State of the art of what can be done with RNA-seq

Marc DELOGER

Gustave Roussy - UMS AMMICA (US 23 INSERM / UMS 3655 CNRS)

DU Bii - March 11th 2020

Outline

- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Outline

- **Experimental and sequencing design**
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Experimental and sequencing design

The sequencing of mRNA can be used to address many different biological questions : expression, alternative splicing, RNA editing, fusion gene, repeats profiling, etc...

These questions have to be clearly defined **BEFORE** running the sequencing part of the project => Experiment "design"

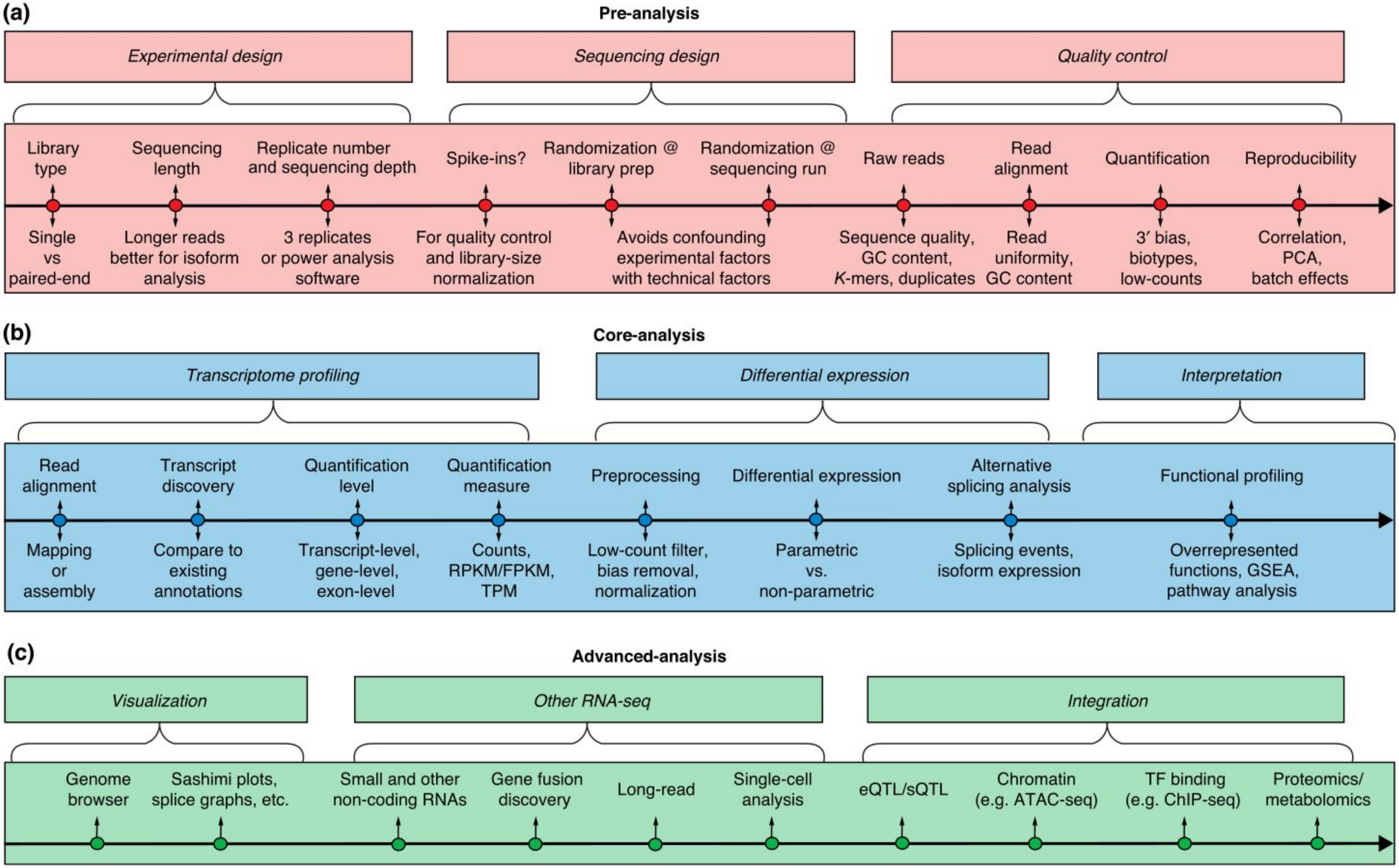
For each of them, **dedicated** sequencing "design" and bioinformatics pipeline should be used

Experimental and sequencing design

- Citation 1 : "To consult a statistician after an experiment is finished is often merely to ask him to conduct **a post-mortem examination**. He can perhaps say what the experiment died of." (Ronald A. Fisher, Indian Statistical Congress, 1938, vol. 4, p 17)

- Citation 2 : “While **a good design does not guarantee a successful experiment**, a suitably bad design guarantees a failed experiment” (Kathleen Kerr, Atelier Inserm 145, 2003)

Experimental and sequencing design



Conesa et al. 2016. Genome Biology, 17:13

Experimental and sequencing design

- Samples collection :
    - Number of replicates per condition : achieve enough statistical power to control experiment variability, and so, to be able to answer to your biological question

Replicates : what does that mean ?

- Replicates are mandatory to estimate the biological variability
- The higher the better !
- A biological replicate is not a technical replicate :
  - Technical = Several extractions of the same RNA or Several libraries built from the same RNA extraction or A library sequenced several times

(…) With three biological replicates, nine of the 11 tools evaluated found only 20%–40% of the significantly differentially expressed (SDE) genes identified with the full set of 42 clean replicates (…)

Recommandations for RNA-seq experimental designs :

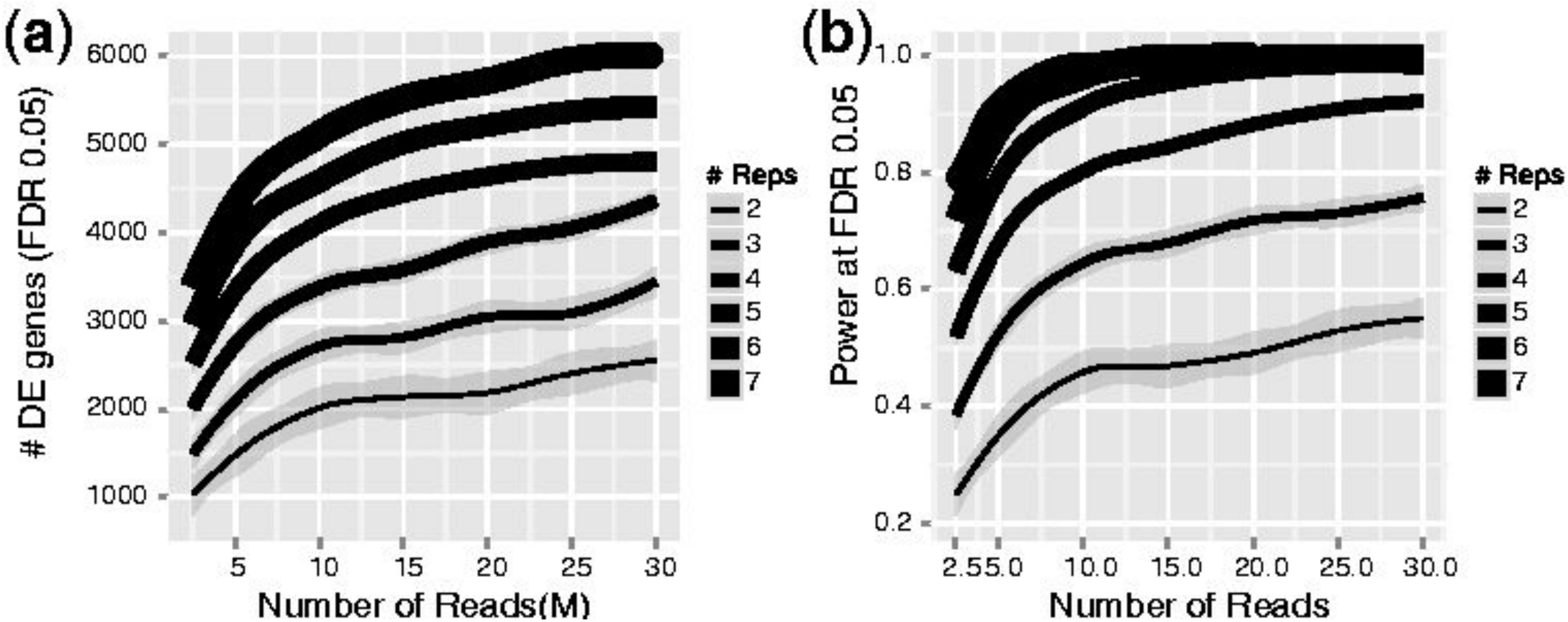
At least 6 replicates per condition for all experiments.

At least 12 replicates per condition for experiments where identifying the majority of all DE genes is important.

(…)

Schurch et al. 2016. RNA, 22(6)

More reads or more replicates



- (…) Increase in biological replication significantly increases the number of DE genes identified (…)
- (…) Power of detecting DE genes increases with both sequencing depth and biological replication (…)

Liu al. 2014. Bioinformatics, 17:13

Experimental and sequencing design

- Samples collection :
    - Number of replicates per condition : achieve enough statistical power to control experiment variability, and so, to be able to answer to your biological question
    - Avoid confounding effects : eg. all the samples for condition 1 processed by technician A (or in year/gender A) and all the samples for condition 2 processed by technician B (or in year/gender B)

Experimental and sequencing design

- Library preparation :
  - Poly-A selection
    - High quantity/quality of input RNA (100-300ng and RIN > 6)
    - Loss of all non-polyadenylated features (majority of non-coding and small RNAs)
  - rRNA depletion
    - Variable depletion procedures efficiency (Petrova OE et al. Sci Rep. 2017)
    - More expensive (+33%)
  - Small RNA / miRNA
  - Unstranded / Stranded protocol => strand-effect control
  - UMI (Unique Molecule Identifier) => PCR-effect control
  - Spike-ins => sensitivity and accuracy of RNA-seq experiments for transcriptome discovery and quantification across different samples (eg. short and/or rare transcripts)

Zhao S. et al. BMC Genomics. 2015 + Kivioja T. et al. Nat. Methods. 2012 + Jiang L. et al. Genome Res. 2011

Experimental and sequencing design

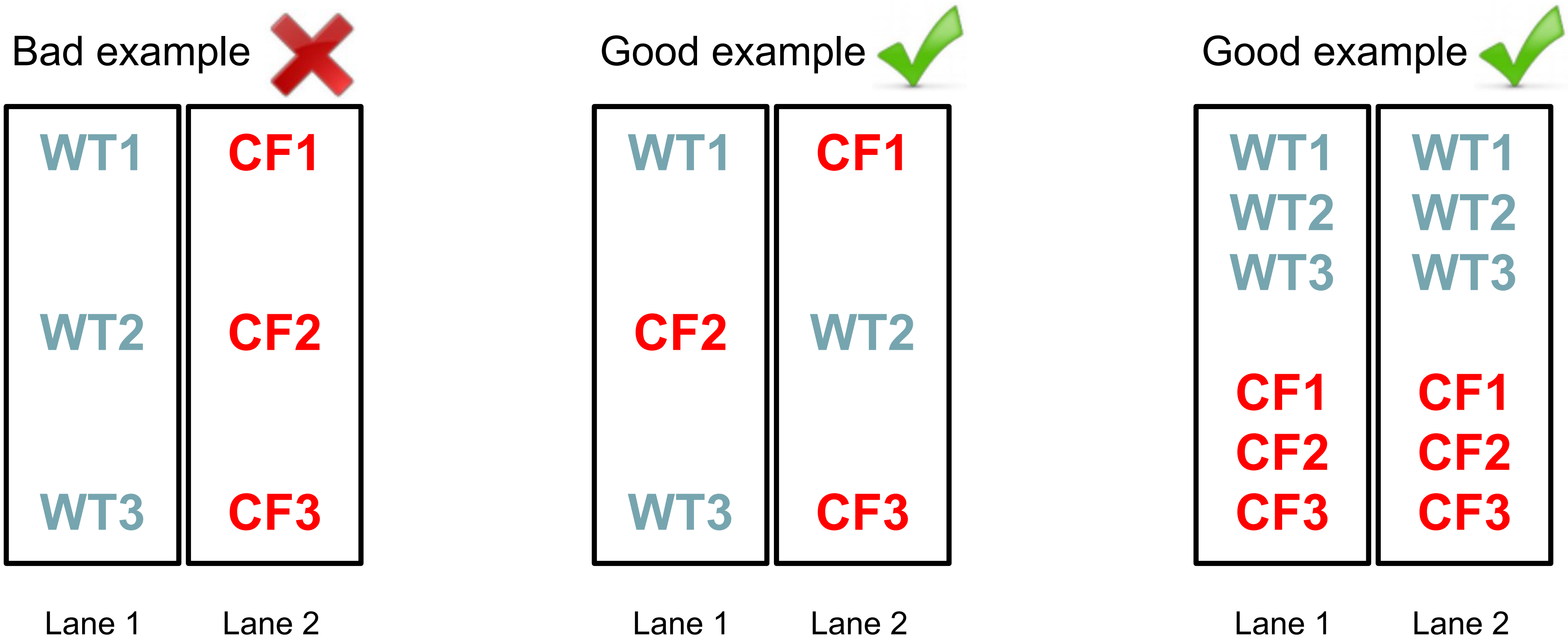
- Sequencing :

  - Paired-End : standard for years now
  - Single-End : no more used except for gene expression only because sequencing price / 2 compared to paired-end and with highly similar gene counts (Pearson R^2=0.9792) **so you can multiply by 2 the number of replicates at same price ;-)**
  - Reads length : 50bp for gene expression, >75bp for the rest
  - Insert length : generally ~300bp fragments, the higher the better for splicing/fusion event discovery
  - Depth : single-cell (1M reads or 50K=highly expressed or 20K=cell types), 30M fragments for gene expression, >50/75M fragments for the rest (75M pairs => 150M reads)
  - Run design

Run design

Goal:

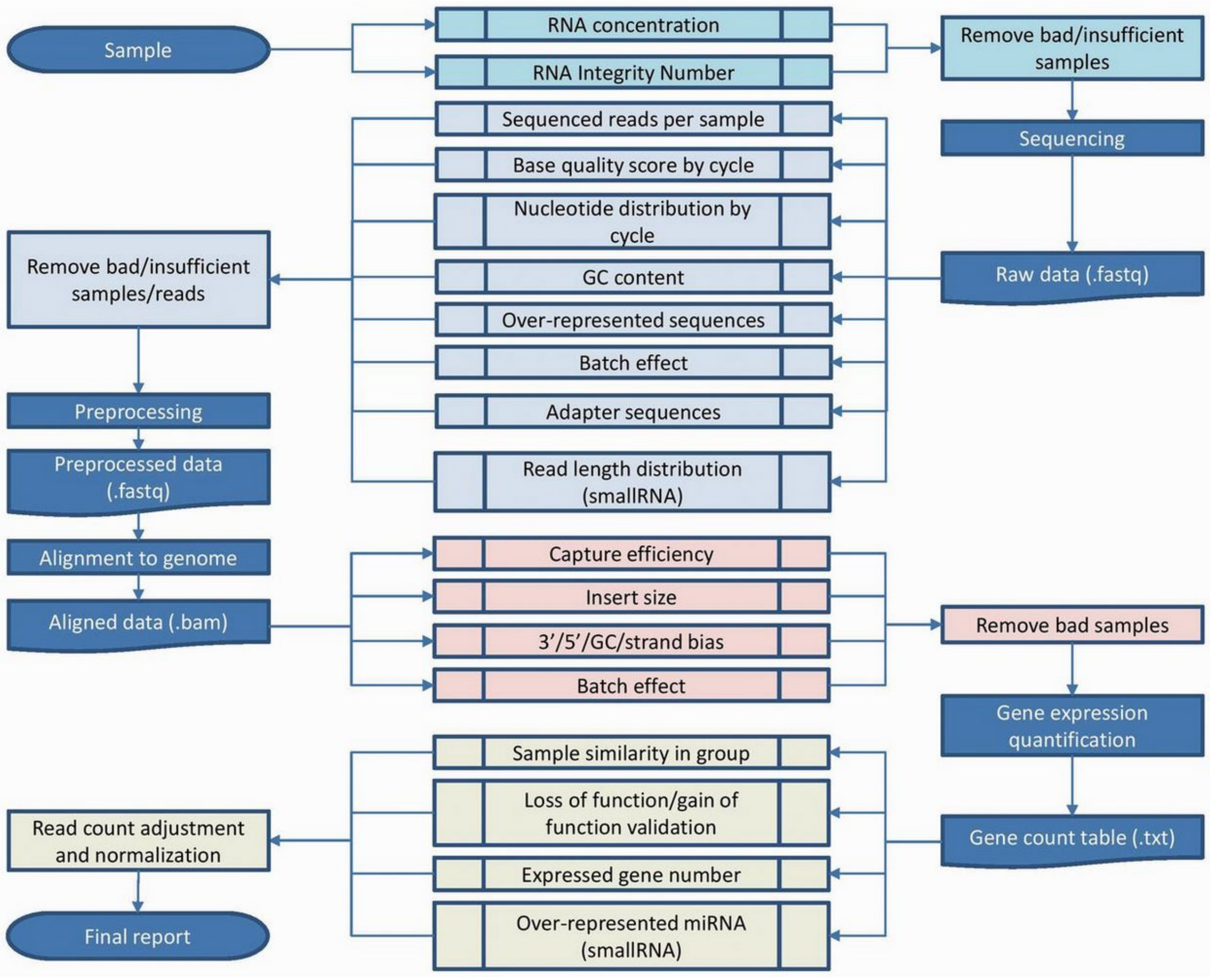
Do not add any confounding technical effect (day, lane, run, etc…) to the factor of interest.



Outline

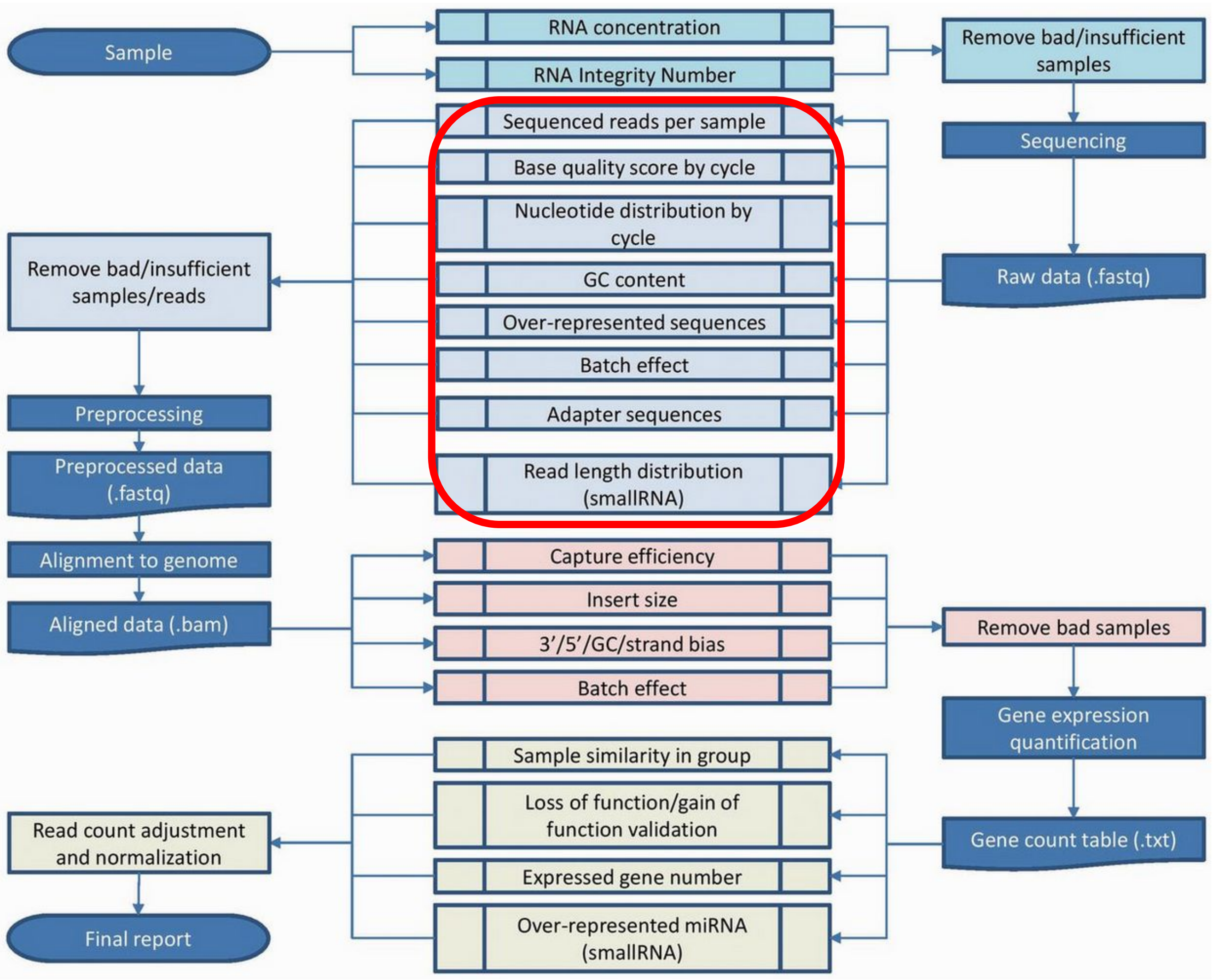
- Experimental and sequencing design
- **Quality control and mapping**
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Quality control and mapping



Sheng Q. et al. Brief Funct Genomics. 2017

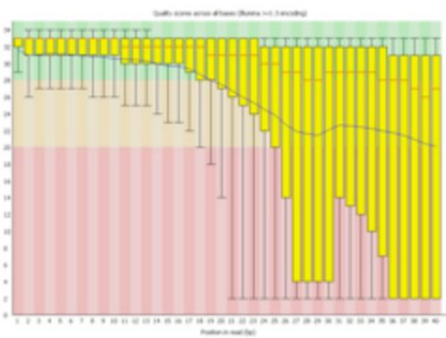
Quality control and mapping



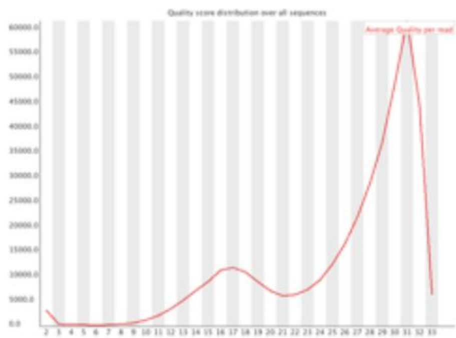
Sheng Q. et al. Brief Funct Genomics. 2017

Quality control and mapping

Per Base Sequence Quality

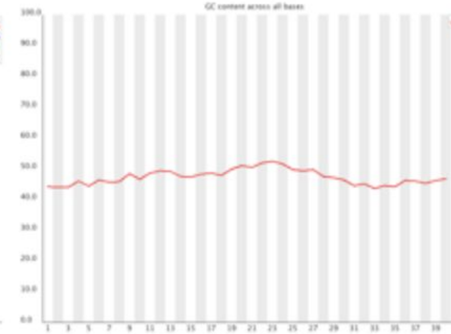


Per Sequence Quality Scores

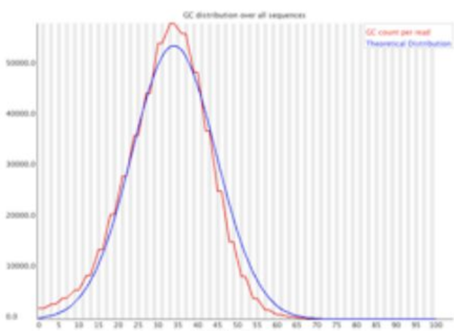


Per Base Sequence Content

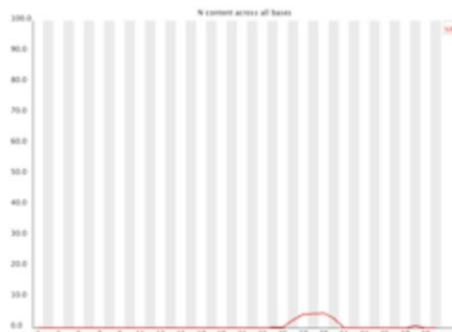
Per Base GC Content



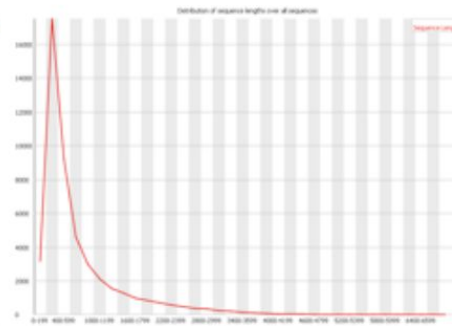
Per Sequence GC Content



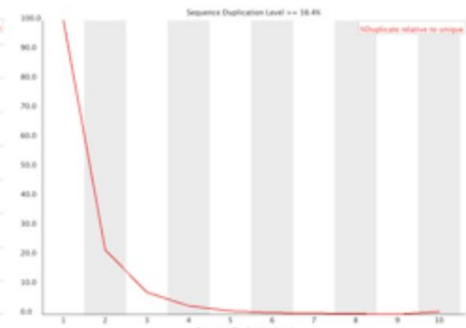
Per Base N Content



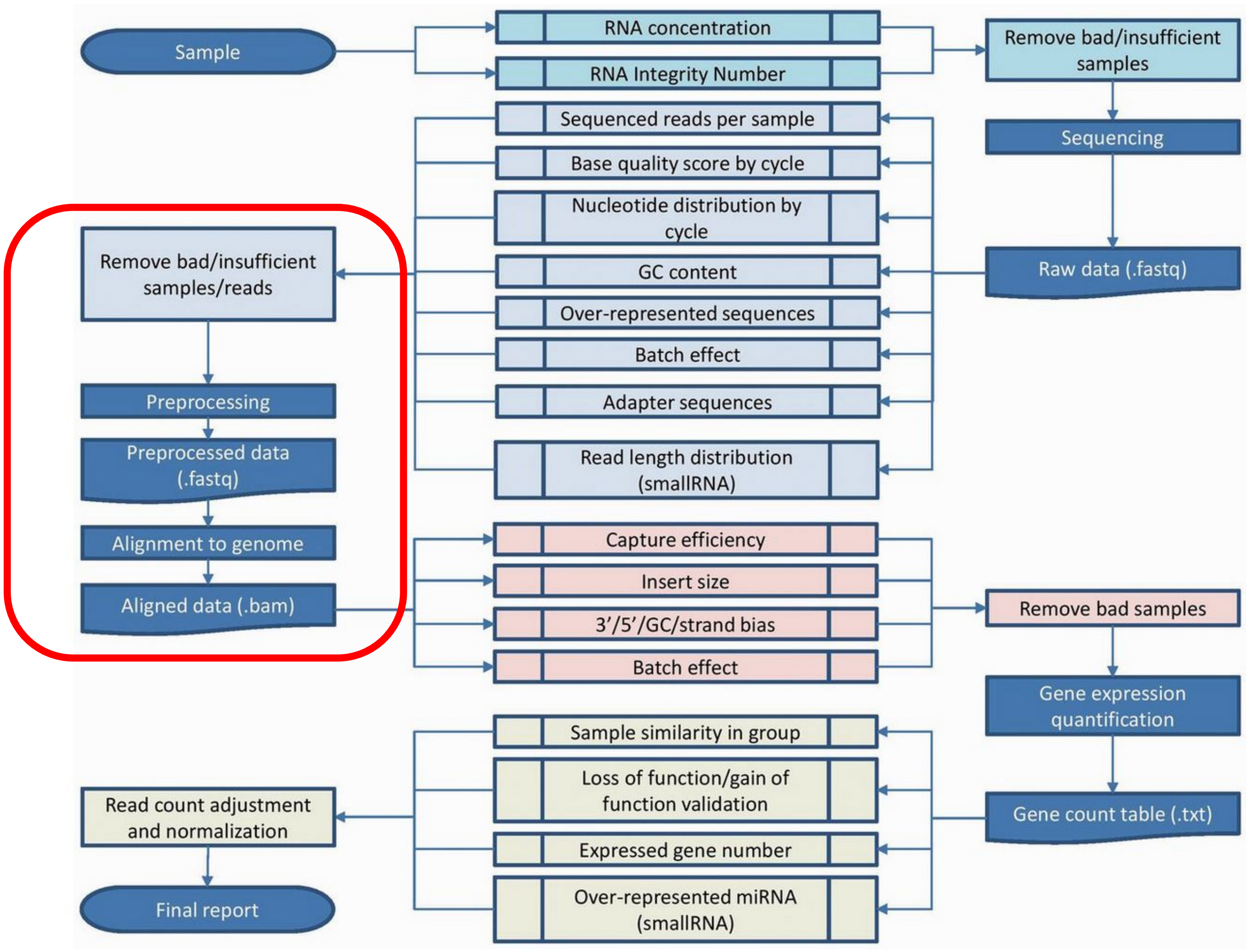
Sequence Length Distribution



Duplicate Sequences

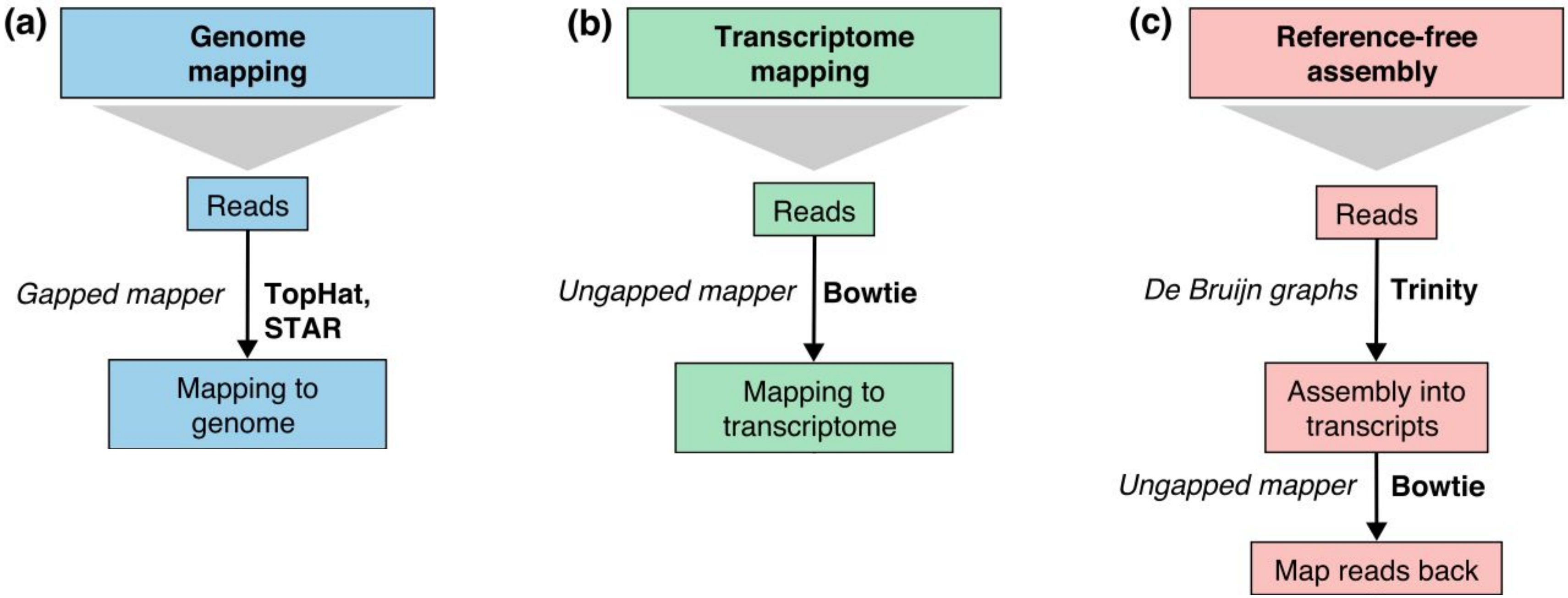


Quality control and mapping



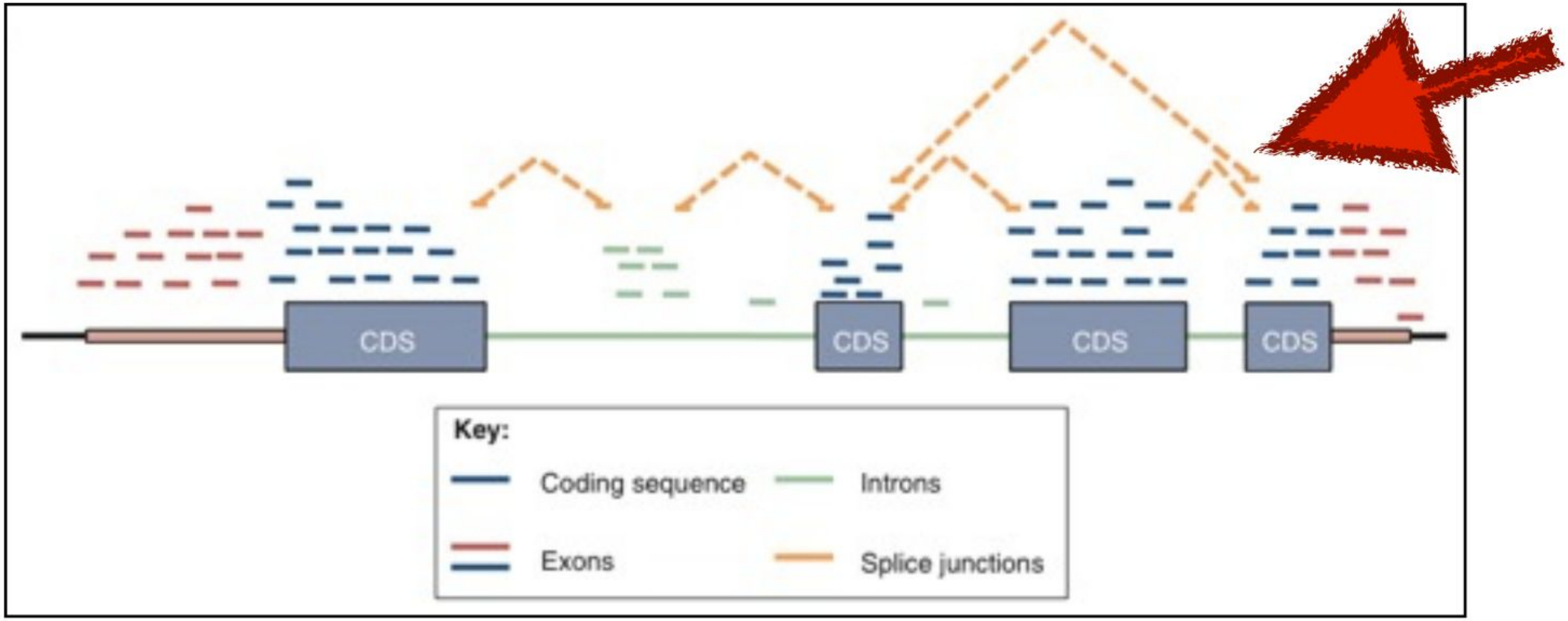
Sheng Q. et al. Brief Funct Genomics. 2017

Quality control and mapping



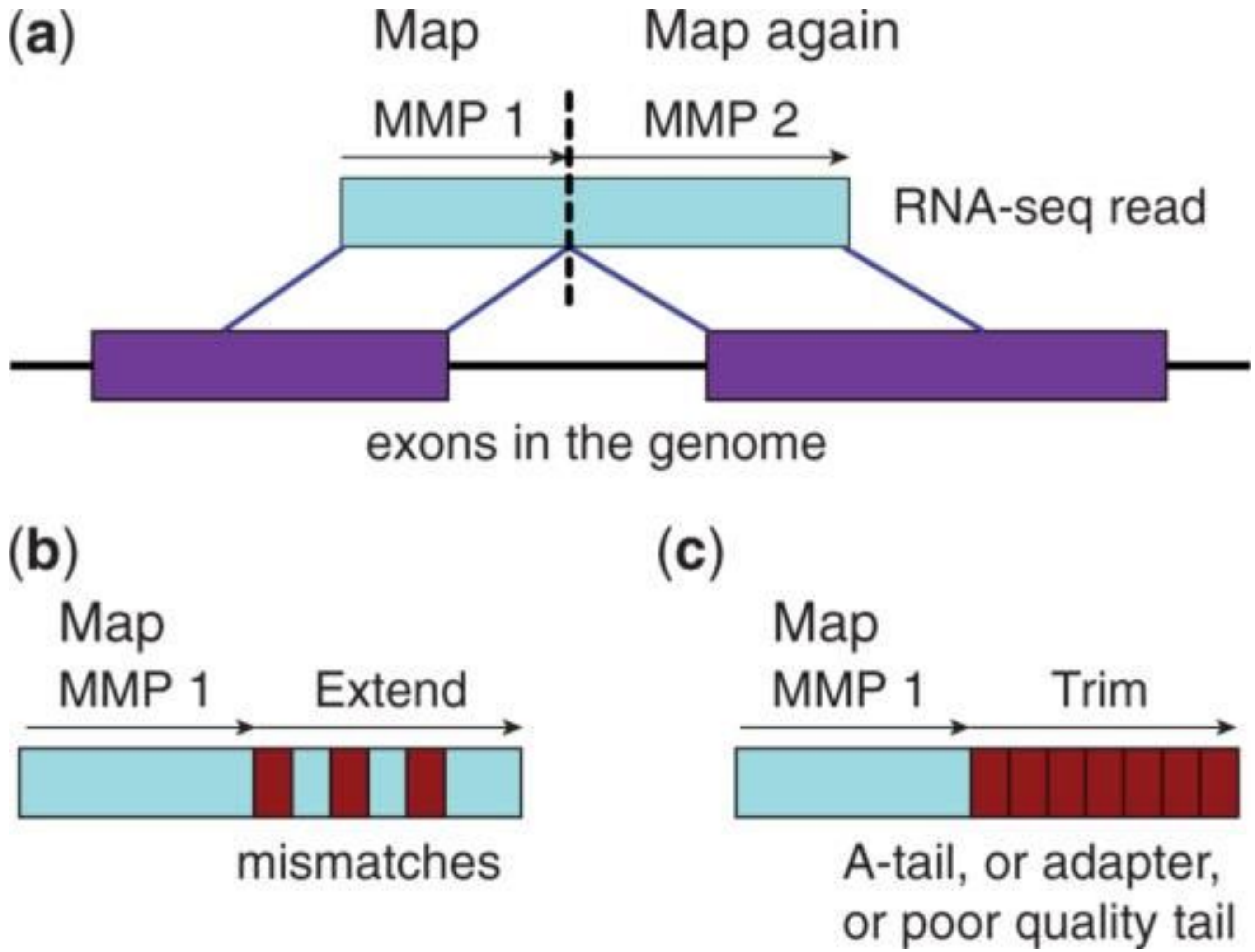
Conesa et al. 2016. Genome Biology, 17:13

Quality control and mapping



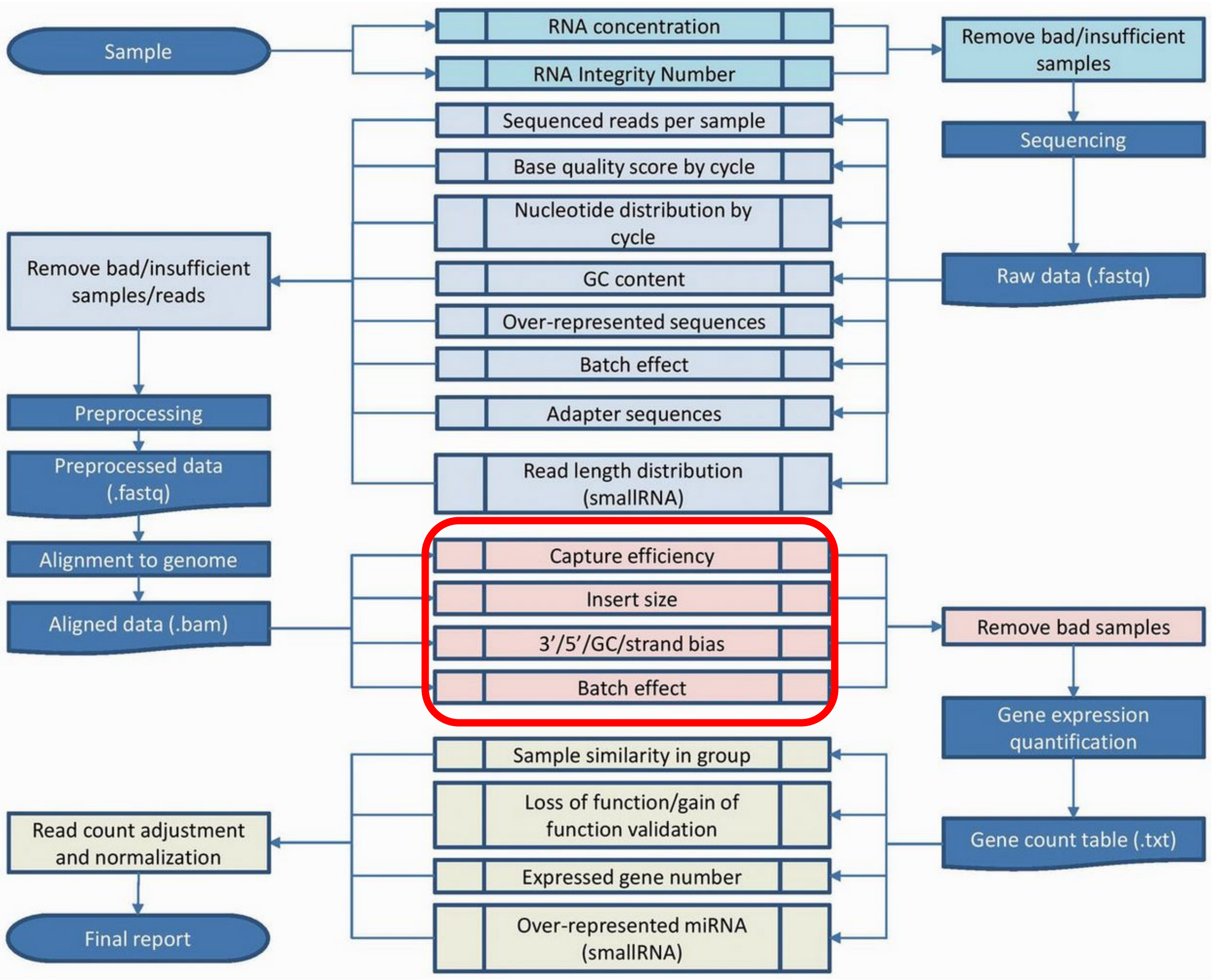
Trapnell C. Nat Biotechnol. 2009

Quality control and mapping



Alexander Dobin et al. Bioinformatics 2013;29:15-21

Quality control and mapping



Sheng Q. et al. Brief Funct Genomics. 2017

Quality control and mapping

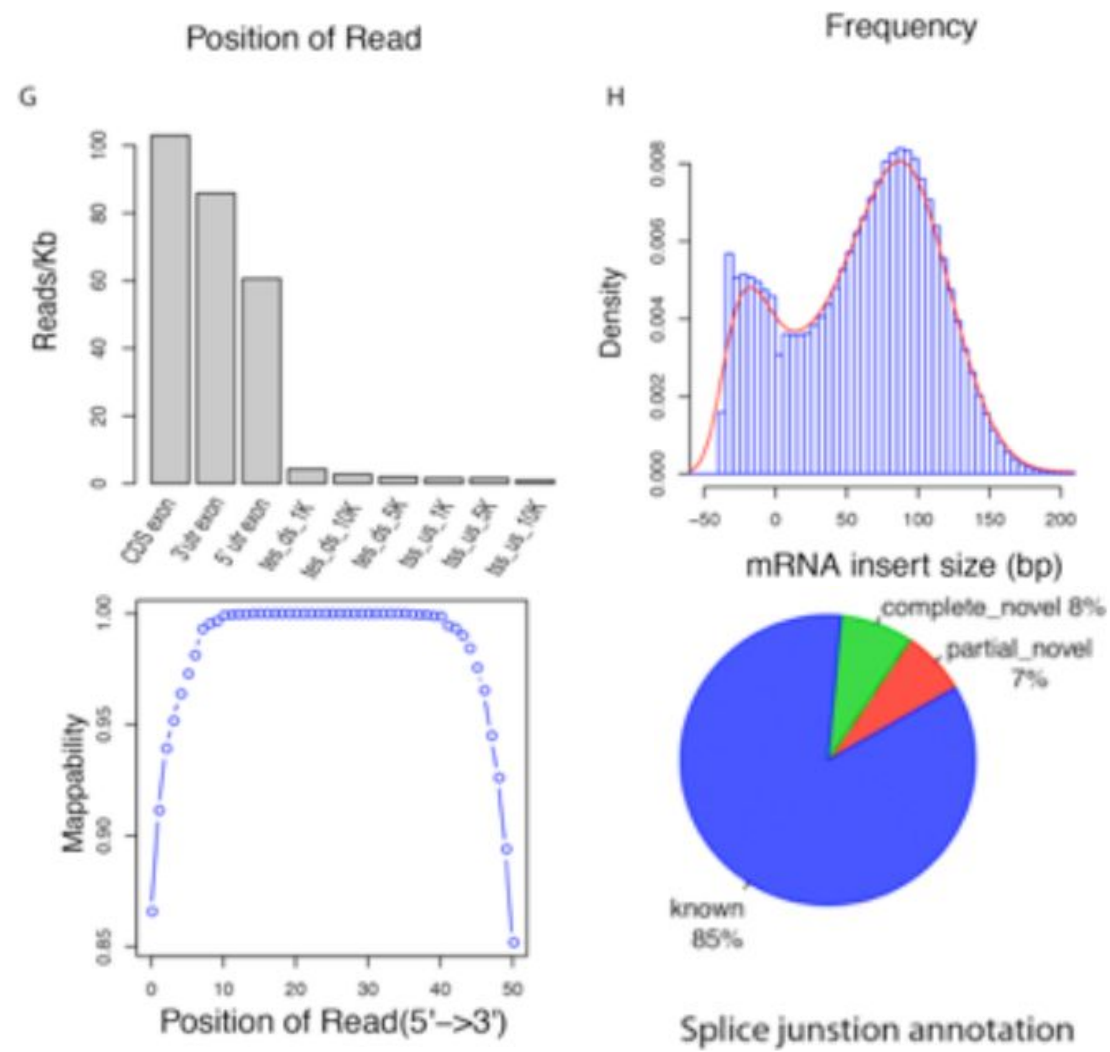
Number of input reads
Average input read length

UNIQUE READS:
Uniquely mapped reads number
Uniquely mapped reads %
Average mapped length
Number of splices: Total
Number of splices: Annotated (sjdb)
Number of splices: GT/AG
Number of splices: GC/AG
Number of splices: AT/AC
Number of splices: Non-canonical
Mismatch rate per base, %
Deletion rate per base
Deletion average length
Insertion rate per base
Insertion average length

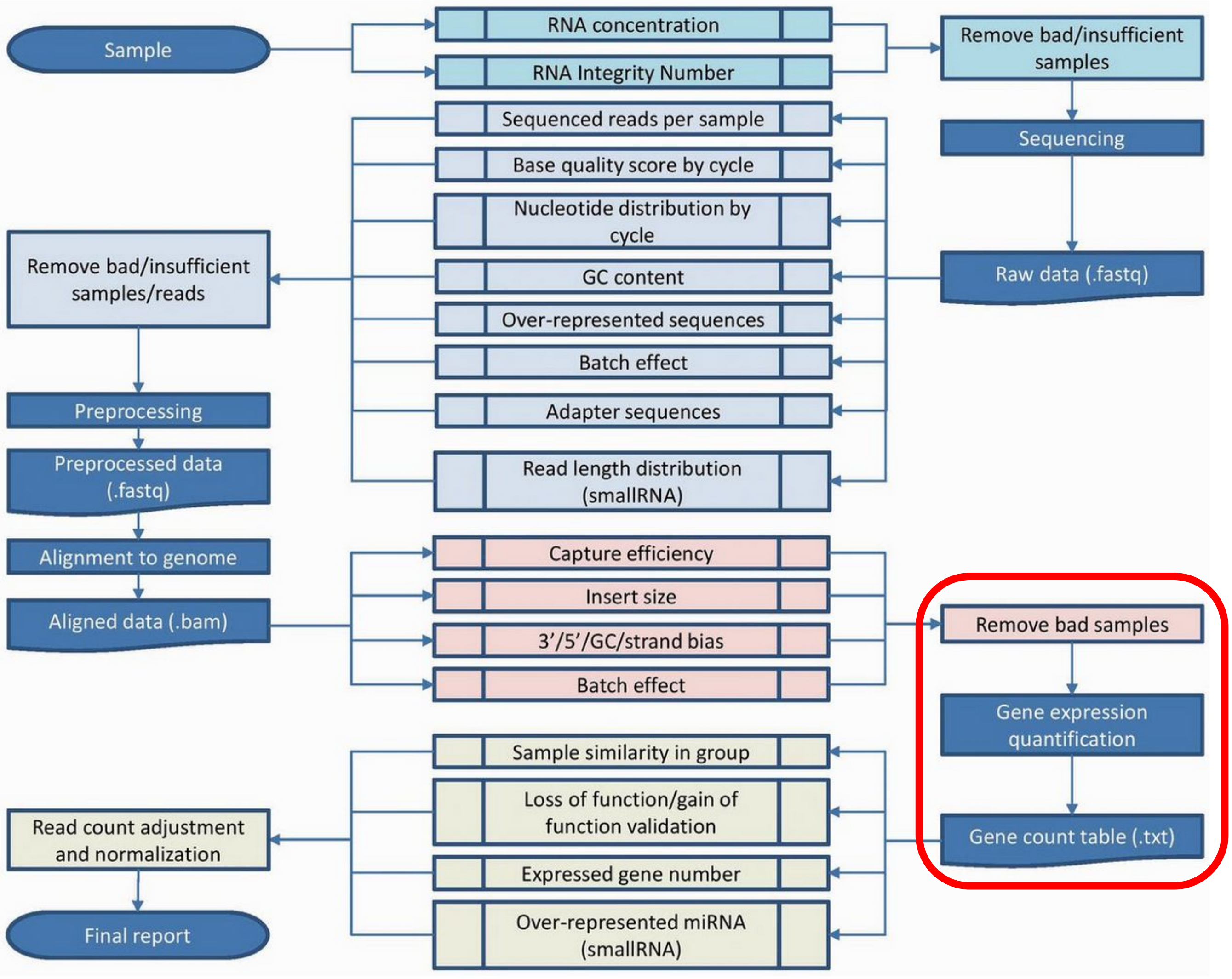
MULTI-MAPPING READS:
Number of reads mapped to multiple loci
% of reads mapped to multiple loci
Number of reads mapped to too many loci
% of reads mapped to too many loci

UNMAPPED READS:
% of reads unmapped: too many mismatches
% of reads unmapped: too short
% of reads unmapped: other

CHIMERIC READS:
Number of chimeric reads
% of chimeric reads



Quality control and mapping

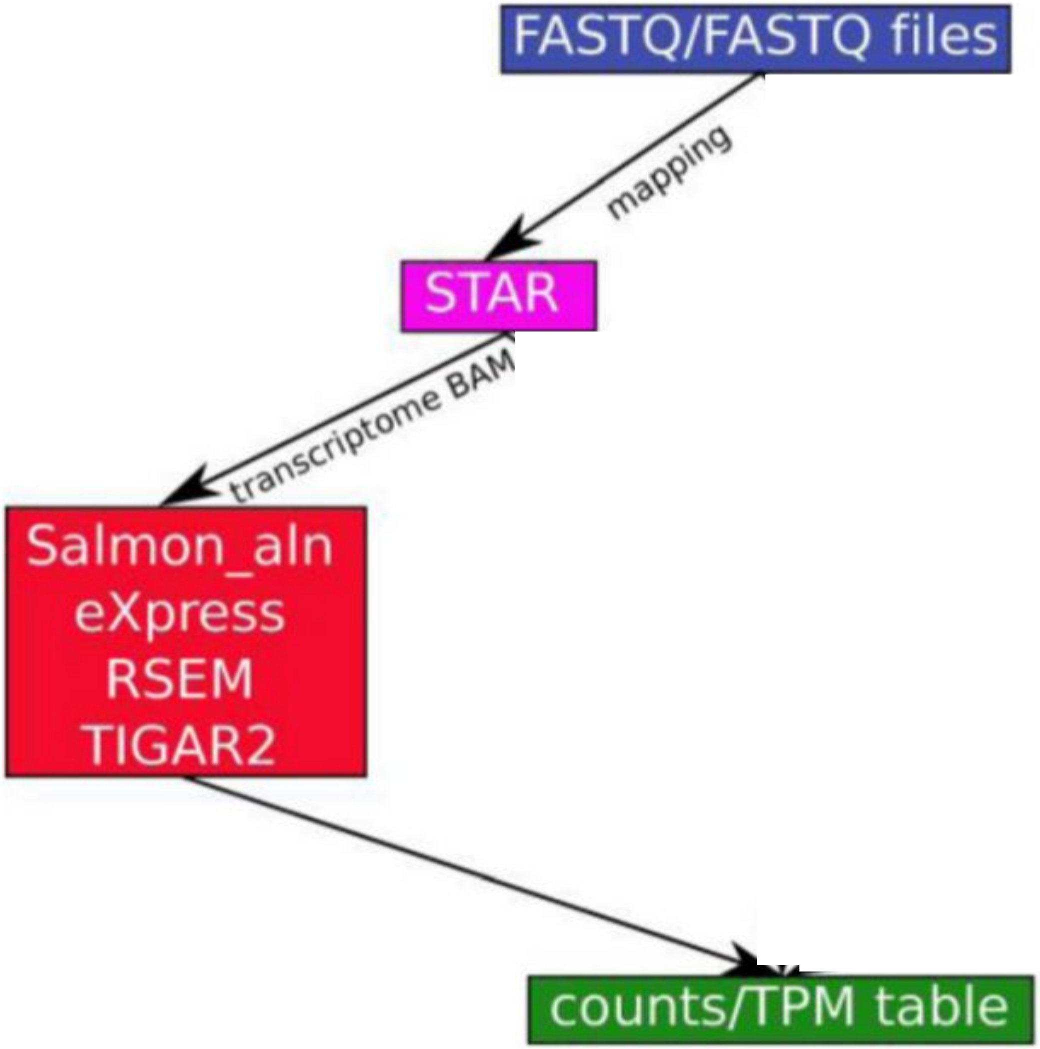


Sheng Q. et al. Brief Funct Genomics. 2017

Outline

- Experimental and sequencing design
- Quality control and mapping
- **Genes/transcripts/exons quantification**
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Genes/Transcripts/Exons quantification

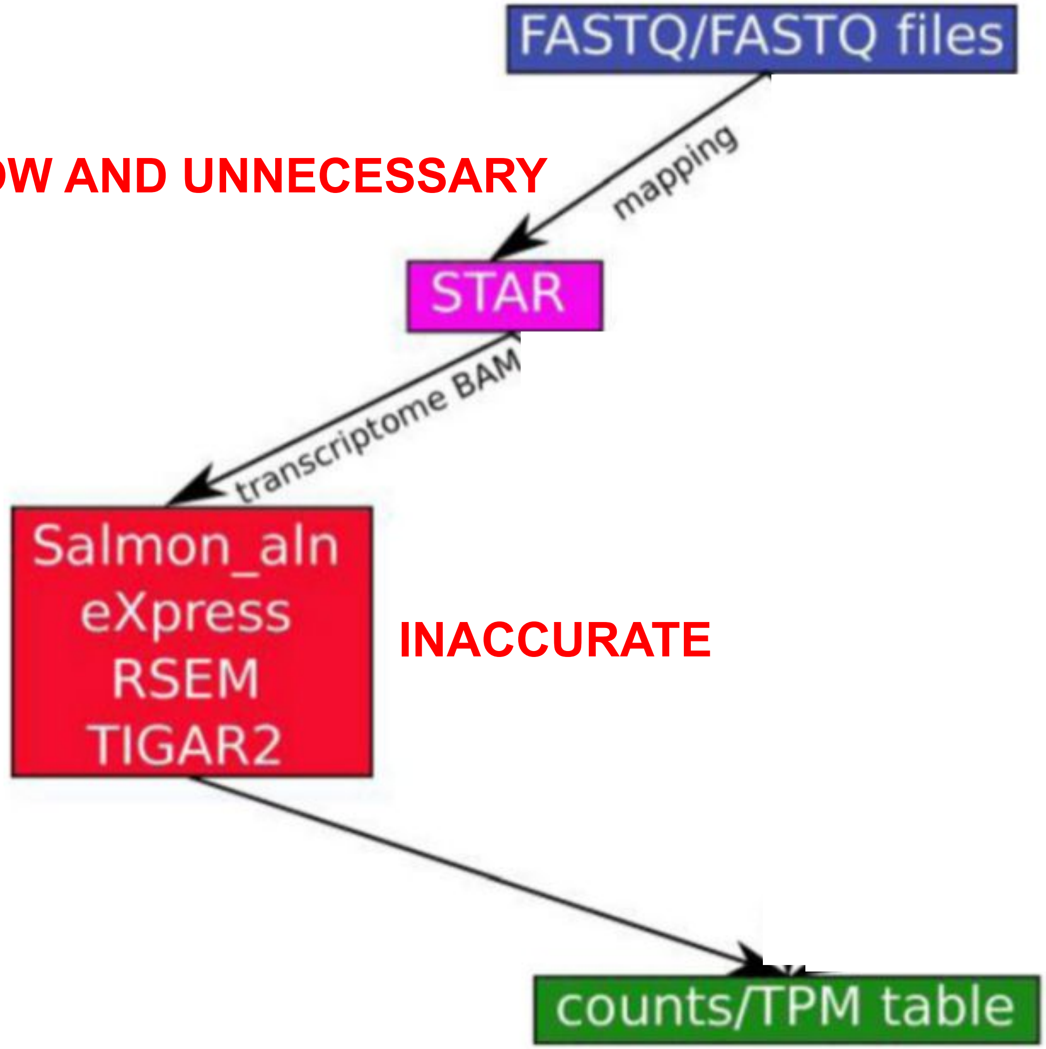


HTSeq
FeatureCounts

Zhang et al. BMC Genomics. 2017

Genes/Transcripts/Exons quantification

SLOW AND UNNECESSARY



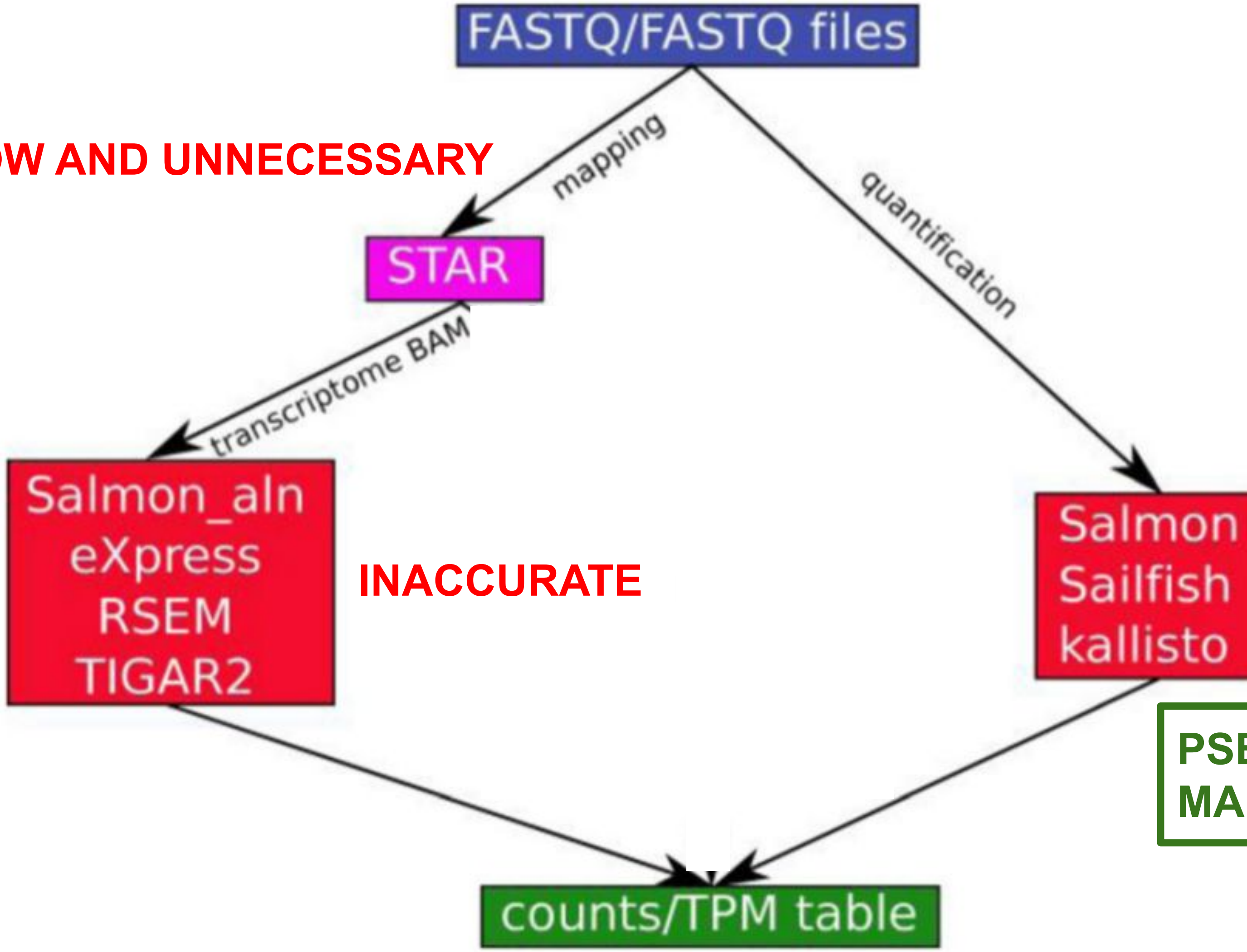
HTSeq
FeatureCounts

INACCURATE

Zhang et al. BMC Genomics. 2017

Genes/Transcripts/Exons quantification

SLOW AND UNNECESSARY



HTSeq
FeatureCounts

INACCURATE

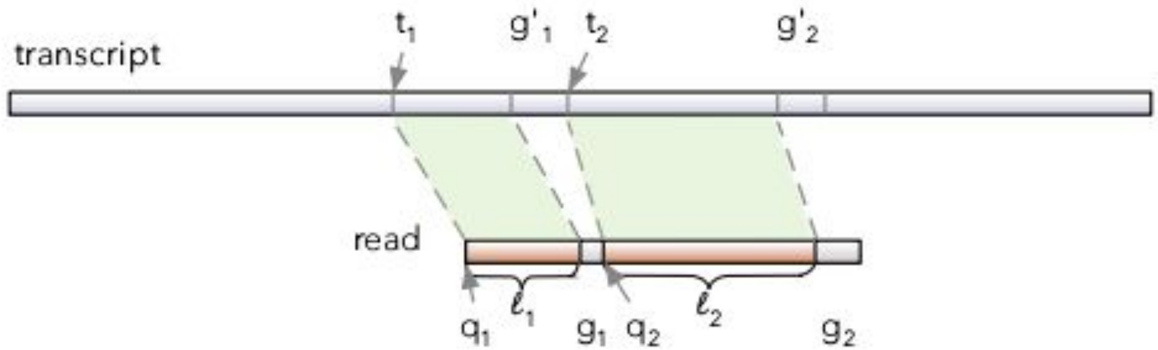
PSEUDO-MAPPING

Zhang et al. BMC Genomics. 2017

Genes/Transcripts/Exons quantification

A simple concept :

Accurate quantification of transcript abundance from RNA-seq data does not require knowing the optimal alignment for every potential locus of origin. Rather, simply knowing which transcripts (and positions within these transcripts) match the fragments reasonably well is sufficient.

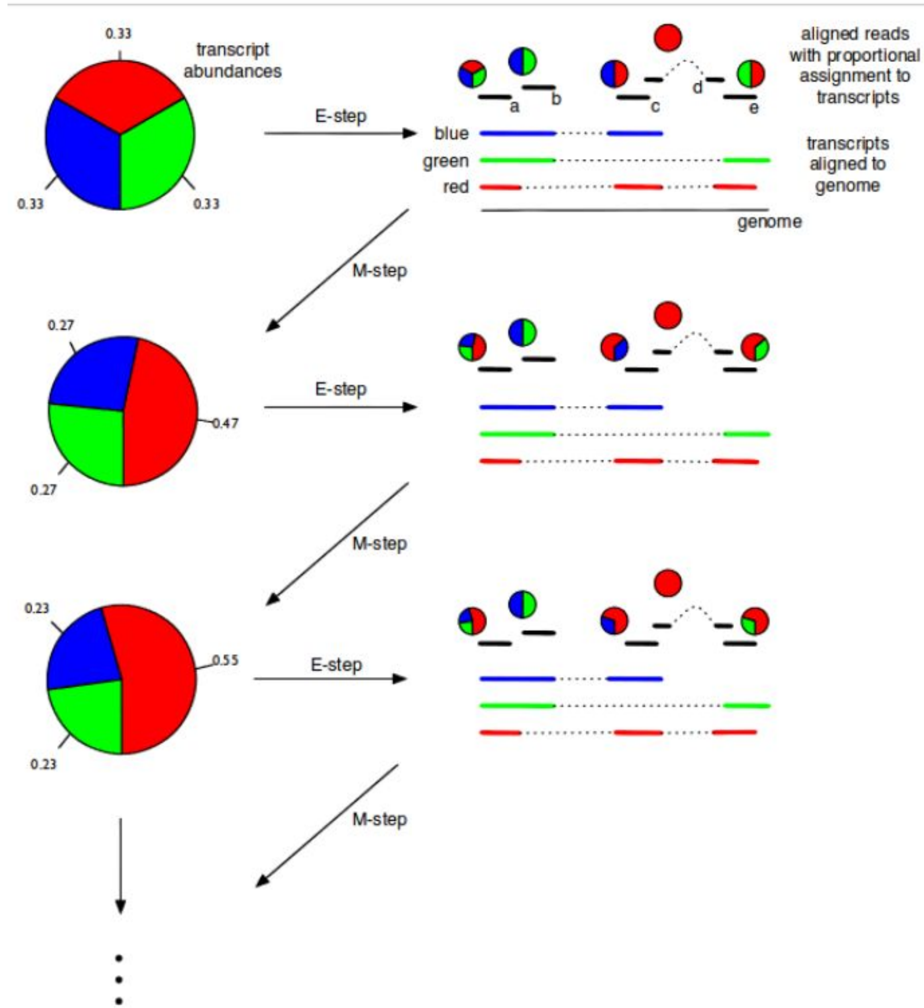


Bray et al. 2016 + Patro et al. 2014 + Patro et al. 2015 + Srivastava et al. 2016

Genes/Transcripts/Exons quantification

First 3 cycles of EM algorithm. Abundance of **red** isoform estimated after the 1st M-step: (⅓ read a + ½ read c + 1 read d + ½ read e)/(total read number), i.e. 0.47 ((0.33+0.5+1+0.5)/5)

- ➢ proved to converge
- ➢ stop criterion: when all probabilities that a fragment is derived from a transcript >= 10^{-7} have a relative change <= than 10^{-3}



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

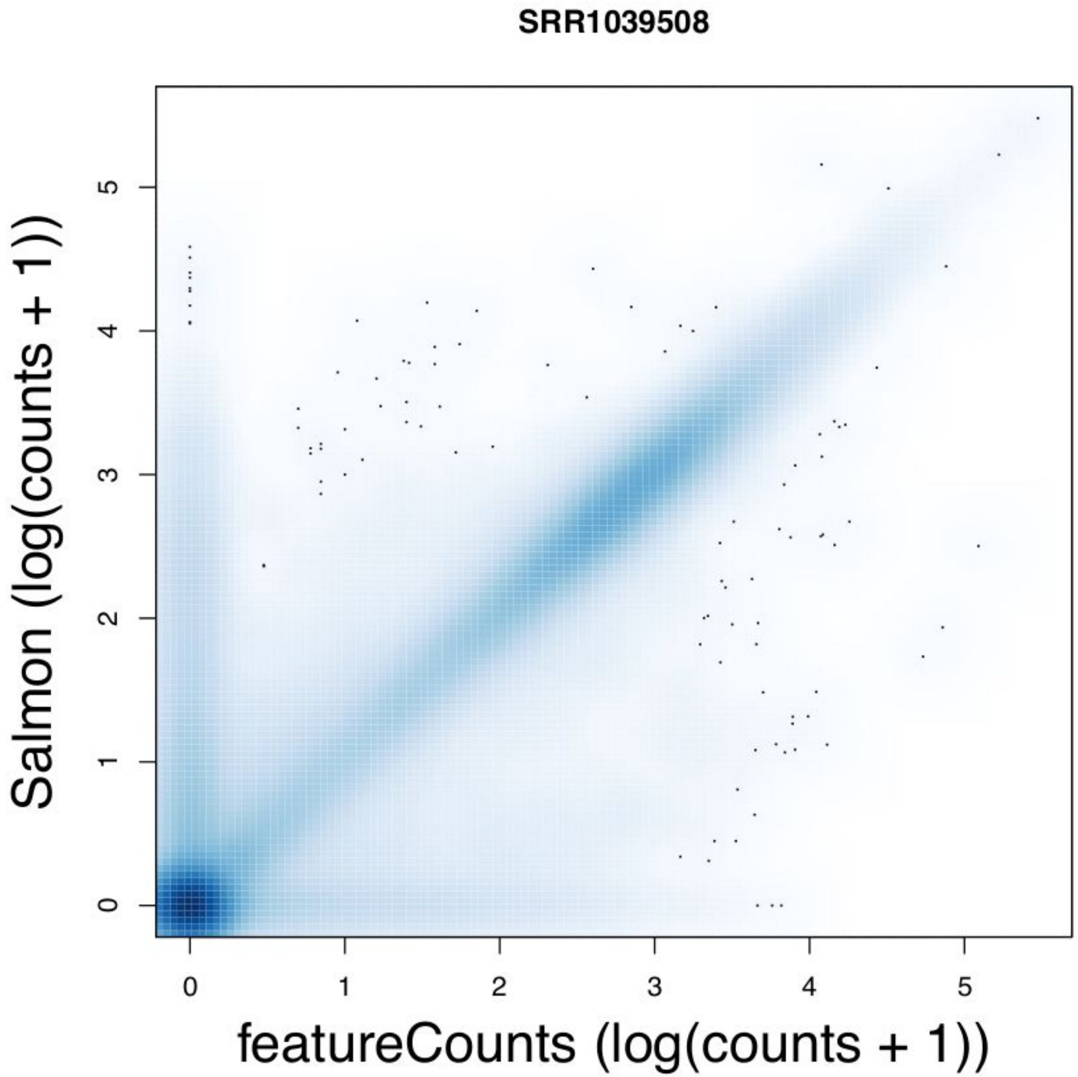
Genes/Transcripts/Exons quantification

Name	Length	Effective Length	TPM	NumReads
ENST00000424770.1	586	408.443	0	0
ENST00000448070.1	121	12.6016	0	0
ENST00000413156.1	578	400.459	24.2572	5
ENST00000420638.1	685	507.217	151.164	39.465
ENST00000398242.2	1049	869.633	973.655	435.825

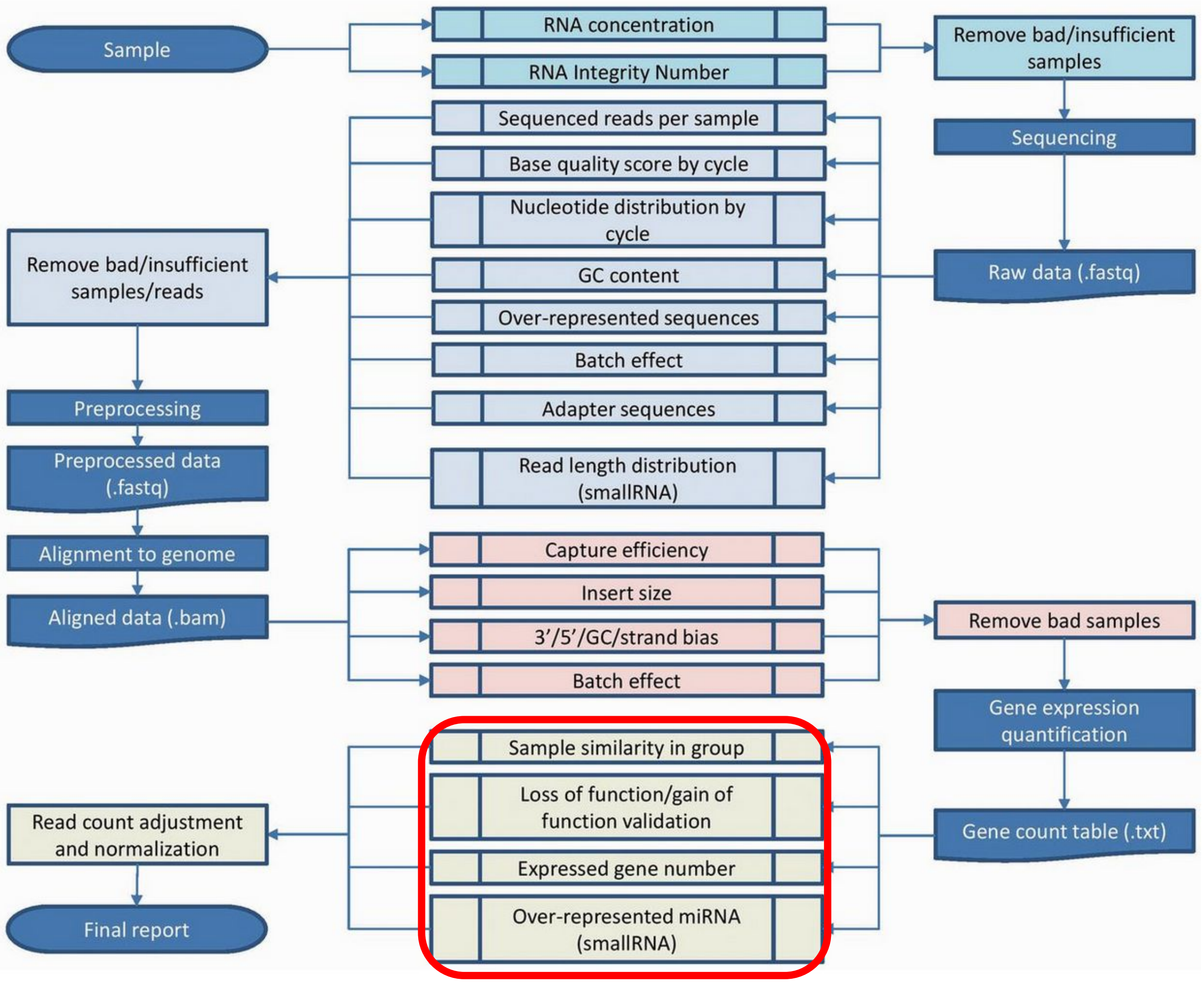
OR

Name	Length	Effective Length	TPM	NumReads
ENSG00000079974.1	176.5	159.427	413.683	356.123
ENSG00000213683.3	308.999	237.082	1808.74	546
ENSG00000254499.1	1606	1432.33	0	0
ENSG00000225929.1	2172.5	1993.13	0	0
ENSG00000212569.1	98	14.2568	0	0

Genes/Transcripts/Exons quantification



Quality control and mapping



Sheng Q. et al. Brief Funct Genomics. 2017

Genes/Transcripts/Exons quantification

RPKM = Reads Per Kilobase per Million
- (Num_sample_total_reads / 1,000,000) = “per million” scaling factor (SF)
- (Read_counts / SF) = This normalizes for sequencing depth, giving you reads per million (RPM)
- (RPM / Gene_Length_in_kb) = This gives you RPKM.

TPM = Transcripts per kilobase Per Million
- (Read_Counts / Gene_Length_in_kb) = This gives you reads per kilobase (RPK).

(Sum_all_sample_RPK / 1,000,000) = This is your “per million” scaling factor (SF).
3- (RPK / SF) = This gives you TPM.

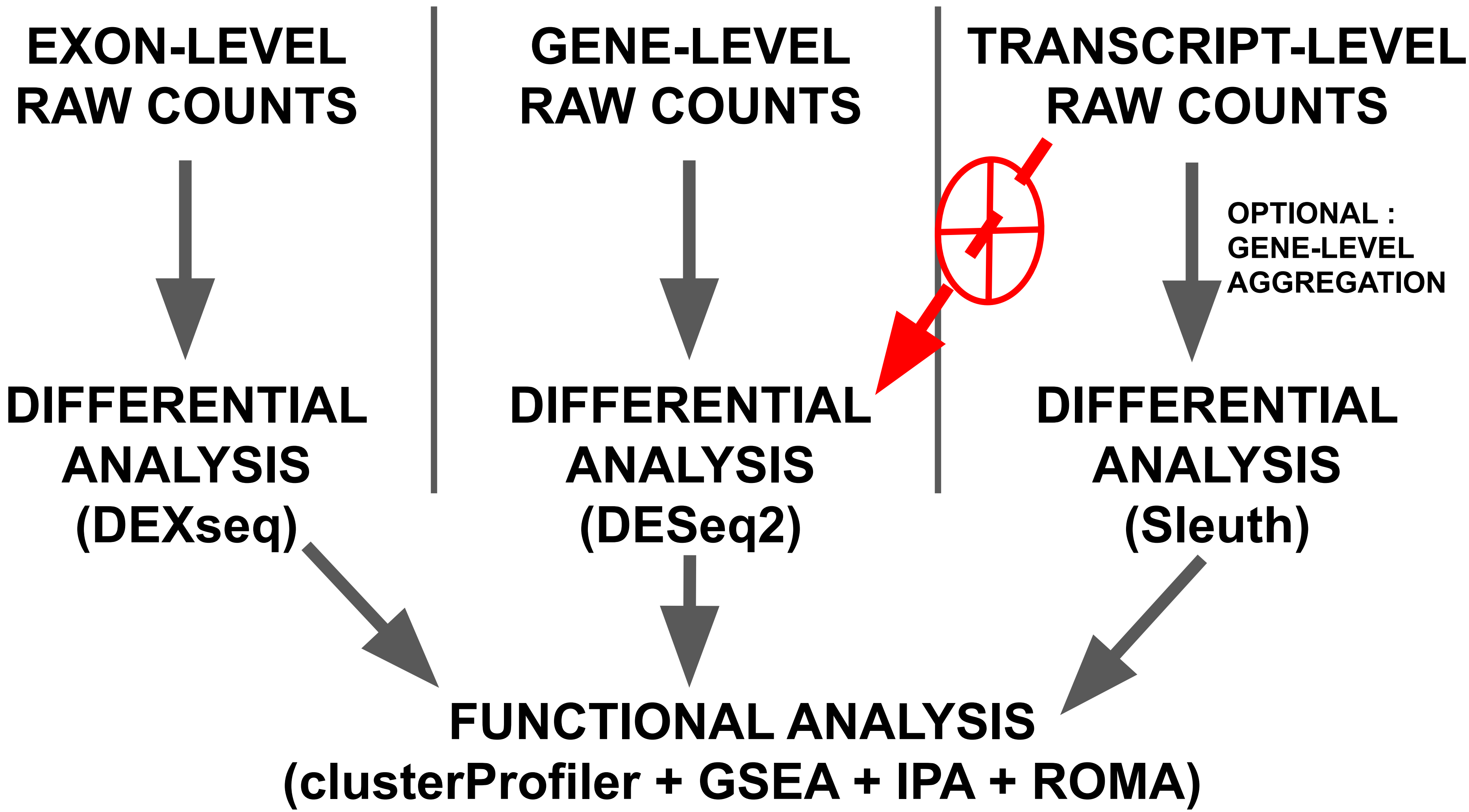
http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

Genes/Transcripts/Exons quantification

RPKM = Relative Expression Level **WITHIN** Sample

TPM = Normalized Expression Level comparable **BETWEEN** Samples

Genes/Transcripts/Exons quantification



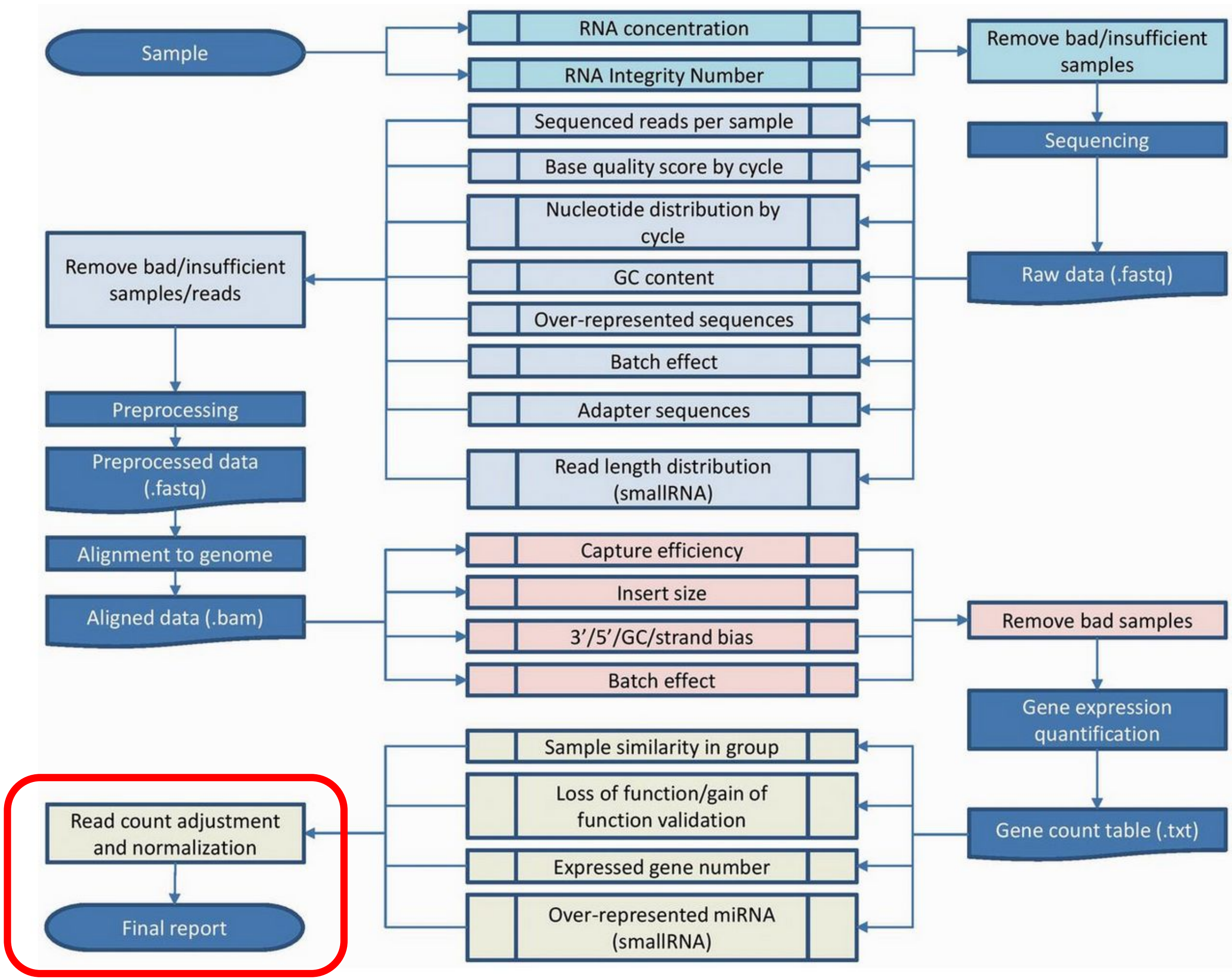
Genes/Transcripts/Exons quantification

TABLE 1. RNA-seq differential gene expression tools and statistical tests

Name	Assumed distribution	Normalization	Description	Version	Citations[d]	Reference
t-test	Normal	DEseq[a]	Two-sample *t*-test for equal variances	–	–	–
log *t*-test	Log-normal	DEseq[a]	Log-ratio *t*-test	–	–	–
Mann-Whitney	None	DEseq[a]	Mann-Whitney test	–	–	Mann and Whitney (1947)
Permutation	None	DEseq[a]	Permutation test	–	–	Efron and Tibshirani (1993a)
Bootstrap	Normal	DEseq[a]	Bootstrap test	–	–	Efron and Tibshirani (1993a)
baySeq[c]	Negative binomial	Internal	Empirical Bayesian estimate of posterior likelihood	2.2.0	159	Hardcastle and Kelly (2010)
Cuffdiff	Negative binomial	Internal	Unknown	2.1.1	918	Trapnell et al. (2012)
DEGseq[c]	Binomial	None	Random sampling model using Fisher's exact test and the likelihood ratio test	1.22.0	325	Wang et al. (2010)
DESeq[c]	Negative binomial	DEseq[a]	Shrinkage variance	1.20.0	1889	Anders and Huber (2010)
DESeq2[c]	Negative binomial	DEseq[a]	Shrinkage variance with variance based and Cook's distance pre-filtering	1.8.2	197	Love et al. (2014)
EBSeq[c]	Negative binomial	DEseq[a] (median)	Empirical Bayesian estimate of posterior likelihood	1.8.0	80	Leng et al. (2013)
edgeR[c]	Negative binomial	TMM[b]	Empirical Bayes estimation and either an exact test analogous to Fisher's exact test but adapted to over-dispersed data or a generalized linear model	3.10.5	1483	Robinson et al. (2010)
Limma[c]	Log-normal	TMM[b]	Generalized linear model	3.24.15	97	Law et al. (2014)
NOISeq[c]	None	RPKM	Nonparametric test based on signal-to-noise ratio	2.14.0	177	Tarazona et al. (2011)
PoissonSeq[c]	Poisson log-linear model	Internal	Score statistic	1.1.2	37	Li et al. (2012)
SAMSeq[c]	None	Internal	Mann-Whitney test with Poisson resampling	2.0	54	Li and Tibshirani (2013)

Schurch et al. 2016. RNA, 22(6)

Quality control and mapping



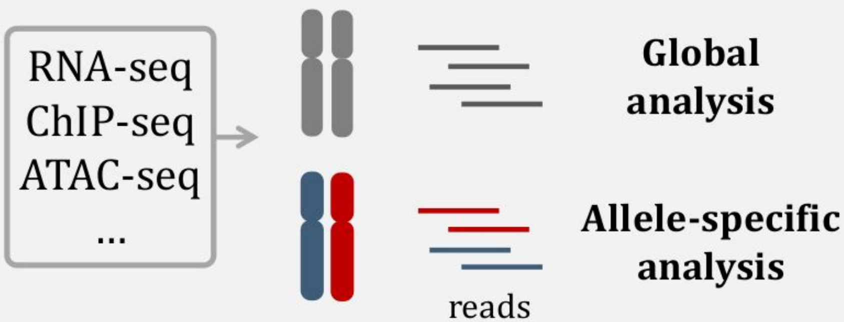
Sheng Q. et al. Brief Funct Genomics. 2017

Outline

- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- **Allele-specific quantification**
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Allele-specific analysis

Principle

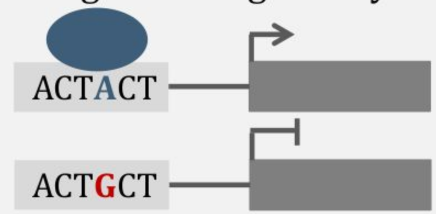


Difficulties

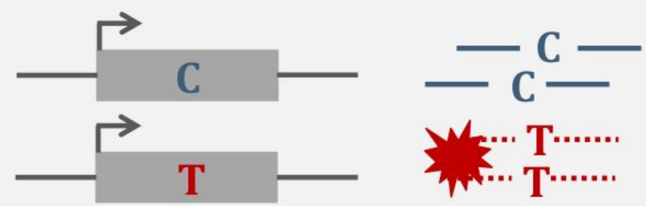
- Need to know haplotypes
- Limited by the number of differential SNPs
- Require dedicated bioinformatics strategy

Domains of application

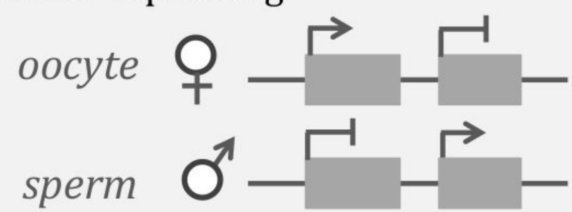
- Effects of genetic regulatory variants



- Non sense mediated decay SNP

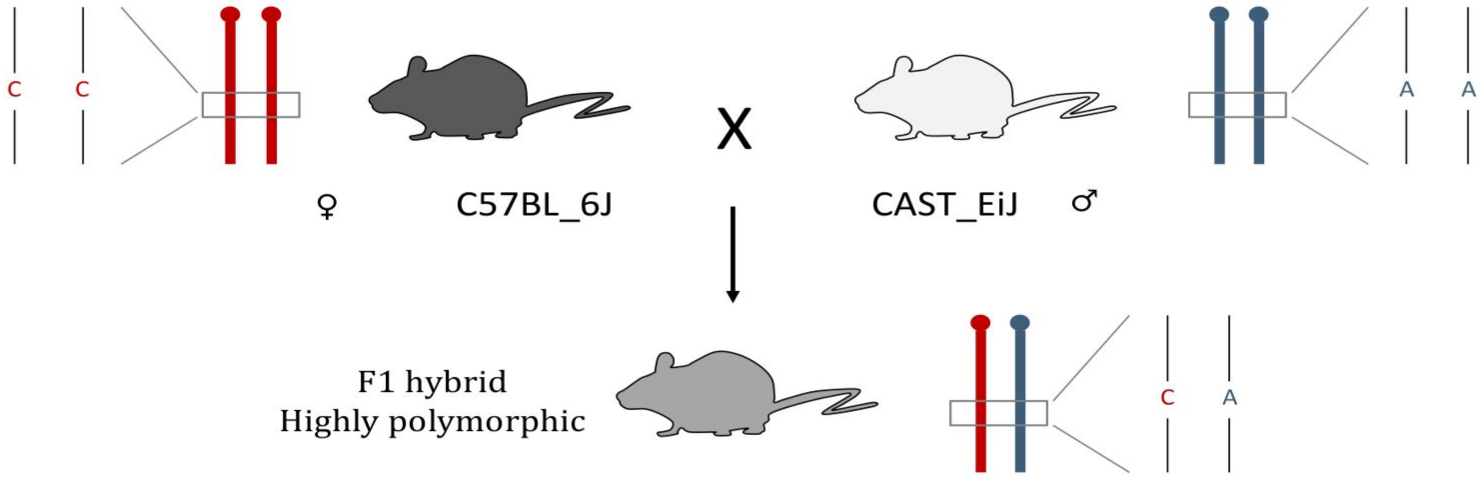


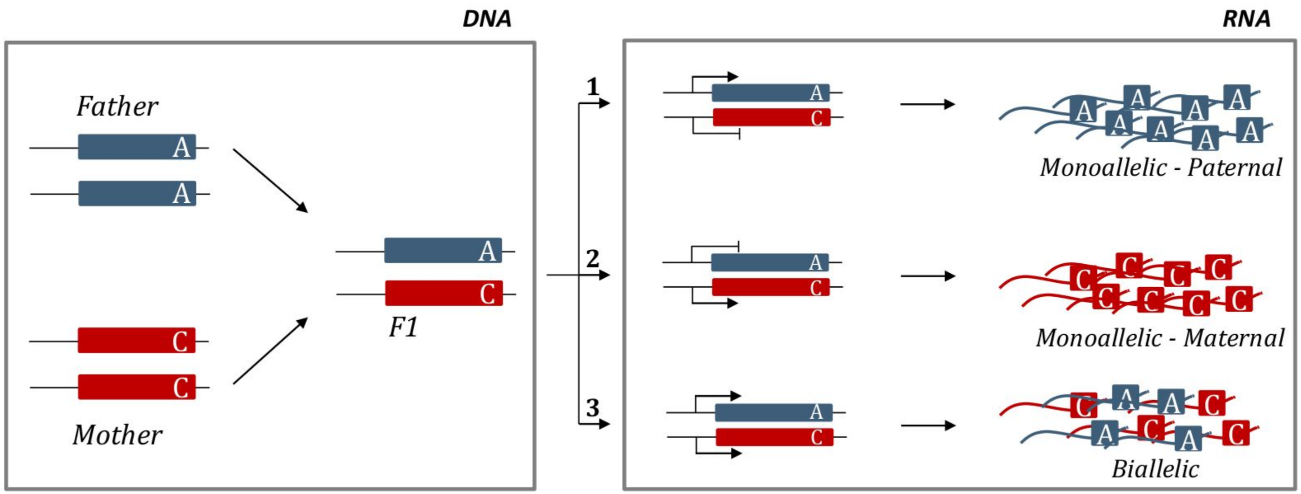
- Parental imprinting



- X chromosome inactivation

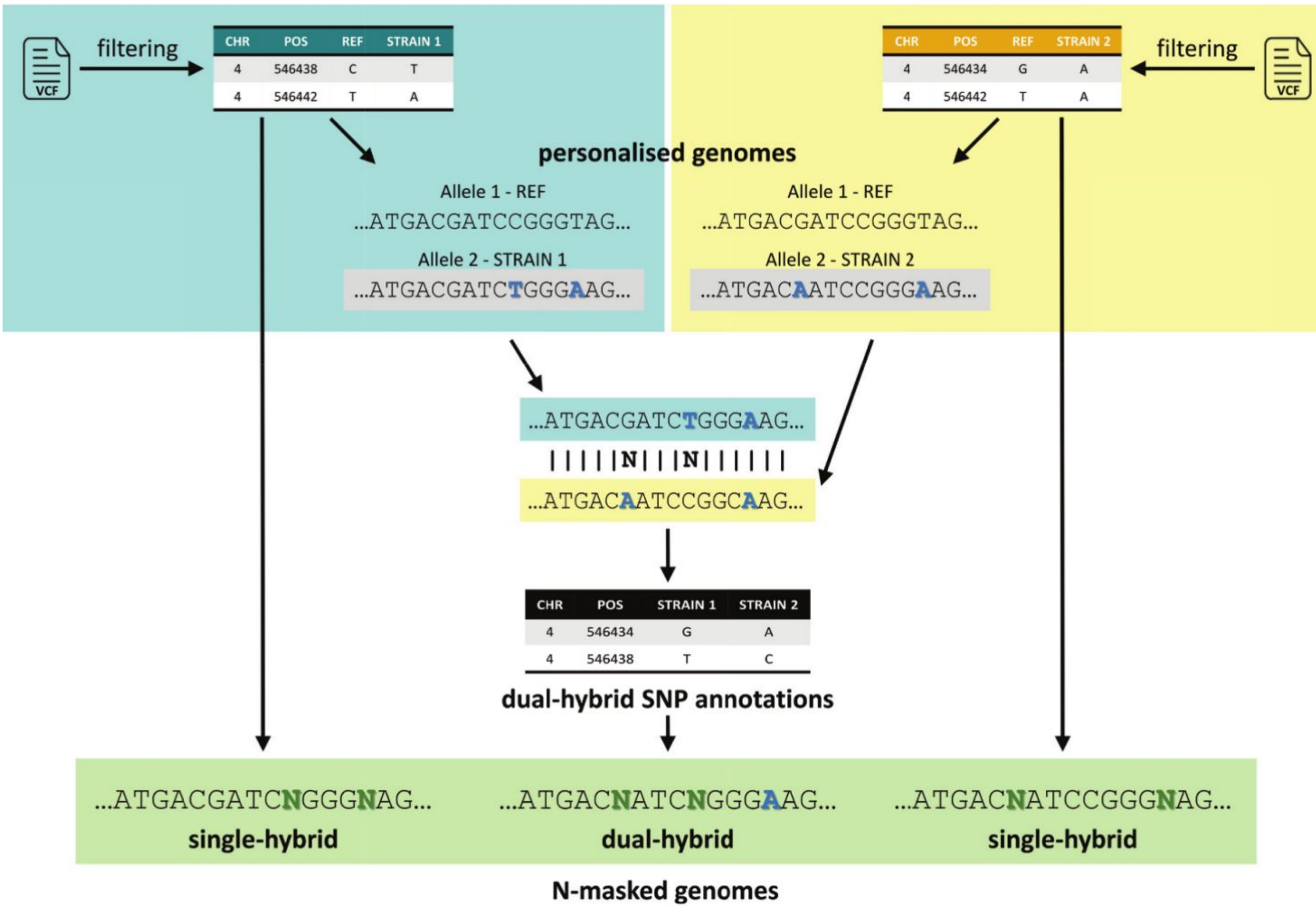
Allele-specific analysis





Allele-specific analysis

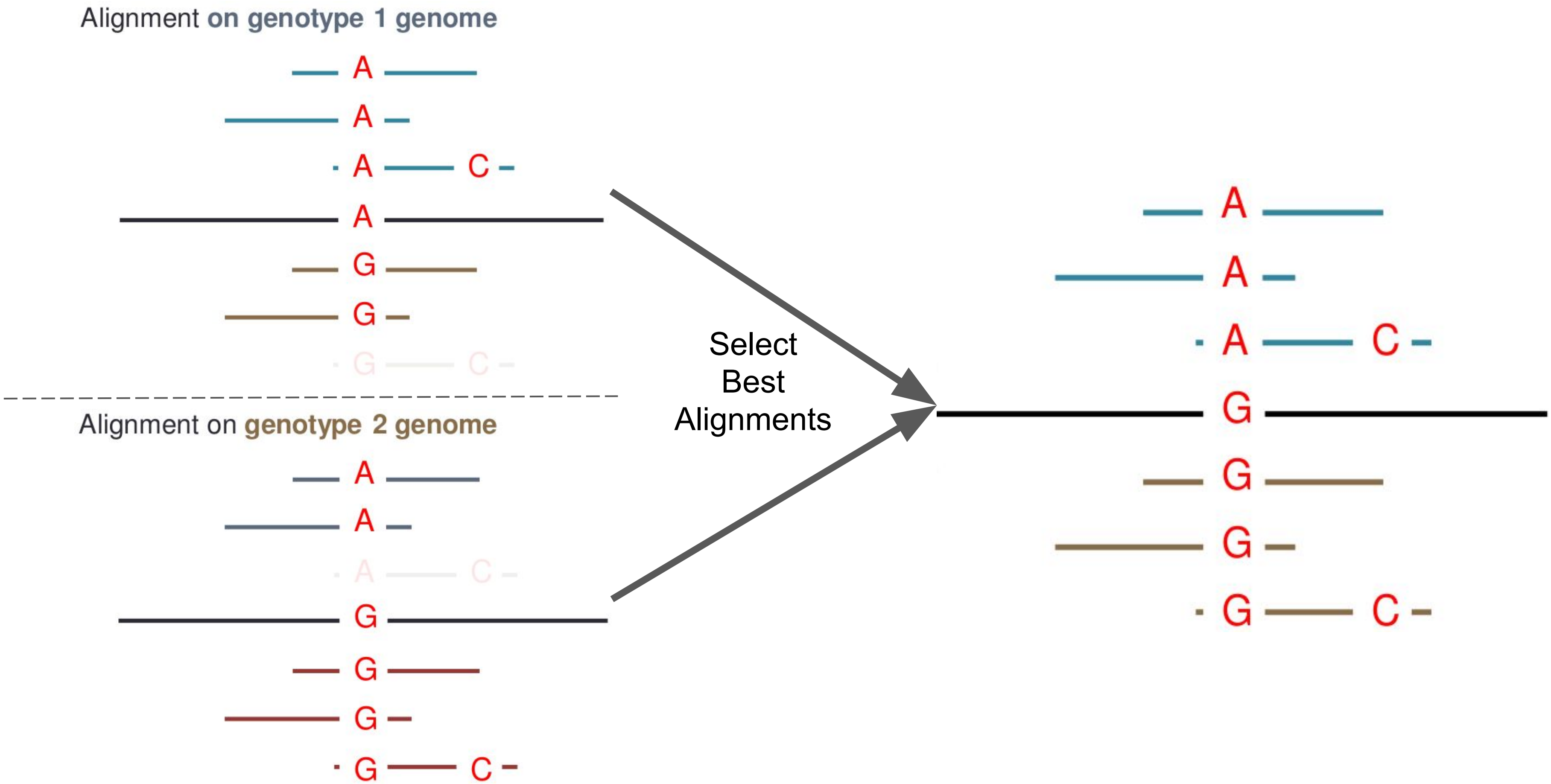
Method 1 : N-masked alignment strategy



Krueger F and Andrews SR. 2016. F1000Res

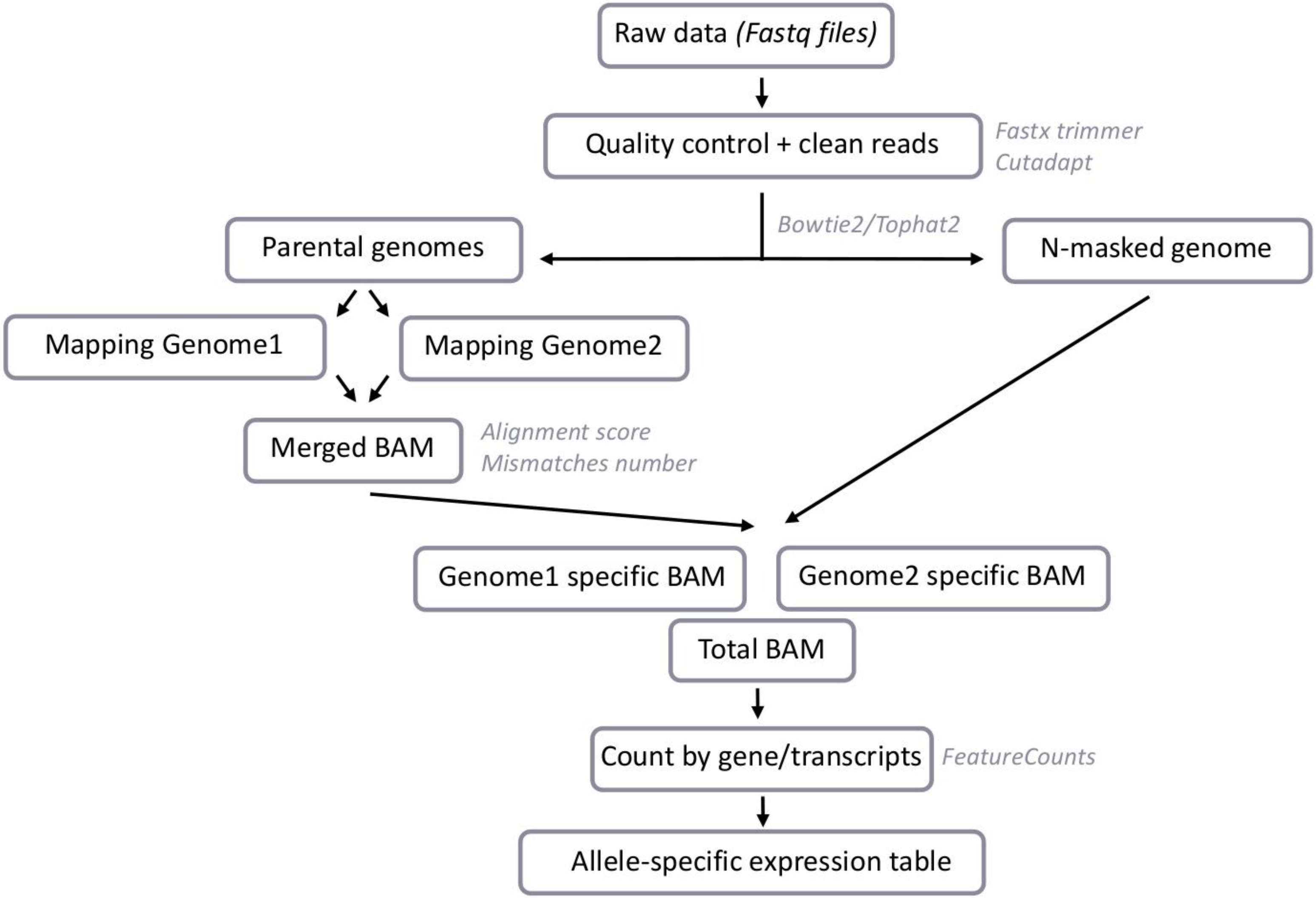
Allele-specific quantification

Method 2 : Alignment to parental genome strategy



Castel S et al. 2015. Genome Biology

Allele-specific quantification



Laurène SYX

Outline

- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- **Functional Analysis**
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Functional analysis

There are a number of questions that you may answer using your set of DEGs. These include:

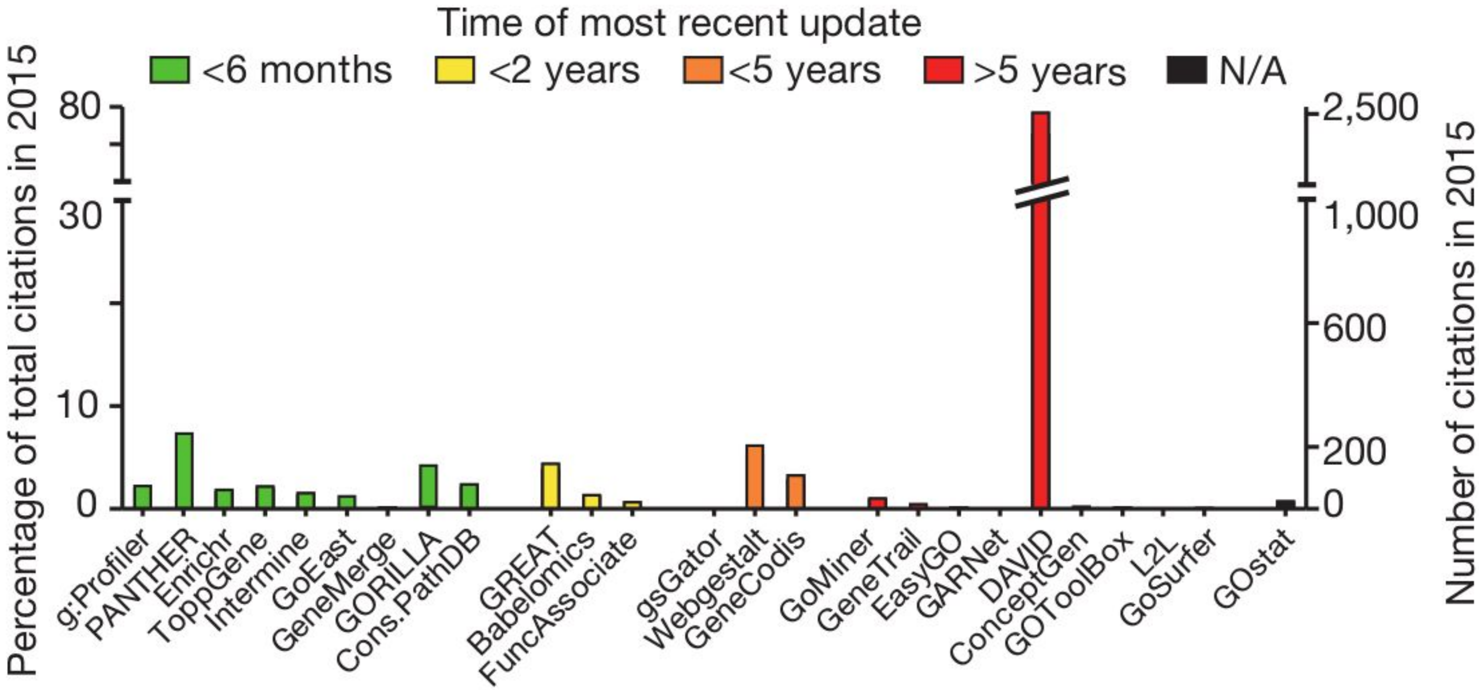
1) What are the biological process, cellular locations and molecular functions that are particularly over- or under-represented in your set of genes ?

2) What are the pathways that are significantly impacted in your condition ?

3) Which are the putative regulatory elements in the promoters of genes that show similar expression patterns (i.e. those which have similar fold-change patterns between different samples) ?

Functional analysis

GO Analysis : There are many tools that allow you to do this...



Wadi L. et al. 2016. Nature Methods

...but there are also many, many mistakes related to the way the p-values are calculated and corrected for multiple comparisons, the association of genes to multiple GO terms and the implicit redundancy, the choice of the background set of genes, etc... (Rhee et al. 2008. Nat Rev Genet)

Functional analysis

DAVID 6.8 Oct. 2016

-- The DAVID Knowledgebase completely rebuilt
-- Entrez Gene integrated as the central identifier to allow for more timely updates while still incorporating Ensembl and Uniprot as integral data sources
-- New GO category (GO Direct) provides GO mappings directly annotated by the source database (no parent terms included)
-- New annotation categories
-- New list identifier systems added for list uploading and conversion
-- A few bugs fixed

DAVID 6.7 Jan. 2010

-- The DAVID Knowledgebase completely rebuilt, including the central DAVID id system
-- Ensembl Gene included as an integral data source
-- DAVID engine completely rebuilt to facilitate future updates and development
-- New GO category (GO FAT) filters out very broad GO terms based on a measured specificity of each term (not level-specificity)
-- New annotation categories
-- New list identifier systems added for list uploading and conversion
-- Automatic list naming based on uploaded file name
-- Ability to upload expression/other values (some display, but otherwise not used in the analysis at this point)
-- A few bugs fixed
-- and more

Functional analysis

ShinyGO: a graphical gene-set enrichment tool for animals and plants

Steven Xijin Ge ✉, Dongmin Jung, Runan Yao

Bioinformatics, btz931, https://doi.org/10.1093/bioinformatics/btz931

Published: 27 December 2019 **Article history ▾**

http://bioinformatics.sdstate.edu/go/

Just paste your gene list to get enriched GO terms and other pathways for over 315 plant and animal species, based on annotation from Ensembl (Release 96), Ensembl plants (R. 43) and Ensembl Metazoa (R. 43). An additional 2031 genomes (including bacteria and fungi) are annotated based on STRING-db (v.10). In addition, it also produces KEGG pathway diagrams with your genes highlighted, hierarchical clustering trees and networks summarizing overlapping terms/pathways, protein-protein interaction networks, gene characteristics plots, and enriched promoter motifs.

Functional analysis

There are a number of questions that you may answer using your set of DEGs. These include:

1) What are the biological process, cellular locations and molecular functions that are particularly over- or under-represented in your set of genes ?

2) What are the pathways that are significantly impacted in your condition ?

3) Which are the putative regulatory elements in the promoters of genes that show similar expression patterns (i.e. those which have similar fold-change patterns between different samples) ?

Functional analysis

- Pathway databases (eg. Reactome, KEGG, etc.) : Here your set of DEG will be mapped on pathways. No real analysis is performed but you see what DEGs are on each pathway. The limitations are obvious: no p-values are calculated, no idea about which pathways are affected beyond random chance, etc.

- Pathway enrichment analysis : Here, pathways are considered as simple sets of genes and an enrichment p-value is calculated for each (e.g. DAVID, GSEA, Ingenuity, etc.). The limitations include the fact that the p-values are calculated based on the assumption that all variables (genes) are independent while the pathways are there precisely to tell you how these genes influence each other. Another limitation is that the pathways are treated as simple bags of genes, disregarding all the phenomena and interactions between genes that they describe. This analysis approach only looks at the number of DE genes and makes no difference between a situation in which a pathway has 3 entry points and all 3 are severely down-regulated thus effectively shutting down the entire pathway and a situation in which 3 other random genes are down regulated on the same pathway.

Reimand J et al. 2019. Nat Protoc

Functional analysis

There are a number of questions that you may answer using your set of DEGs. These include:

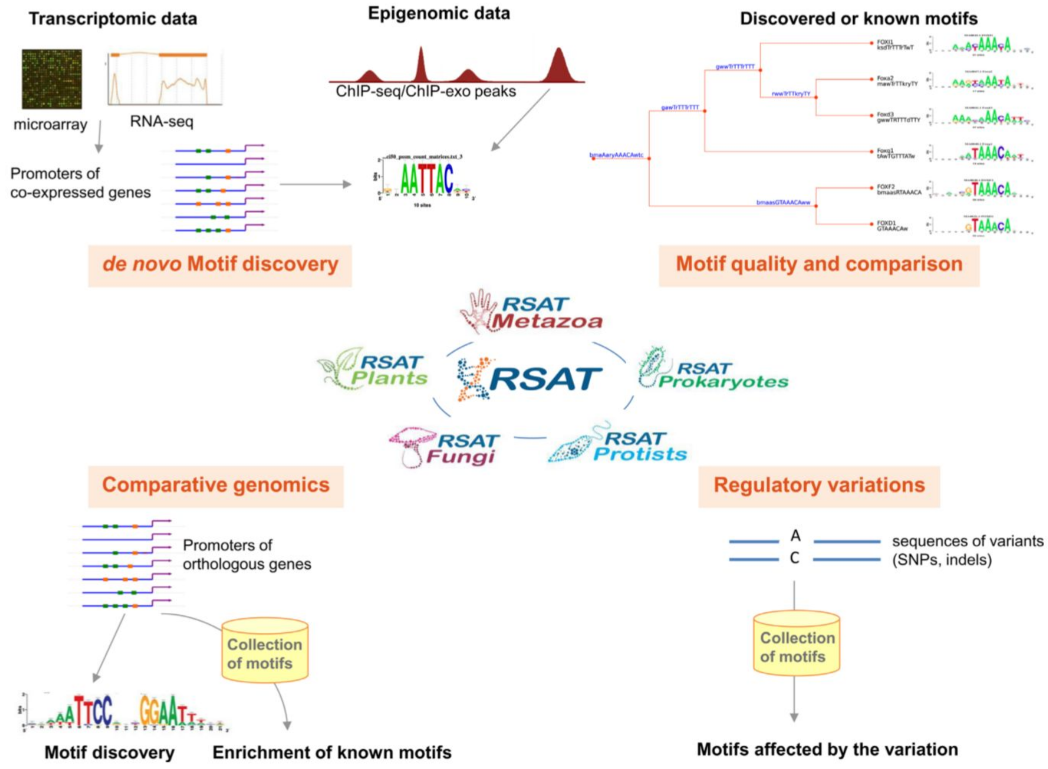
1) What are the biological process, cellular locations and molecular functions that are particularly over- or under-represented in your set of genes ?

2) What are the pathways that are significantly impacted in your condition ?

3) Which are the putative regulatory elements in the promoters of genes that show similar expression patterns (i.e. those which have similar fold-change patterns between different samples) ?

Functional analysis

- You can download the putative promoters (usually 1kb upstream of the transcriptional start site is a good starting point) of all the genes you are interested in and search for enriched DNA motifs - you can use the online programs RSAT (http://www.rsat.eu/) or MEME (http://meme.nbcr.net/meme/) . These DNA motifs may be recognised by sequence-specific transcription factors, thereby giving you insight into how this particular transcription profile may be regulated.

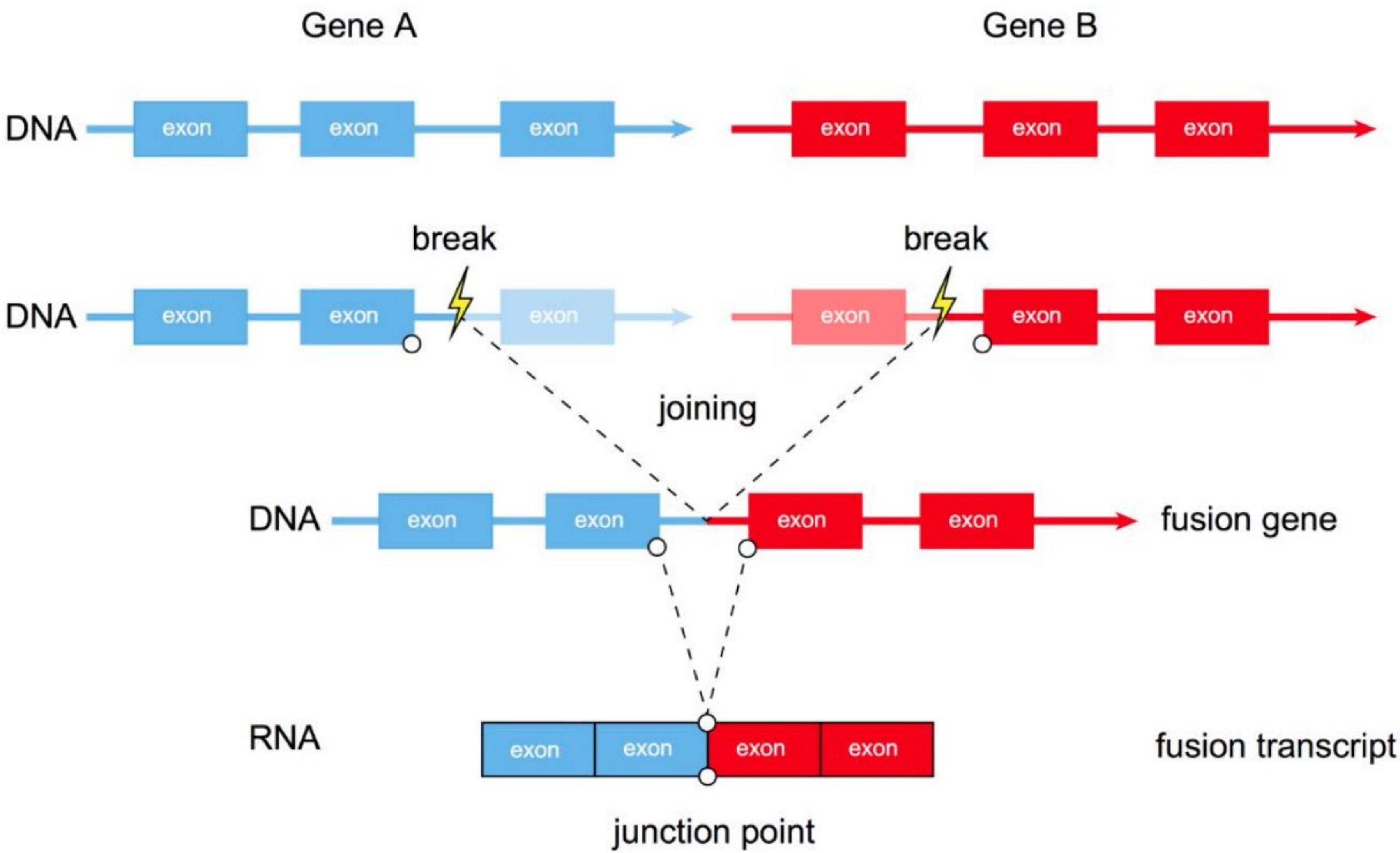


Nguyen NTT et al. 2018. NAR

Outline

- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- **Fusion transcripts / chimera detection**
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Fusion transcripts

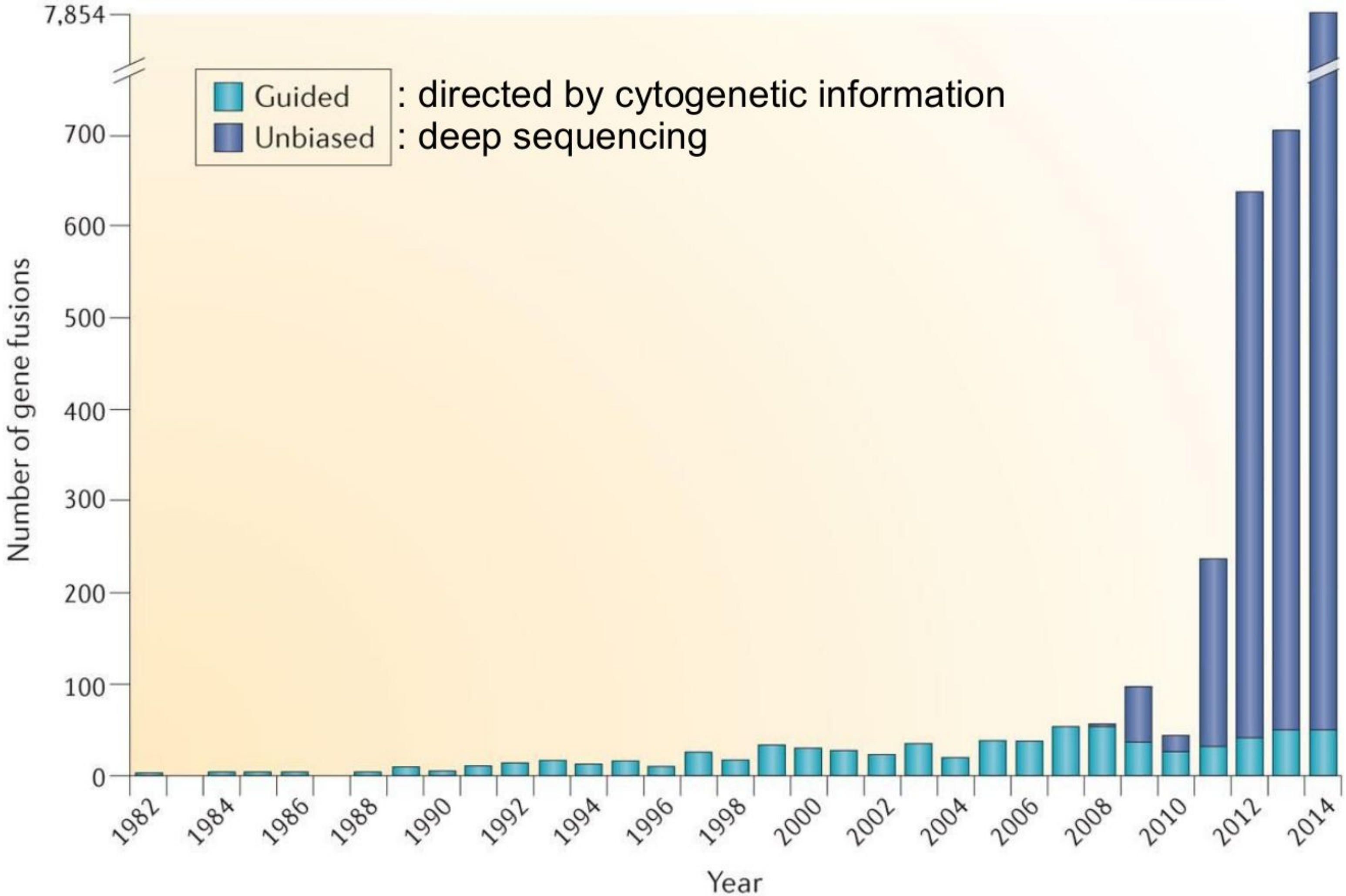


Torres-García et al. 2014. Bioinformatics

Fusion transcripts

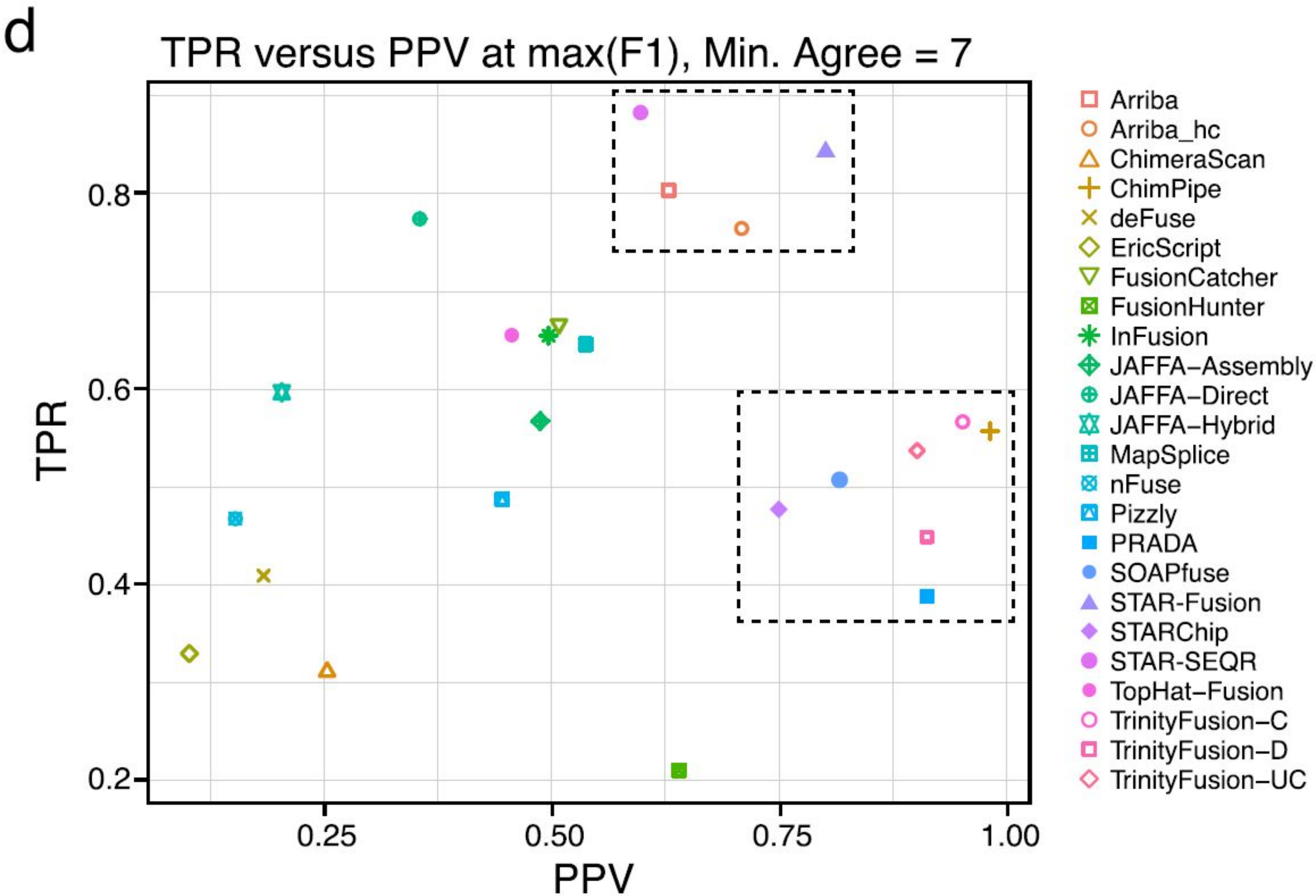
- An important class of cancer-contributing somatic alterations because they can drive the development of cancer
- Attractive as both therapeutic targets and diagnostic tools due to their tumor-specific expression
  - ex. BCR-ABL1 is associated with chronic myeloid leukemia and used as biomarker
- Other classes of chimeric transcripts :
  - Read-through
  - Trans-splicing

Fusion transcripts



Mertens et al. 2015. Nat. Rev. Cancer

Fusion transcripts



Haas BJ et al. 2019. Genome Biol.

Fusion transcripts

Liu et al., 2016 demonstrated the potential of a meta-caller algorithm

\supporting reads\	Tool 1	Tool 2	Tool 3	Tool 4
Fusion 1	23	100	20	22
Fusion 2	66	130	-	34
Fusion 3	17	-	-	-
Fusion 4	4	-	-	7
...				

Step 1: Keep the fusion transcripts detected by at least 2 tools

\supporting reads\	Tool 1	Tool 2	Tool 3	Tool 4
Fusion 1	23	100	20	22
Fusion 2	66	130	-	34
Fusion 4	4	-	-	7
...				

Step 2: Rank within each tool (small to large) , calculate the sum rank and order it from large to small

\rank\	Tool 1	Tool 2	Tool 3	Tool 4	Rank sum	Order of rank sum
Fusion 1	2	1	1	2	6	2
Fusion 2	3	2	0	3	8	1
Fusion 4	1	0	0	1	2	3
...						

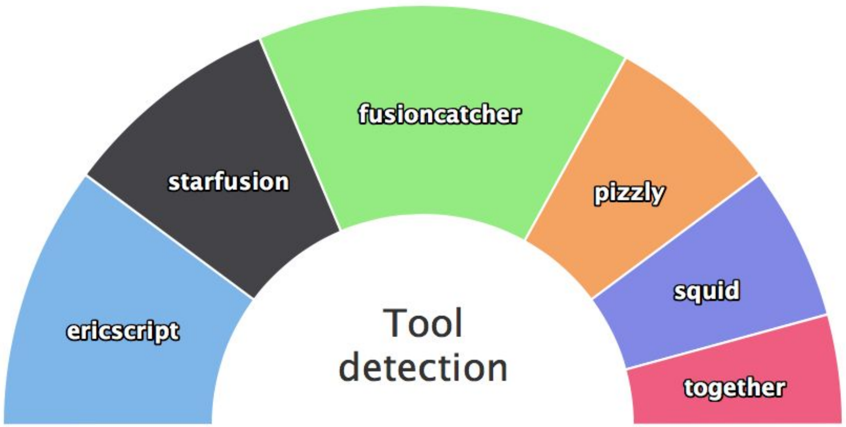
nf-core/ rnafusion

Pipeline overview (v1.1.0 - 2020/02/10)

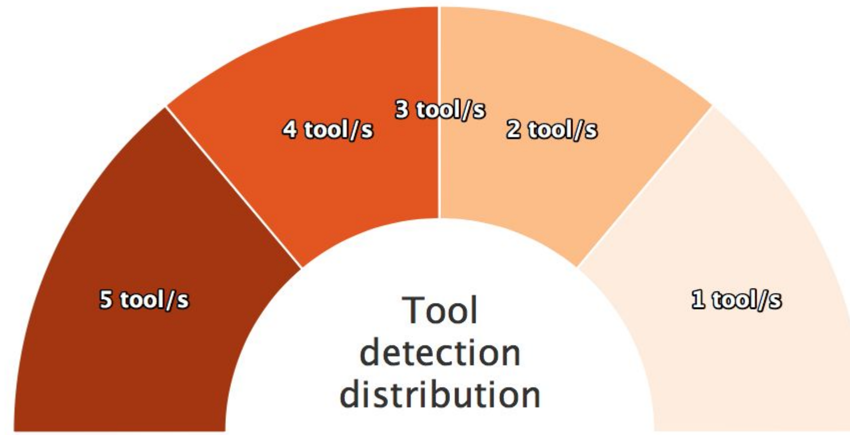
The pipeline is built using Nextflow and processes data using the following steps:

- Arriba
- EricScript
- FastQC - read quality control
- FusionCatcher
- FusionInspector
- fusion-report
- MultiQC - aggregate report, describing results of the whole pipeline
- Pizzly
- Squid
- Star-Fusion

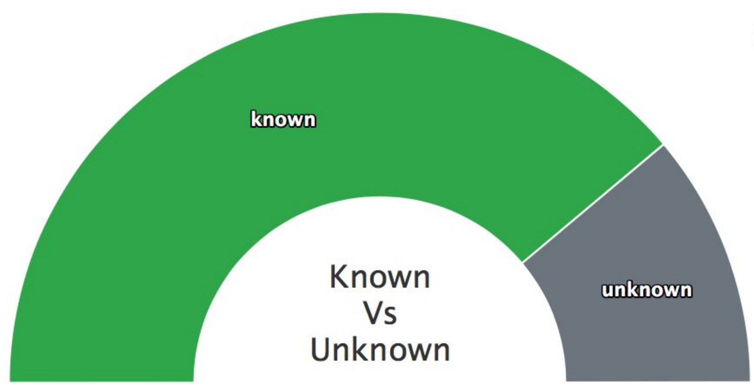
Fusion transcripts



Displays number of found fusions per tool.



Sum of counts detected by different tools per fusion.

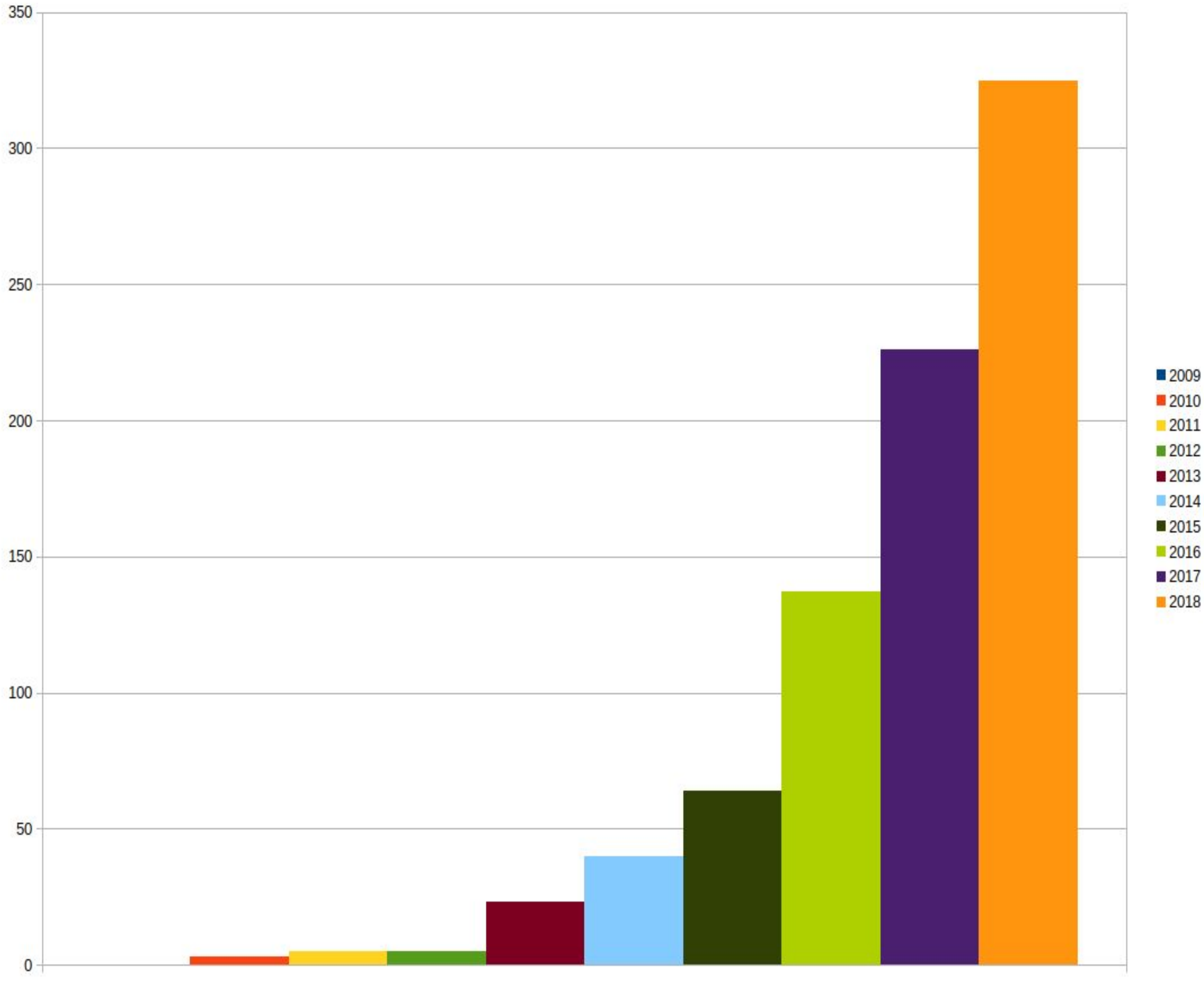


Shows the ration between found and unknown missing fusions in the local database.

Outline

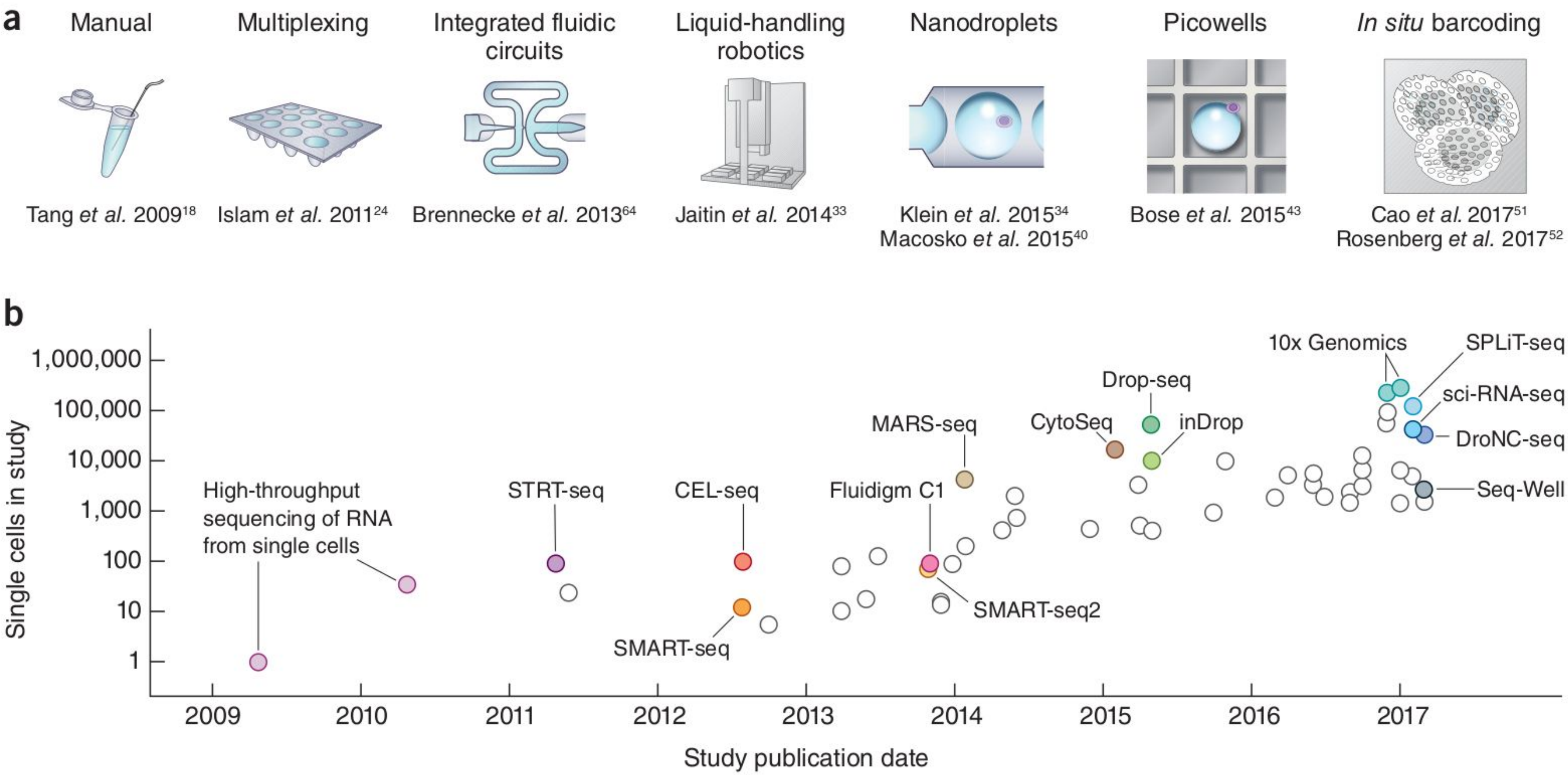
- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- **Single-cell RNA-seq**
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Single-cell RNA-seq



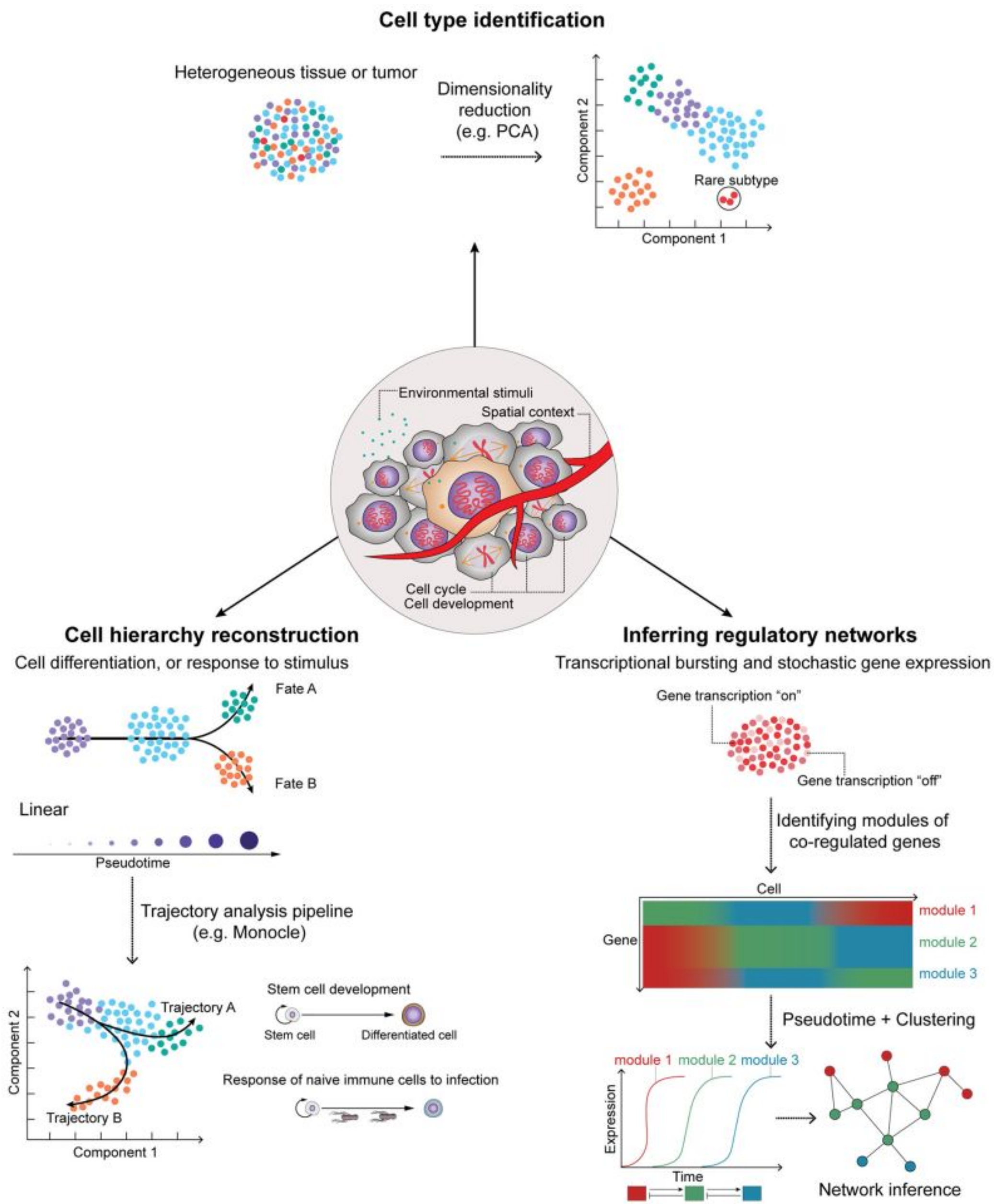
http://dan.corlan.net/medline-trend.html

Single-cell RNA-seq



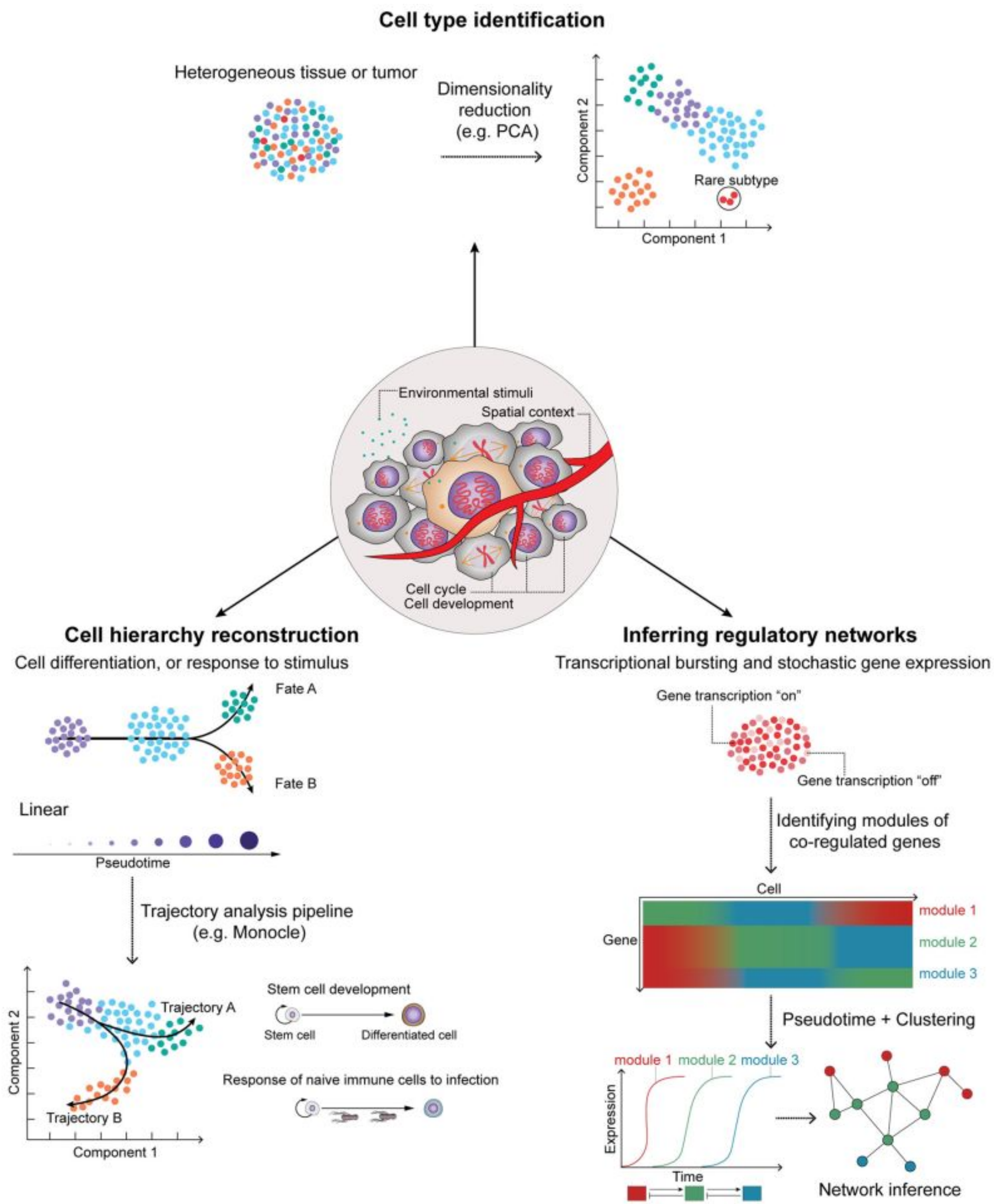
Svensson et al. Nature protocols (2018)

Single-cell RNA-seq



Hwang et al. Experimental & Molecular Medicine (2018)

Single-cell RNA-seq



Hwang et al. Experimental & Molecular Medicine (2018)

Outline

- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- **Full-length transcripts sequencing using long reads**
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Why ?

Korf, I. Genomics: the state of the art in RNA-seq analysis. Nature Publishing Group 10, 1165-1166 (2013)

SHORT READ TRANSCRIPT ASSEMBLY IS INSUFFICIENT

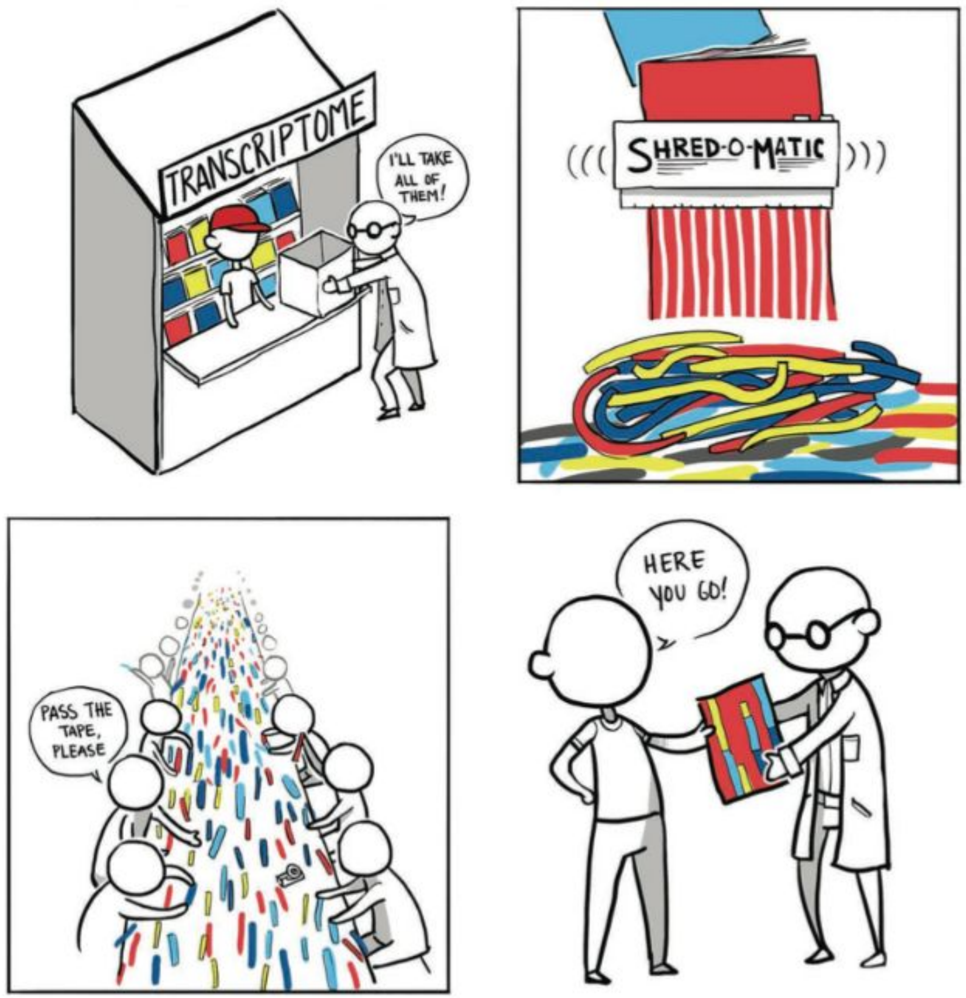


Figure 1 | Transcriptome reconstruction—akin to reassembling magazine articles after they have been through a paper shredder.

How insufficient ?

STEIJGER (2013) VS. ANGELINI (2014) VS. CHANG (2014)

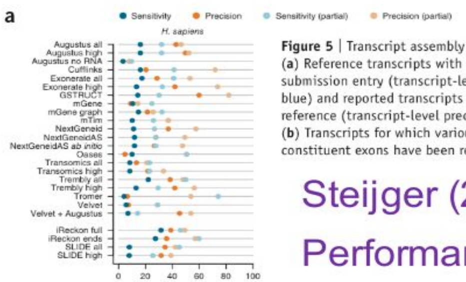


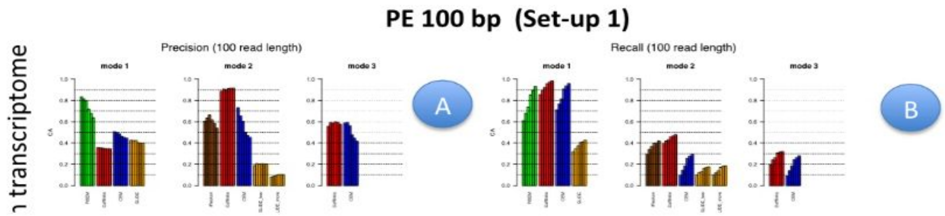
Figure 5 | Transcript assembly performance. (**a**) Reference transcripts with a matching submission entry (transcript-level sensitivity, blue) and reported transcripts that match the reference (transcript-level precision, orange). (**b**) Transcripts for which various subsets of constituent exons have been reported.

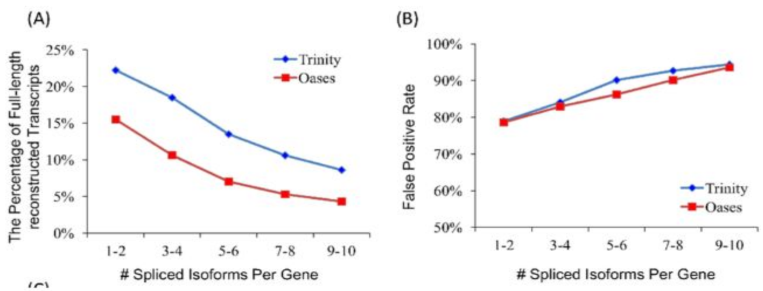
Steijger (2013): recall best at 20%, precision best at 40%

Performance worse for non-coding transcripts

Angelini (2014): recall best < 30%, precision best ~60%

Simulated coding transcripts





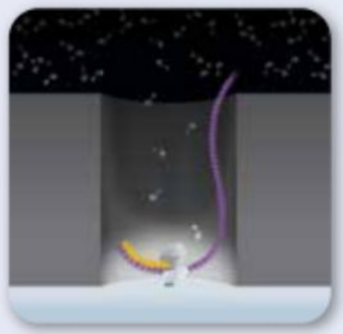
Chang (2014): recall best < 25%, precision best ~20%

Simulated coding+non-coding transcripts

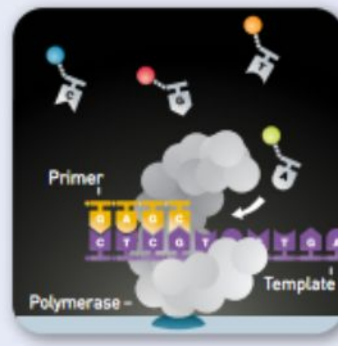
Steijger, T. et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Meth 10, 1177–1184 (2013).
Angelini, C., Canditiis, D. & Feis, I. Computational approaches for isoform detection and estimation: good and bad news. BMC Bioinformatics 15, 135–43 (2014).
Chang, Z., Wang, Z. & Li, G. The Impacts of Read Length and Transcriptome Complexity for De Novo Assembly: A Simulation Study. PLoS ONE 9, e94825–8 (2014).

SMRT-sequencing

SMRT Cells



Zero-Mode
Waveguides

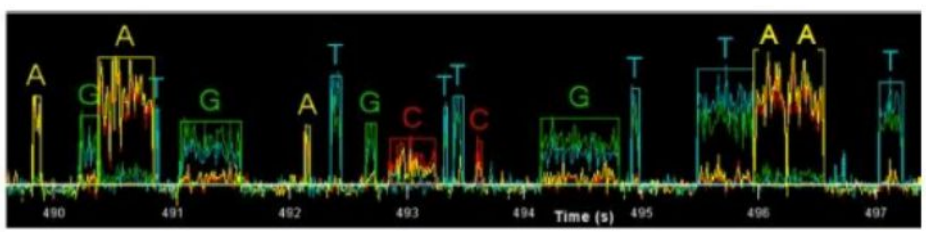


Phospholinked
Nucleotides

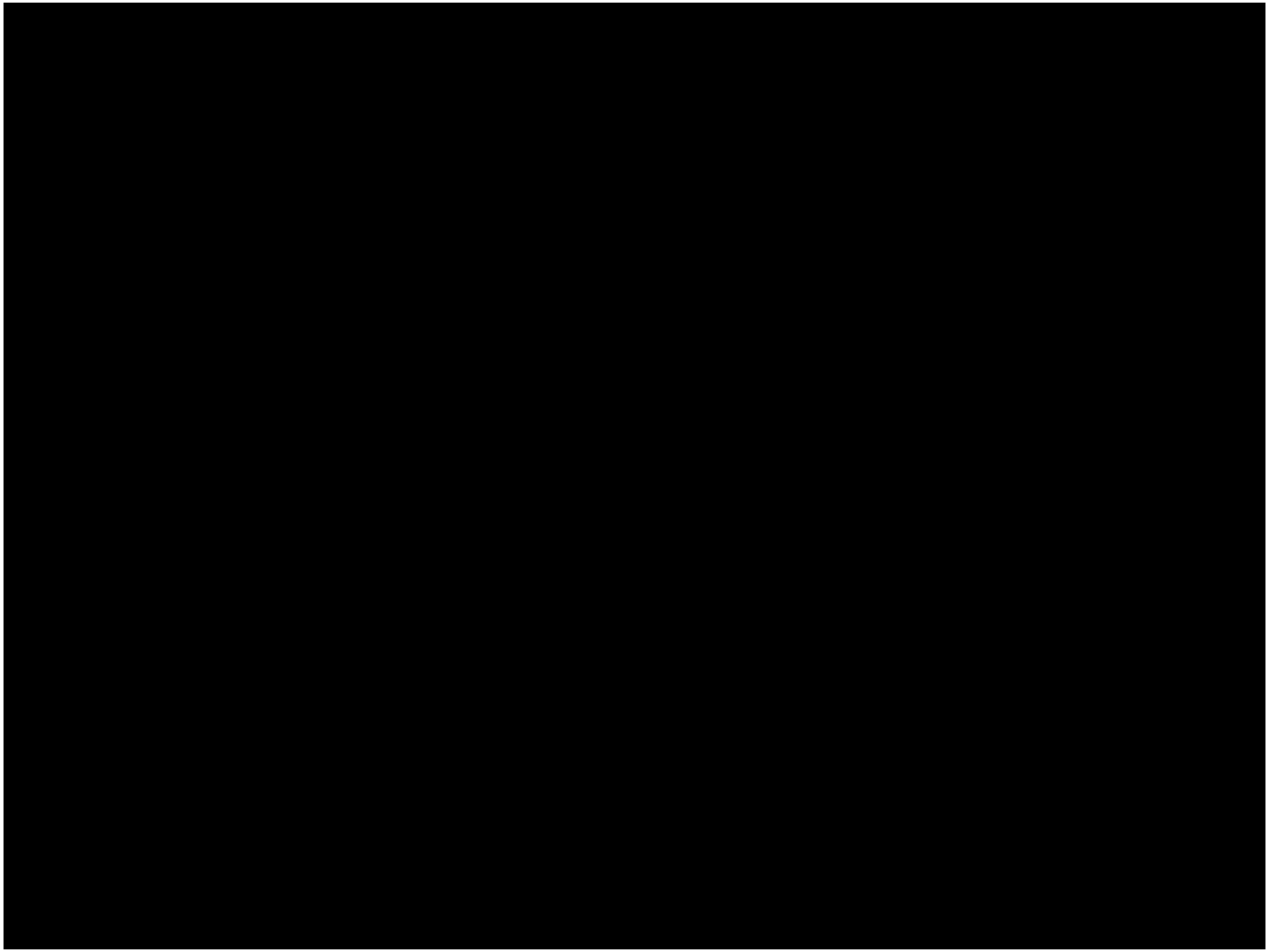
Sequel System

1M ZMW / SMRT cell

5-8Gb



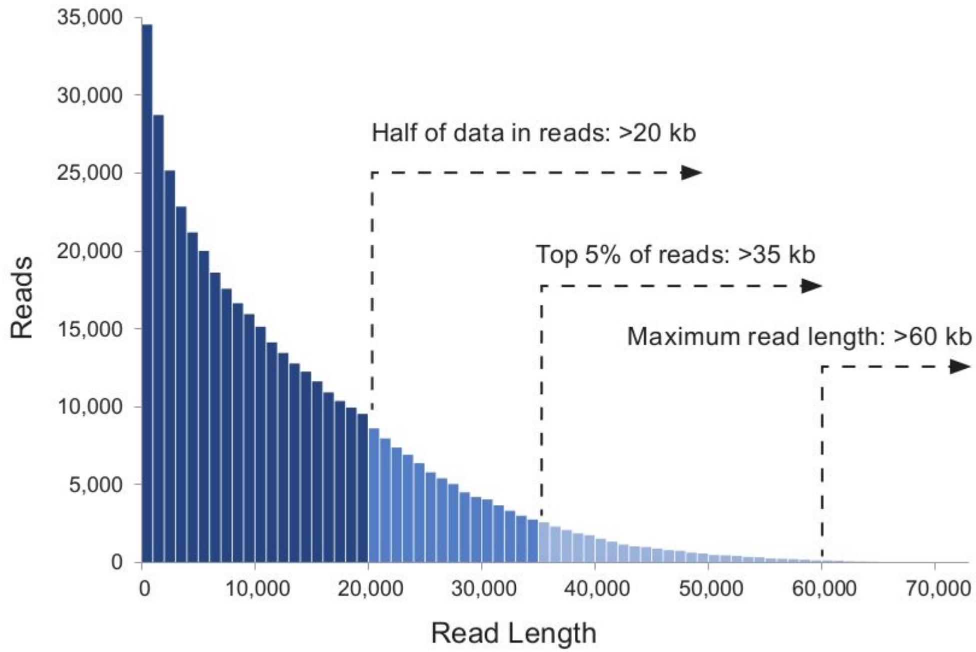
SMRTbell templates



SMRT-sequencing

Long Read Lengths

Read lengths >20 kb
Data per SMRT Cell: 5-8 Gb



Read length data shown above from a 30 kb size-selected human library on the Sequel System (10-hour movie, 2.0 chemistry) with a total output of 7.6 Gb. Sequel System SMRT Cells 1M typically generate ~365,000 reads each. Read lengths, reads/data per SMRT Cell and other sequencing performance results vary based on sample quality/type and insert size among other factors.

Applications :

- Genome sequencing (structural variations)
- Full-length RNA-sequencing (Iso-Seq)
- Haplotyping and variant phasing
- Methylation analysis

Advantages :

- Random error (no GC% bias, no homopolymer stretch bias)
- Long reads > 20kb
- No amplification bias
- Fast run time (0,5 to 10 hours/SMRT cell)
- Reproducibility
- Capacity to detect modified nucleotides

Weaknesses :

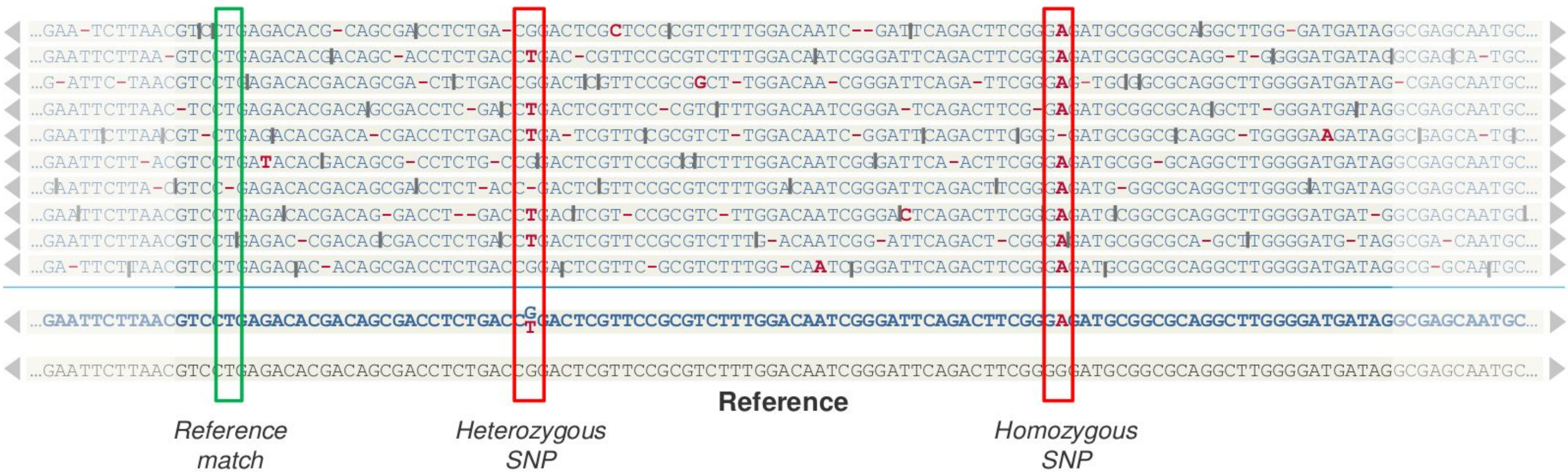
- High error rate : 10-15% (compensated by consensus analysis)
- High cost
- Low output (compared to Illumina)

Errors are random

i. Generate sequence read:

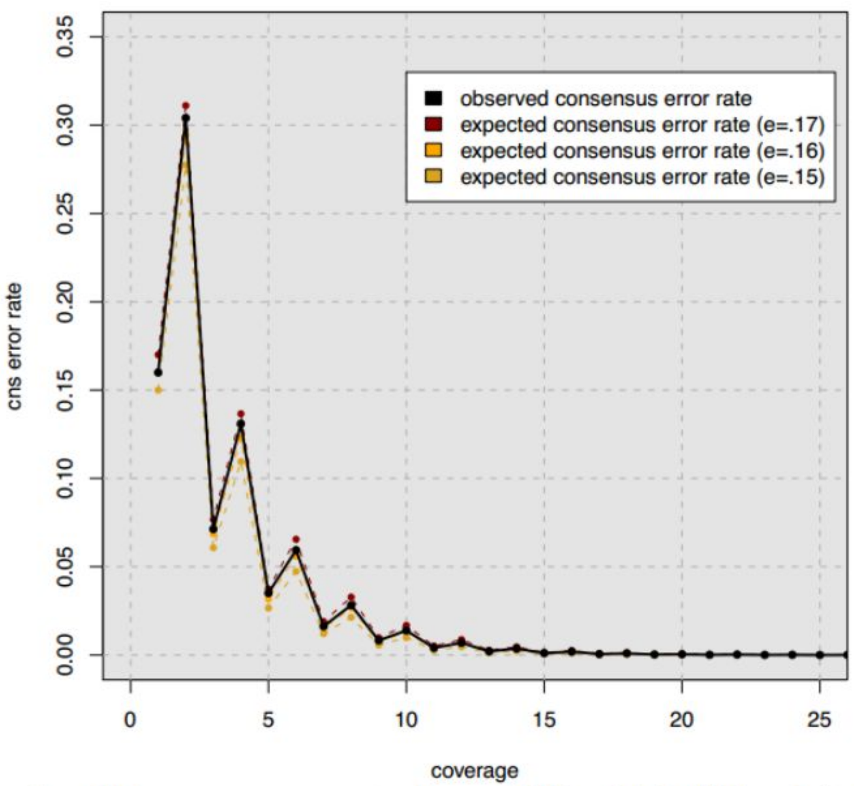
ii. Map to reference:

iii. Generate consensus (10x coverage):

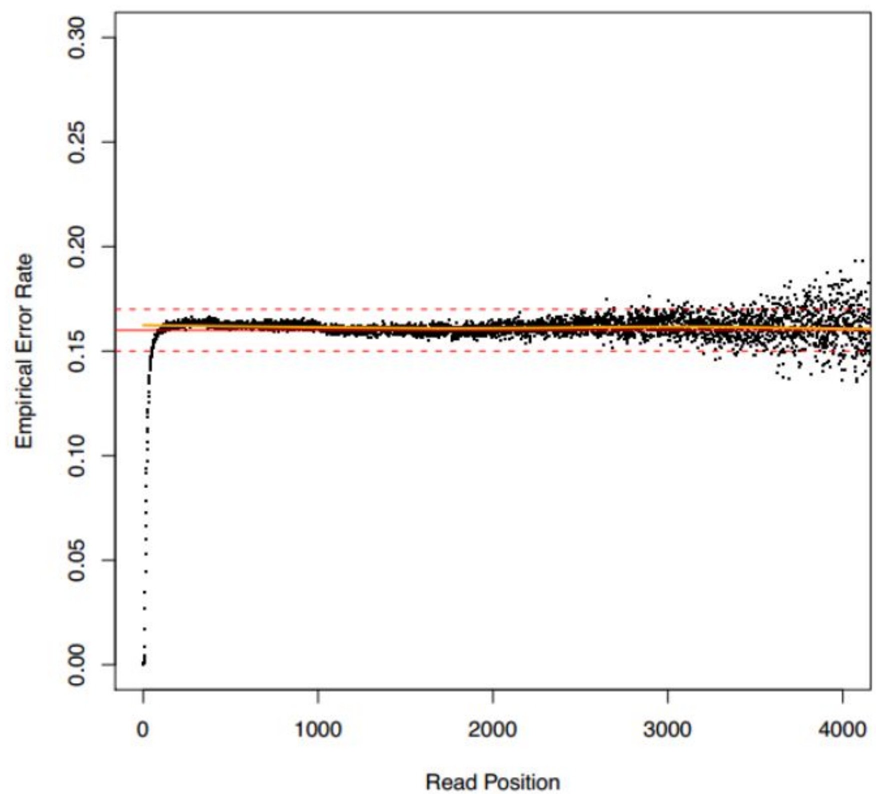


Errors are random

Also confirmed through simulation

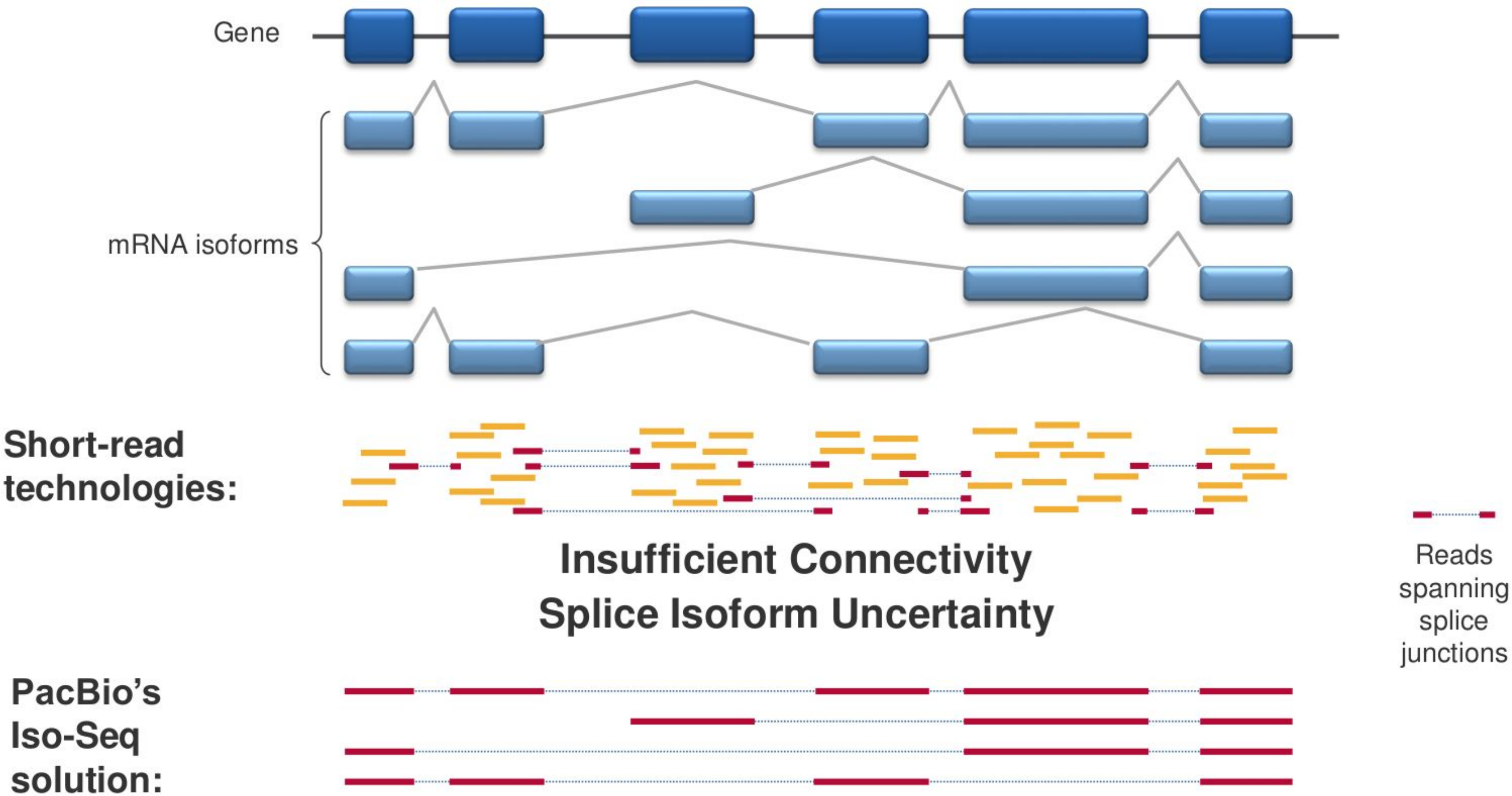


Accuracy is independent from read length



Koren et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology 30, 693-700

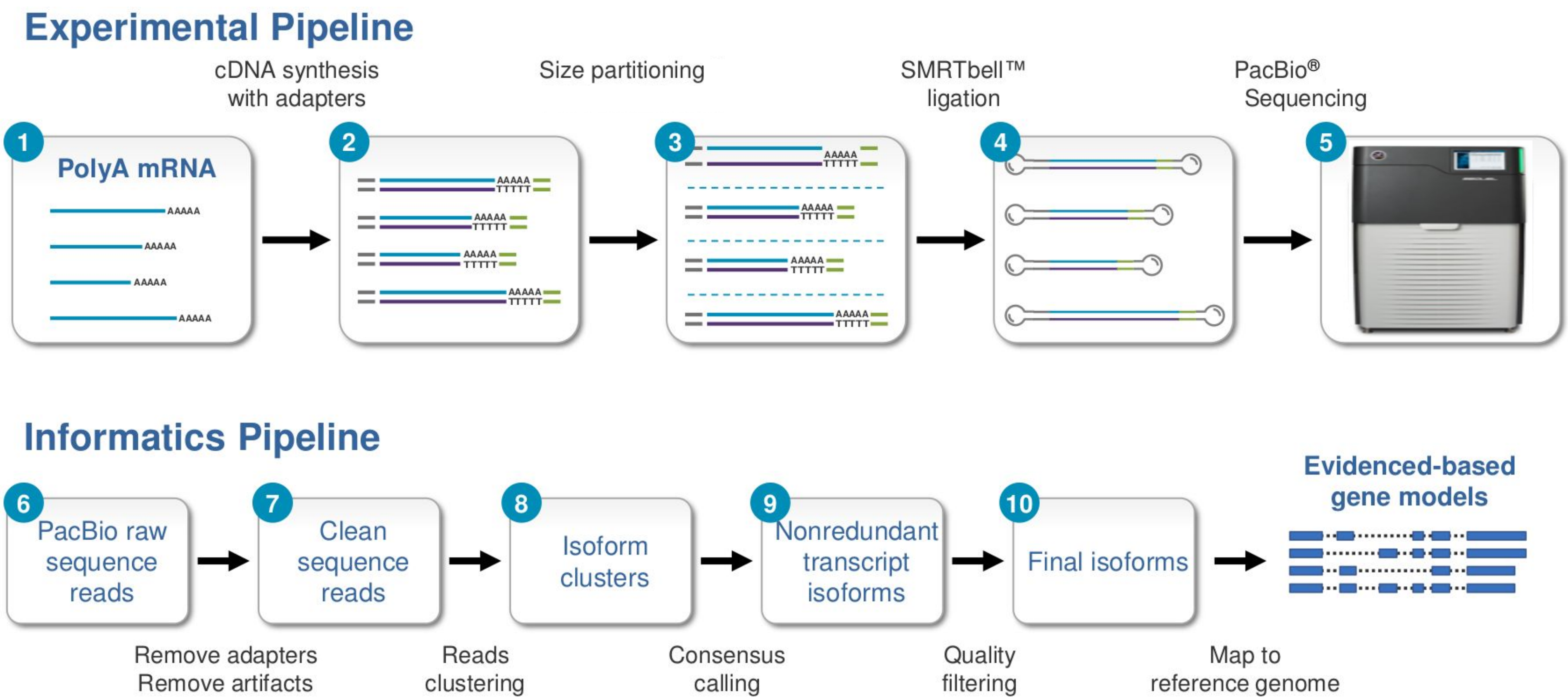
Iso-Seq



Full-length cDNA Sequence Reads

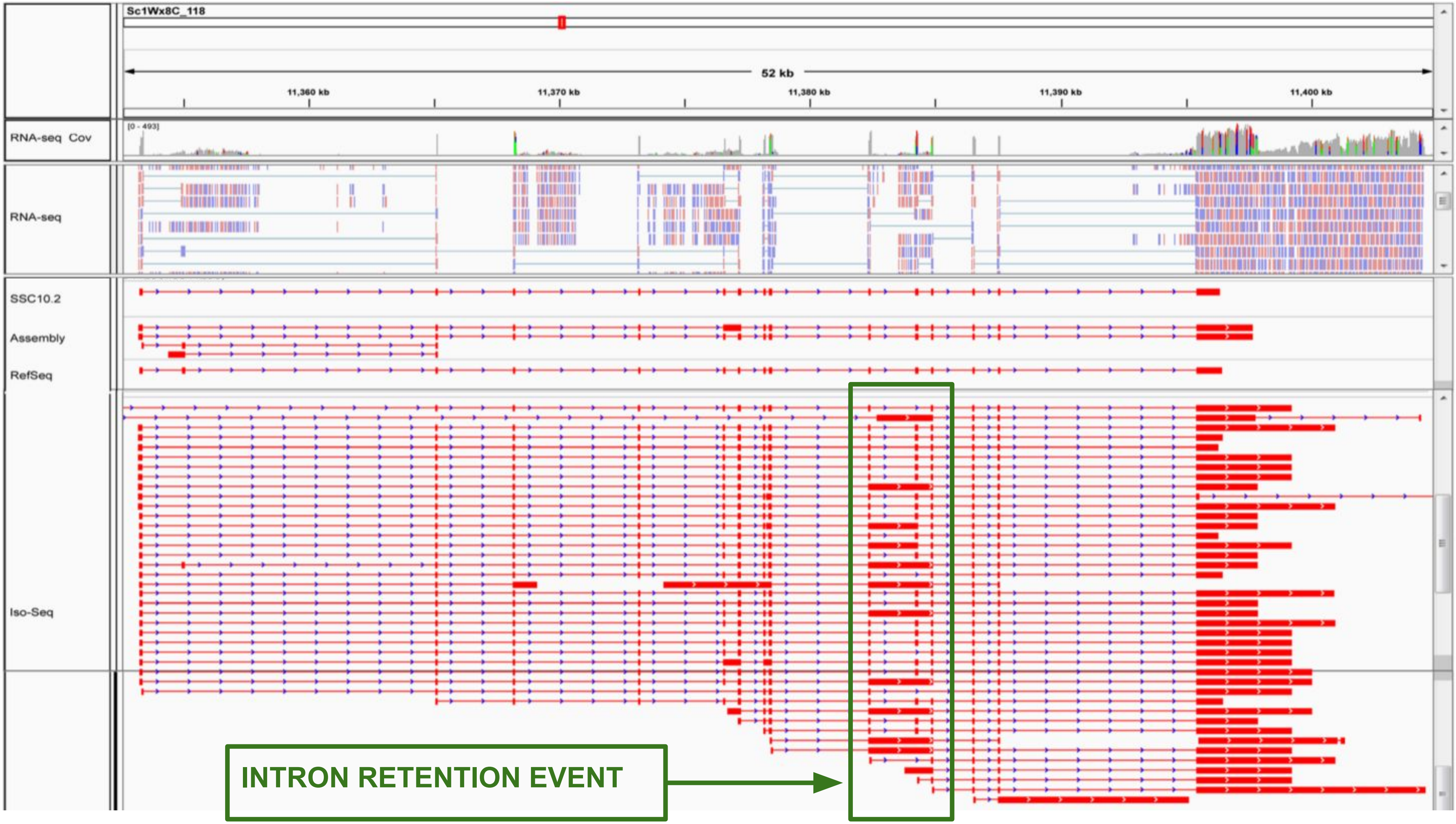
Splice Isoform Certainty – <u>No Assembly Required</u>

Iso-Seq

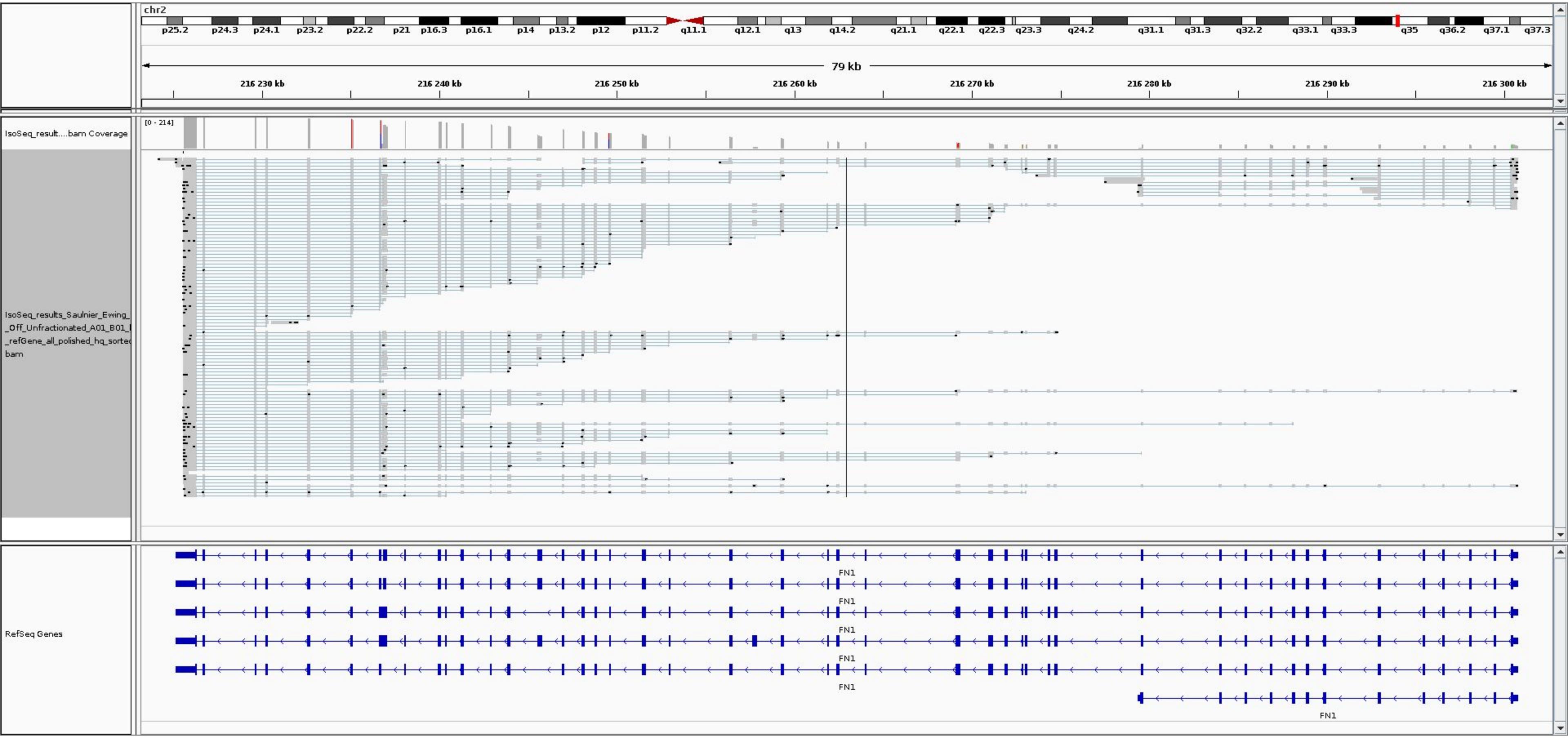


Experimental Pipeline
cDNA synthesis with adapters
Size partitioning
SMRTbell™ ligation
PacBio® Sequencing
1
PolyA mRNA
AAAAA
2
AAAAA
TTTTT
3
4
5
Informatics Pipeline
6
PacBio raw sequence reads
7
Clean sequence reads
8
Isoform clusters
9
Nonredundant transcript isoforms
10
Final isoforms
Evidenced-based gene models
Remove adapters
Remove artifacts
Reads clustering
Consensus calling
Quality filtering
Map to reference genome

Concrete example 1



Concrete example 2

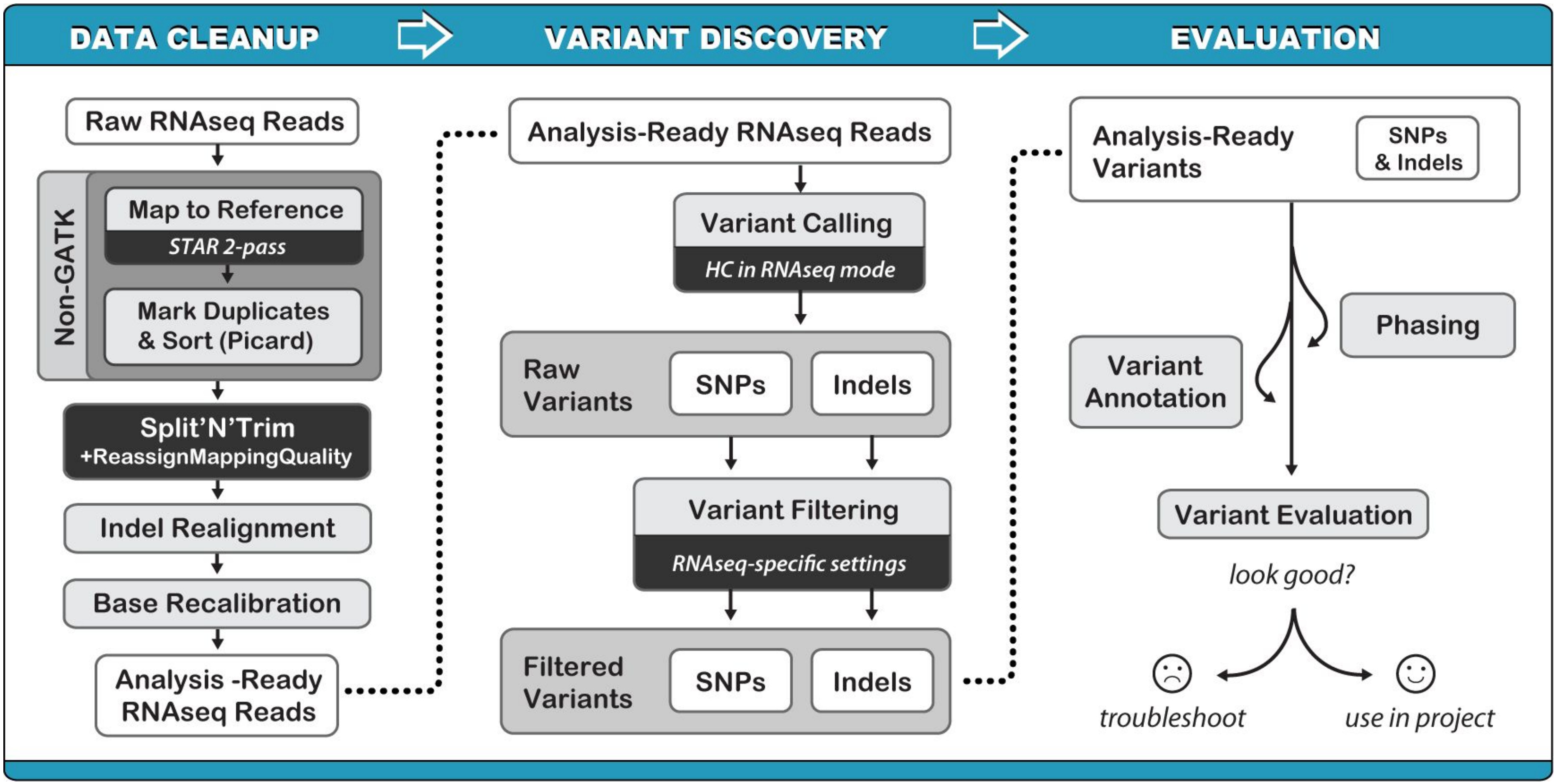


FN1 = Fibronectine1 = 8103 bp

Outline

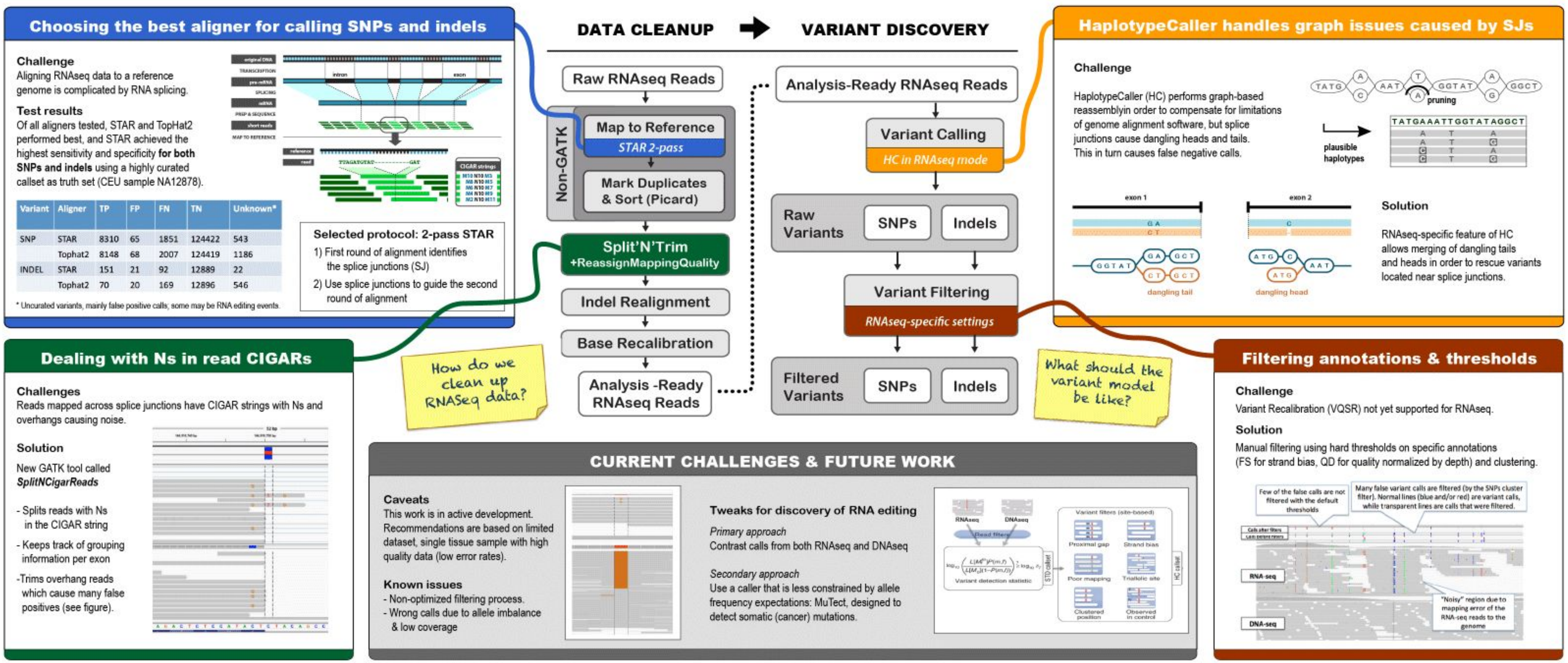
- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- **Expressed SNVs/indels variants detection**
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Expressed SNVs/indels



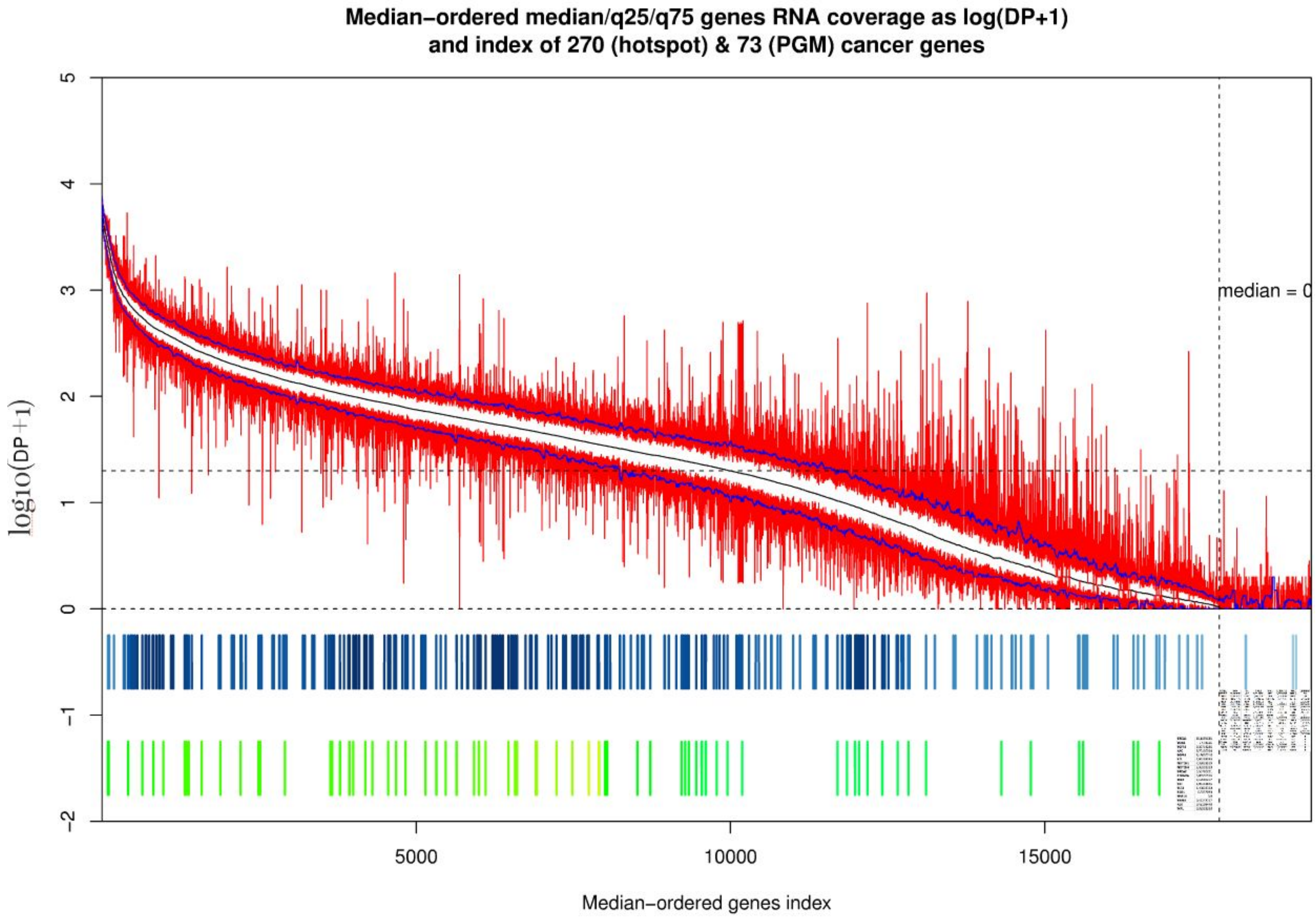
https://software.broadinstitute.org/gatk/documentation/article.php?id=3891

Expressed SNVs/indels



https://software.broadinstitute.org/gatk/documentation/article.php?id=3891

Expressed SNVs/indels



Expressed SNVs/indels

BRCA1	19,1435185
DDR2	17,78125
FGFR3	9,35714286
APC	8,76136364
FGFR4	8,18656716
KIT	7,95238095
NOTCH1	7,33823529
NOTCH4	6,38333333
BRCA2	5,55769231
CDKN2A	5,05555556
JAK3	4,19565217
RET	1,96153846
FLT3	1,45833333
ROS1	0,8372093
HNF1A	0,8
ERBB4	0,47272727
ALK	0,43103448
MPL	0,33333333

BCORL1	19,54	CBL	9,0625	MYCL	5,28571429	FGF10	0,83333333
KMT2A	19,1805556	APC	8,76136364	FLT4	5,06666667	HNF1A	0,8
BRCA1	19,1435185	PTCH1	8,61437909	FLT4	5,06666667	WISP3	0,8
RPTOR	18,328125	EPHA3	8,47916667	CDKN2A	5,05555556	NTRK3	0,77692308
BRD3	18,0454546	FGFR4	8,18656716	BCL2	5	WT1	0,76041667
SUFU	17,9347826	BRIP1	8,10526316	JAK3	4,19565217	GRIN2A	0,56756757
DDR2	17,78125	MITF	8,06521739	GATA3	3,8	NTRK1	0,52941177
BTK	17,2307692	CDK6	8	CEBPA	3,125	ERBB4	0,47272727
CCND2	16,5	KIT	7,95238095	IKZF1	3,06097561	ALK	0,43103448
FCHSD1	16,225	CARD11	7,79166667	ZNF703	2,5	GATA1	0,4
BLM	15,6410256	NTRK2	7,67441861	PAX8	2,27777778	MPL	0,33333333
BARD1	15,0512821	KIAA1549	7,6375	AR	2,21875	FGF19	0,33333333
INHBA	15	HGF	7,5877193	EPHB1	2,09375	NTRK1	0,30106952
RICTOR	14,3962264	NOTCH1	7,33823529	RET	1,96153846	LRP1B	0,24725275
ETV1	14,1643836	CD79A	6,95	IRF4	1,75	NKX2-1	0,2
TET2	13,1	IL7R	6,9375	MYCN	1,75	EPHA5	0,16091954
IRS2	12,5	NOTCH4	6,38333333	ESR1	1,6875	NUTM1	0,13636364
CD79B	11,3235294	PIK3CG	6,25	PLAG1	1,6	FGF23	0
STAT4	11,1521739	GATA2	6,03333333	FLT3	1,45833333	FGF3	0
PRDM1	10,25	SLC34A2	5,625	FGF14	1,4	FGF4	0
DOT1L	10,1785714	BRCA2	5,55769231	CDKN2B	1,16666667	FGF6	0
FGFR3	9,35714286	SS18L1	5,55555556	PAK3	0,85227273	SSX1	0
RNF43	9,1	SLC45A3	5,375	ROS1	0,8372093	SOX10	0

Expressed SNVs/indels

The Author(s) *BMC Genomics* 2017, **18**(Suppl 6):690
DOI 10.1186/s12864-017-4022-x

BMC Genomics

RESEARCH Open Access

CrossMark

The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data

Yan Guo[1*†], Shilin Zhao[1†], Quanhu Sheng[1], David C Samuels[2] and Yu Shyr[3*]

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2016
Houston, TX, USA. 08-10 December 2016

Expressed SNVs/indels

OXFORD

Briefings in Bioinformatics, 18(6), 2017, 973–983

doi: 10.1093/bib/bbw069
Advance Access Publication Date: 26 July 2016
Paper

Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations

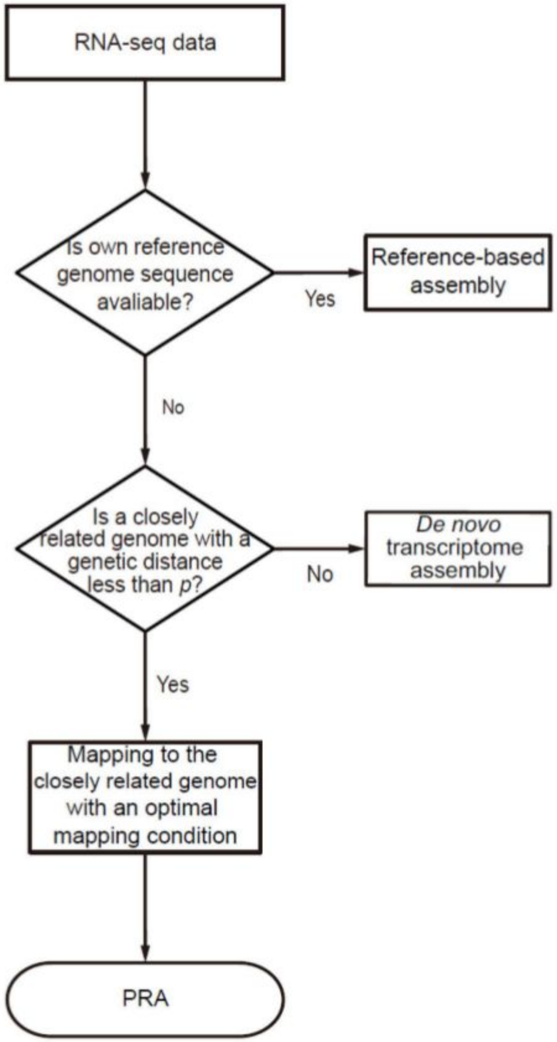
Zhifu Sun, Aditya Bhagwate, Naresh Prodduturi, Ping Yang and Jean-Pierre A. Kocher

Outline

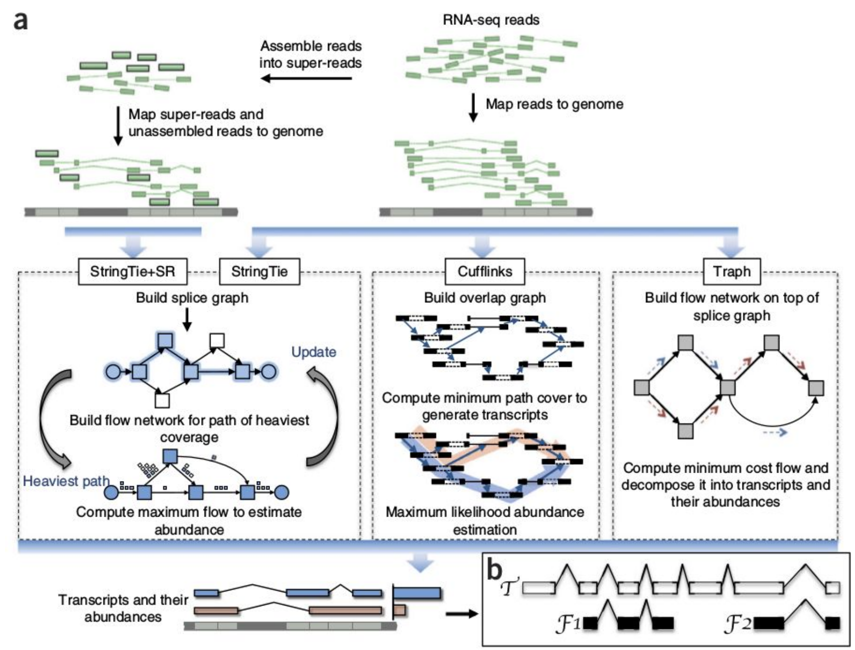
- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- **Transcripts reconstruction (assembly)**
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- RNA editing events detection

Transcripts reconstruction

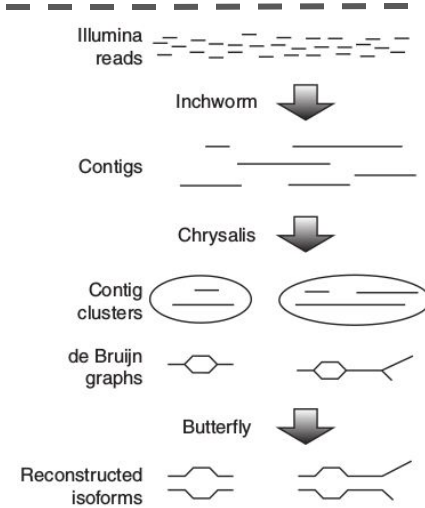
Nam K. et al. 2016. Genes



StringTie (Pertea M. et al. 2015. Nature biotechnology)



Trinity (Haas BJ. et al. 2013. Nature protocols)



Transcripts reconstruction

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 17 2014, pages 2447–2455
doi:10.1093/bioinformatics/btu317

Gene expression

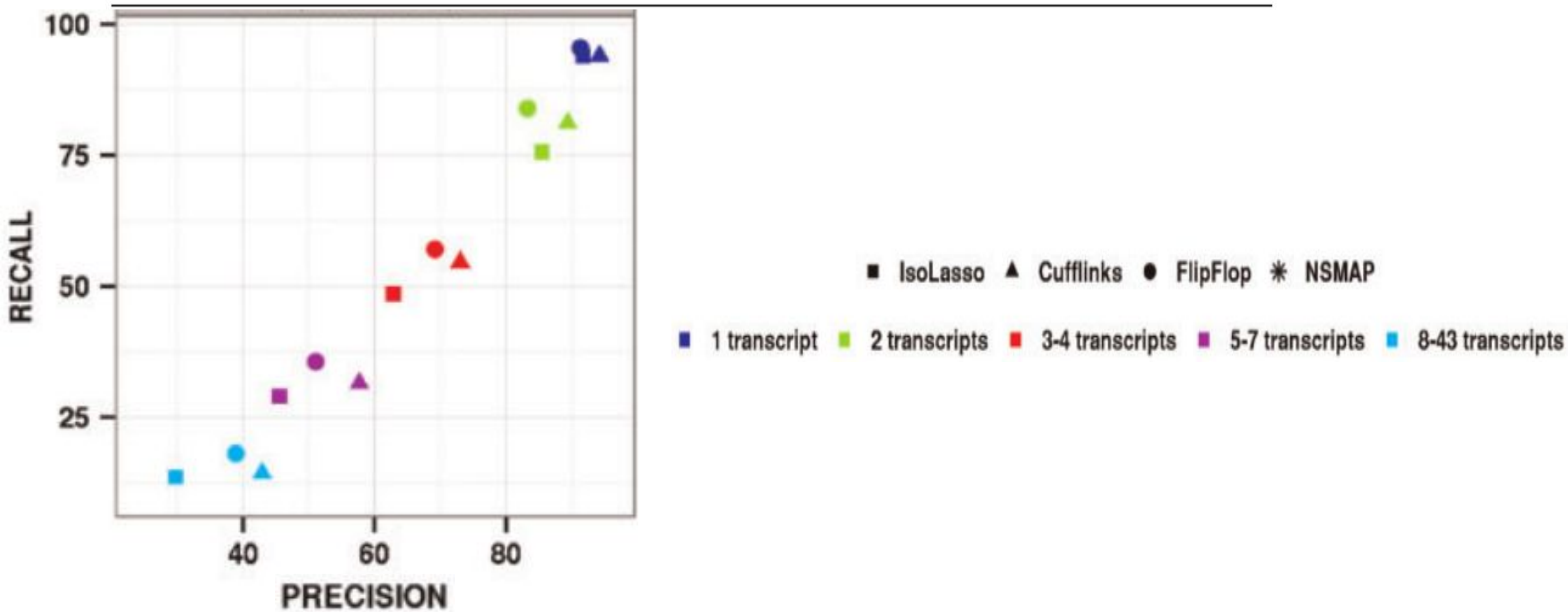
Advance Access publication May 9, 2014

Efficient RNA isoform identification and quantification from RNA-Seq data with network flows

Elsa Bernard[1,2,3], Laurent Jacob[4], Julien Mairal[5] and Jean-Philippe Vert[1,2,3,*]

[1]Mines ParisTech, Centre for Computational Biology, 77300 Fontainebleau, [2]Institut Curie, 26 rue d'Ulm, 75248 Paris Cedex 05, [3]INSERM U900, Paris F-75248, France, [4]Laboratoire Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, INRA, UMR5558, Villeurbanne, France and [5]LEAR Project-Team, INRIA Grenoble Rhône Alpes, 38330 Montbonnot, France

Associate Editor: Janet Kelso



How insufficient

STEIJGER (2013) VS. ANGELINI (2014) VS. CHANG (2014)

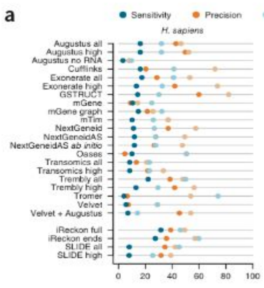


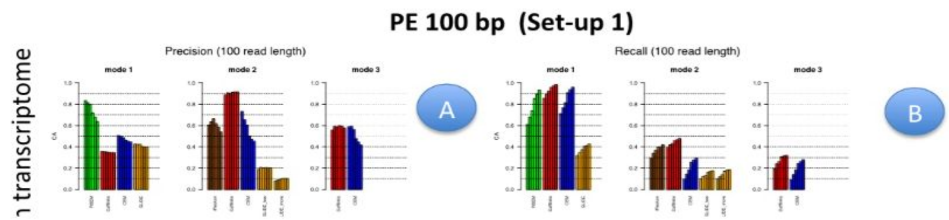
Figure 5 | Transcript assembly performance. (**a**) Reference transcripts with a matching submission entry (transcript-level sensitivity, blue) and reported transcripts that match the reference (transcript-level precision, orange). (**b**) Transcripts for which various subsets of constituent exons have been reported.

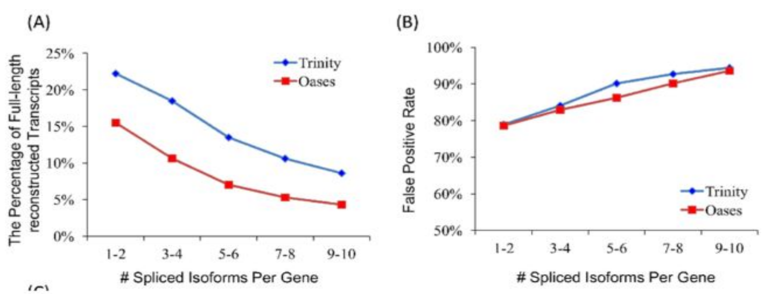
Steijger (2013): recall best at 20%, precision best at 40%

Performance worse for non-coding transcripts

Angelini (2014): recall best < 30%, precision best ~60%

Simulated coding transcripts





Chang (2014): recall best < 25%, precision best ~20%

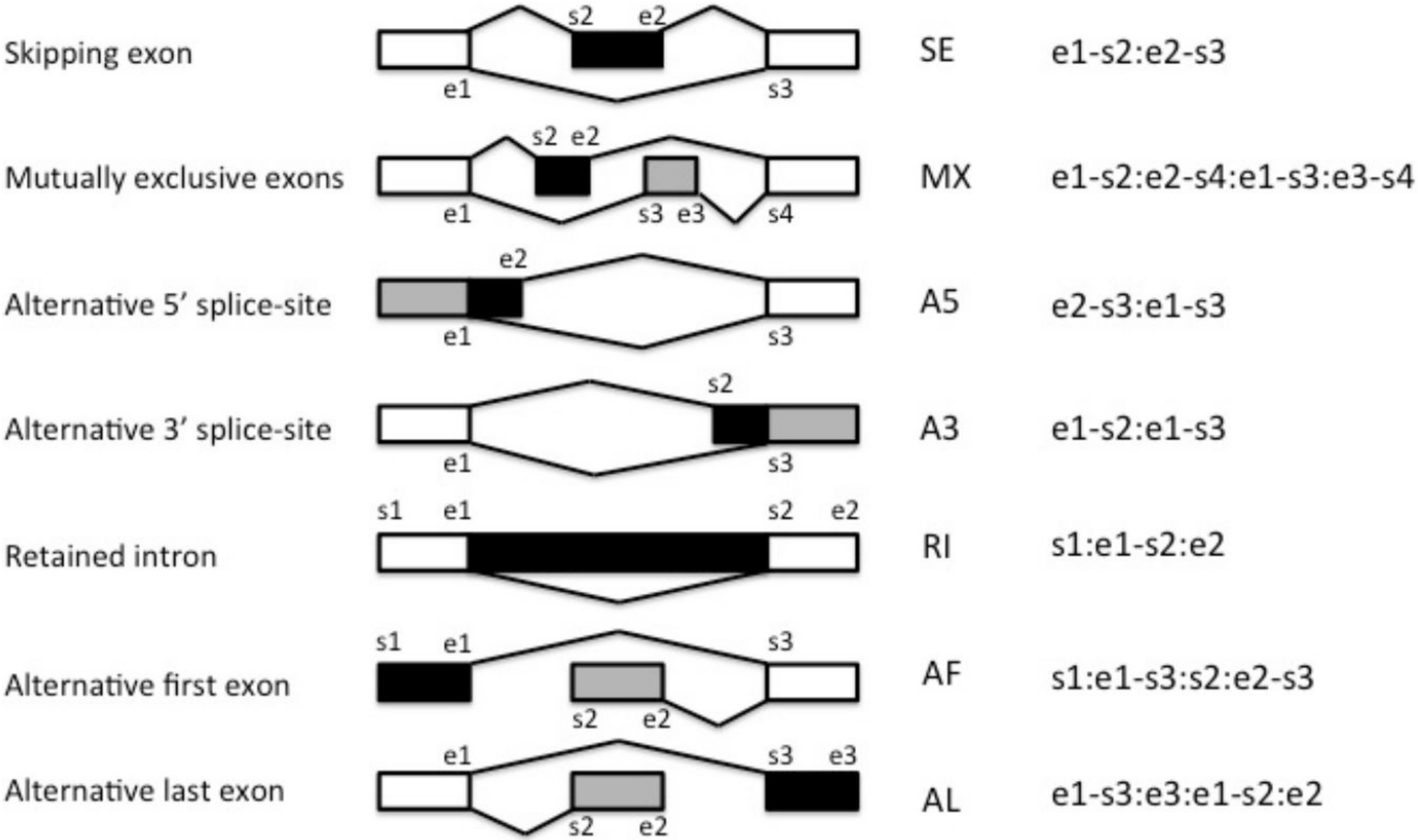
Simulated coding+non-coding transcripts

Steijger, T. et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Meth 10, 1177–1184 (2013).
Angelini, C., Canditiis, D. & Feis, I. Computational approaches for isoform detection and estimation: good and bad news. BMC Bioinformatics 15, 135–43 (2014).
Chang, Z., Wang, Z. & Li, G. The Impacts of Read Length and Transcriptome Complexity for De Novo Assembly: A Simulation Study. PLoS ONE 9, e94825–8 (2014).

Outline

- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- **Alternative splicing / polyadenylation events detection**
- Virus/phages expression detection
- RNA editing events detection

Alternative splicing / polyA



Alamancos GP. et al. 2015. RNA

Alternative splicing / polyA

Table 5 Summary of the main observation for selected methods

	Class	Novel AS	Detection region	Comments
DiffSplice	IR	Any type	ASM	Assembles transcriptome based on graph theory. Does not rely on annotation but does not use annotation either. The goodness of ASM is questionable. Generally low AUC. Performs poorly when detecting SE events.
Cufflinks	IR	Any type	Gene	Assembled transcripts merge with annotation to provide a more confident reference. Is least affected by incomplete annotation. Model is designed for pair-end data. Performs better for medium read depth than both low and high read depth. Performs better when detecting A3SS and A5SS events than other types of AS events. Computationally slow, but allows parallelization.
DEXSeq	CB	Only SE	Exon	Uses a generalized linear NB model. Achieves the highest AUC in many cases using accurate annotation. However, incomplete annotation can impose considerable problems for it. Poor FDR control.
MATS	CB	NS	AS event	Uses a Bayesian model. Solely based on junction reads. Can not detect complex AS events. Annotates splicing events with corresponding event types. Good FDR control in many simulation studies. Performs the best for real data.
rDiff-param	CB	NS	Gene	Conservative with default settings. Good FDR control but low AUC in many cases. Computationally fast.
SplicingCompass	CB	Only SE	Gene	Compares geometry angles of read count vectors. Generally poor FDR control and Medium AUC. Performs well when detecting SE events.
DSGseq	CB	Only SE	Gene	No p-value reported. Generally medium AUC. Performs well when detecting IR events and when using incomplete annotation. Computationally fast.
SeqGSEA	CB	Only SE	Gene	Integrates DE analysis with DS analysis. Generally high AUC. Requires a sample size around 5 to claim significance at a reasonable FDR level, i.e. $FDR = 0.05$. Computation time increases dramatically as permutation times increases.

IR: Isoform resolution models.
CB: Count based models.
NS: Not Supported.
ASM: Alternative Spliced Module.

Liu R. et al. 2014. BMC Bioinformatics

Alternative splicing / polyA

A

Gene Expression data

Gene	Median Expression	# Exon Skipping	# Alt. '3 Splice border	# Alt. '5 Splice border	# Intron Retention
TTN	0.732	134	0	1	31
TNXB	7.364	43	8	5	1
NBPF14	7.716	42	2	0	1
NEB	0.497	32	0	1	1
NBPF10	4.685	29	0	0	0
ABI3BP	7.869	29	1	1	0
NEBL	1.989	22	1	1	0
UTY	0.089	22	5	3	0
OBSCN	0.885	22	4	2	1

Page 1 of 705 View 1 - 30 of 21,128

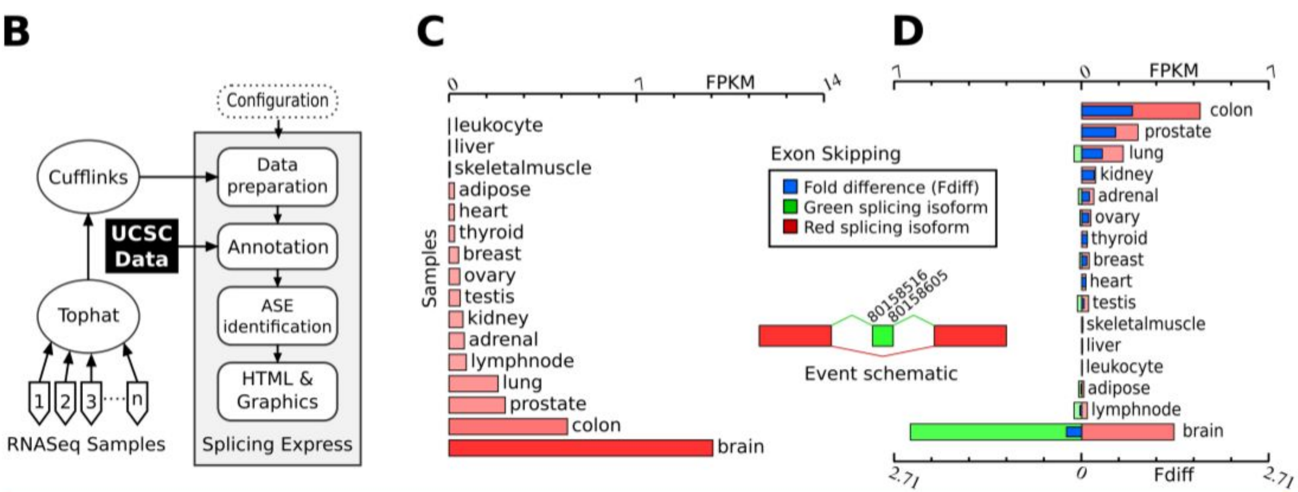


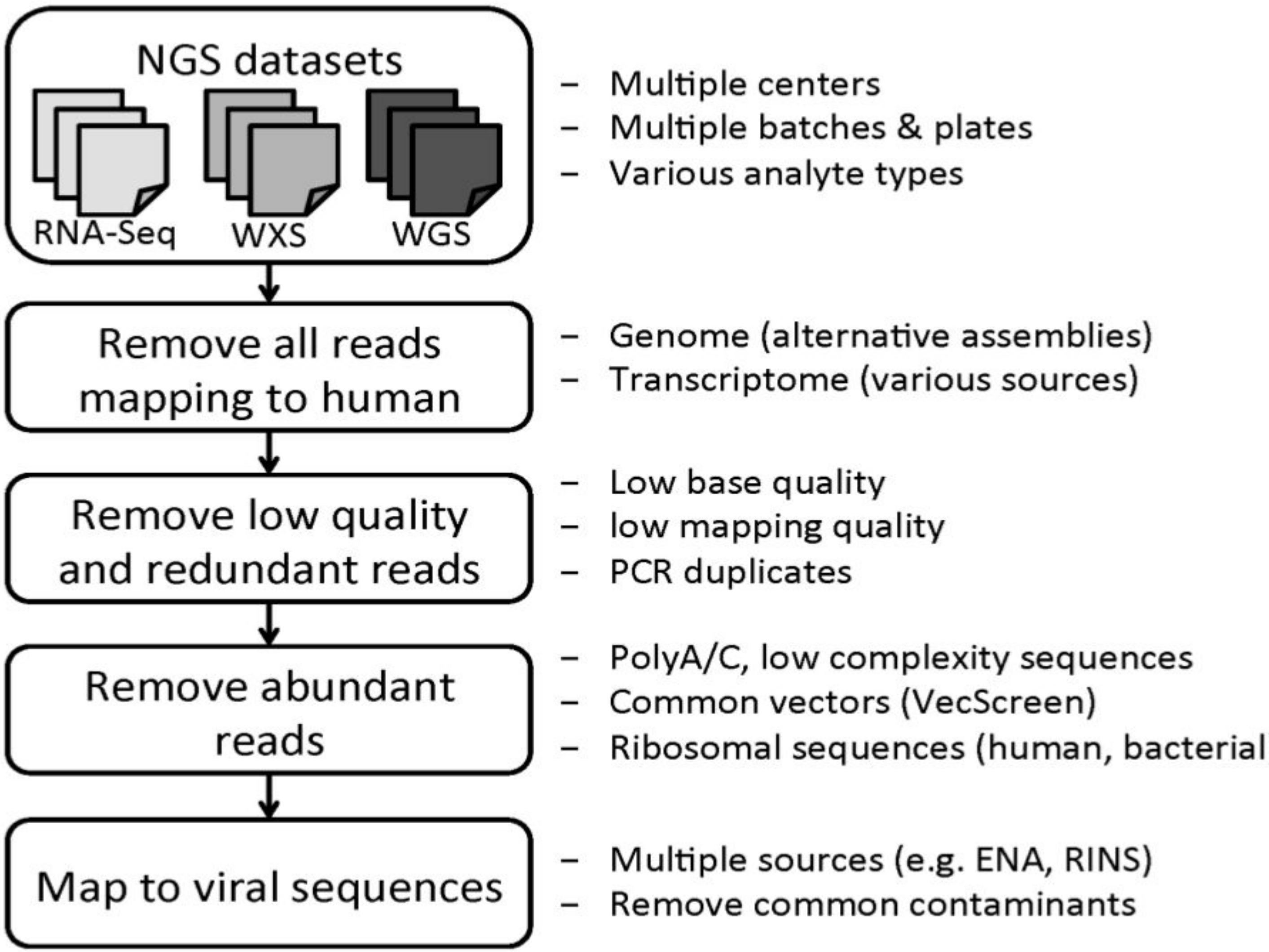
Figure 1 **Overview of *Splicing Express* features.** (A) Table generated by *Splicing Express* showing expression level and number of variants in a gene-by-gene basis. (B) Schematic representation of the computational pipeline used by *Splicing Express*. (C) Graphic represention of the expression pattern of NRXN3. (D) Graphic represention of the differential expression of one isoform (exon skipping) of NRXN3. The exon skipping isoform is represented in red, while the isoform including the respective exon is colored in green.

Kroll JE et al. 2015. PeerJ

Outline

- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- **Virus/phages expression detection**
- RNA editing events detection

Virus/Phages expression



Kazemian M. et al. 2015. JVI

Virus/Phages expression

Table 1. Association of viruses with human cancer.

Cancer	#Samples	HBV	HCV	EBV	CMV	HHV6B	HPV16	HPV38	Other viruses
ACC - adrenocortical carcinoma	79								
BLCA - bladder urothelial carcinoma	119			1	3	1	1		3(HPV6)
BRCA - breast invasive carcinoma	125					1			
CESC - cervical squamous cell carcinoma	91				1	1	59		11(HPV18), 9(HPV45), 3(HPV59), 13(HPV^{other})
COAD - colon adenocarcinoma	44			2	4	2			2(HPyV6)
DLBC - diffuse large B-cell lymphoma	28								
GBM - glioblastoma multiforme	95								
HNSC - head & neck squamous cell carcinoma	123			2	1		14		2(HPV33), 2(HPV35), 1(HPV56), 1(HHV1)
KICH - kidney chromophobe	66								
KIRC - kidney renal clear cell carcinoma	67								1(HPV18), 1(HPV94)
KIRP - kidney renal papillary cell carcinoma	120								
LGG - brain lower grade glioma	100	1							
LIHC - liver hepatocellular carcinoma	115	22	6		1		1		1(HPV18), 1(AAV2)
LUAD - lung adenocarcinoma	125								
LUSC - lung squamous cell carcinoma	125			3			1		1(HPV30)
PRAD - prostate adenocarcinoma	124			1	1				
READ - rectum adenocarcinoma	36			5	4				
SKCM - skin cutaneous melanoma	82				1		1		
THCA - thyroid carcinoma	123								
UCEC - uterine corpus endometrioid carcinoma	168							30*	
UCS - uterine carcinosarcoma	57				1				1(HPV5), 1(HPV38b), 1(HPV133)
Grand Total	2012	23	6	14	17	5	77	30	55

Kazemian M. et al. 2015. JVI

Virus/Phages expression

Table 4. Quantification of three types of HPV from PCR (number of copies of HPV per ng of DNA) vs. RNAseq (counts and percentage of total).

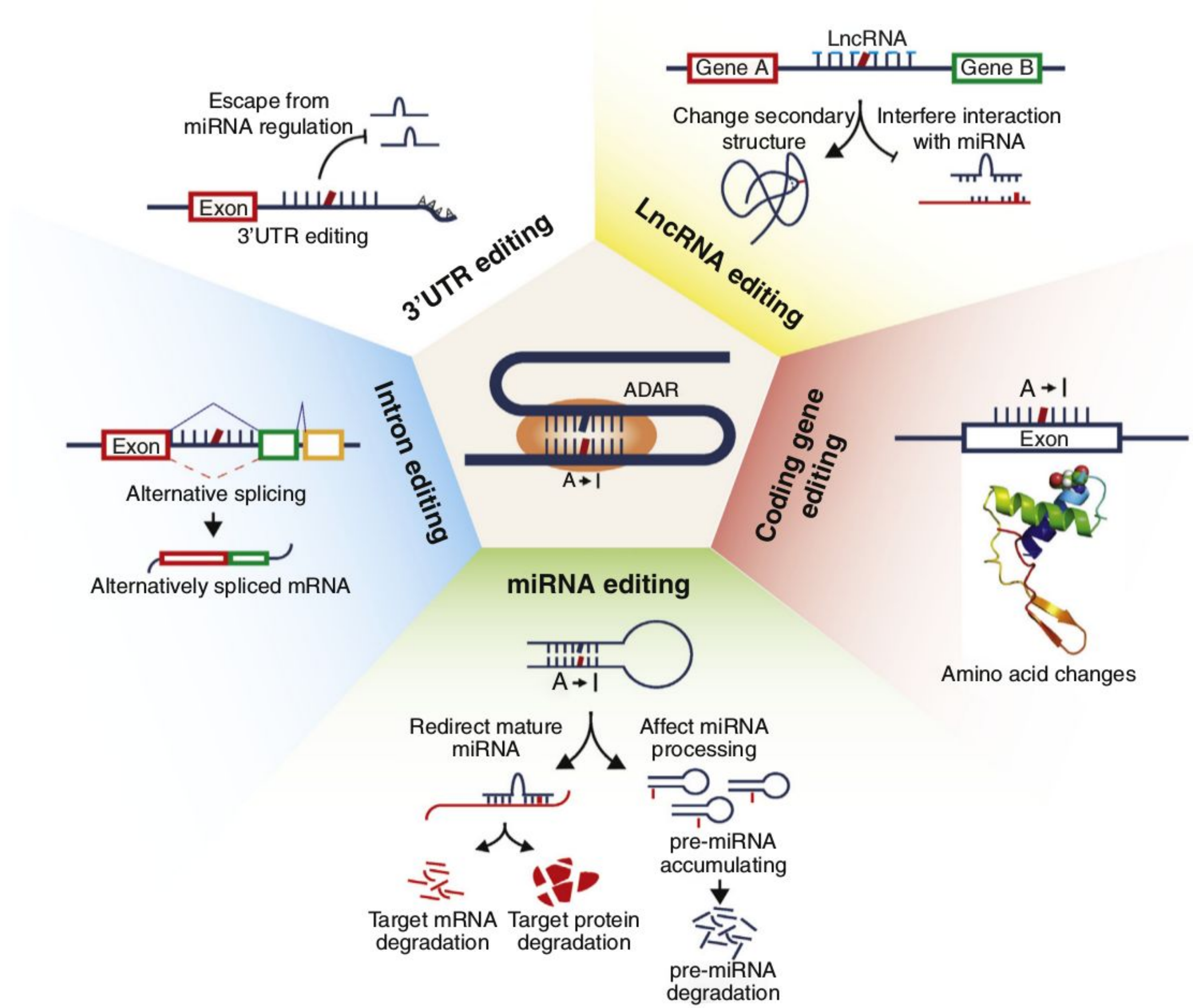
		PCR			RNAseq		
ID	**Cancer type**	**HPV16**	**HPV18**	**HPV33**	**HPV16**	**HPV18**	**HPV33**
370	H&N	-	-	-	-	-	-
536	H&N	-	-	-	-	-	-
646	H&N	-	-	-	-	-	-
653	H&N	379	-	-	14482 (0.01%)	-	-
666	H&N	842	-	-	30421 (0.02%)	-	-
670	H&N	-	-	-	-	-	-
683	H&N	-	-	-	-	-	-
708	H&N	-	-	-	-	-	-
721	H&N	-	-	-	-	-	-

Koeppel F. et al. 2018. Submitted to International Journal of Genomics

Outline

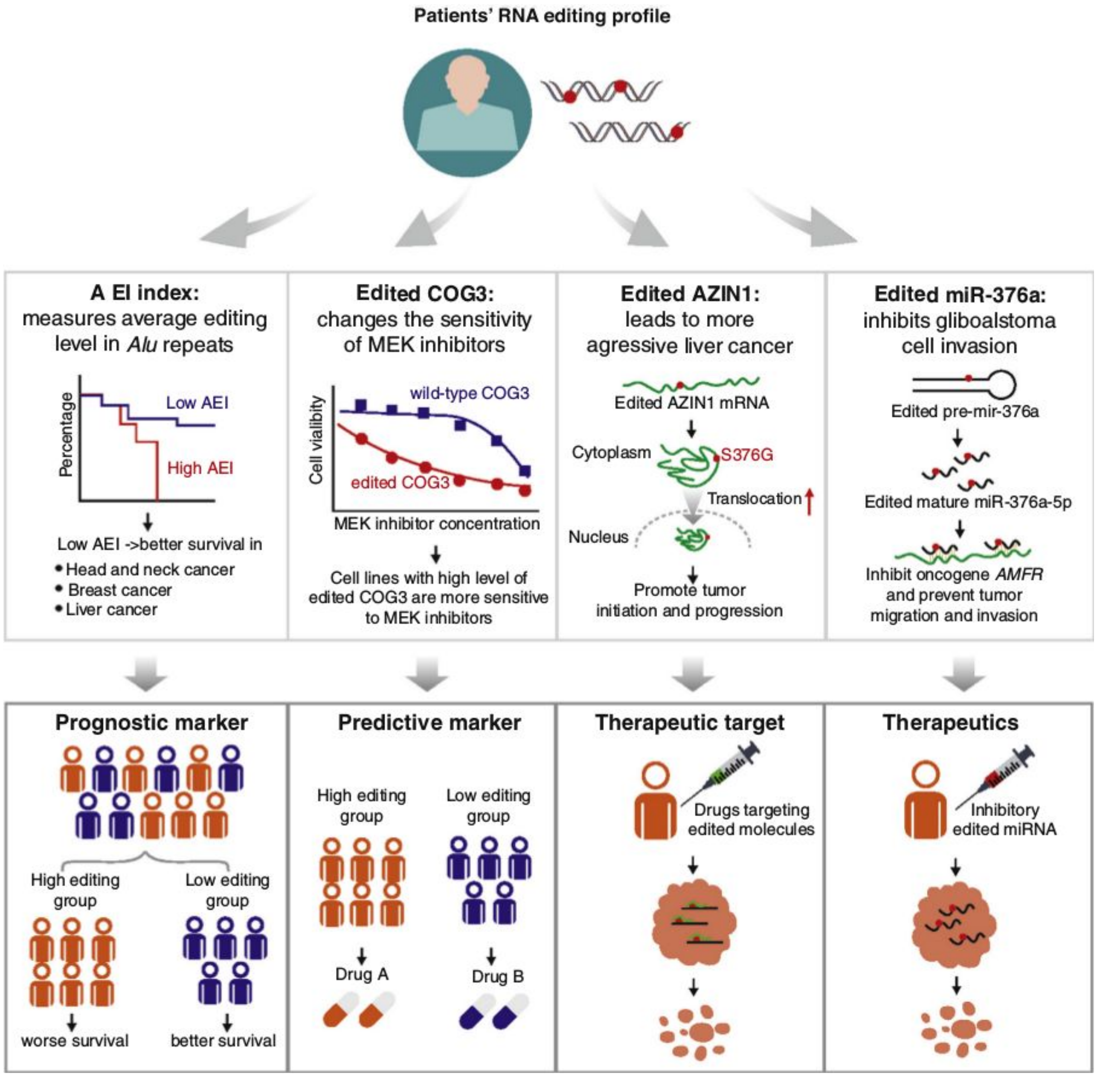
- Experimental and sequencing design
- Quality control and mapping
- Genes/transcripts/exons quantification
- Allele-specific quantification
- Functional Analysis
- Fusion transcripts / chimera detection
- Single-cell RNA-seq
- Full-length transcripts sequencing using long reads
- Expressed SNVs/indels variants detection
- Transcripts reconstruction (assembly)
- Alternative splicing / polyadenylation events detection
- Virus/phages expression detection
- **RNA editing events detection**

RNA editing



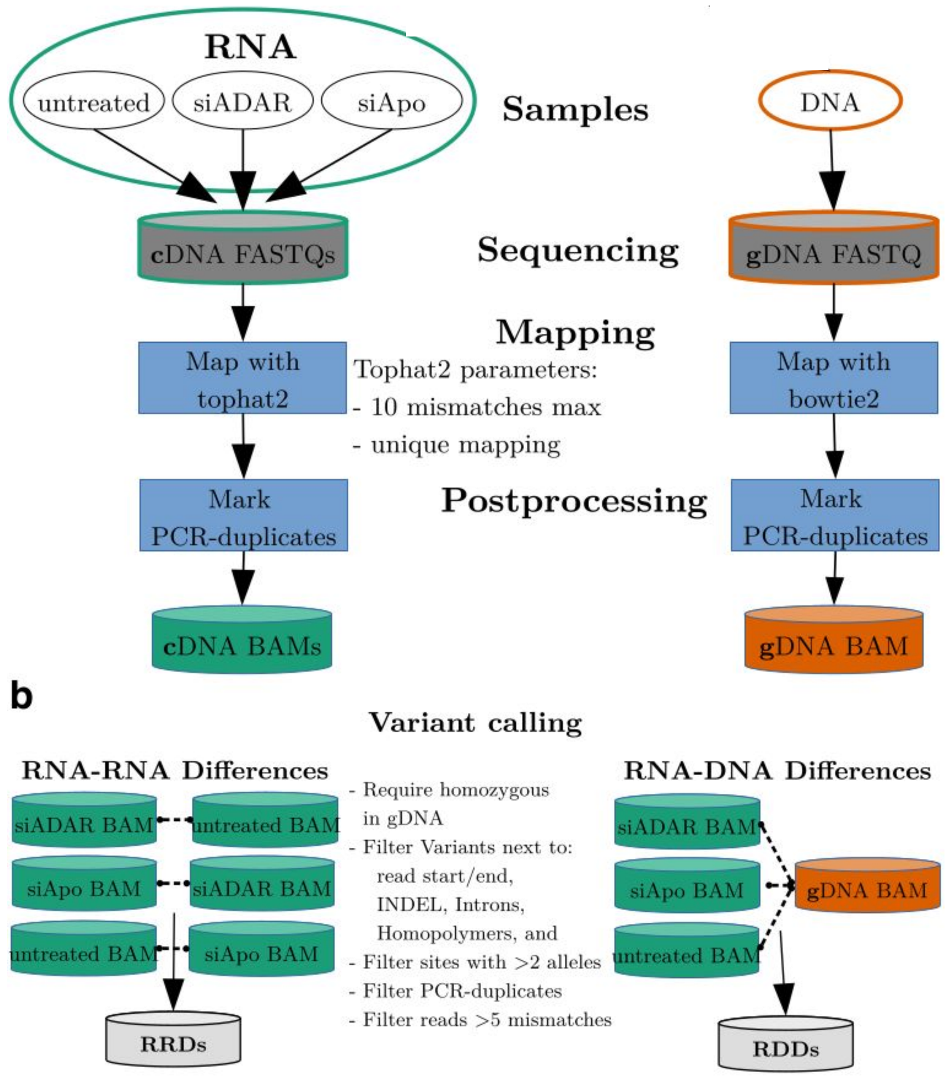
Xu X. et al. Oct 2017. Current Opinion in Genetics & Development

RNA editing



Xu X. et al. Oct 2017. Current Opinion in Genetics & Development

RNA editing



Diroma MA. et al. 2017. Briefings in Bioinformatics

RNA editing

Table 1. Main features of tested tools for RNA editing detection

Name	GIREMI	JACUSA	REDItools	RES-Scanner	RNAEditor
Required dependencies	HTSlib, SAMtools, R	R for JacusHelper	pysam, BLAT, SAMtools	Perl modules, BWA, SAMtools, BLAT	pysam, pyqt4, matplotlib, numpy, BWA, Picard Tools, GATK, BLAT, BEDtools
Required input files	BAM+ filtered SNVs	BAM	BAM	Fastq or BAM	Fastq
Stranded-oriented samples	Yes	Only single-end reads	Yes	Only FR-stranded (dUTP-protocol)	Yes, but no specified strand
RNA replicates accepted	Yes	Yes	No	No	No
De novo detection of RNA editing sites	Yes	Yes	Yes	Yes	Yes
DNA-RNA comparison	No	Yes	Yes	Yes	No
Works with RNA solo	Yes	Only with replicates	Yes	No	Yes
Multiple types of RNA editing	Yes	Yes	Yes	Yes	Yes
Mapping included	No	No	No	Optional (BWA)	Mandatory (BWA)
SNV calling	No	Yes	Yes (Samtools)	Yes (Samtools)	Yes (GATK)
Annotation of editing sites	No	No	Yes	Yes	Yes
Filtering or P-value provided	P-value	Filtering	Filtering	Filtering+P-value	Filtering

Diroma MA. et al. 2017. Briefings in Bioinformatics


That's all Folks!