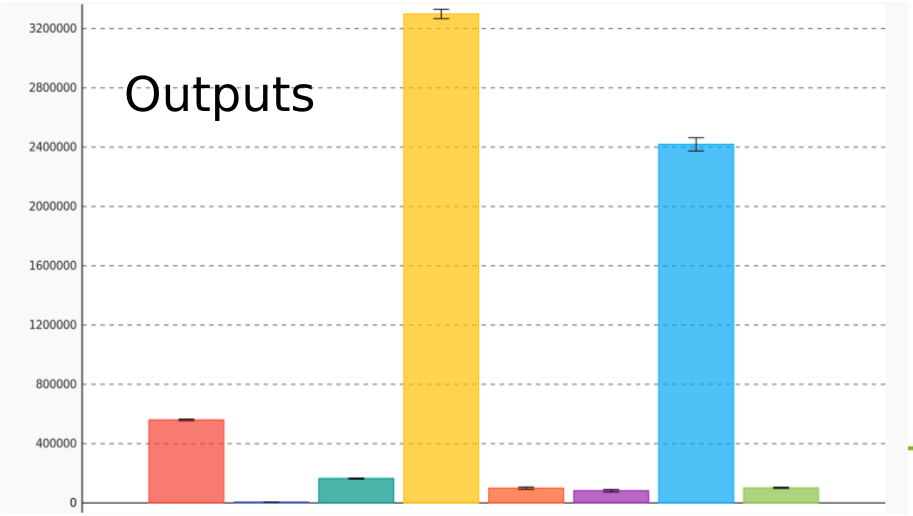
Max RAM usage (Ko)



CPU time (seconds)



Number of outputs



- ■ Bowtie2
- ■ DEXseq
- ■ MMSeq/MMDiff
- ■ RSEM/EBseq
- ■ Sailfish/DESeq2
- ■ Salmon/Sleuth
- ■ TopHat2
- ■ Cufflinks/CuffDiff

Tuxedo : bowtie2+Tophat2+cufflinks/cuffdiff

Choice of the method

For our benchmark

human chr22, 10 % of ON/OFF transcripts with 1 ASE between 2 conditions, 3 replicates, 100 reads/genes

Salmon/Sleuth

(RSEM/EBSeq, Sailfish)

developped pipeline: fastq ⇒ DE transcripts

appliance for IFB cloud, https://cloud.france-bioinformatique.fr

ASE Benchmarking conclusions

A benchmark is always limited:

- Software

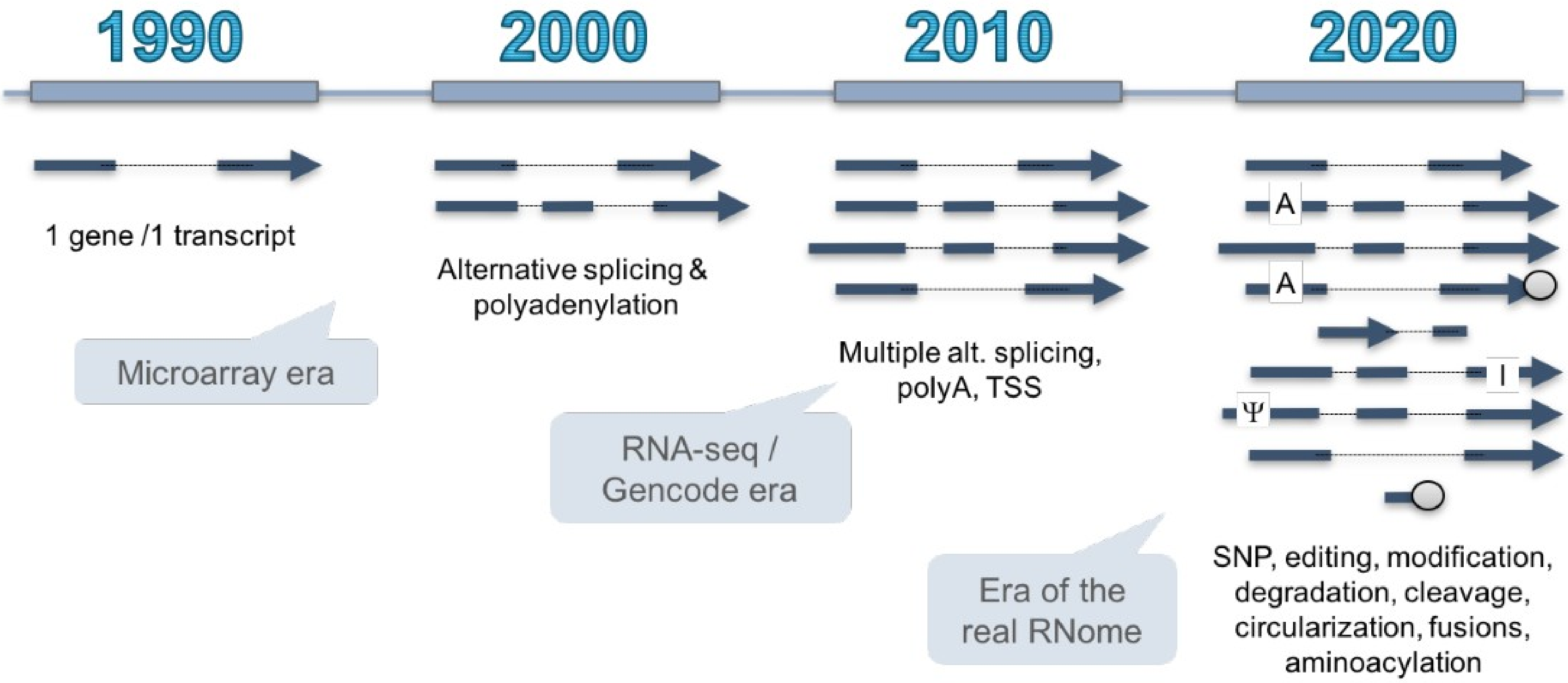
- Simulation (human, plant, …)
  - Skipped exon, Retained Intron, Alternative First/Last Exon, Alternative 3’/5’ Splicing Site
  - 100 reads / gene (3-4 days, 6 cores, 13 G RAM)
    - 1000 reads / gene (to check if the wrong predictions result from the coverage deepness or from the algorithm)
  - On/off condition : 1 ASE / transcript / gene :
    - more than 1 ASE / transcript
    - combination of ASE in the same transcript

Several benchmark studies should be conducted to have a global overview

The conclusions should be regularly update

RNomics evolution

Part1-smallRNA: search for the method corresponding to the smallRNA
Part2-isoform level: will be easier with full length RNAseq technology


1990
2000
2010
2020
1 gene /1 transcript
Alternative splicing & polyadenylation
Multiple alt. splicing, polyA, TSS
A
A
I
Ψ
SNP, editing, modification, degradation, cleavage, circularization, fusions, aminoacylation
Microarray era
RNA-seq / Gencode era
Era of the real RNome

The more there is technical advances, the more we are going towards the unknown biology

Thanks

Marie-Laure Martin-Magniette
Etienne Delannoy
Véronique Brunaud

AVIESAN-IFB

Eukaryotic small RNA
P. Bardou, C. Gaspin, S. Maman, J. Mariette, O. Rué, M. Zytnicki
http://www.france-bioinformatique.fr/sites/default/files/sRNA-Seq.pdf

Isoforms

Bioinformatics and Biostatistics HUB
Marie-Agnès Dillies
Rachel Legendre
Hugo Varet

The SSFA team:
Daniel Gautheret
Fabrice Leclerc
Jean Lehmann

Coline Billerey
Thibault Dayris
Marc Gabriel