An RNA-seq workflow

Charlotte Soneson

2017-11-13

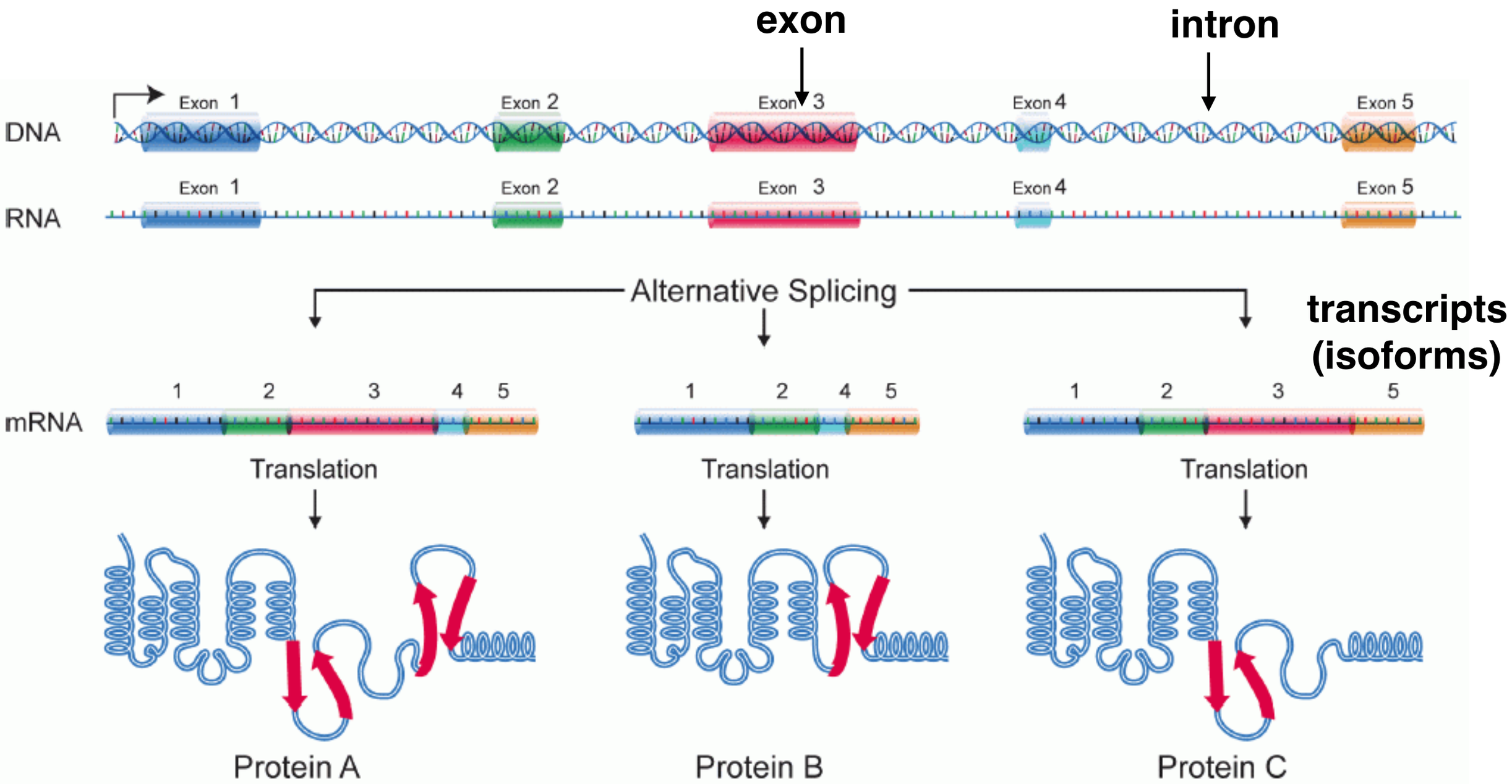
imlspenticton.uzh.ch/dump/STA426_2017/

Apps UZH UZH - Institute of M... Bioconductor build/... Angebote | ASVZ Profiles Altmetric it English - German D

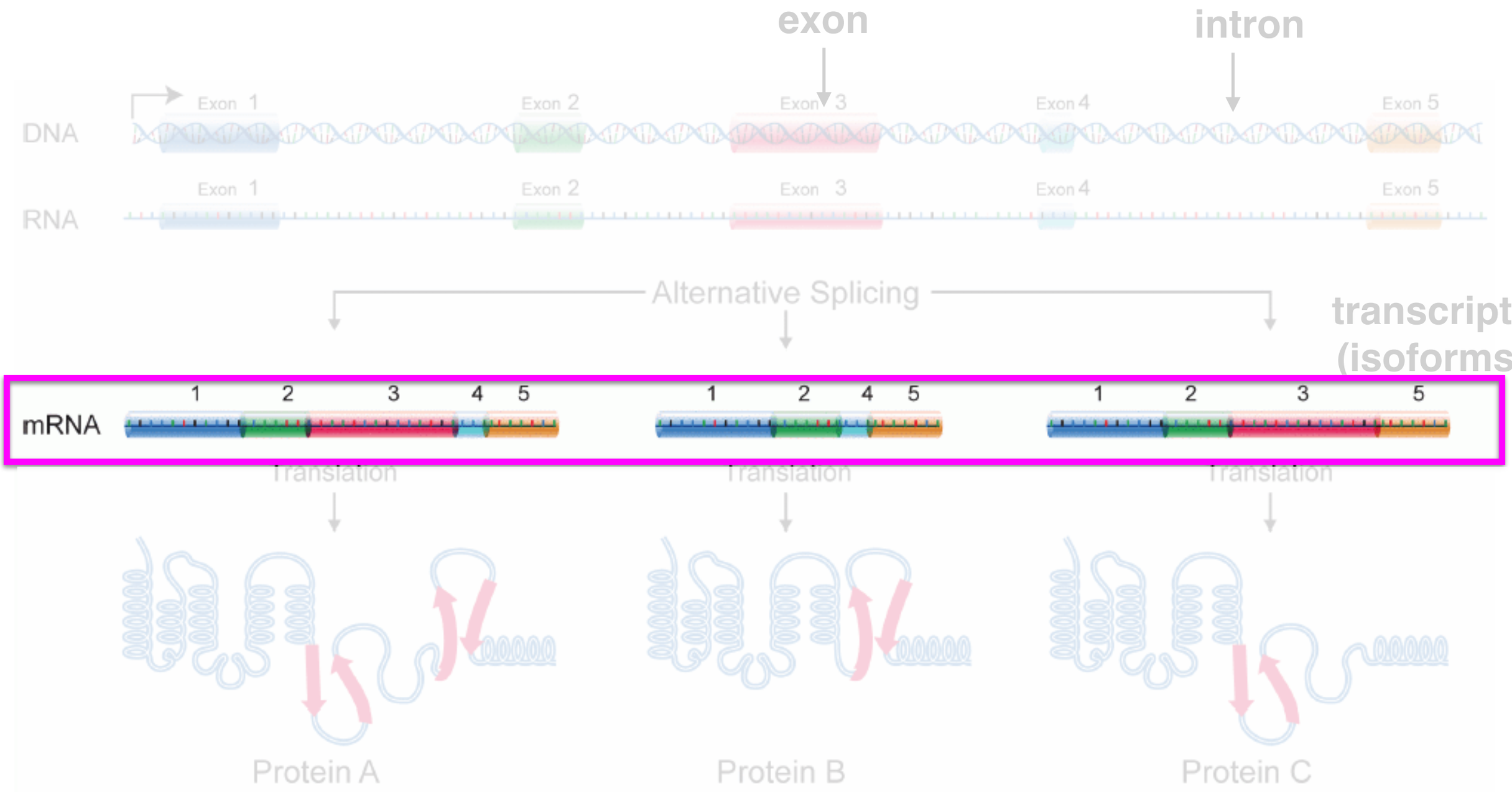
Index of /dump/STA426_2017

Name	Last modified	Size	Description
Parent Directory		-	
STA426_2017_RNAseq_input.zip	2017-11-06 19:20	70M	
STA426_2017_RNAseq_output.zip	2017-11-06 19:20	199M	

Apache Server at imlspenticton.uzh.ch Port 80


exon
intron
DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
transcripts (isoforms)
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

Wikipedia


exon
intron
DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
transcripts (isoforms)
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

Wikipedia

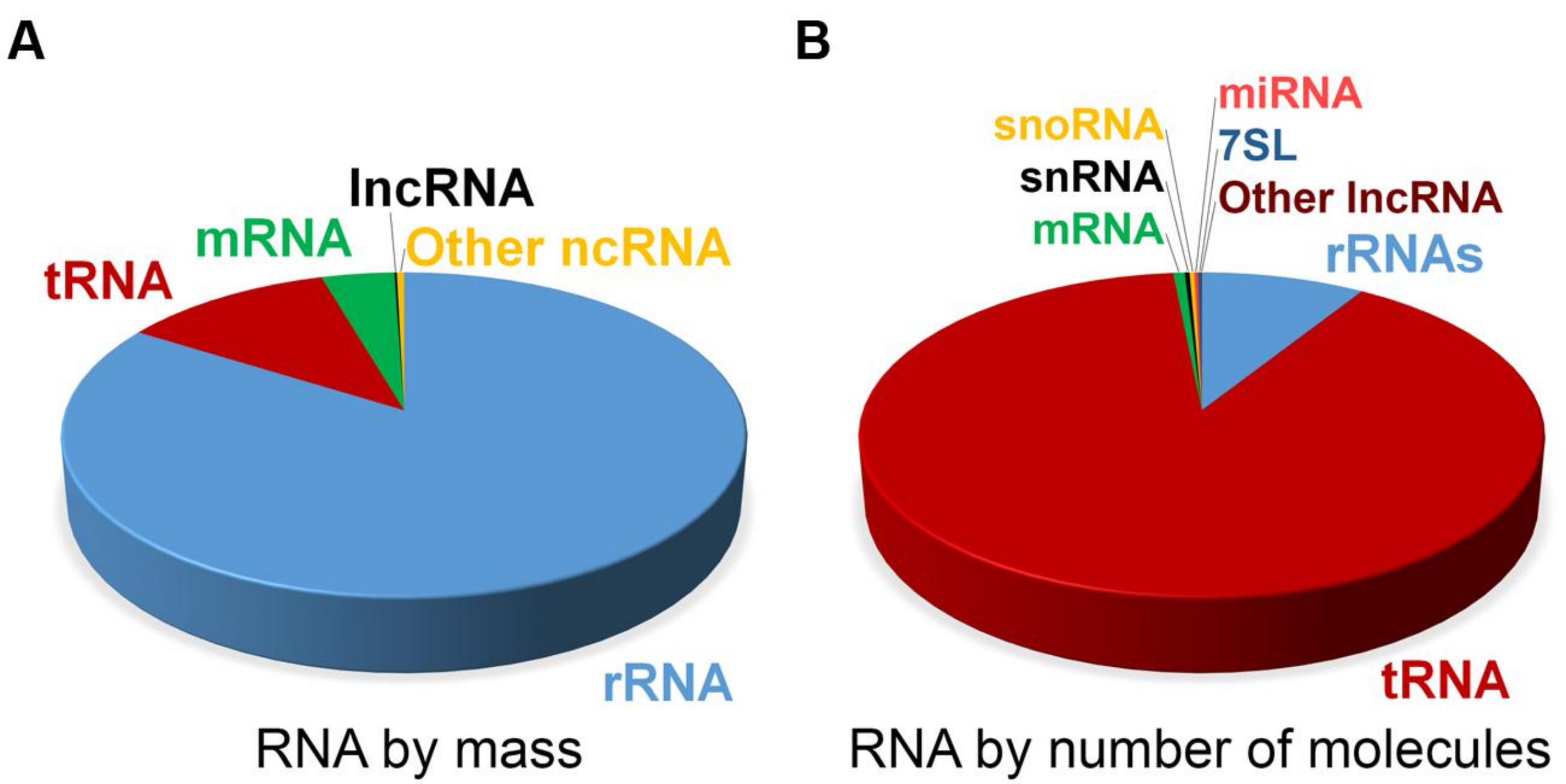


FIGURE 1. Estimate of RNA levels in a typical mammalian cell. Proportion of the various classes of RNA in mammalian somatic cells by total mass (A) and by absolute number of molecules (B). Total number of RNA molecules is estimated at roughly 10^7 per cell. Other ncRNAs in (A) include snRNA, snoRNA, and miRNA. Note that due to their relatively large sizes, rRNA, mRNA, and lncRNAs make up a larger proportion of the mass as compared to the overall number of molecules.

Palazzo & Lee, Front. Genet. 2015

Differential analysis types for RNA-seq

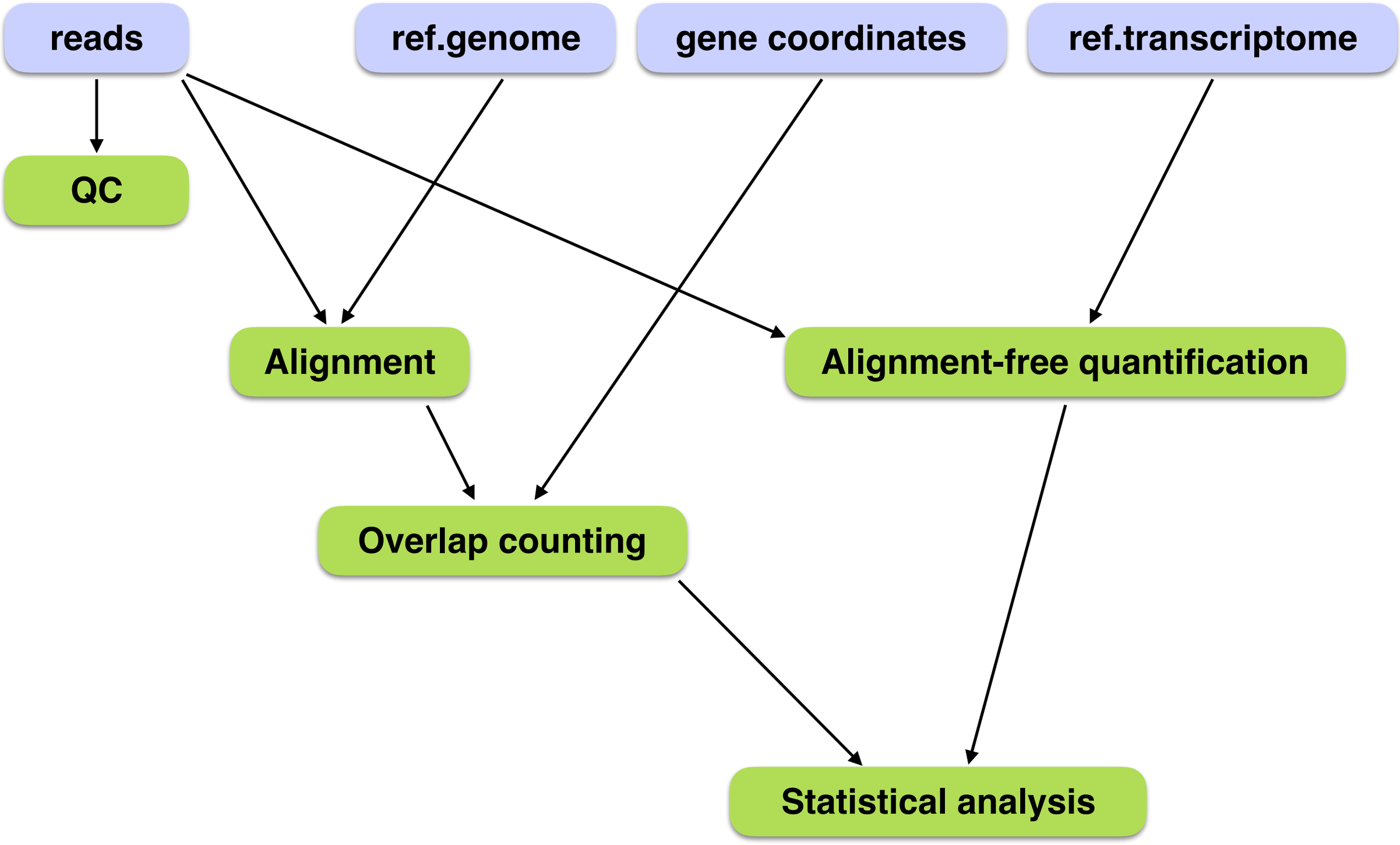
- Does the total output of a gene change between conditions? **D**ifferential **G**ene **E**xpression

- Does the expression of individual transcripts change? **D**ifferential **T**ranscript **E**xpression

- Does *any* isoform of a given gene change? **DTE+G**

- Does the isoform composition for a given gene change? **D**ifferential **T**ranscript **U**sage**/D**ifferential **E**xon **U**sage

- need **different** computational approaches (quantifications + tests)


reads
ref.genome
gene coordinates
ref.transcriptome
QC
Alignment
Alignment-free quantification
Overlap counting
Statistical analysis

Raw data

- FASTQ files: sequence + base quality (phred score)
- quality encoding:

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0.2......................26...31.......41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

DEMO

Reference files

- Ensembl: http://www.ensembl.org/info/data/ftp/index.html
- UCSC: http://hgdownload.cse.ucsc.edu/downloads.html
- iGenome: http://support.illumina.com/sequencing/sequencing_software/igenome.html
- **Be consistent!**
- Different chromosome identifiers

Reference files

- Reference genomes and annotations are continuously refined, extended and improved

- Keep track of version and be consistent!

SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
MAMMALS				
Human	hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI Build 34	Available
	hg15	Apr. 2003	NCBI Build 33	Archived
	hg13	Nov. 2002	NCBI Build 31	Archived
	hg12	Jun. 2002	NCBI Build 30	Archived
	hg11	Apr. 2002	NCBI Build 29	Archived (data only)
	hg10	Dec. 2001	NCBI Build 28	Archived (data only)
	hg8	Aug. 2001	UCSC-assembled	Archived (data only)
	hg7	Apr. 2001	UCSC-assembled	Archived (data only)
	hg6	Dec. 2000	UCSC-assembled	Archived (data only)
	hg5	Oct. 2000	UCSC-assembled	Archived (data only)
	hg4	Sep. 2000	UCSC-assembled	Archived (data only)
	hg3	Jul. 2000	UCSC-assembled	Archived (data only)

Ex: Ensembl, GRCh38.86

- ftp://ftp.ensembl.org/pub/release-86/fasta/homo_sapiens/dna/

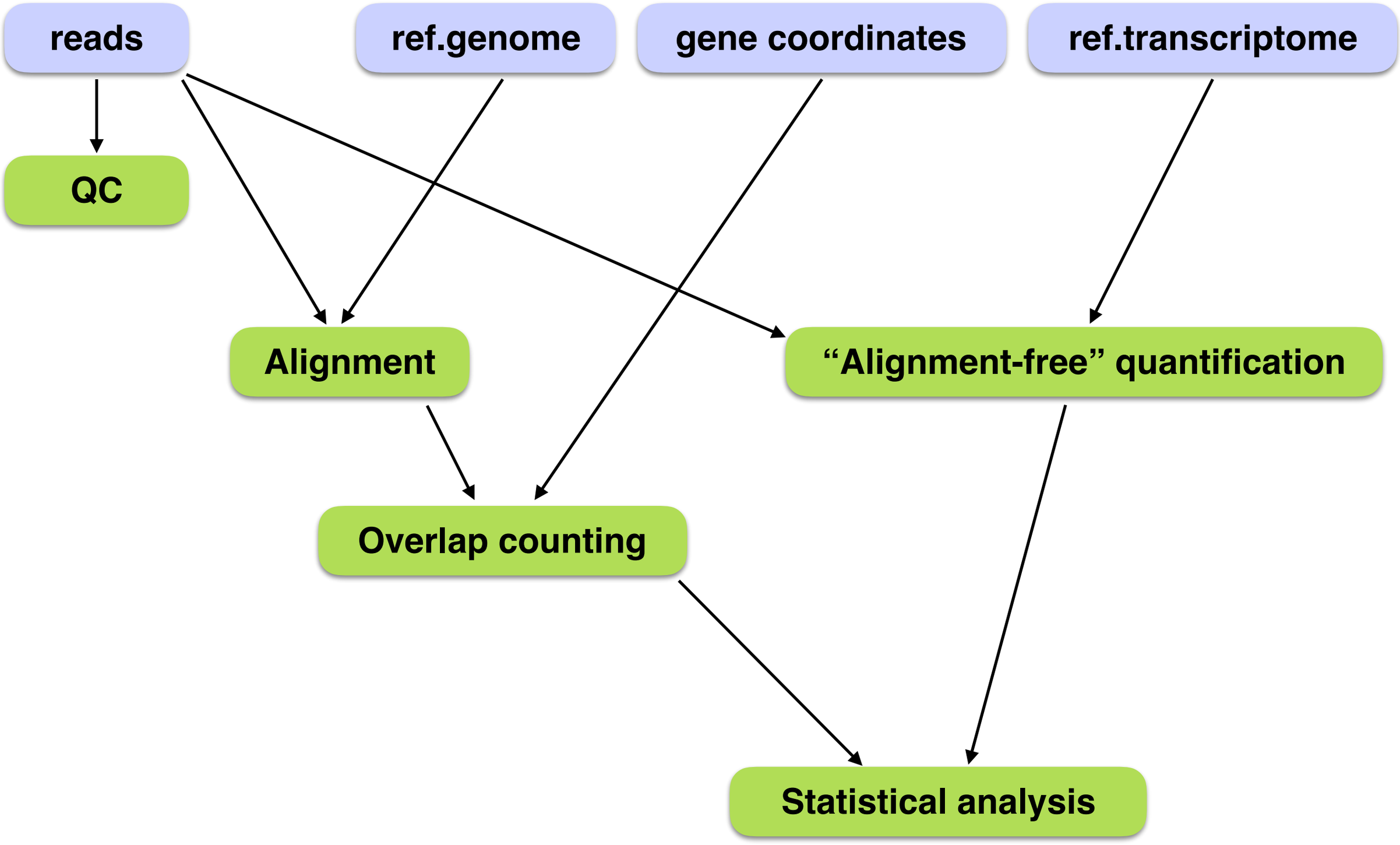
```
---------
TOPLEVEL
---------
These files contains all sequence regions flagged as toplevel in an Ensembl
schema. This includes chromsomes, regions not assembled into chromosomes and
N padded haplotype/patch regions.


-----------------
PRIMARY ASSEMBLY
-----------------
Primary assembly contains all toplevel sequence regions excluding haplotypes
and patches. This file is best used for performing sequence similarity searches
where patch and haplotype sequences would confuse analysis.
```

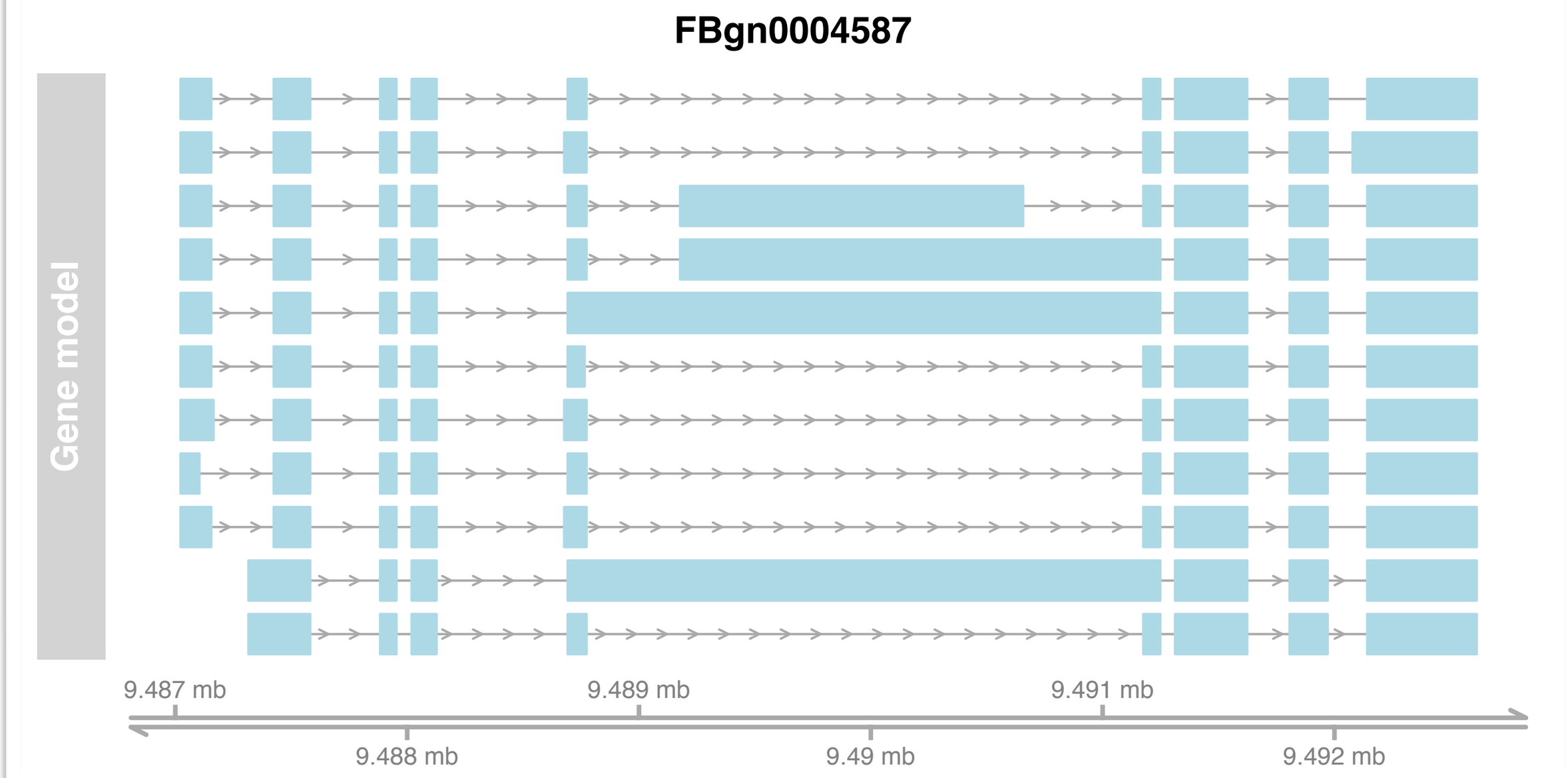
DEMO

Preprocessing

- We often want to compare abundance (expression) of genes or other features between conditions
- Data come as sequencing reads
- Preprocessing turns these into an abundance table


reads
ref.genome
gene coordinates
ref.transcriptome
QC
Alignment
“Alignment-free” quantification
Overlap counting
Statistical analysis

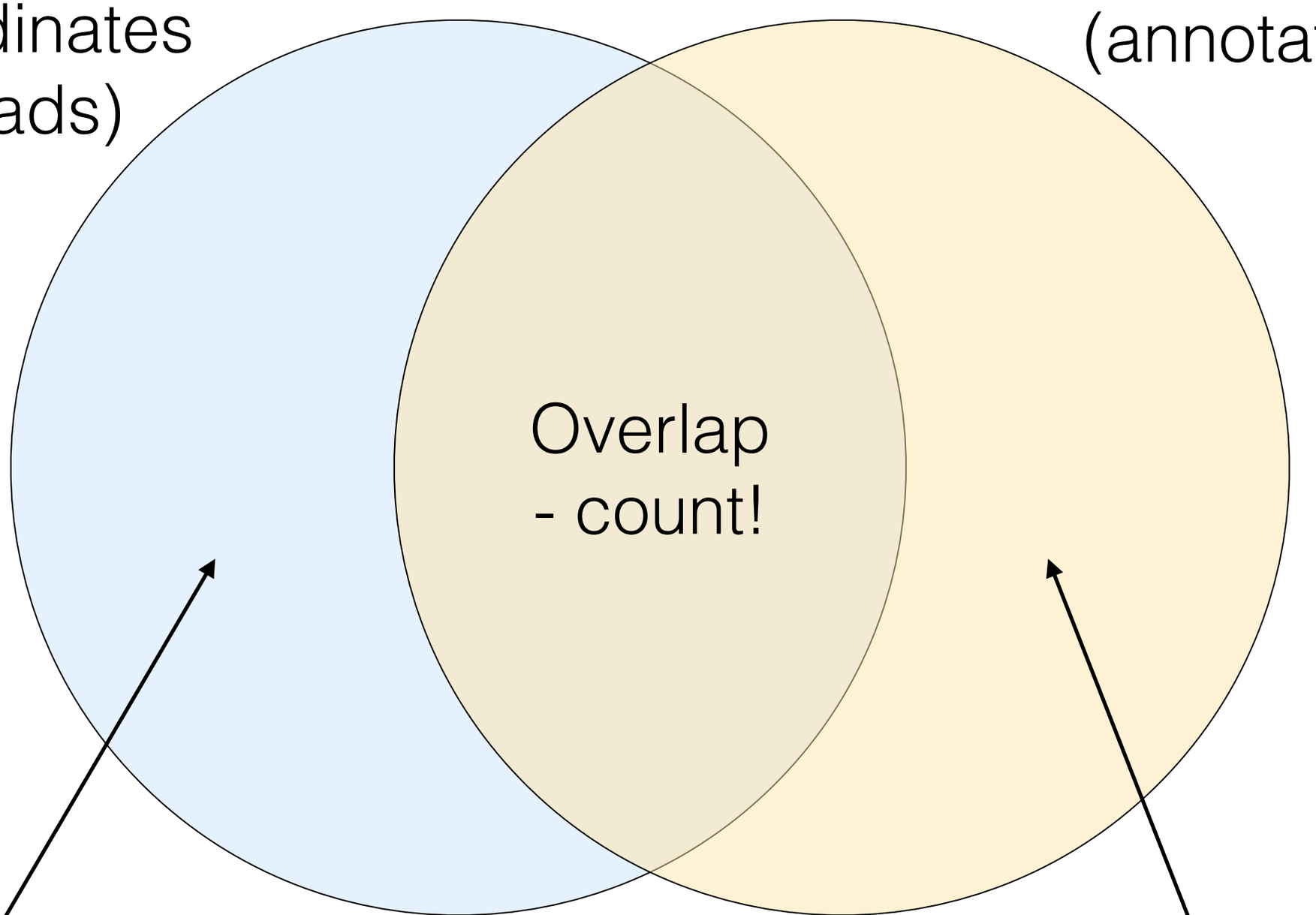
Genome vs transcriptome alignment



DEMO

Alignment coordinates (reads)

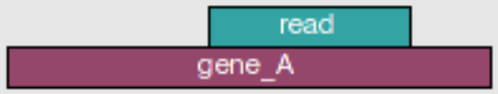
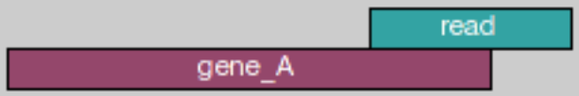
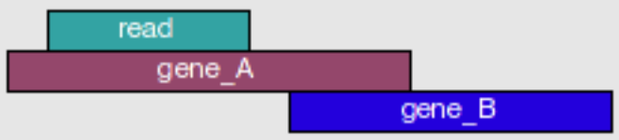
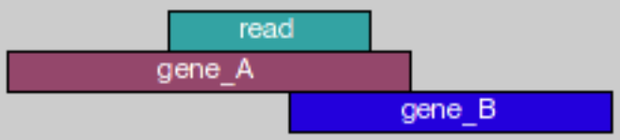
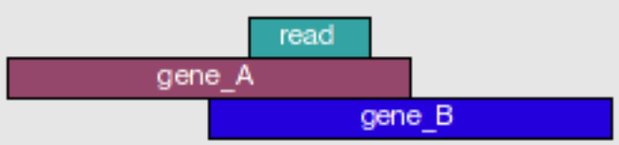
Gene coordinates (annotation)



Reads from “unannotated” genes

Non-expressed genes

The annotation matters!

	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

The annotation matters!

DEMO

http://bioconductor.org/

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1211 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

News

- Bioconductor 3.3 is available.
- Bioconductor F1000 Research Channel launched.
- Orchestrating high-throughput genomic analysis with *Bioconductor* (abstract) and other recent literature.
- Read our latest newsletter and course material.
- Use the support site to get help installing, learning and using Bioconductor.

Install »

Get started with *Bioconductor*

- Install *Bioconductor*
- Explore packages
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

Use »

Create bioinformatic solutions with *Bioconductor*

- Software, Annotation, and Experiment packages
- Amazon Machine Image
- Latest release annoucement
- Support site

Develop »

Contribute to *Bioconductor*

- Developer resources
- Use Bioc 'devel'
- 'Devel' Software, Annotation and Experiment packages
- Package guidelines
- New package submission
- Build reports

https://support.bioconductor.org/

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ASK QUESTION | LATEST | NEWS | JOBS | TUTORIALS | TAGS | USERS

Limit ▾ Sort ▾ Search

1 vote	1 answer	84 views	**How to remove duplicated overlap hit index in IntegerList efficiently [Updated] ?** R genomicranges iranges integerlist written 4 days ago by Jurat Shahidin • 50
1 vote	1 answer	20 views	**blastSequences queries failing due to BLAST switch to https?** blastsequences annotate blast https written 11 hours ago by ariel.hecht • 0 • updated 10 hours ago by Martin Morgan ♦♦ 18k
0 votes	1 answer	30 views	**DataFrame showAsCell function** s4vectors written 7 days ago by vedran.franke • 0 • updated 12 hours ago by Michael Lawrence ♦ 8.5k
0 votes	0 answers	17 views	**Problem using sva with limma for microarray DE analysis** sva limma written 12 hours ago by Stane • 0
0 votes	1 answer	34 views	**modeling heteroscedasticity in limma-voom for RNA-Seq data analysis** limma written 16 hours ago by Yanzhu Lin • 110 • updated 13 hours ago by Gordon Smyth ♦ 28k
0 votes	0 answers	17 views	**Limma DE analysis using all microarray or subset of interest microarray** limma written 13 hours ago by Stane • 0
3 votes	3 answers	37 views	**Biomart getLDS error** biomart getlds orthlog written 19 hours ago by mohamed.diwan • 0 • updated 14 hours ago by Thomas Maurel • 530
1 vote	4 answers	47 views	**Error in biomaRt User Guide example Task 11** biomart written 3 days ago by cring • 0 • updated 14 hours ago by Thomas Maurel • 530
0	2	31	**biomaRt error while querying homo sapien structural variants**

Recent...

Replies
- C: Table export from R by Michael Love ♦ 9.2k
- C: blastSequences queries f... by ariel.hecht • 0
- A: blastSequences queries f... by Martin Morgan ♦♦ 18k
- C: How to remove duplicated... by Jurat Shahidin • 50
- C: How to remove duplicated... by Jurat Shahidin • 50

Votes
- A: blastSequences queries f...
- C: How to remove duplicated...
- Fisher's method of combinin...
- A: Fisher's method of combi...
- qRT-PCR - reading tab-delim...

Awards • All »
- Scholar to Aaron Lun • 11k
- Scholar to Michael Lawrence ♦ 8.5k
- Commentator to Aaron Lun • 11k
- Scholar to Steve Lianoglou ♦ 11k
- Appreciated to Aaron Lun • 11k
- Scholar to Dan Tenenbaum ♦♦ 8.1k

Locations • All »
- Italy, 14 minutes ago
- Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, 35

https://bioconductor.org/packages/release/bioc/html/Rsubread.html

Rsubread

platforms some | downloads top 5% | posts 12 / 2 / 3 / 2 | in Bioc 5.5 years
build ok | commits 4.17 | test coverage unknown

Subread sequence alignment for R

Bioconductor version: Release (3.3)

Provides powerful and easy-to-use tools for analyzing next-gen sequencing read data. Includes quality assessment of sequence reads, read alignment, read summarization, exon-exon junction detection, fusion detection, detection of short and long indels, absolute expression calling and SNP calling. Can be used with reads generated from any of the major sequencing platforms including Illumina GA/HiSeq/MiSeq, Roche GS-FLX, ABI SOLiD and LifeTech Ion PGM/Proton sequencers.

Author: Wei Shi and Yang Liao with contributions from Jenny Zhiyin Dai and Timothy Triche, Jr.

Maintainer: Wei Shi <shi at wehi.edu.au>

Citation (from within R, enter `citation("Rsubread")`):

> Liao Y, Smyth GK and Shi W (2013). "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote." *Nucleic Acids Research*, **41**, pp. e108.

Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("Rsubread")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("Rsubread")
```

PDF	R Script	Rsubread Vignette
PDF		Reference Manual
Text		NEWS

DEMO

Homo_sapiens.GRC... | 18 | 18:47,006,705-47,019,802 | Go

13 kb

47,008 kb | 47,010 kb | 47,012 kb | 47,014 kb | 47,016 kb | 47,018 kb

SRR1039508_chr18_hisat2.bam
rage

[0 - 6912]

SRR1039508_chr18_hisat2.bam

Homo_sapiens.GRCh37.75.chr18

ENST00000579820
ENST00000581232
RPL17
ENST00000318240
ENST00000410758
ENST00000584364
ENST00000582782
ENST00000584895
ENST00000607313
ENST00000582392
C18orf32
ENST00000580261
ENST00000332968
SNORD58B
ENST00000577910

DEMO

Gene identifiers (ex: BCL2)

- Ensembl ID: ENSG00000171791
- Entrez ID: 596
- Vega ID: OTTHUMG00000132791
- HGNC ID: 990
- RefSeq ID: NG_009361.1
- UCSC ID: uc002liu.2
- Official symbol: BCL2
- Synonyms: PPP1R50, Bcl-2
- …

Typically, no 1-1 mapping between different ID types

Gene symbols can change over time

ensembl_gene_id	may2009_hgnc_symbol	may2012_hgnc_symbol	dec2013_hgnc_symbol
ENSG00000162825	KIAA1245		NBPF8

feb2014_hgnc_symbol	aug2014_hgnc_symbol	oct2014_hgnc_symbol	dec2014_hgnc_symbol
NBPF8	NBPF8	NBPF8	NBPF8

mar2015_hgnc_symbol	may2015_hgnc_symbol	jul2015_hgnc_symbol	sep2015_hgnc_symbol
NBPF8	NBPF8	NBPF20	NBPF20

dec2015_hgnc_symbol	mar2016_hgnc_symbol	jul2016_hgnc_symbol	oct2016_hgnc_symbol
NBPF20	NBPF20	NBPF20	NBPF20

ensembl_gene_id	may2009_hgnc_symbol	may2012_hgnc_symbol	dec2013_hgnc_symbol
ENSG00000179412			HNRNPCP5

feb2014_hgnc_symbol	aug2014_hgnc_symbol	oct2014_hgnc_symbol	dec2014_hgnc_symbol
HNRNPCP5	HNRNPCL2	HNRNPCL2	HNRNPCL4

mar2015_hgnc_symbol	may2015_hgnc_symbol	jul2015_hgnc_symbol	sep2015_hgnc_symbol
HNRNPCL4	HNRNPCL4	HNRNPCL4	HNRNPCL4

dec2015_hgnc_symbol	mar2016_hgnc_symbol	jul2016_hgnc_symbol	oct2016_hgnc_symbol
HNRNPCL4	HNRNPCL4	HNRNPCL4	HNRNPCL4

DEMO

Normalization

- Observed counts depend on:
  - abundance
  - gene length
  - sequencing depth
  - sequencing biases
  - …
- “As-is”, not directly comparable across samples

Normalization

raw count for gene *i* in sample *j*

relative abundance

normalization factor

$$C_{ij} \sim NB(\mu_{ij} = s_{ij} q_{ij}, \theta_i)$$

dispersion

- s_{ij} is a normalization factor (or offset) in the model
- counts are not explicitly scaled
  - important exception: voom/limma (followed by explicit modeling of mean-variance association)

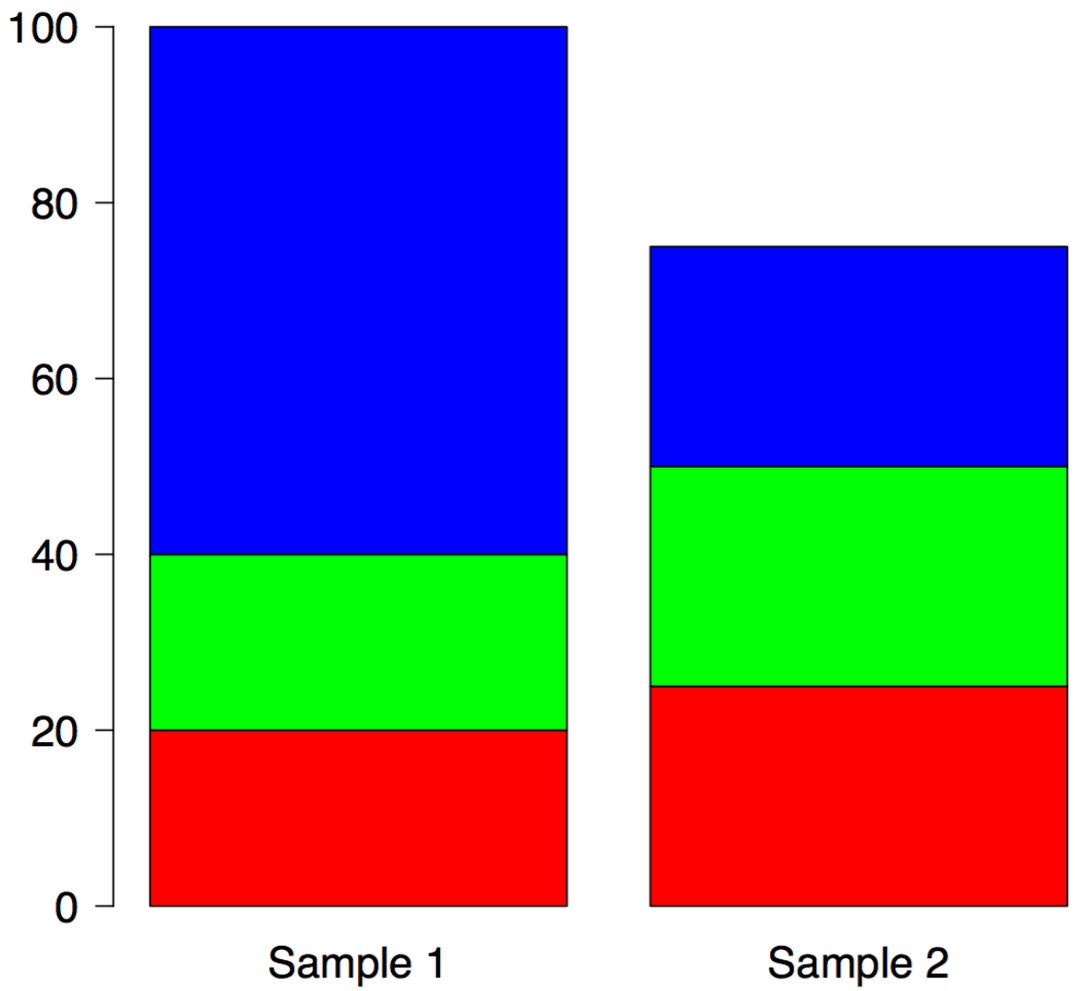
How to calculate normalization factors?

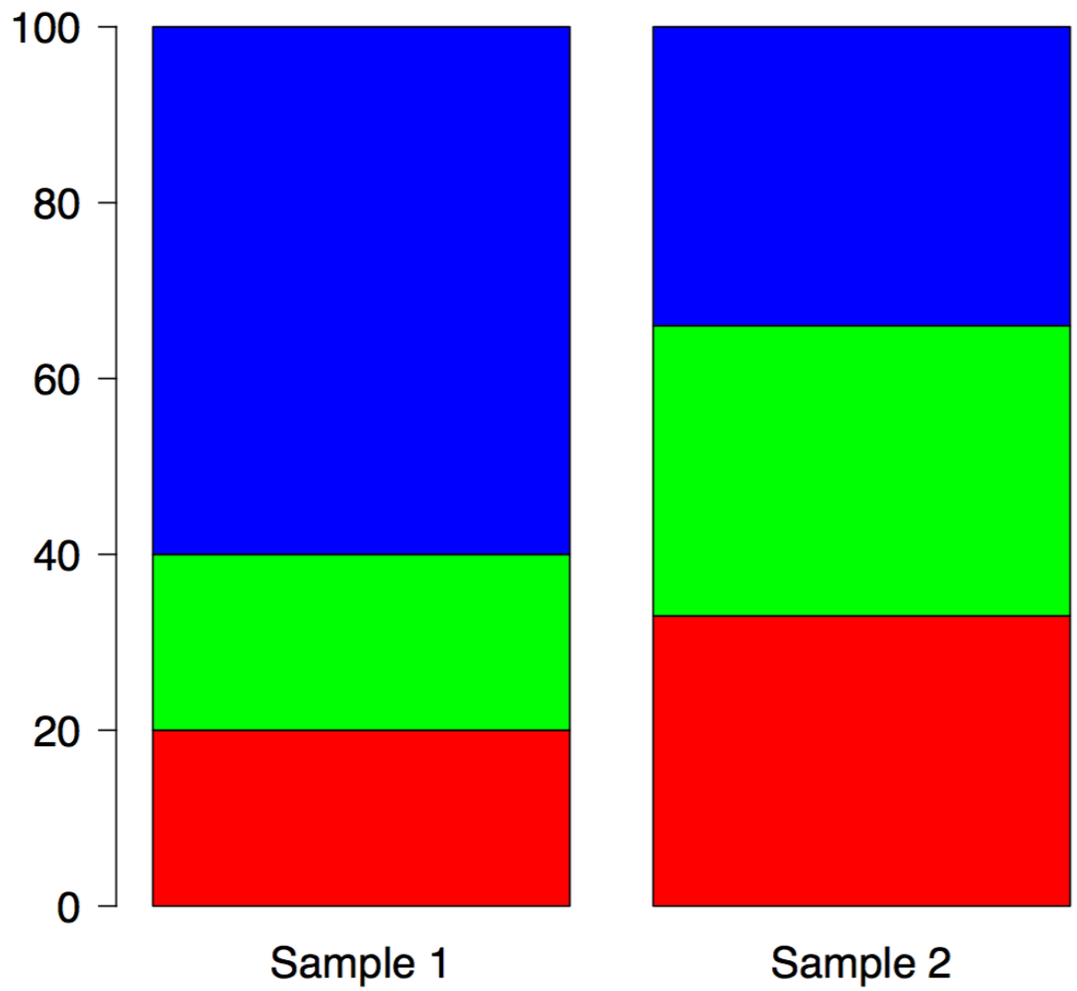
- Attempt 1: **total count** (library size)
  - Define a reference sample (one of the observed samples or a “pseudo-sample”) - gives a “target library size”
- Normalization factor for sample *j* is defined by

$$\frac{\text{total count in sample } j}{\text{total count in reference sample}}$$

The influence of RNA composition

- Observed counts are relative
- High counts for some genes are “compensated” by low counts for other genes





How to calculate normalization factors?

- Attempt 2: total count (library size) * compensation for differences in composition

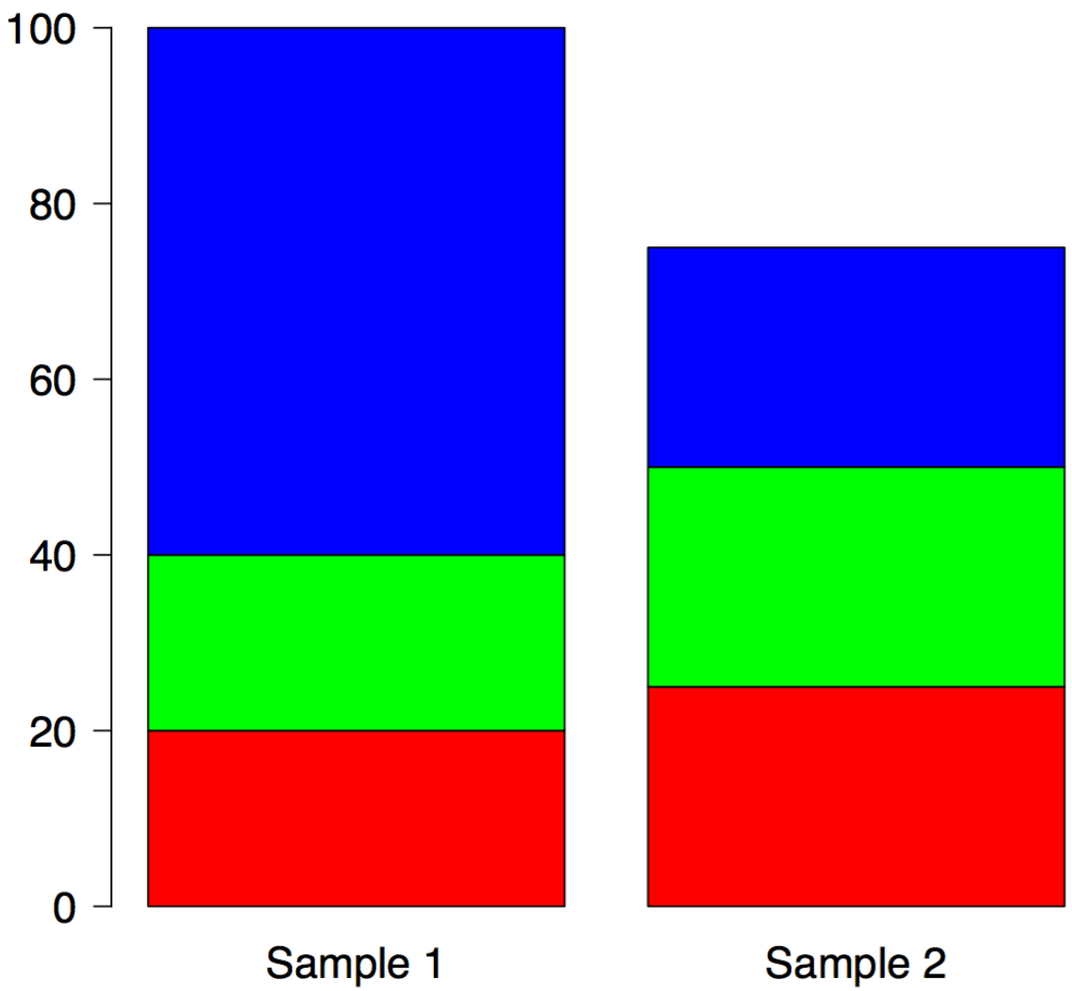
- Idea: use only non-differentially expressed genes to compute the normalization factor

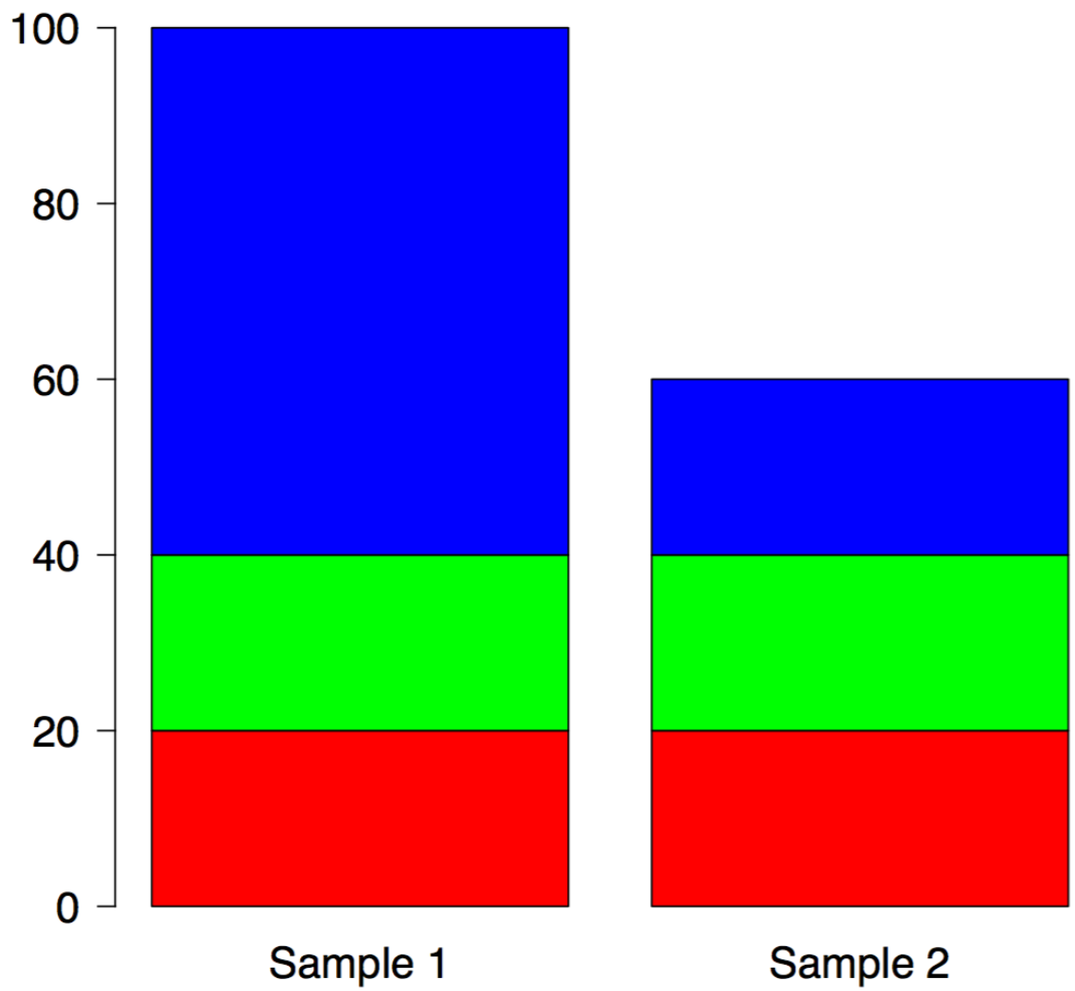
- Implemented by both edgeR (TMM) and DESeq2 (median count ratio)

- Both these methods assume that most genes are not differentially expressed

How to calculate normalization factors?

- Attempt 2: total count (library size) * compensation for differences in composition





DEMO

Be careful when importing identifiers in Excel

gene names | internal date format | default date format | gene names | internal date format | default date format | gene names | internal date format | default date format

excel.gene2date.xls

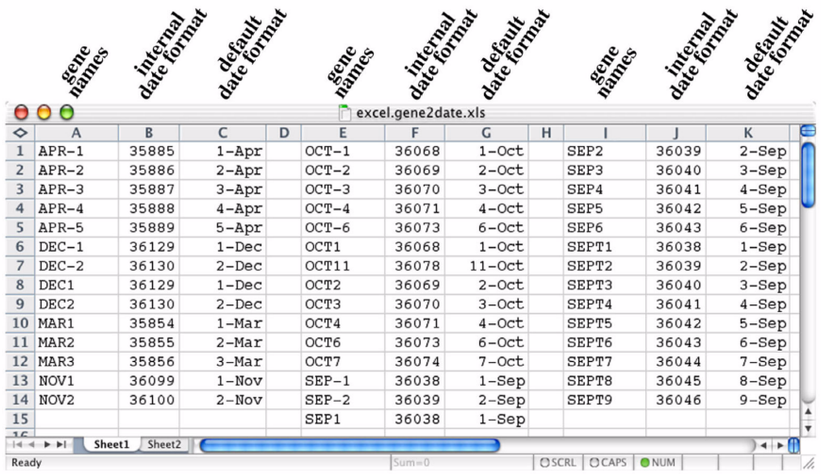
	A	B	C	D	E	F	G	H	I	J	K
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

Sheet1 Sheet2

Ready Sum=0 SCRL CAPS NUM

Zeeberg et al., BMC Bioinformatics 2004

There is a solution



	A	B	C	D	E	F	G	H	I	J	K
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

- Import properly: http://www.genenames.org/help/importing-gene-symbol-data-into-excel-correctly

- The *HGNChelper* R package can help identify misrepresented gene symbols

DEMO

Downloading public data

- Search for data sets by keywords in GEO (https://www.ncbi.nlm.nih.gov/geo/) or ArrayExpress (https://www.ebi.ac.uk/arrayexpress/)

- Download fastq files from the European Nucleotide Archive (ENA): http://www.ebi.ac.uk/ena

- Note the number of files per sample (single- vs paired-end, one sample can have multiple *runs*)

Assignment

- Choose **one** of the following:

  - Download the fastq file(s) for one publicly available sample. Briefly describe the sample, run FastQC on the file(s) and comment on the results.

  - Download the fastq file(s) for one publicly available sample. Briefly describe the sample, determine the appropriate parameters for running Salmon or kallisto, build a transcriptome index and quantify the transcript abundances.

  - Choose one organism and characterize its annotated transcripts. E.g., give the total number of transcripts and genes, plot the distribution of transcripts per gene, plot the distribution of transcript lengths, … [tip: look at the R packages rtracklayer and Biostrings].

Some suggestions for further reading

- Robinson et al.: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139-140 (2010) - **edgeR**
- Love et al.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:550 (2014) - **DESeq2**
- Law et al.: voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15:R29 (2014) - **voom**
- Patro et al.: Accurate, fast, and model-aware transcript expression quantification with Salmon. bioRxiv http://dx.doi.org/10.1101/021592 (2015) - **Salmon**
- Bray et al.: Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34(5):525-527 (2016) - **kallisto**
- Patro et al.: Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology 32:462-464 (2014) - **Sailfish**
- Pimentel et al.: Differential analysis of RNA-Seq incorporating quantification uncertainty. bioRxiv http://dx.doi.org/10.1101/058164 (2016) - **sleuth**
- Wagner et al.: Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences 131:281-285 (2012) - **TPM vs FPKM**
- Soneson et al.: Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4:1521 (2016) - **ATL offsets (tximport package)**
- Li et al.: RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics 26(4):493-500 (2010) - **TPM, RSEM**
- Soneson, Matthes et al.: Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. Genome Biology 17:12 (2016)
- Schurch et al.: How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22:839-851 (2016)
- Dillies et al.: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics 14(6):671-683 (2013)
- Soneson & Delorenzi: A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14:91 (2013)
- Anders et al.: Detecting differential usage of exons from RNA-seq data. Genome Research 22(10):2008-2017 (2012) - **DEXSeq**
- Goeman & Bühlmann: Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23(8): 980-987 (2007) - **competitive vs self-contained gene set tests**