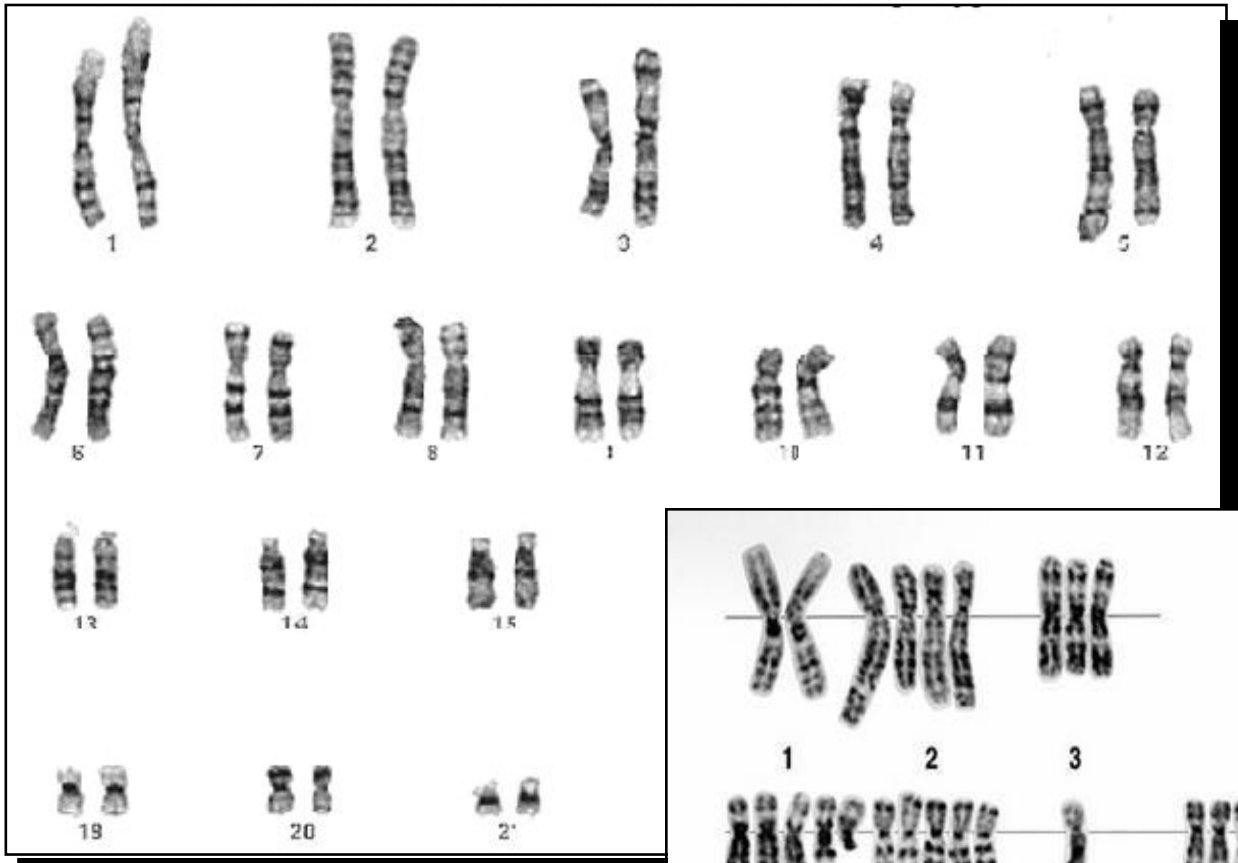
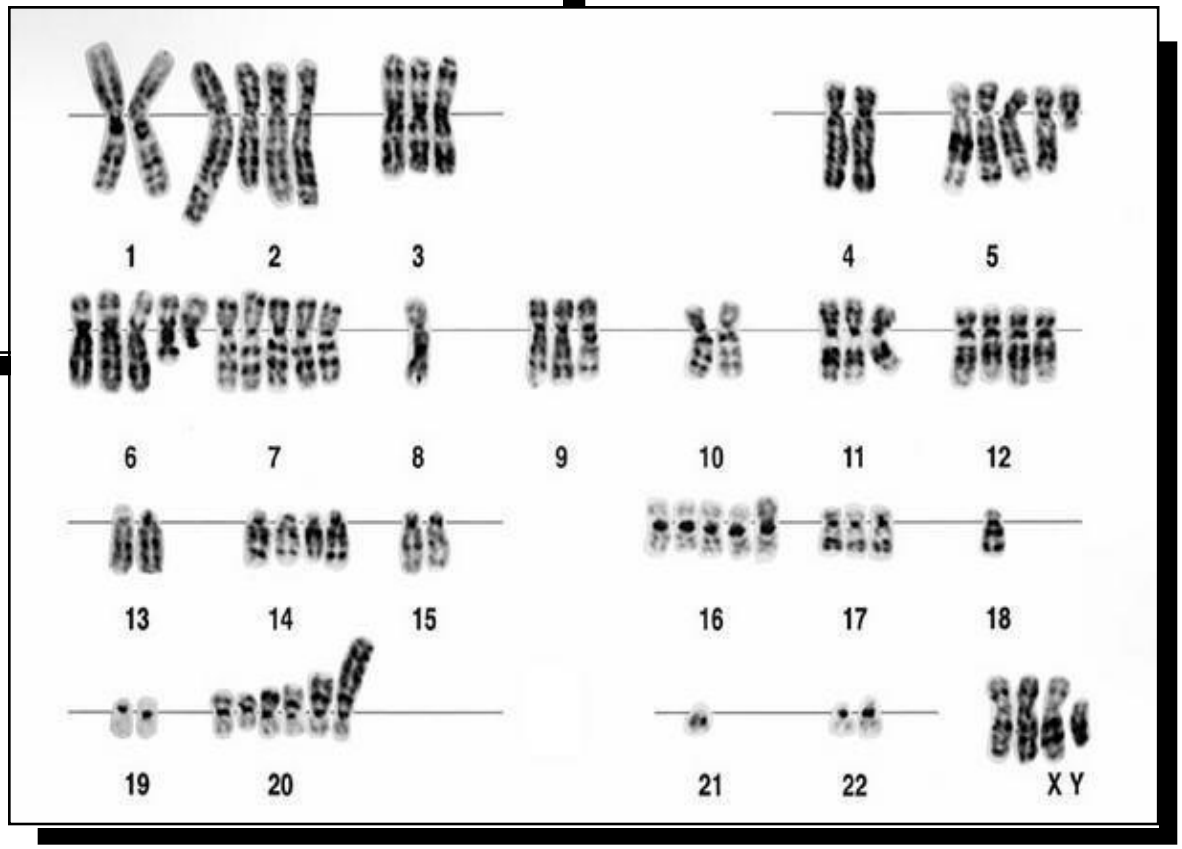
RNA-seq: A high-resolution View of the Transcriptome

Sean Davis, M.D., Ph.D.
Univ of Colorado Anschutz School of Medicine
2022-3-28

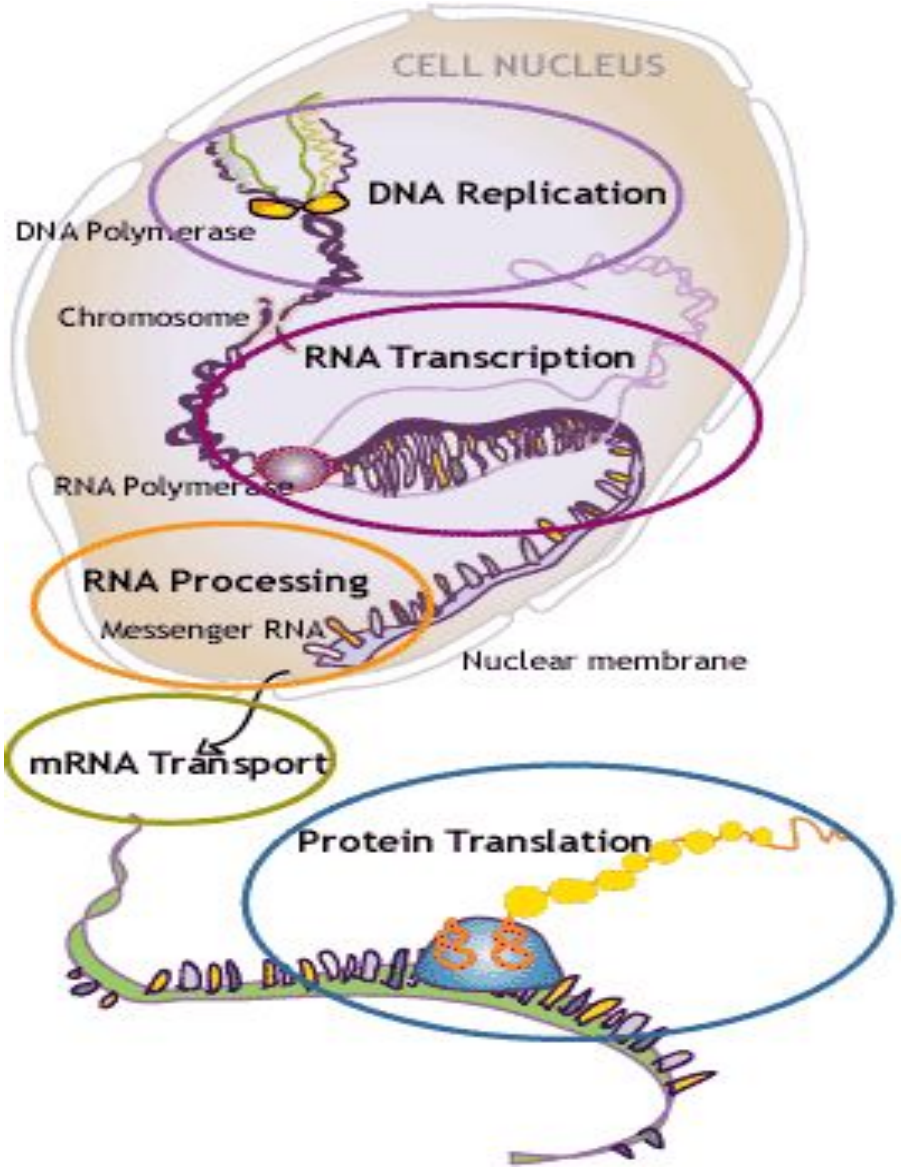
Normal Karyotype

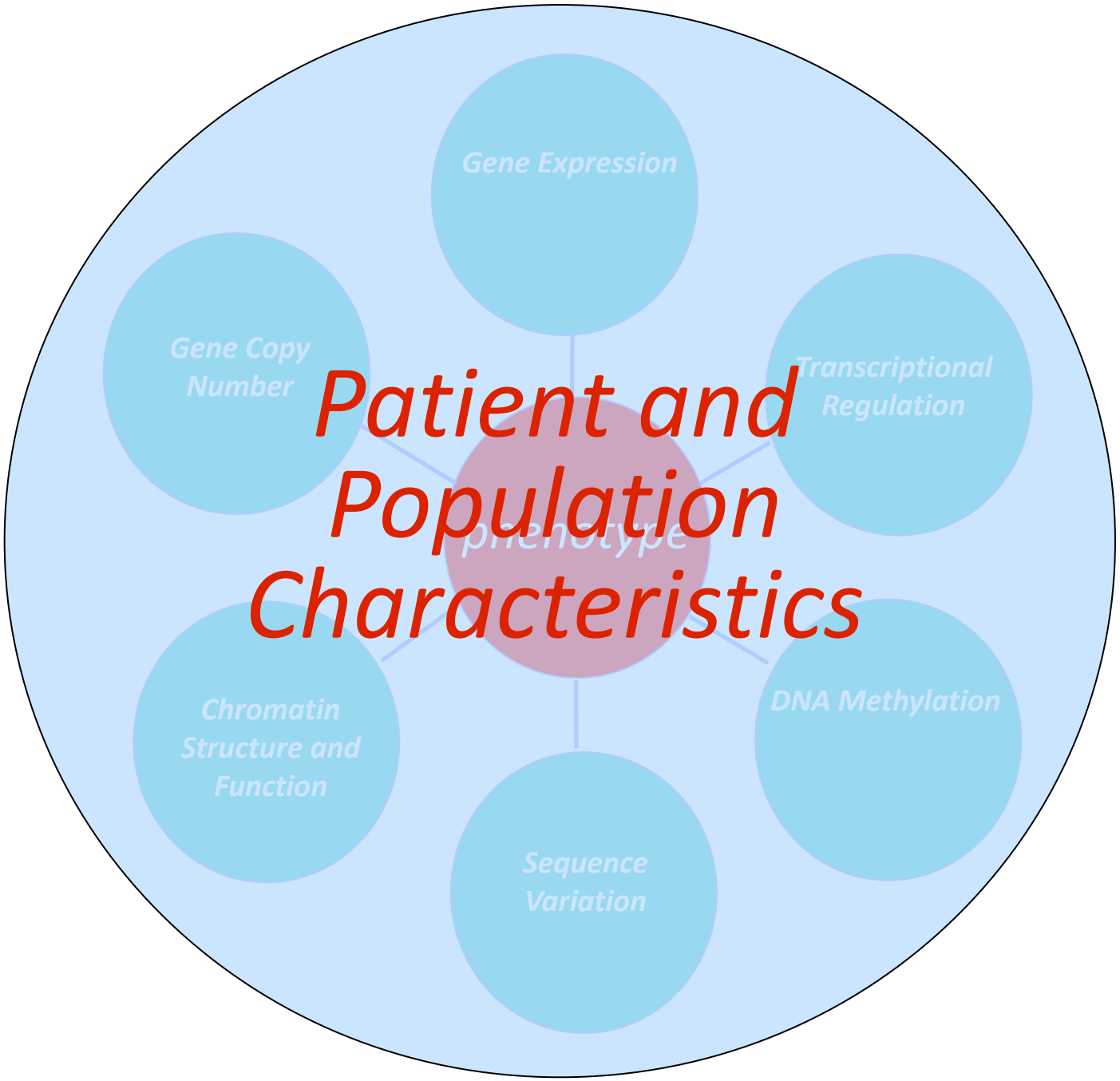


Tumor Karyotype

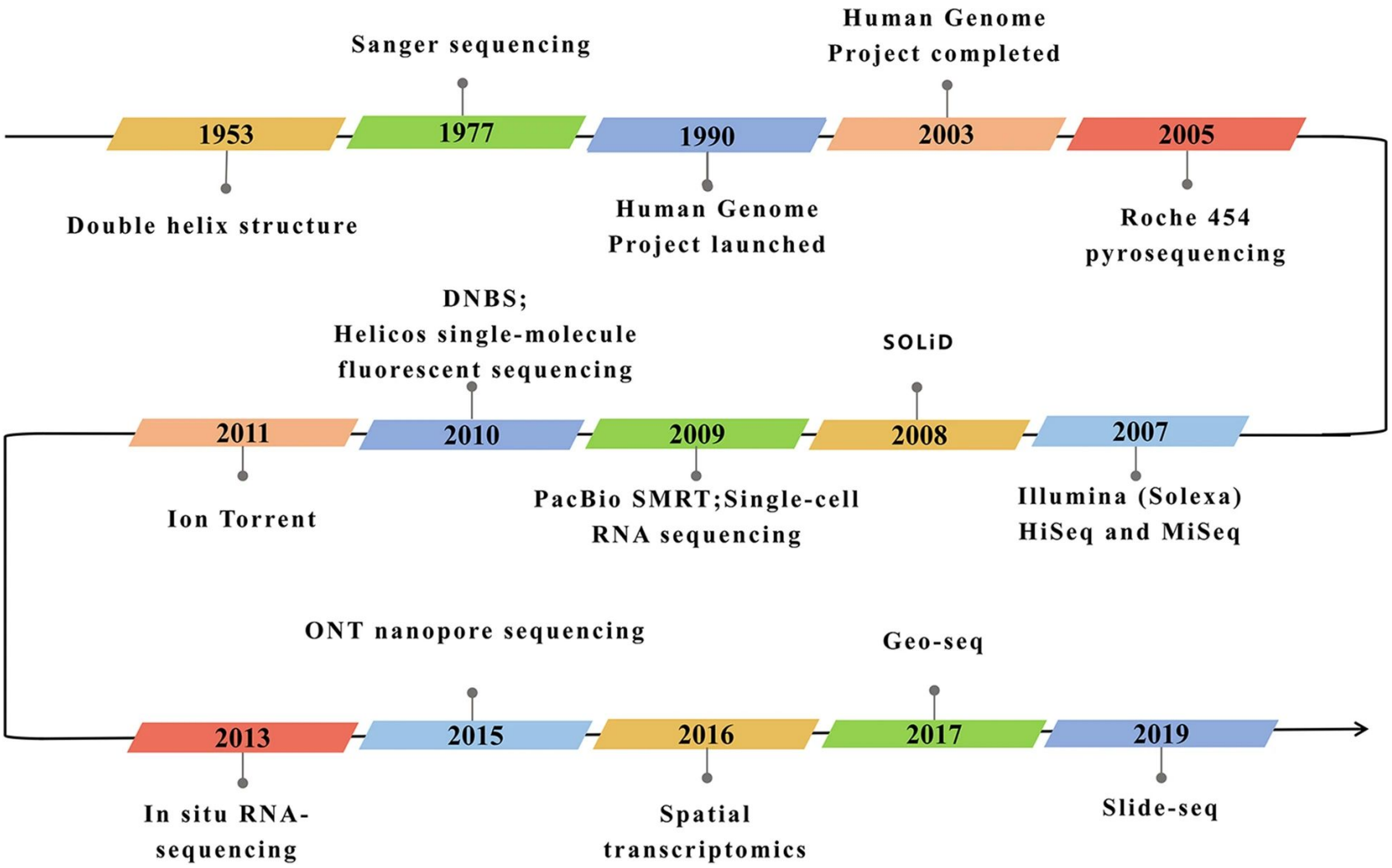


The Central Dogma

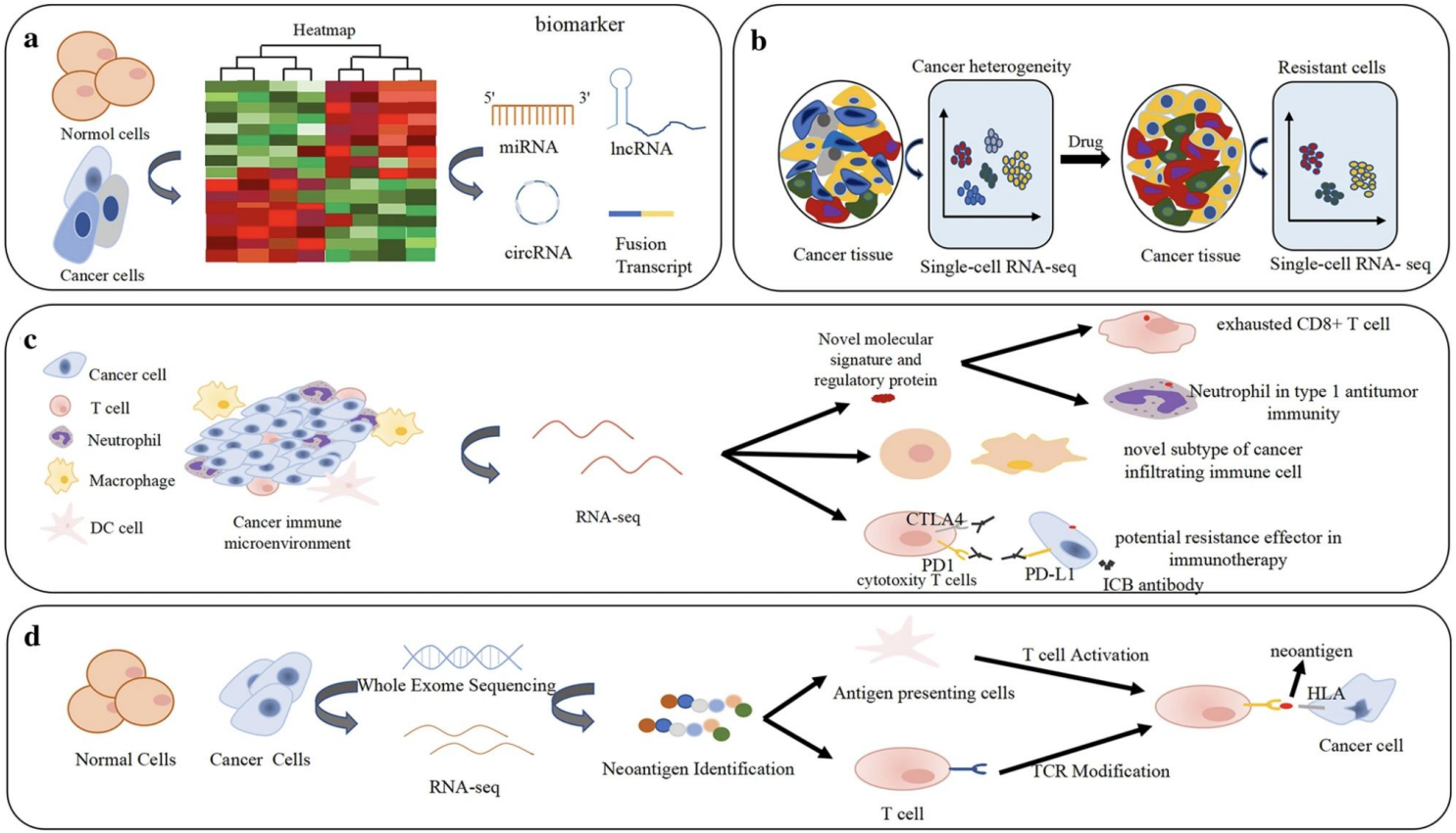



Gene Expression
Gene Copy Number
Transcriptional Regulation
phenotype
Chromatin Structure and Function
DNA Methylation
Sequence Variation
Patient and Population Characteristics

Development of RNA-Seq



Applications of RNA-Seq

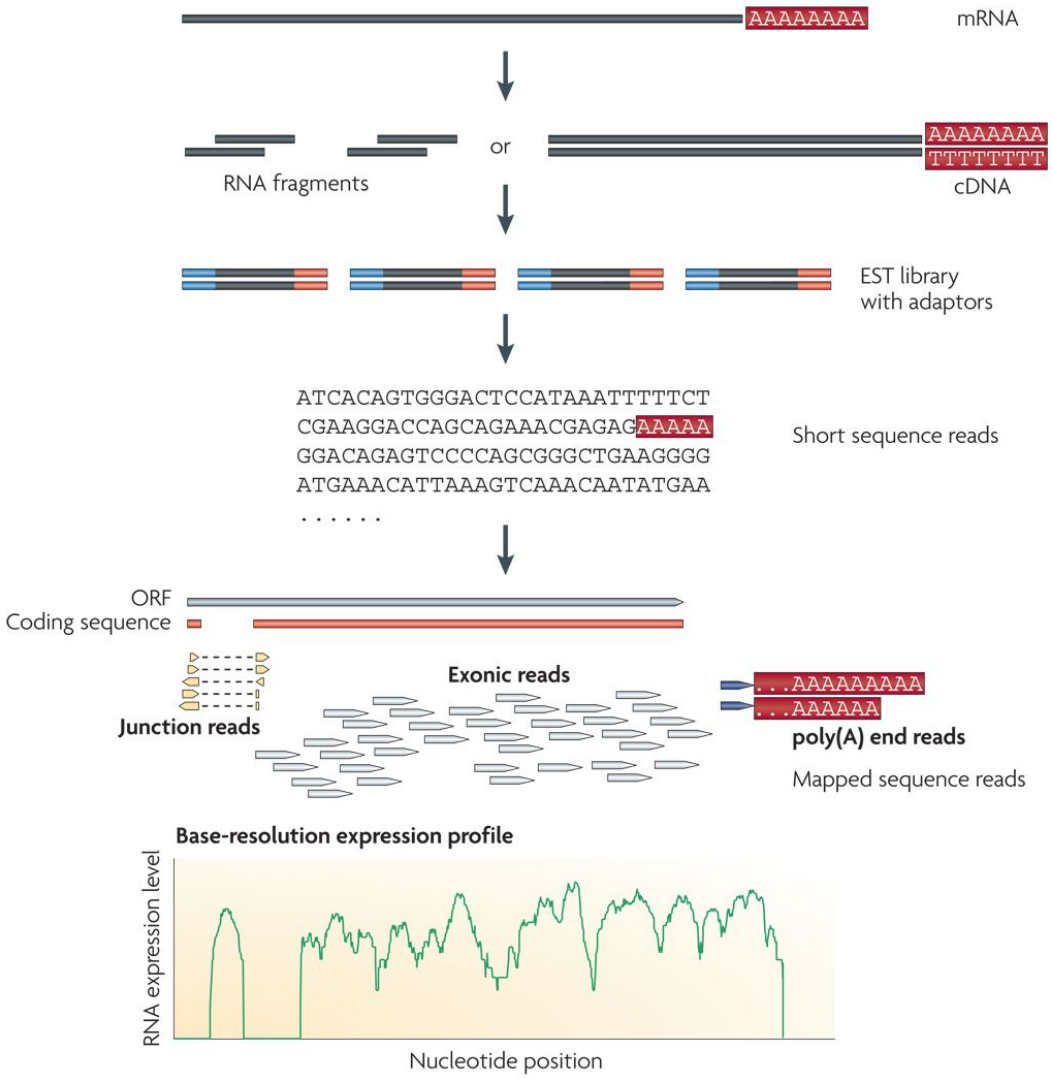


Applications of RNA-seq in differential expression analysis and cancer biomarkers, cancer heterogeneity and drug resistance, cancer immune microenvironment, immunotherapy and neoantigen. **a** Differential expression analysis by RNA sequencing can identify potential biomarkers, including fusion transcript, lncRNA, miRNA and circRNA. **b** The heterogeneity and drug resistance of cancer cells identified by RNA-seq. **c** Novel molecular signature, regulatory protein and unknown subtypes in cancer infiltrating immune cells and potential resistance effector in immunotherapy can be identified by RNA-seq; **d** Neoantigen profiling by RNA-seq and TCR modification targeted neoantigens

https://jhoonline.biomedcentral.com/articles/10.1186/s13045-020-01005-x

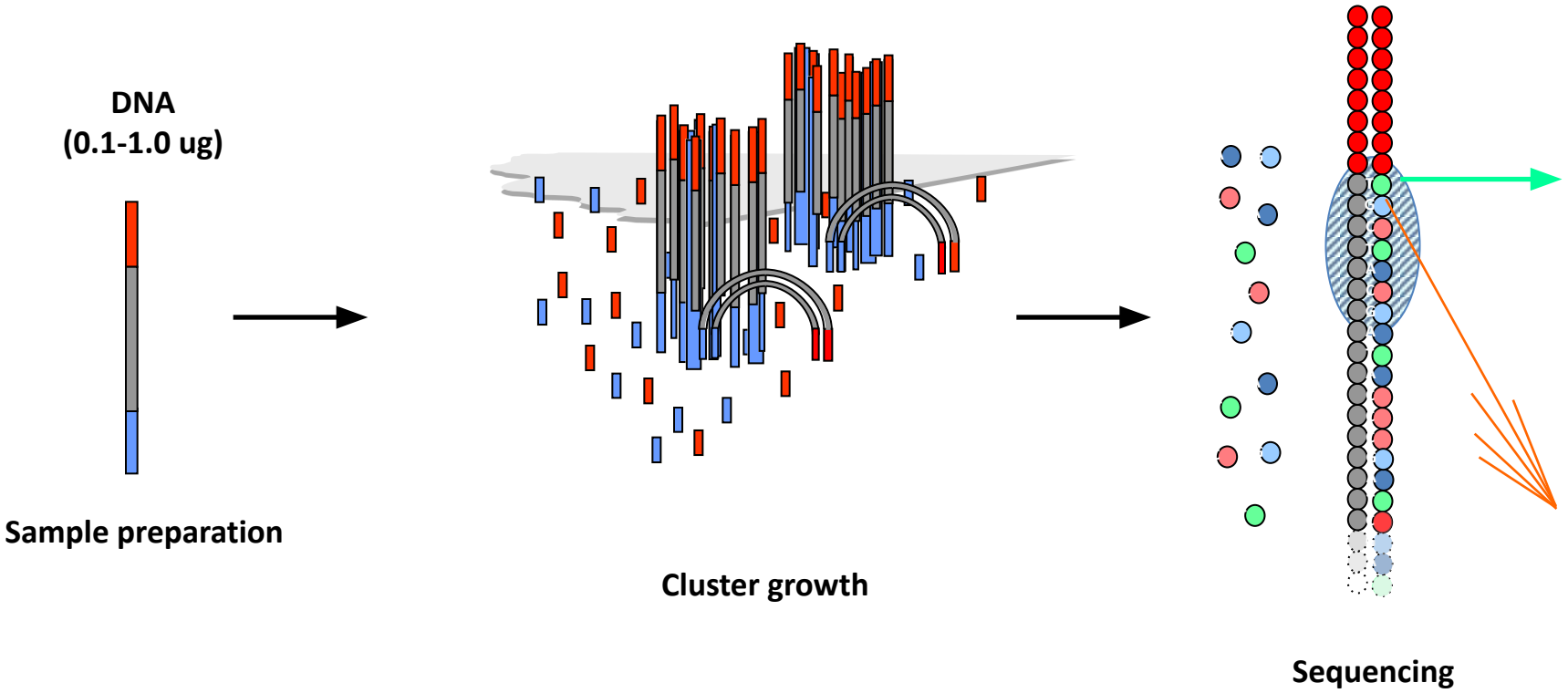
High Throughput Sequencing
AKA, HTS

RNA-seq protocol schematic



Illumina SBS Technology

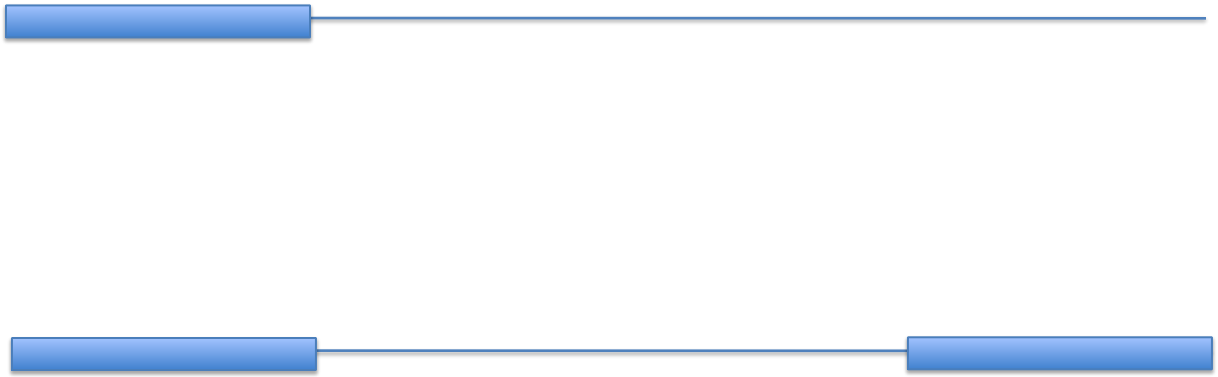
Reversible Terminator Chemistry Foundation

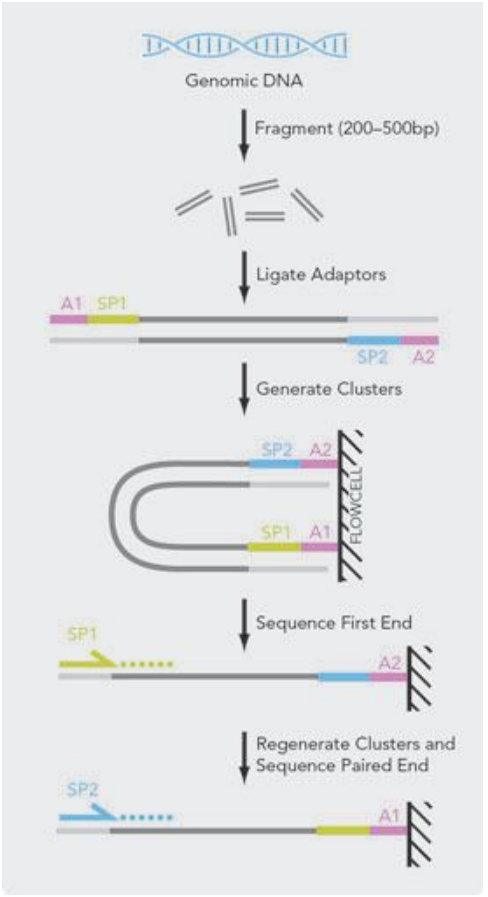


http://seqanswers.com/forums/showthread.php?t=21

http://www.illumina.com/technology/sequencing_technology.ilmn

Single end *vs* paired end sequencing





Illumina Paired-end sequencing

Paired-end: useful for RRBS, essential for RNA-seq, not useful for ChIP-seq

What comes out of the machine:
short reads in fastq format

```
@D3B4KKQ1_0166:8:1101:1960:2190#CGATGT/1
CTCCTGGAAAACGCTTTGGTAGATTTGGCCAGGAGCTTTCTTTTATGTAAATTG
+D3B4KKQ1_0166:8:1101:1960:2190#CGATGT/1
[^^cedeefee`cghhhfcRX`_gfghf^bZbecg^eeb[caef`ef^a_`eXa
@D3B4KKQ1_0166:8:1101:2154:2137#CGATGT/1
TCCANCCATGGCAAATTCCATGGCACCGTCAAGGCTGAGAACGGGAAGCTTGTC
+D3B4KKQ1_0166:8:1101:2154:2137#CGATGT/1
ab_eBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D3B4KKQ1_0166:8:1101:2249:2171#CGATGT/1
TACAAGTGCAGCATCAAGGAGCGAATGCTCTACTCCAGCTGCAAGAGCCGCCTC
+D3B4KKQ1_0166:8:1101:2249:2171#CGATGT/1
_[_ceeec[^eeghdffffhh^efh_egfhfgeec_fbafhhhhd`caegfheh
@D3B4KKQ1_0166:8:1101:2043:2187#CGATGT/1
GAAGGAGAGAAGGGGAGGAGGGCGGGGGGCACCTACTACATCGCCCTCCACATC
+D3B4KKQ1_0166:8:1101:2043:2187#CGATGT/1
\^_accceg`gga`f[fgcb`Ucgfaa_LVV^[bbbbbRWW`W^Y[_[^bbbbb
@D3B4KKQ1_0166:8:1101:2188:2232#CGATGT/1
GTGGCCGATTCCTGAGCTGTGTTTGAGGAGAGGGCGGAGTGCCATCTGGGTAGC
+D3B4KKQ1_0166:8:1101:2188:2232#CGATGT/1
```

QS to int In R:
as.integer(ch
arToRaw('e'))
-33

Pair end sequencing

s_8_1_sequence.txt.gz

```
@D3B4KKQ1_0166:8:1101:1960:2190#CGATGT/1
CTCCTGGAAAACGCTTTGGTAGATTTGGCCAGGAGCTTTCTTTTATGTAAATTG
+D3B4KKQ1_0166:8:1101:1960:2190#CGATGT/1
[^^cedeefee`cghhhfcRX`_gfghf^bZbecg^eeb[caef`ef^a_`eXa
@D3B4KKQ1_0166:8:1101:2154:2137#CGATGT/1
TCCANCCATGGCAAATTCCATGGCACCGTCAAGGCTGAGAACGGGAAGCTTGTC
+D3B4KKQ1_0166:8:1101:2154:2137#CGATGT/1
ab_eBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D3B4KKQ1_0166:8:1101:2249:2171#CGATGT/1
TACAAGTGCAGCATCAAGGAGCGAATGCTCTACTCCAGCTGCAAGAGCCGCCTC
+D3B4KKQ1_0166:8:1101:2249:2171#CGATGT/1
_[_ceeec[^eeghdffffhh^efh_egfhfgeec_fbafhhhhd`caegfheh
@D3B4KKQ1_0166:8:1101:2043:2187#CGATGT/1
GAAGGAGAGAAGGGGAGGAGGGCGGGGGGCACCTACTACATCGCCCTCCACATC
+D3B4KKQ1_0166:8:1101:2043:2187#CGATGT/1
\^_accceg`gga`f[fgcb`Ucgfaa_LVV^[bbbbbRWW`W^Y[_[^bbbbb
@D3B4KKQ1_0166:8:1101:2188:2232#CGATGT/1
GTGGCCGATTCCTGAGCTGTGTTTGAGGAGAGGGCGGAGTGCCATCTGGGTAGC
+D3B4KKQ1_0166:8:1101:2188:2232#CGATGT/1
aa_eeeeeggggihhiiifgeghfeghbgcghifiidg^dbgggeeeee`dcd
```

...

s_8_2_sequence.txt.gz

```
@D3B4KKQ1_0166:8:1101:1960:2190#CGATGT/2
GGCATATTTAACAGCATTGAACAGAATTCTGTGTCCTGTAAAAAAATTAGCTTA
+D3B4KKQ1_0166:8:1101:1960:2190#CGATGT/2
a__aaa`ce`cgcffdf_acda^ea]befffbeged`g[a`e_caaac]cb`gb
@D3B4KKQ1_0166:8:1101:2154:2137#CGATGT/2
TTGAGGCTGTTGTCATACTTCTCATGGTTCACACCCATGACGAACATGGGGGCG
+D3B4KKQ1_0166:8:1101:2154:2137#CGATGT/2
a__eeeeeggegefhhhiiihhhhhiieghhhghhiiffhiififhhiihegic
@D3B4KKQ1_0166:8:1101:2249:2171#CGATGT/2
CGGGGTGCACCTCGTCGTAGAGGAACTCTGCCGTCAGCTCTGCCCCATCGCCAA
+D3B4KKQ1_0166:8:1101:2249:2171#CGATGT/2
^__ee__cge`cghghhfgddgfgi]ehhffffff^ec[beegidffhhfhadba
@D3B4KKQ1_0166:8:1101:2043:2187#CGATGT/2
CTTAGTCTCAGTTTTCCTCCAGCAGCCTGAGGAAACTCAAAGGCACAGTTCCCA
+D3B4KKQ1_0166:8:1101:2043:2187#CGATGT/2
_abeaaacg^g^eghhhhgafghhdfghfedeghfiiicfbgdHYagfeecggf
@D3B4KKQ1_0166:8:1101:2188:2232#CGATGT/2
TAGGCTCAAAGTCTAACGCCAATCCCGAACCTGGGCATCTGTACACACACACAC
+D3B4KKQ1_0166:8:1101:2188:2232#CGATGT/2
abbeceeegggcghiihiihhhhiifhiiiiihiiiiiiihegh`eggfebfhg
```

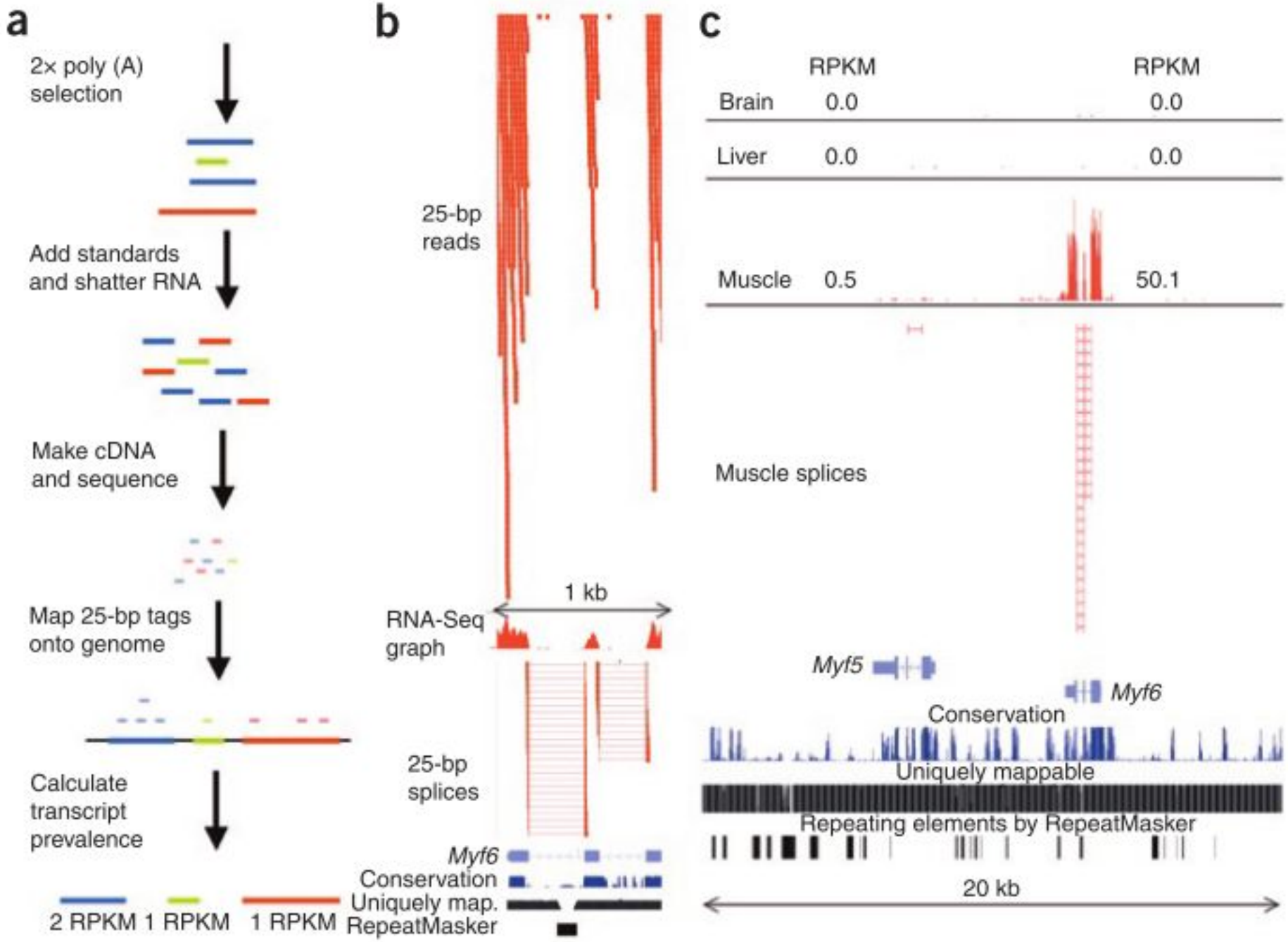
...

Mapping and quantifying mammalian transcriptomes by RNA-Seq

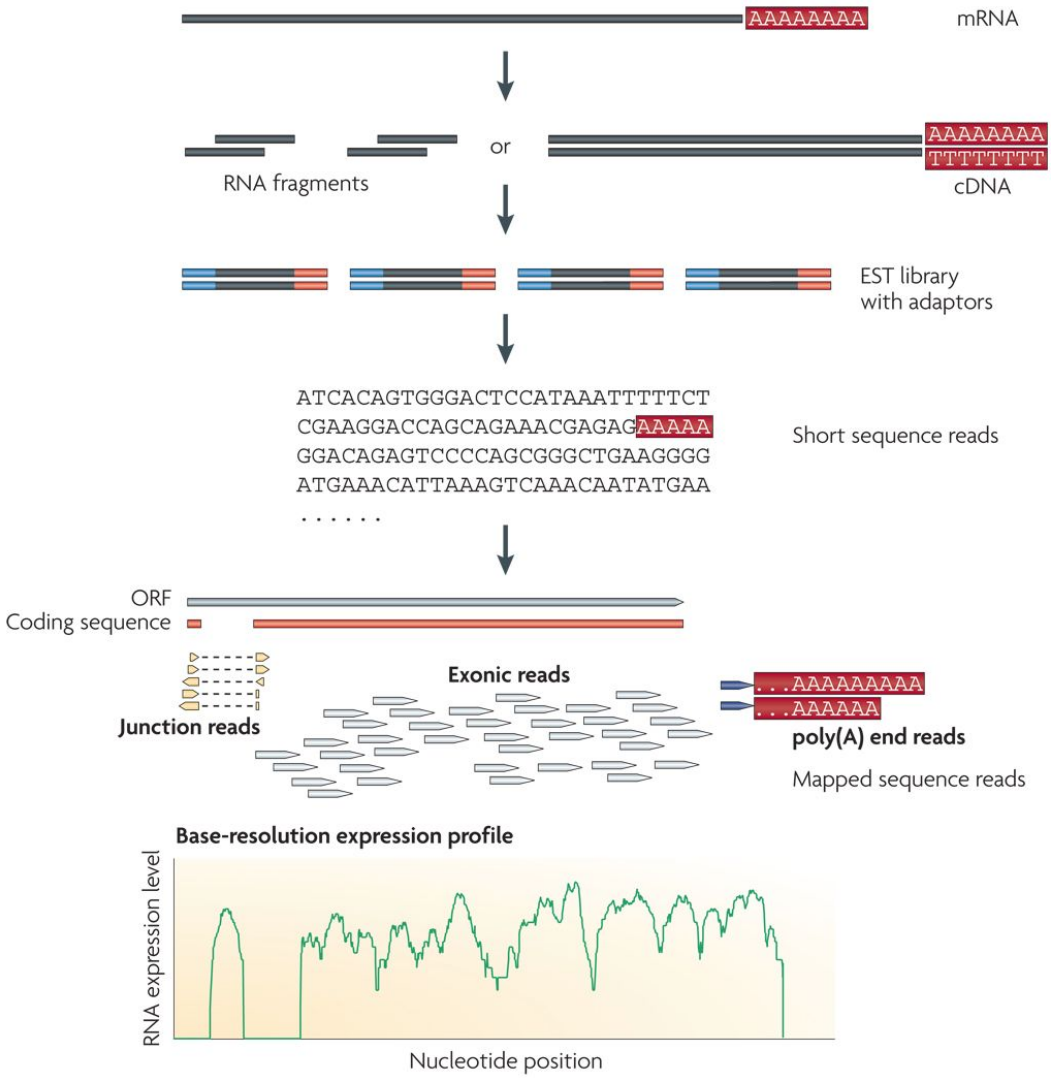
Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

We have mapped and quantified mouse transcriptomes by deeply sequencing them and recording how frequently each gene is represented in the sequence sample (RNA-Seq). This provides a digital measure of the presence and prevalence of transcripts from known and previously unknown genes. We

approaches to large-scale RNA analysis are serial analysis of gene expression (SAGE)[4,5] and related methods such as massively parallel signature sequencing (MPSS)[6], which use DNA sequencing of previously cloned tags 17–25 base pairs (bp) from terminal 3′ (or 5′) sequence tags. These sequence tags are then identified by


a
2× poly (A) selection
Add standards and shatter RNA
Make cDNA and sequence
Map 25-bp tags onto genome
Calculate transcript prevalence
2 RPKM
1 RPKM
1 RPKM
b
25-bp reads
1 kb
RNA-Seq graph
25-bp splices
Myf6
Conservation
Uniquely map.
RepeatMasker
c
RPKM
RPKM
Brain 0.0 0.0
Liver 0.0 0.0
Muscle 0.5 50.1
Muscle splices
Myf5
Myf6
Conservation
Uniquely mappable
Repeating elements by RepeatMasker
20 kb

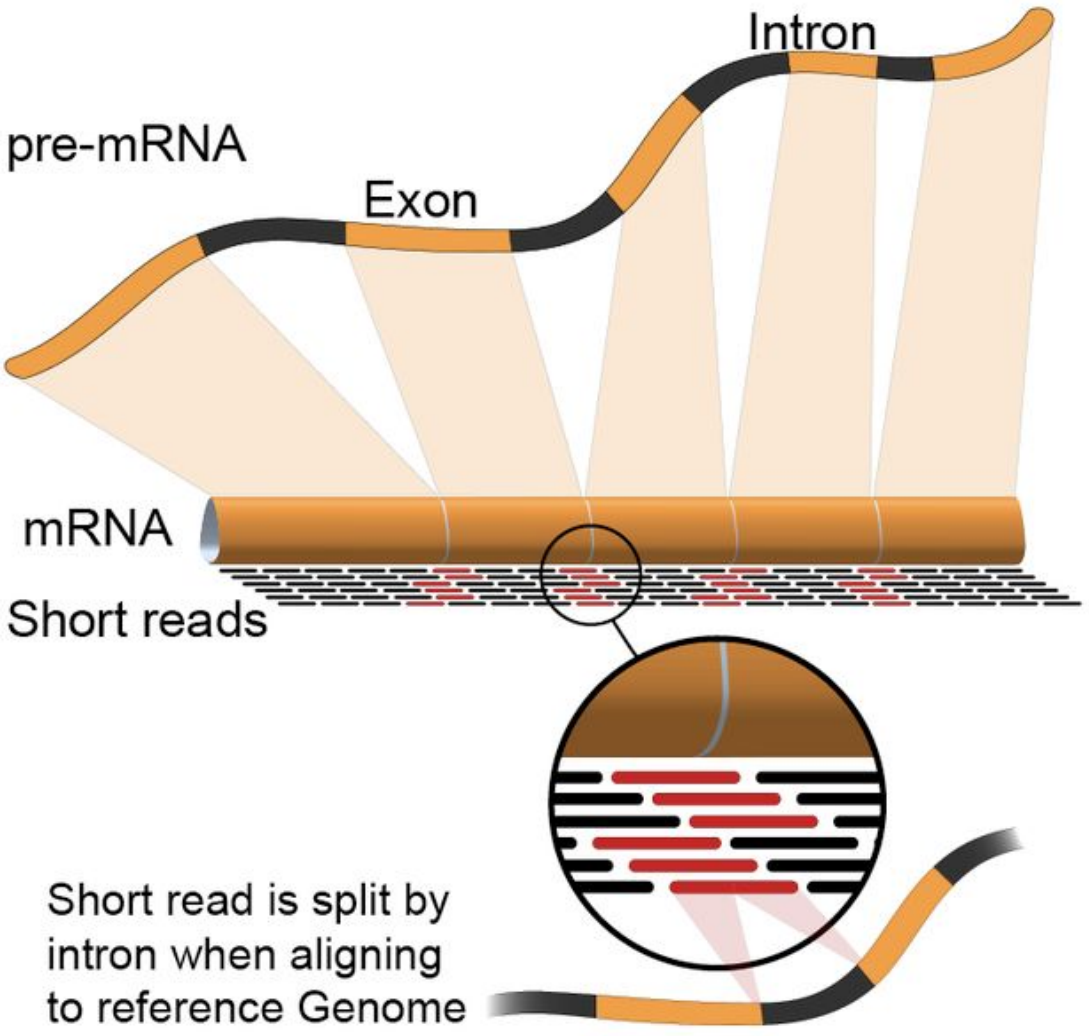
RNA-seq protocol schematic



Alignment

Overview of RNA-seq

RNA-seq Alignment



BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 18 2011, pages 2518–2528
doi:10.1093/bioinformatics/btr427

Sequence analysis

Advance Access publication July 19, 2011

Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)

Gregory R. Grant[1,2,3,*], Michael H. Farkas[4], Angel D. Pizarro[2], Nicholas F. Lahens[5], Jonathan Schug[3], Brian P. Brunk[1], Christian J. Stoeckert[1,3], John B. Hogenesch[1,2,5] and Eric A. Pierce[4,*]

[1]Penn Center for Bioinformatics, [2]Institute for Translational Medicine and Therapeutics, [3]Department of Genetics, [4]F.M. Kirby Center for Molecular Ophthalmology and [5]Department of Pharmacology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

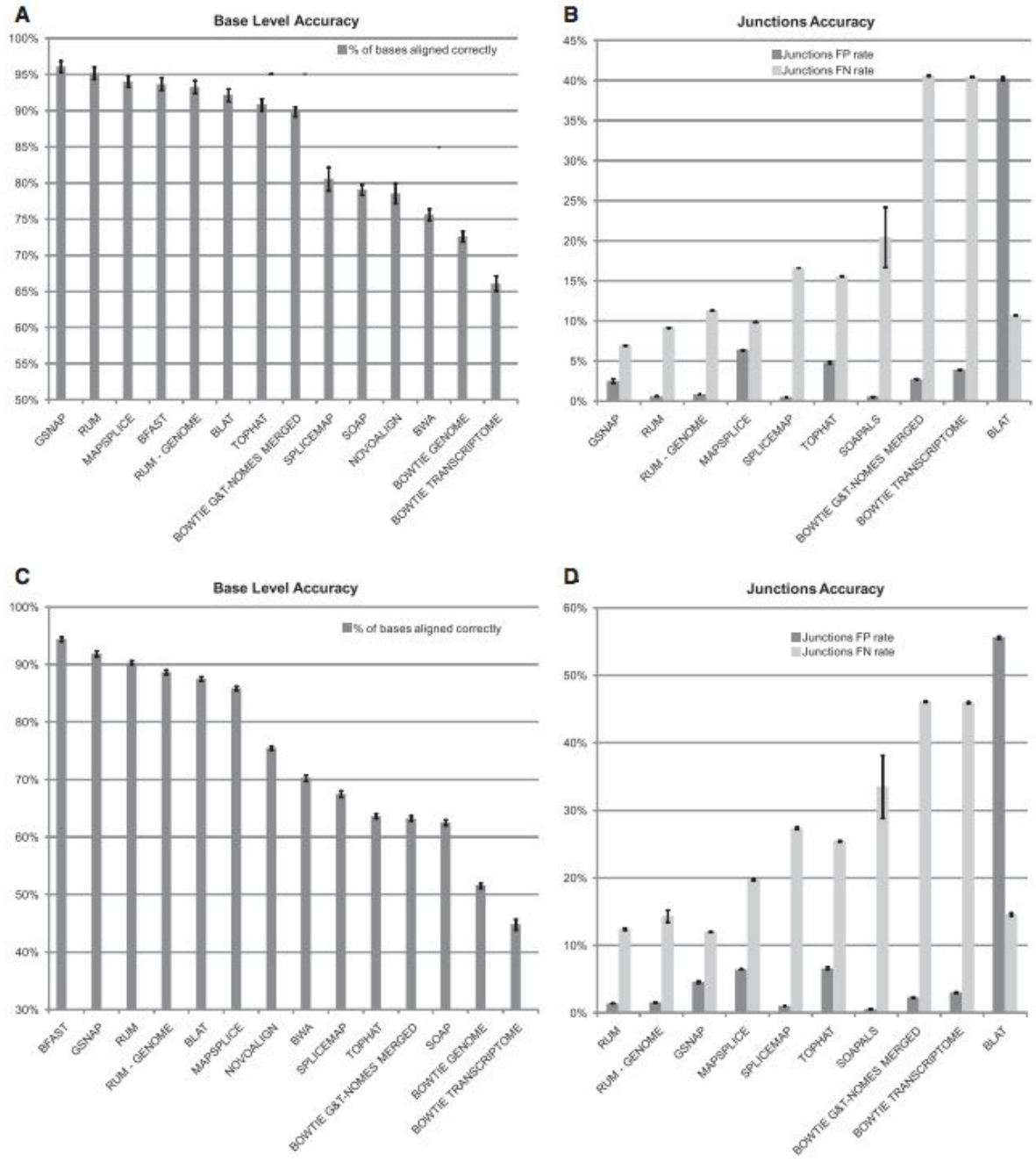
Associate editor: Ivo Hofacker

ANALYSIS

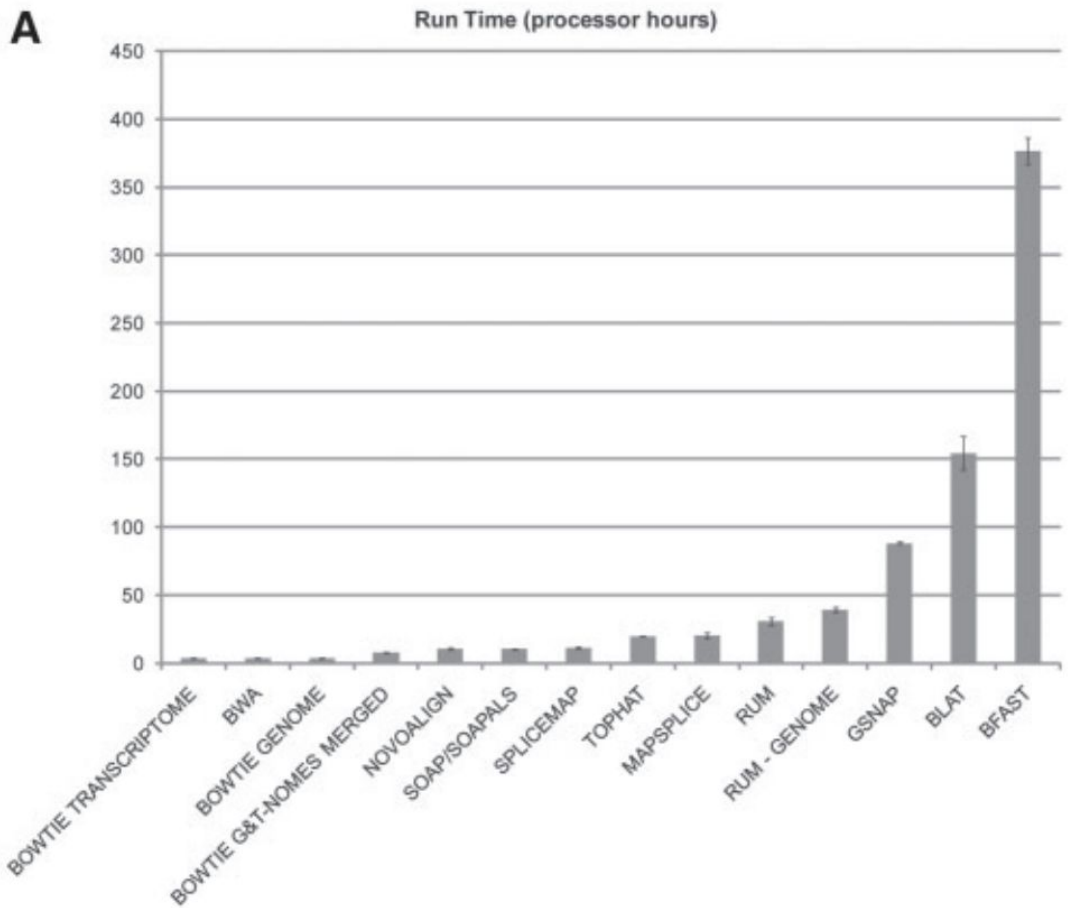
OPEN

Systematic evaluation of spliced alignment programs for RNA-seq data

Pär G Engström[1,13], Tamara Steijger[1], Botond Sipos[1], Gregory R Grant[2,3], André Kahles[4,5], The RGASP Consortium[6], Gunnar Rätsch[4,5], Nick Goldman[1], Tim J Hubbard[7], Jennifer Harrow[7], Roderic Guigó[8,9] & Paul Bertone[1,10–12]


A
Base Level Accuracy
% of bases aligned correctly
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
GSNAP
RUM
MAPSPLICE
BFAST
RUM - GENOME
BLAT
TOPHAT
BOWTIE G&T-NOMES MERGED
SPLICEMAP
SOAP
NOVOALIGN
BWA
BOWTIE GENOME
BOWTIE TRANSCRIPTOME
B
Junctions Accuracy
Junctions FP rate
Junctions FN rate
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
GSNAP
RUM
RUM - GENOME
MAPSPLICE
SPLICEMAP
TOPHAT
SOAPALS
BOWTIE G&T-NOMES MERGED
BOWTIE TRANSCRIPTOME
BLAT
C
Base Level Accuracy
% of bases aligned correctly
100%
90%
80%
70%
60%
50%
40%
30%
BFAST
GSNAP
RUM
RUM - GENOME
BLAT
MAPSPLICE
NOVOALIGN
BWA
SPLICEMAP
TOPHAT
BOWTIE G&T-NOMES MERGED
SOAP
BOWTIE GENOME
BOWTIE TRANSCRIPTOME
D
Junctions Accuracy
Junctions FP rate
Junctions FN rate
60%
50%
40%
30%
20%
10%
0%
RUM
RUM - GENOME
GSNAP
MAPSPLICE
SPLICEMAP
TOPHAT
SOAPALS
BOWTIE G&T-NOMES MERGED
BOWTIE TRANSCRIPTOME
BLAT

Run Time


A
Run Time (processor hours)
450
400
350
300
250
200
150
100
50
0
BOWTIE TRANSCRIPTOME
BWA
BOWTIE GENOME
BOWTIE G&T-NOMES MERGED
NOVOALIGN
SOAP/SOAPALS
SPLICEMAP
TOPHAT
MAPSPLICE
RUM
RUM - GENOME
GSNAP
BLAT
BFAST

Alignment Yield

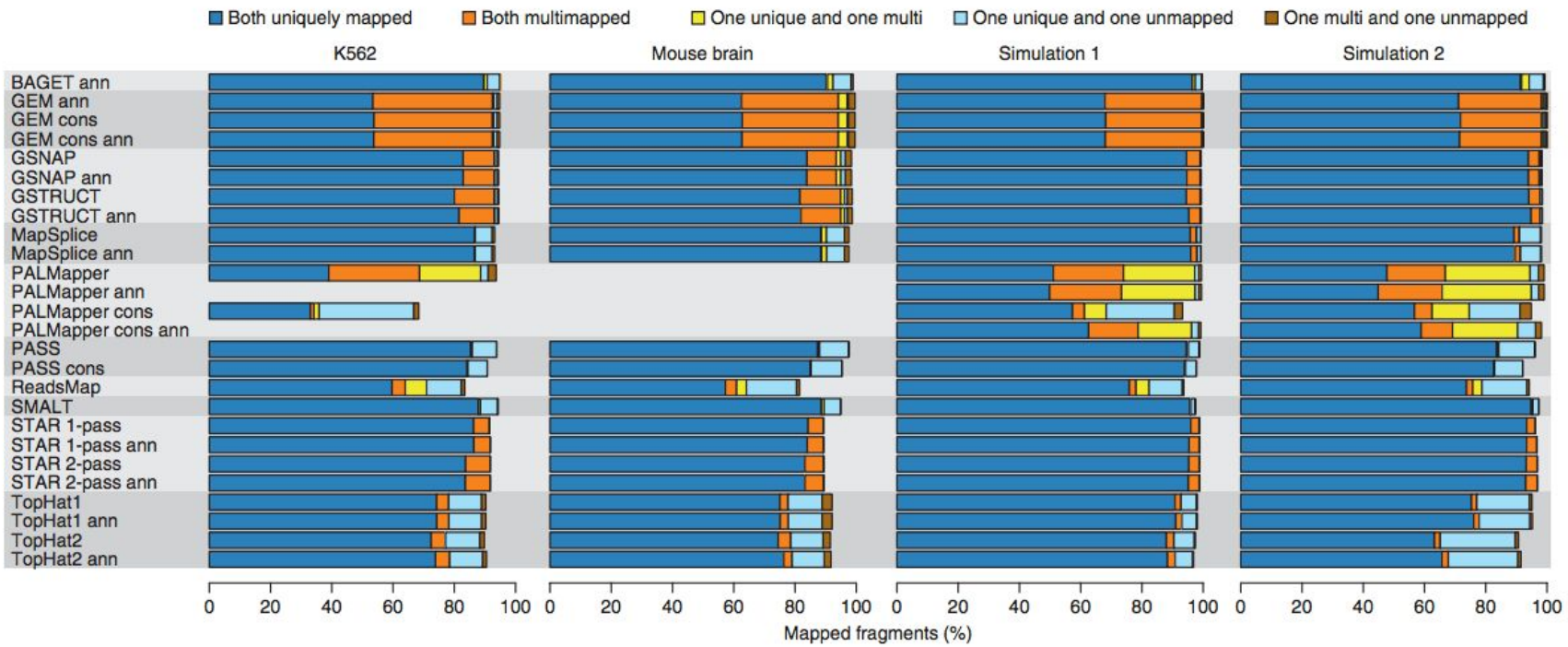


Figure 1 | Alignment yield. Shown is the percentage of sequenced or simulated read pairs (fragments) mapped by each protocol. Protocols are grouped by the underlying alignment program (gray shading). Protocol names contain the suffix "ann" if annotation was used. The suffix "cons" distinguishes more conservative protocols from others based on the same aligner. The K562 data set comprises six samples, and the metrics presented here were averaged over them.

Splice Read Placement Accuracy

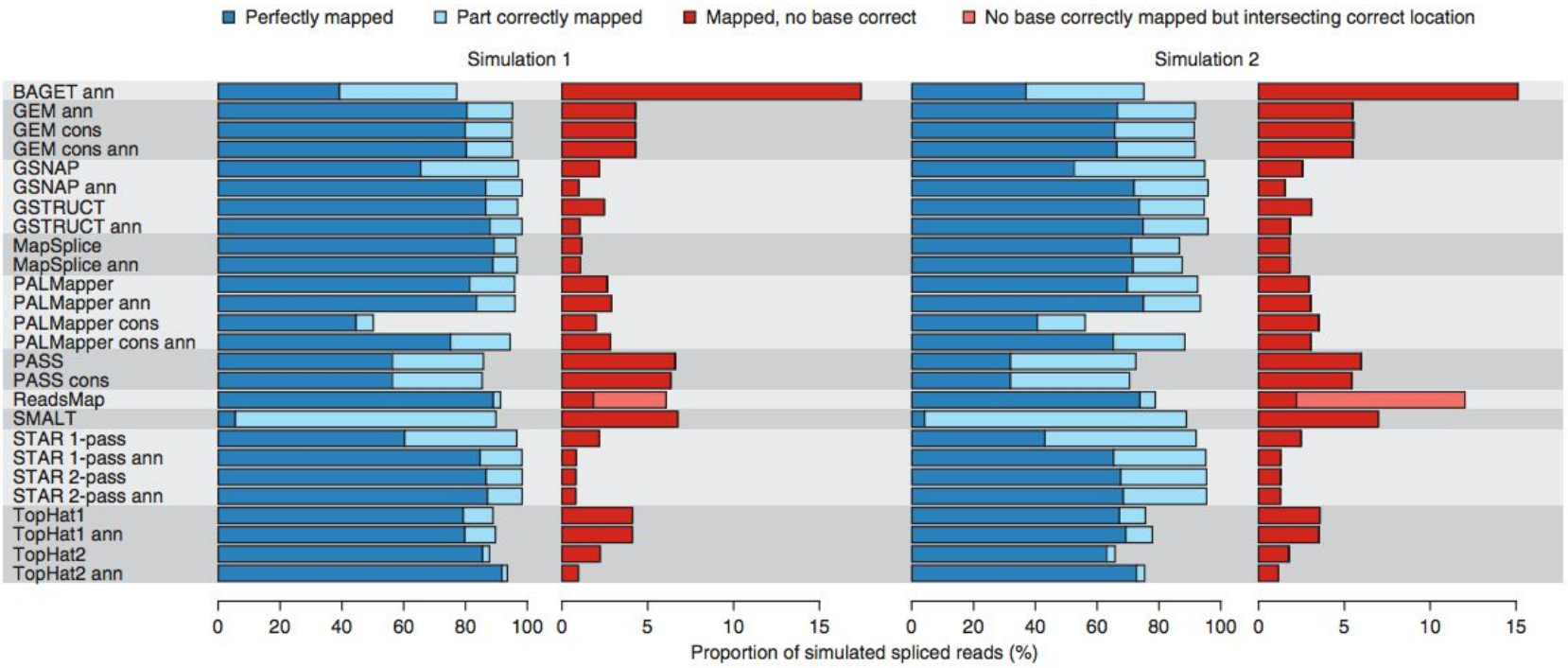
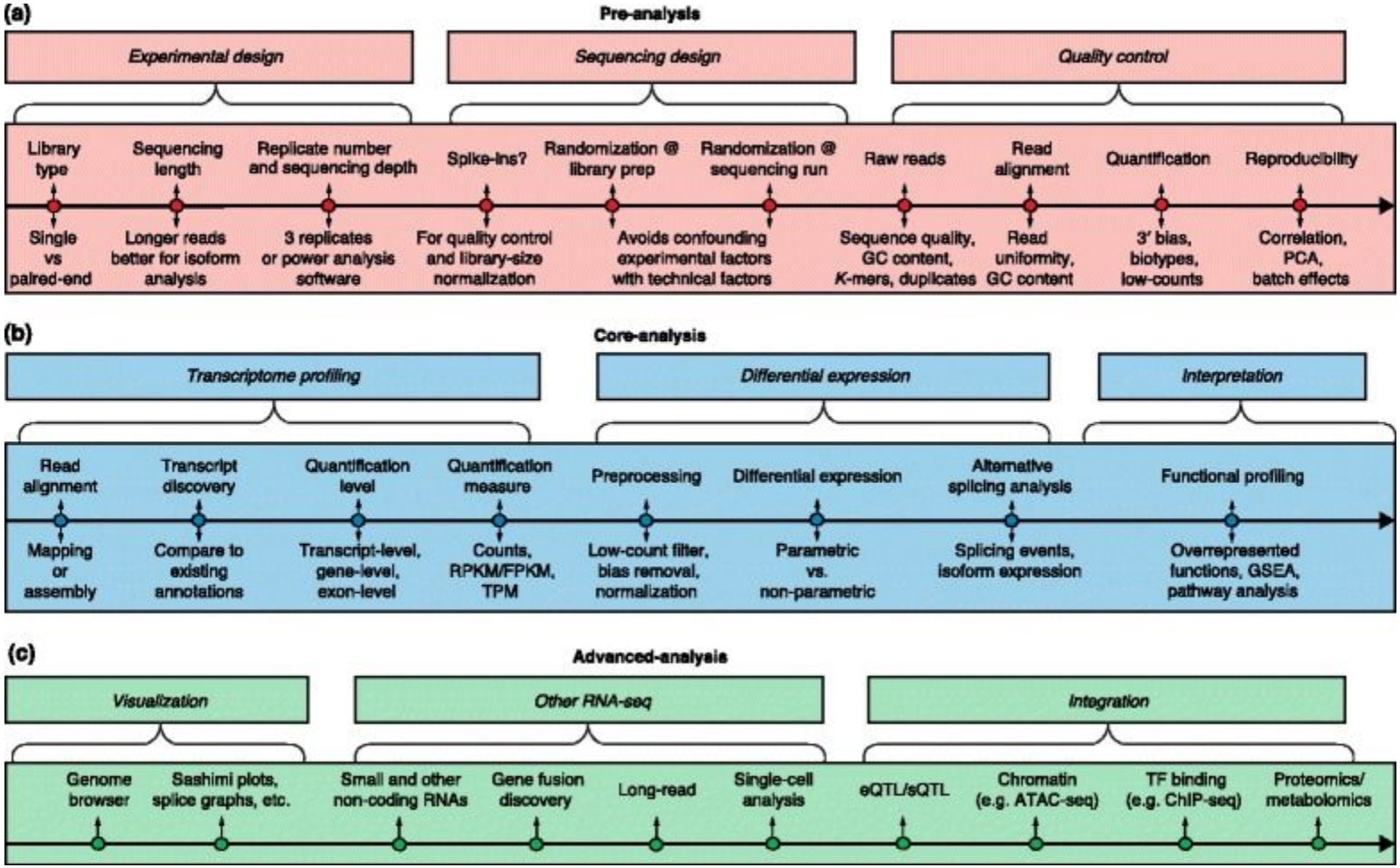


Figure 3 | Read placement accuracy for simulated spliced reads.

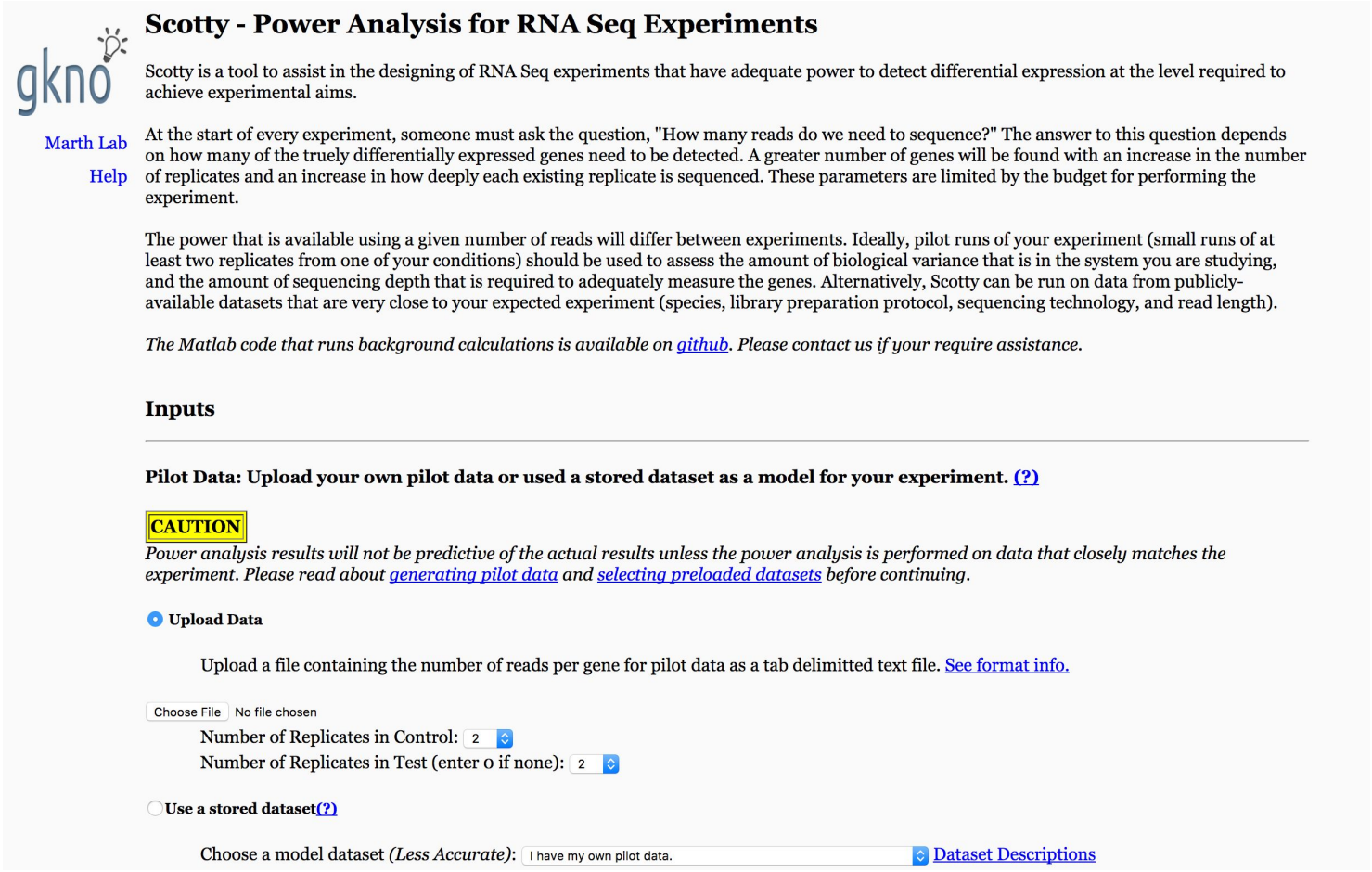
Overview of RNA-seq

Schematic for RNA-seq Design, Analysis, and Interpretation



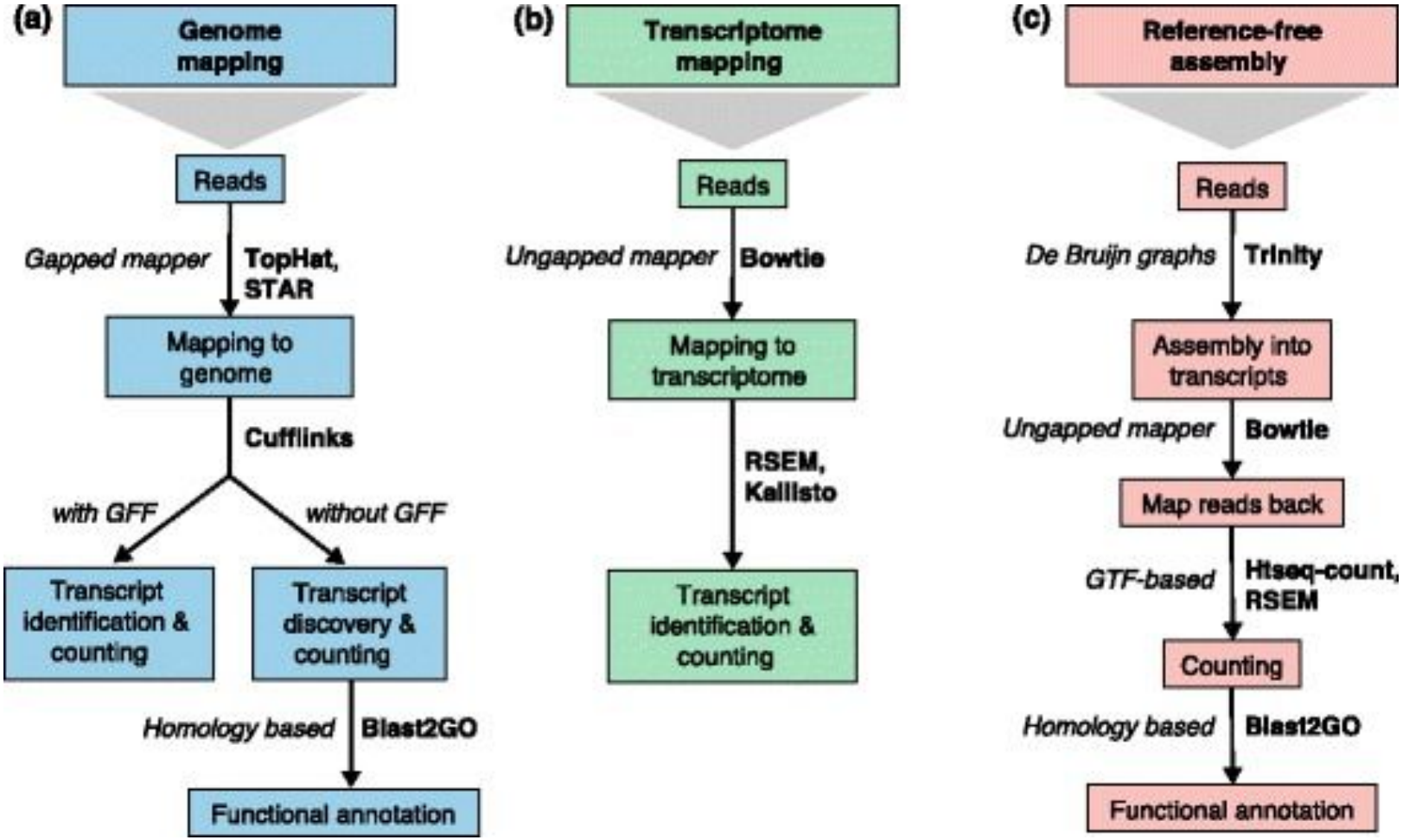
https://dx.doi.org/10.1186%2Fs13059-016-0881-8

Power calculations (Diff Expr)



http://scotty.genetics.utah.edu/

Approaches to RNA-seq Quantification

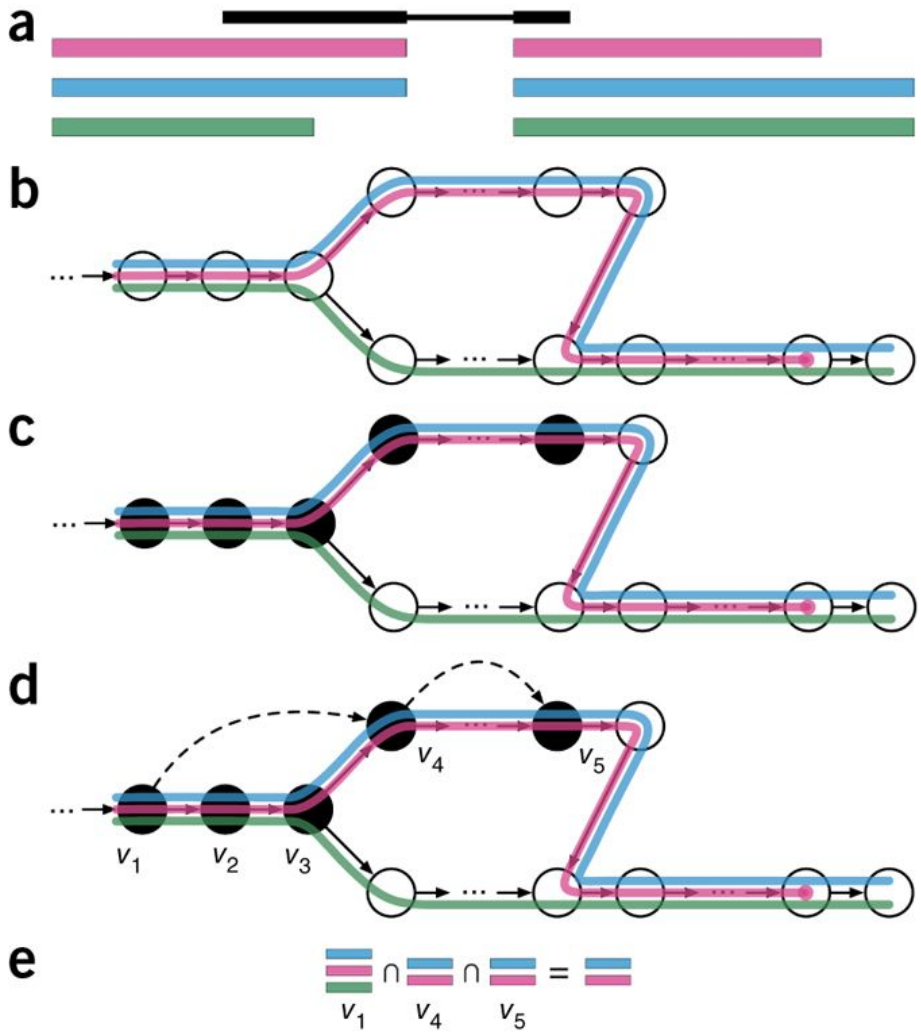


https://dx.doi.org/10.1186%2Fs13059-016-0881-8

Pseudo-mapping

RNA-seq processing

Pseudomapping



The input consists of a reference transcriptome and reads from an RNA-seq experiment.
(a) An example of a read (in black) and three overlapping transcripts with exonic regions as shown.
(b) An index is constructed by creating the transcriptome de Bruijn Graph (T-DBG) where nodes (v_1, v_2, v_3, …) are k-mers, each transcript corresponds to a colored path as shown and the path cover of the transcriptome induces a k-compatibility class for each k-mer.
(c) Conceptually, the k-mers of a read are hashed (black nodes) to find the k-compatibility class of a read.
(d) Skipping (black dashed lines) uses the information stored in the T-DBG to skip k-mers that are redundant because they have the same k-compatibility class.
(e) The k-compatibility class of the read is determined by taking the intersection of the k-compatibility classes of its constituent k-mers.

Transcript Quantification

MODELS FOR TRANSCRIPT QUANTIFICATION FROM RNA-SEQ

LIOR PACHTER

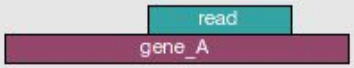
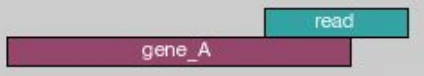
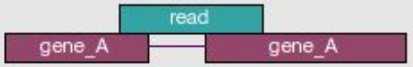
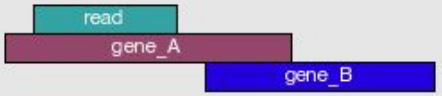
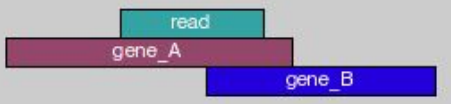
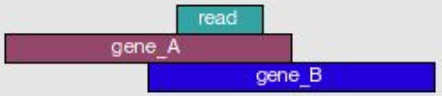
May 16, 2011

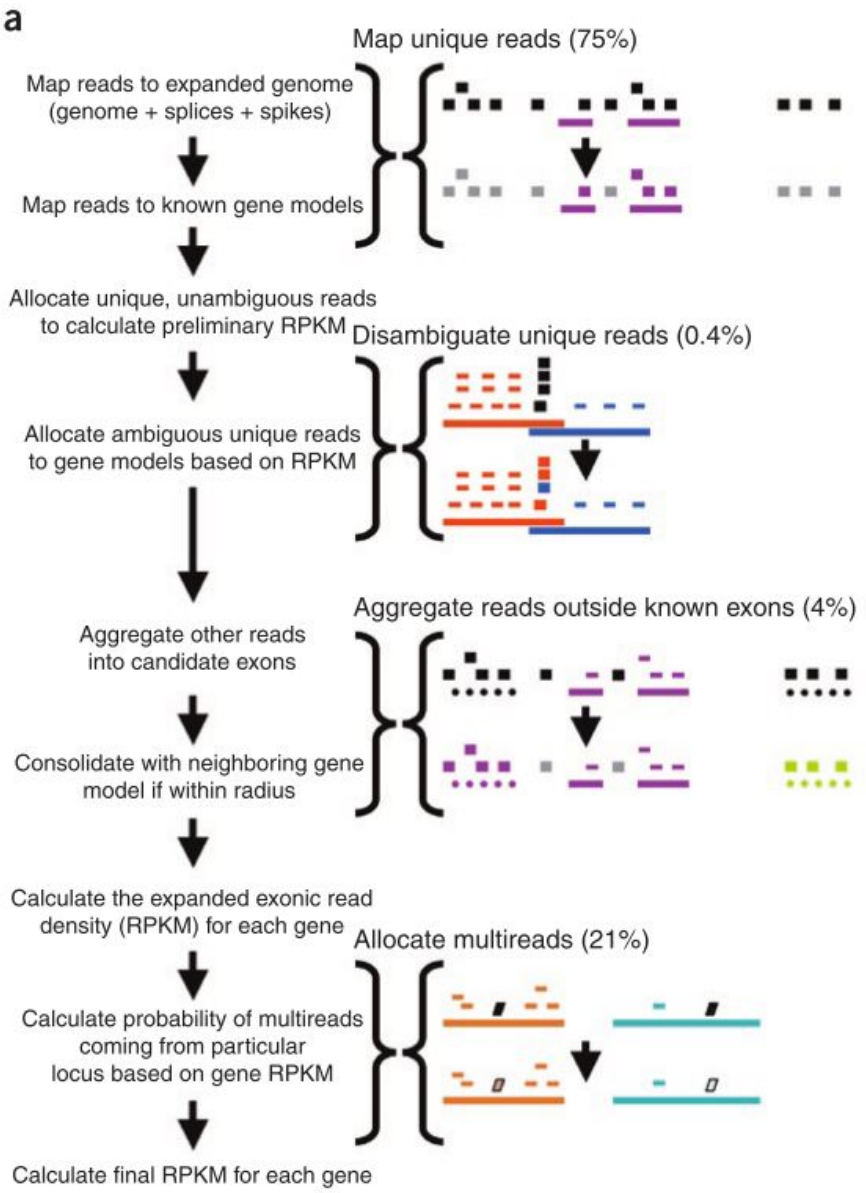
ABSTRACT. RNA-Seq is rapidly becoming the standard technology for transcriptome analysis. Fundamental to many of the applications of RNA-Seq is the quantification problem, which is the accurate measurement of relative transcript abundances from the sequenced reads. We focus on this problem, and review many recently published models that are used to estimate the relative abundances. In addition to describing the models and the different approaches to inference, we also explain how methods are related to each other. A key result is that we show how inference with many of the models results in identical estimates of relative abundances, even though model formulations can be very different. In fact, we are able to show how a single general model captures many of the elements of previously published methods. We also review the applications of RNA-Seq models to differential analysis, and explain why accurate relative transcript abundance estimates are crucial for downstream analyses.

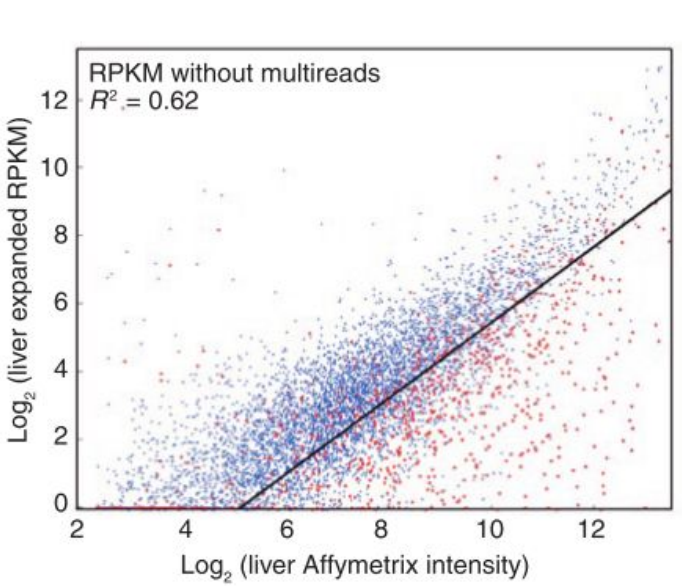
Models for RNA-seq

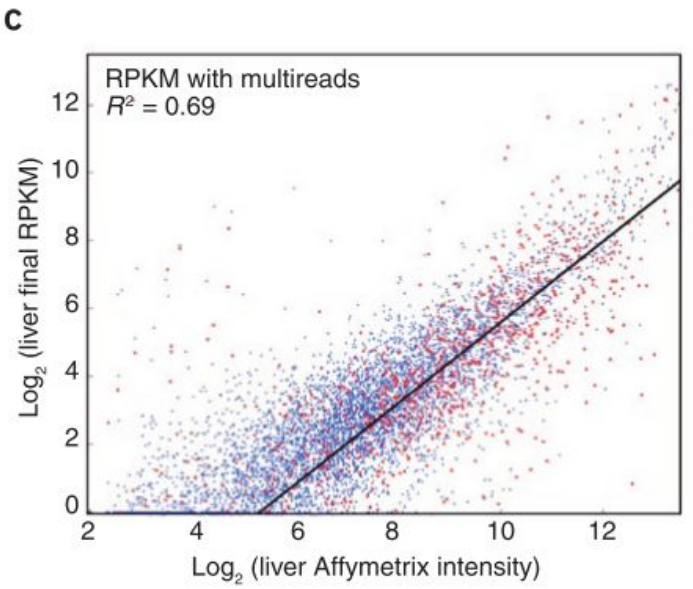
- Count-based models
- Multi-reads (isoform resolution)
- Paired-end reads (include length resolution step)
- Positional bias along transcript length
- Sequence bias
- Fragment-length bias

Read Counting

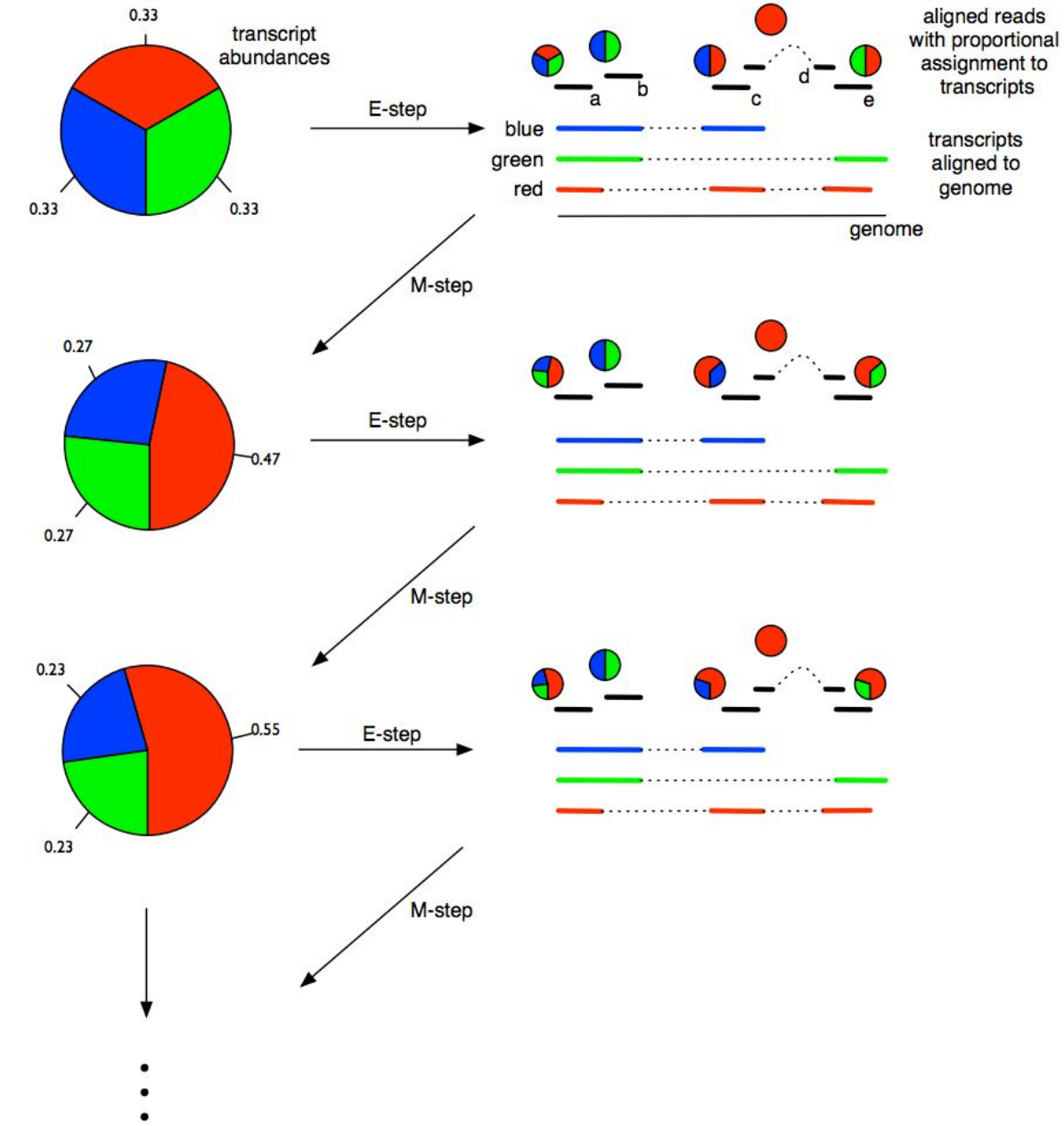
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous







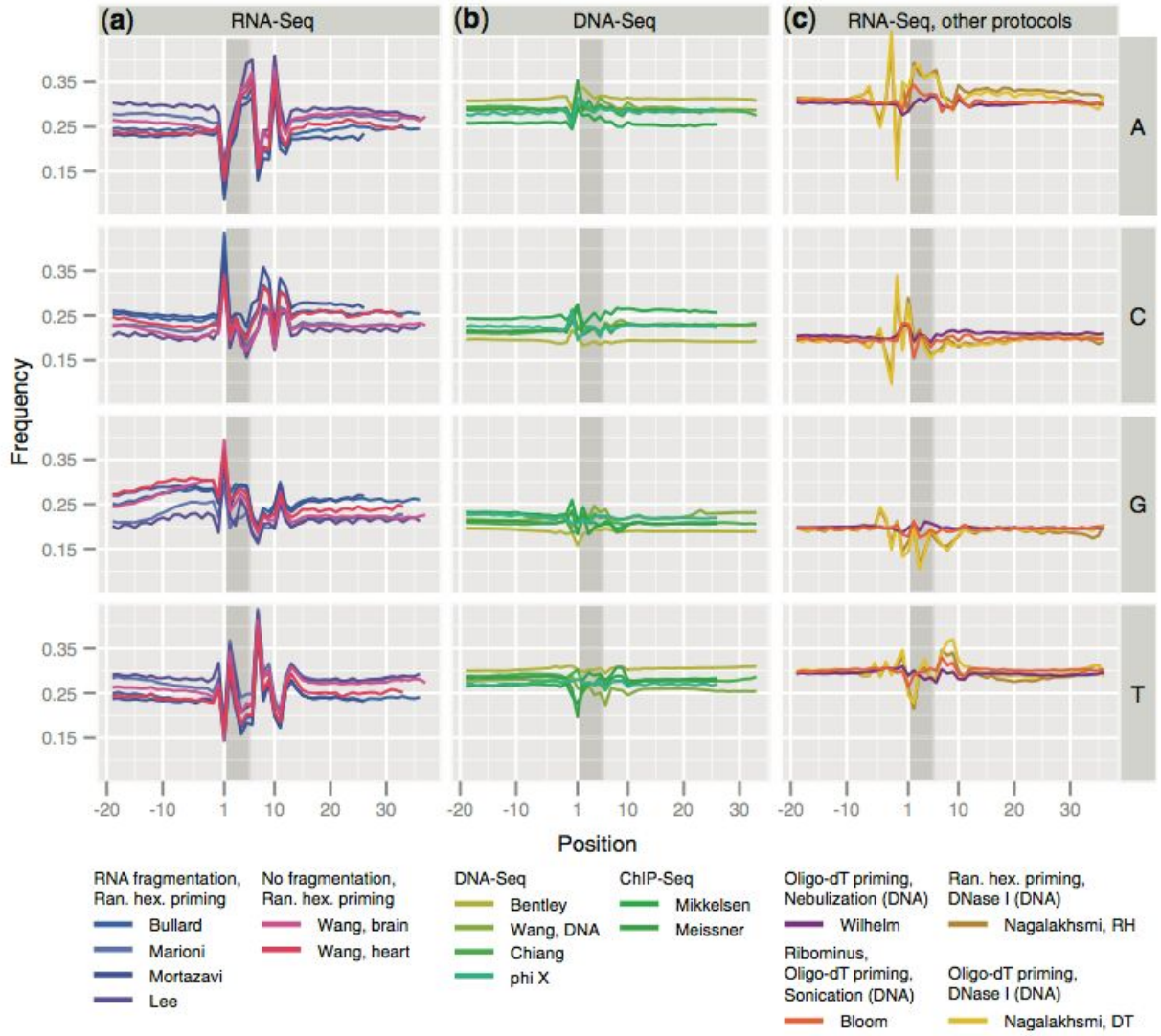
Mortazavi, 2008, NMeth



L. Pachter (2011) arXiv:1104.3889v

RNA-seq bias examples

Sequence Bias--priming



Hansen (2010), NAR

Sample-specific Sequence Bias

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Johns Hopkins University, Dept. of Biostatistics Working Papers

5-24-2011

REMOVING TECHNICAL VARIABILITY IN RNA-SEQ DATA USING CONDITIONAL QUANTILE NORMALIZATION

Kasper D. Hansen
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Rafael A. Irizarry
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Zhijin Wu
Department of Community Health, Section of Biostatistics, Brown University, Zhijin_Wu@brown.edu

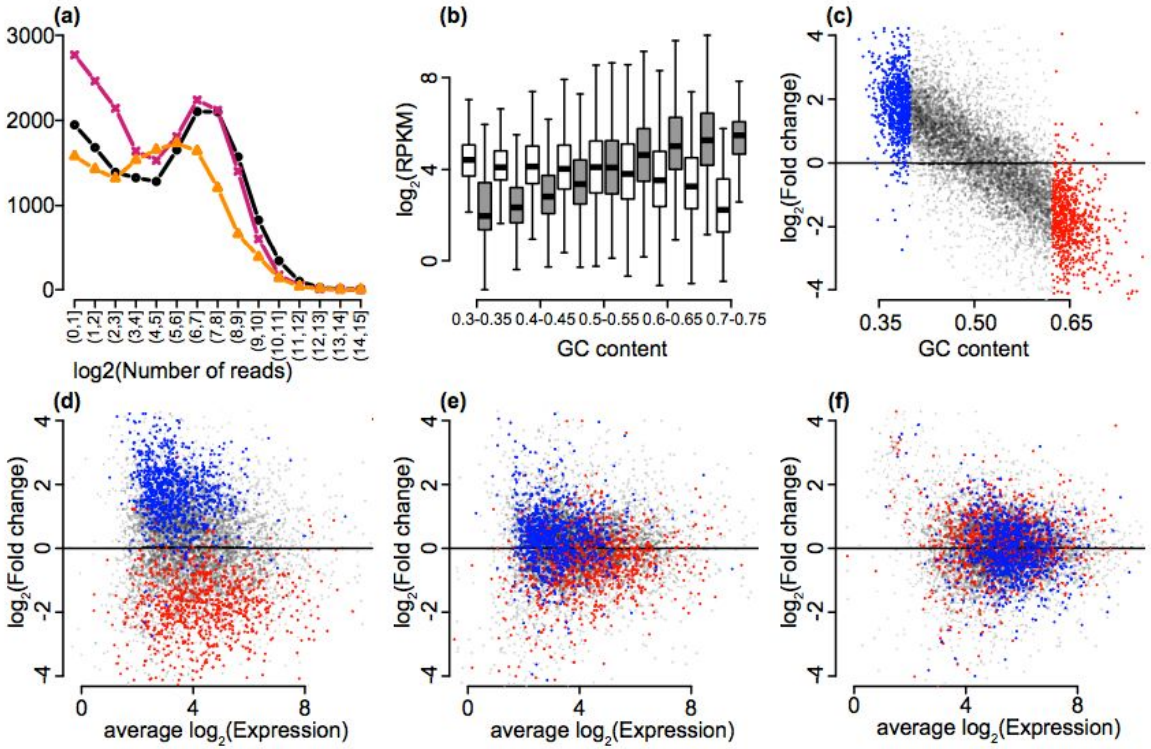
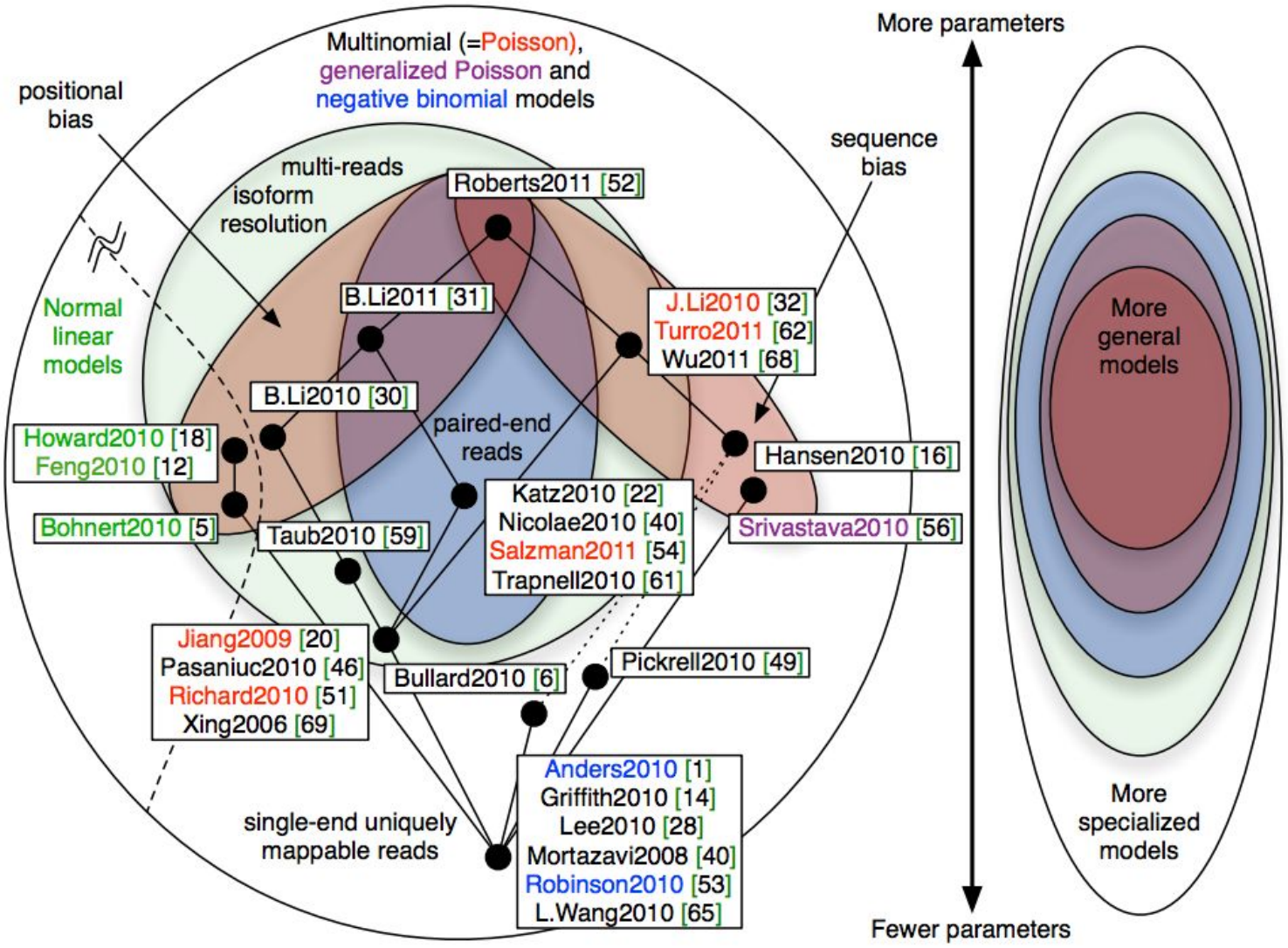
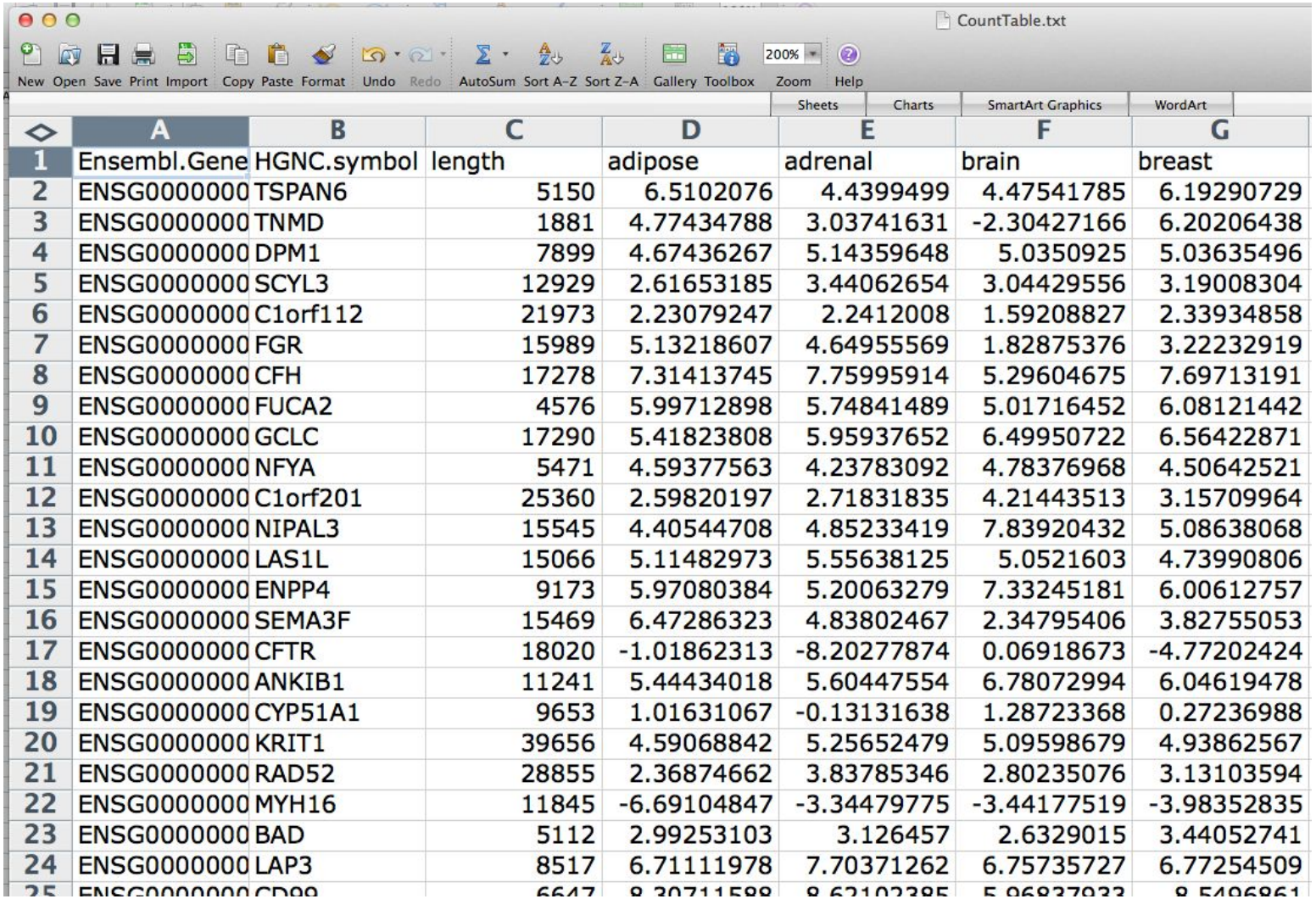


Figure 1. Exploratory plots. (a) The points show the frequency of counts in the bins shown on the x-axis. The three colors represent three samples (NA12812, NA12874, NA11993) from the Montgomery data. (b) $\log_2$ RPKM values are stratified by GC-content for two biological replicates from the Montgomery data (NA11918, NA12761) and are summarized by boxplots. The two samples are distinguished by the two colors. Genes with average (across all 60 samples) $\log_2$ RPKM values below 2 are not shown. (c) Log-fold changes between RPKM values from the two samples and the same genes shown in (b) were computed and are plotted against GC-content. Red is used to show the genes with the 10% highest GC-content and blue is used to show the genes with the 10% lowest GC-content. (d) RPKM log-fold-changes are plotted against average $\log_2$ counts for the samples and genes shown in (b), with the same color coding as in (c). (e) As (d) but from values corrected using the method proposed by Pickrell *and others* (2010). (f) As (d) but for values normalized using our approach (see Methods).

Stat Models for RNA-seq Data



Result of Quantification



	A	B	C	D	E	F	G
1	Ensembl.Gene	HGNC.symbol	length	adipose	adrenal	brain	breast
2	ENSG0000000	TSPAN6	5150	6.5102076	4.4399499	4.47541785	6.19290729
3	ENSG0000000	TNMD	1881	4.77434788	3.03741631	-2.30427166	6.20206438
4	ENSG0000000	DPM1	7899	4.67436267	5.14359648	5.0350925	5.03635496
5	ENSG0000000	SCYL3	12929	2.61653185	3.44062654	3.04429556	3.19008304
6	ENSG0000000	C1orf112	21973	2.23079247	2.2412008	1.59208827	2.33934858
7	ENSG0000000	FGR	15989	5.13218607	4.64955569	1.82875376	3.22232919
8	ENSG0000000	CFH	17278	7.31413745	7.75995914	5.29604675	7.69713191
9	ENSG0000000	FUCA2	4576	5.99712898	5.74841489	5.01716452	6.08121442
10	ENSG0000000	GCLC	17290	5.41823808	5.95937652	6.49950722	6.56422871
11	ENSG0000000	NFYA	5471	4.59377563	4.23783092	4.78376968	4.50642521
12	ENSG0000000	C1orf201	25360	2.59820197	2.71831835	4.21443513	3.15709964
13	ENSG0000000	NIPAL3	15545	4.40544708	4.85233419	7.83920432	5.08638068
14	ENSG0000000	LAS1L	15066	5.11482973	5.55638125	5.0521603	4.73990806
15	ENSG0000000	ENPP4	9173	5.97080384	5.20063279	7.33245181	6.00612757
16	ENSG0000000	SEMA3F	15469	6.47286323	4.83802467	2.34795406	3.82755053
17	ENSG0000000	CFTR	18020	-1.01862313	-8.20277874	0.06918673	-4.77202424
18	ENSG0000000	ANKIB1	11241	5.44434018	5.60447554	6.78072994	6.04619478
19	ENSG0000000	CYP51A1	9653	1.01631067	-0.13131638	1.28723368	0.27236988
20	ENSG0000000	KRIT1	39656	4.59068842	5.25652479	5.09598679	4.93862567
21	ENSG0000000	RAD52	28855	2.36874662	3.83785346	2.80235076	3.13103594
22	ENSG0000000	MYH16	11845	-6.69104847	-3.34479775	-3.44177519	-3.98352835
23	ENSG0000000	BAD	5112	2.99253103	3.126457	2.6329015	3.44052741
24	ENSG0000000	LAP3	8517	6.71111978	7.70371262	6.75735727	6.77254509
25	ENSG0000000	CD99	6647	8.30711588	8.62102385	5.96837933	8.5496861

Soneson and Delorenzi *BMC Bioinformatics* 2013, **14**:91
http://www.biomedcentral.com/1471-2105/14/91

BMC Bioinformatics

RESEARCH ARTICLE Open Access

A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson[1*] and Mauro Delorenzi[1,2]

Abstract

Background: Finding genes that are differentially expressed understanding the molecular basis of phenotypic variation used extensively to quantify the abundance of mRNA corr high-throughput sequencing of cDNA (RNA-seq) has emer sequencing decreases, it is conceivable that the use of RN rapidly. To exploit the possibilities and address the challenges software packages have been developed especially for differer

Results: We conducted an extensive comparison of eleven m data. All methods are freely available within the R framework a reads mapping to each genomic feature of interest in each of on both simulated data and real RNA-seq data.

Conclusions: Very small sample sizes, which are still common evaluated methods and any results obtained under such conc sample sizes, the methods combining a variance-stabilizing tra expression analysis perform well under many different conditi

Keywords: Differential expression, Gene expression, RNA-seq

Rapaport *et al. Genome Biology* 2013, **14**:R95
http://genomebiology.com/2013/14/9/R95

Genome Biology

METHOD Open Access

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4*]

Abstract

A large number of computational methods have been developed for analyzing differential gene expression in
valuation of common methods using the SEQC benchmark dataset
ey features, including normalization, accuracy of differential
analysis when one condition has no detectable expression. We
, but note that array-based methods adapted to RNA-seq data
RNA-seq. Our results demonstrate that increasing the number of
on power over increased sequencing depth.

Briefings in Bioinformatics Advance Access published December 2, 2013

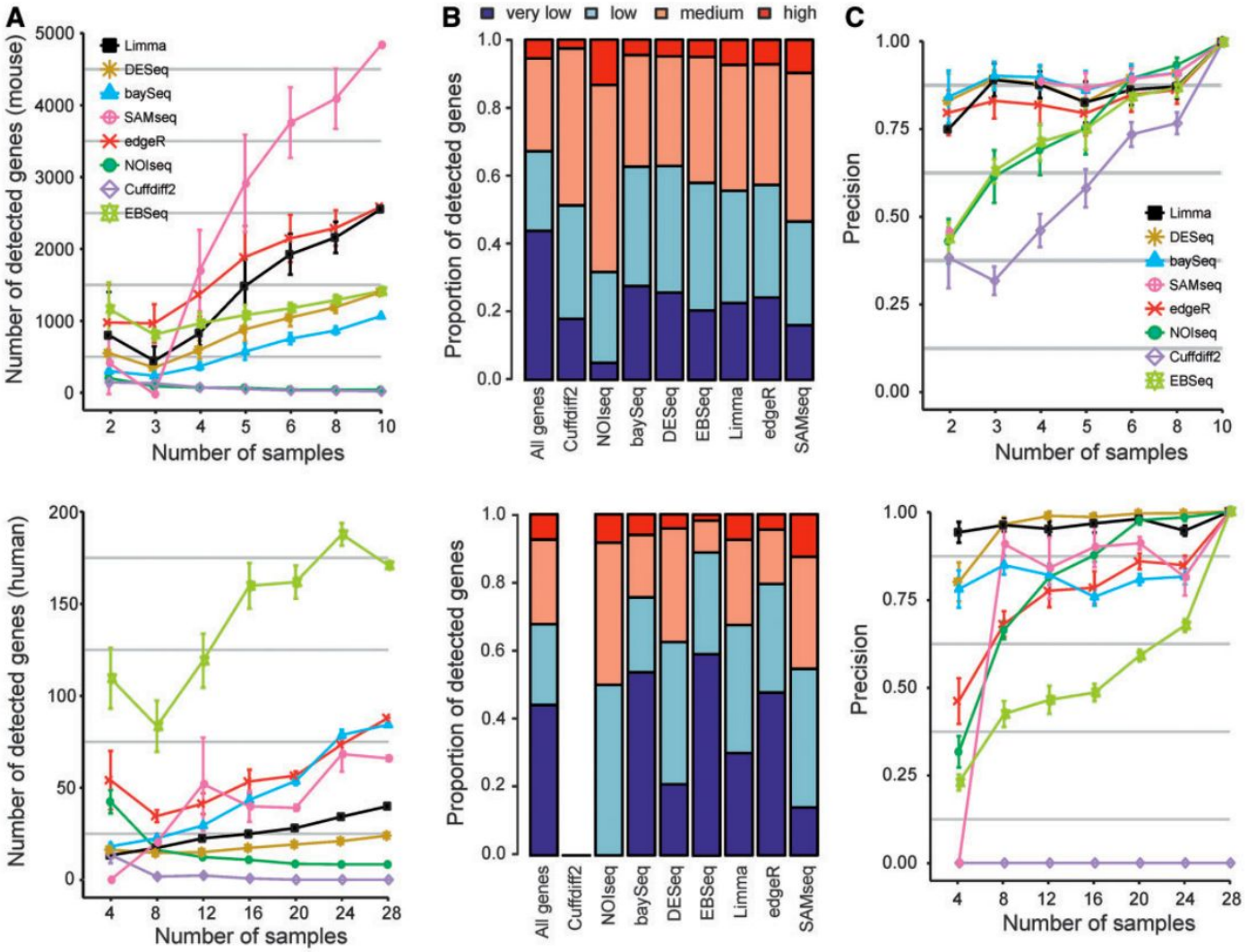
BRIEFINGS IN BIOINFORMATICS. page 1 of 12 doi:10.1093/bib/bbt086

Comparison of software packages for detecting differential expression in RNA-seq studies

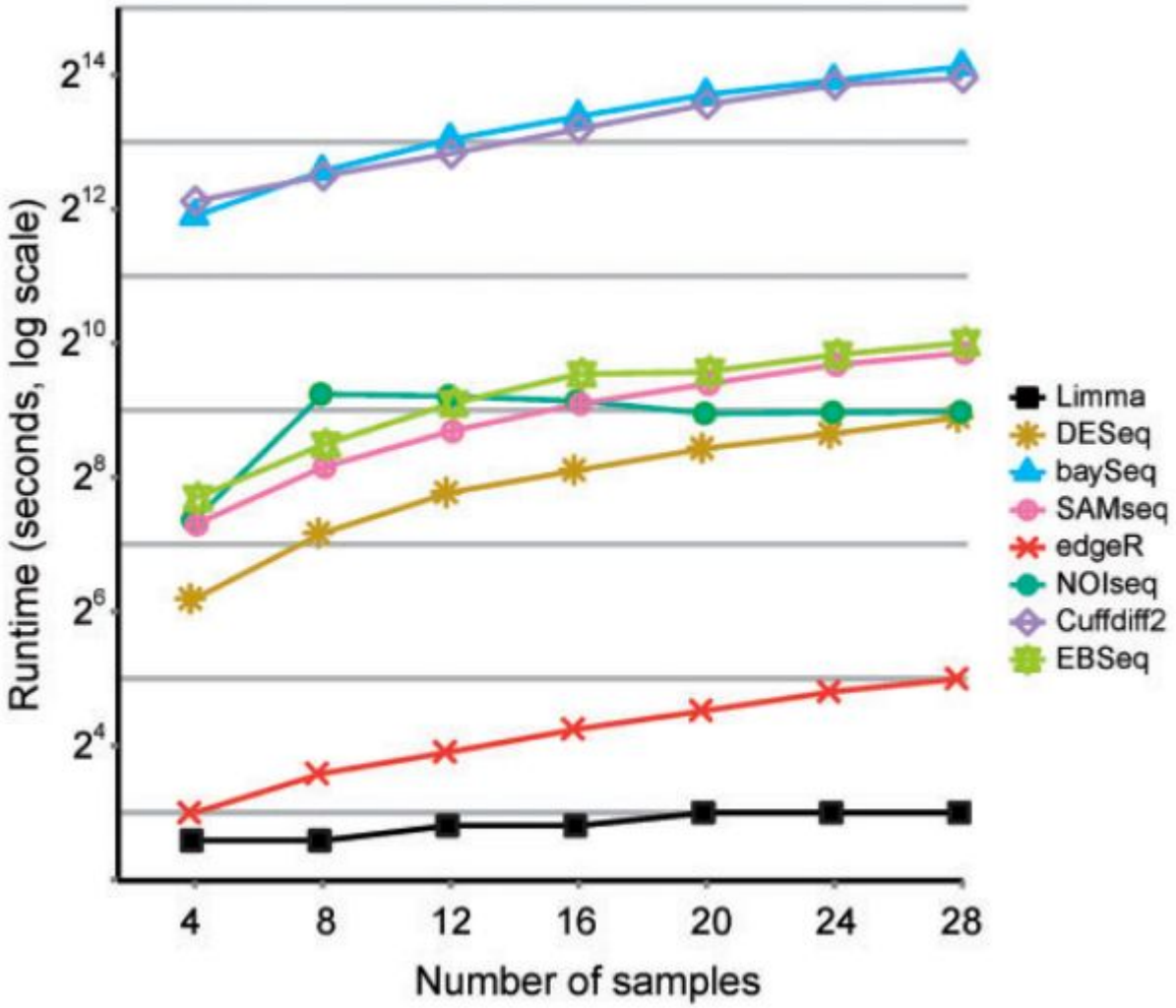
Fatemeh Seyednasrollah, Asta Laiho and Laura L. Elo

Submitted: 20th August 2013; Received (in revised form): 9th October 2013

DE Evaluation



DE Software Runtime



PROTOCOL

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

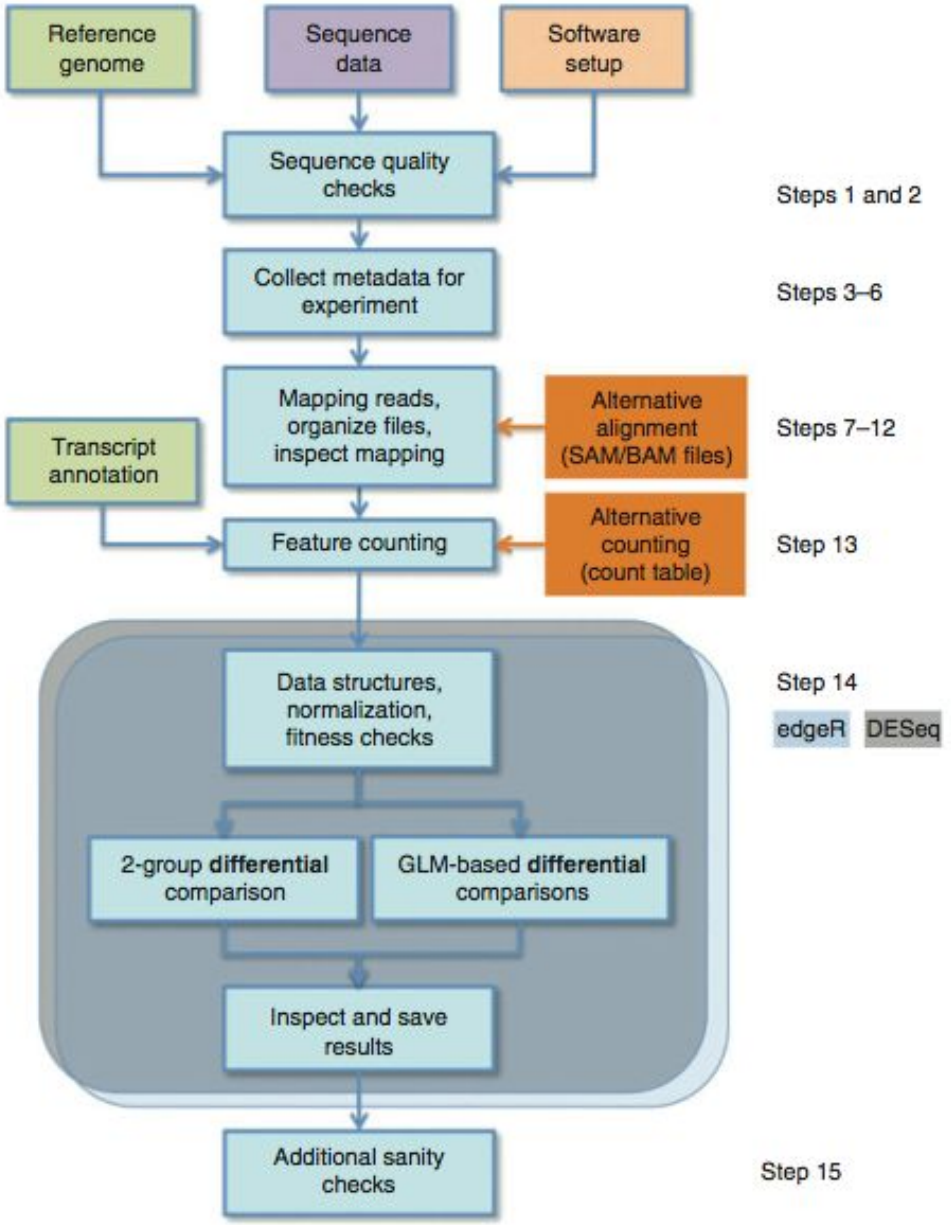
Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]

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

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

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

RNA-seq workflow as proposed by Anders et al. in Nature Protocols



Experimental Design

- What are my goals?
  - Differential expression?
  - Transcriptome assembly?
  - Identify rare, novel trancripts?
- System characteristics?
  - Large, expanded genome?
  - Intron/exon structures complex?
  - No reference genome or transcriptome

Experimental Design

- Technical replicates
  - Probably not needed due to low technical variation
- Biological replicates
  - Not explicitly needed for transcript assembly
  - Essential for differential expression analysis
  - Number of replicates often driven by sample availability for human studies
  - More is almost always better

Further Reading

https://www.rna-seqblog.com/

Further Reading

Analyzing RNA-seq data with DESeq2

Michael I. Love, Simon Anders, and Wolfgang Huber

10/27/2021

Abstract

A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, and mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions. This vignette explains the use of the package and demonstrates typical workflows. An RNA-seq workflow on the Bioconductor website covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files. DESeq2 package version: 1.35.0

http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Further Reading

REVIEWS

RNA sequencing: the teenage years

Rory Stark[1], *Marta Grzelak*[1] *and James Hadfield*[2]*

Abstract | Over the past decade, RNA sequencing (RNA-seq) has become an indispensable tool for transcriptome-wide analysis of differential gene expression and differential splicing of mRNAs. However, as next-generation sequencing technologies have developed, so too has RNA-seq. Now, RNA-seq methods are available for studying many different aspects of RNA biology, including single-cell gene expression, translation (the translatome) and RNA structure (the structurome). Exciting new applications are being explored, such as spatial transcriptomics (spatialomics). Together with new long-read and direct RNA-seq technologies and better computational tools for data analysis, innovations in RNA-seq are contributing to a fuller understanding of RNA biology, from questions such as when and where transcription occurs to the folding and intermolecular interactions that govern RNA function.

https://www.nature.com/articles/s41576-019-0150-2

Further Reading

REVIEW Open Access

RNA sequencing: new technologies and applications in cancer research

Mingye Hong[1†], Shuang Tao[2†], Ling Zhang[3], Li-Ting Diao[2], Xuanmei Huang[1], Shaohui Huang[1], Shu-Juan Xie[2], Zhen-Dong Xiao[2*] and Hua Zhang[1*]

Abstract

Over the past few decades, RNA sequencing has significantly progressed, becoming a paramount approach for transcriptome profiling. The revolution from bulk RNA sequencing to single-molecular, single-cell and spatial transcriptome approaches has enabled increasingly accurate, individual cell resolution incorporated with spatial information. Cancer, a major malignant and heterogeneous lethal disease, remains an enormous challenge in medical research and clinical treatment. As a vital tool, RNA sequencing has been utilized in many aspects of cancer research and therapy, including biomarker discovery and characterization of cancer heterogeneity and evolution, drug resistance, cancer immune microenvironment and immunotherapy, cancer neoantigens and so on. In this review, the latest studies on RNA sequencing technology and their applications in cancer are summarized, and future challenges and opportunities for RNA sequencing technology in cancer applications are discussed.

Keywords: RNA sequencing, Application, Cancer

https://jhoonline.biomedcentral.com/articles/10.1186/s13045-020-01005-x