

Applications of RNA-Seq

– Quantification and Normalization –



seit 1558

Martin Hölzer

RNA Bioinformatics and
High Throughput Analysis

Friedrich-Schiller-University Jena

January 5, 2017

RNA-Seq

- Next-Generation High-Throughput Sequencing of RNA (cDNA)
- Quality control, preprocessing of raw reads
- Assembly (genome, transcriptome, *de novo*)
- Mapping (reference)

⇒ What now?

Raw reads in Fastq format

```
1 | # Show first 4 lines of FASTQ file
2 | head -4 $FASTQ
```

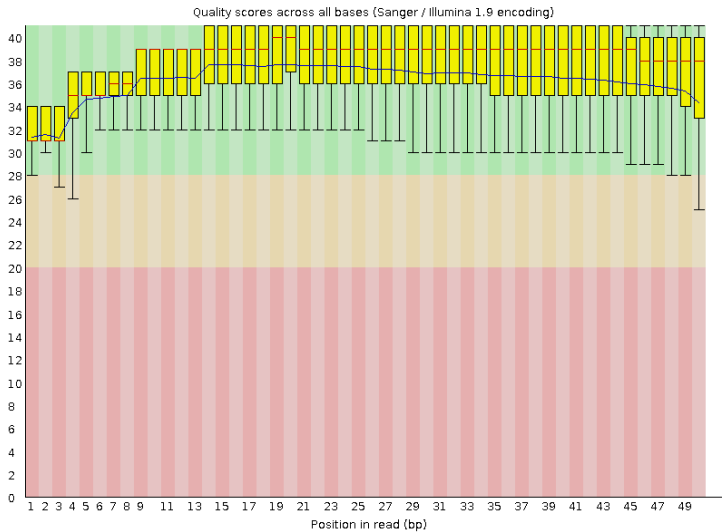
```
1 | # Output on terminal should look like:
2 | @HISEQ2500:386:C65KJACXX:6:1101:1547:2249 1:N:0:TGACCA
3 | CGGGTGTTTGAGAAAGAAGAGGAAGATGAGGATGAAGACGAGGAGGAGGA
4 | +
5 | ???D=ADDDHHHFDGGE>EGHGHHG;GCGGEFIEIIGFGGGGI@FHFGA;
```

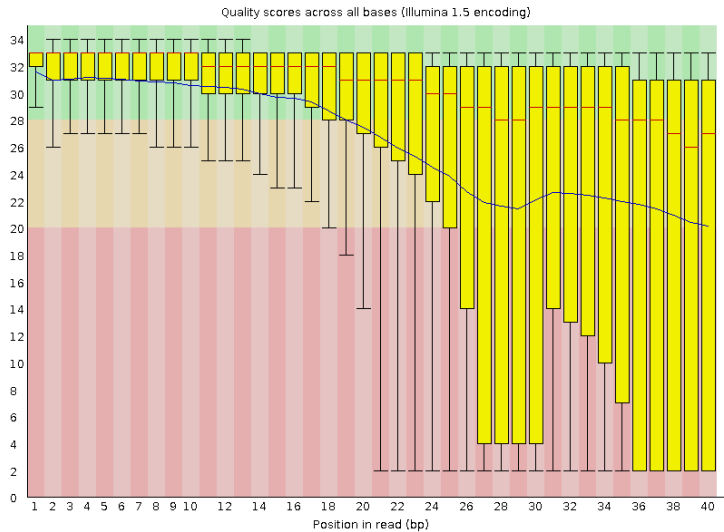
Phred quality score

- is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing
- a quality value Q is an integer mapping of P (i.e., the probability that the corresponding base call is incorrect).
- $Q = -10 \log_{10} P$
- $P = 10^{\frac{-Q}{10}}$
- e.g.: if Phred assigns a quality score Q of 30 to a base, the chances P that this base is called incorrectly are 1 in 1000 (0.001%) \implies base call accuracy is 99.9%

Phred quality score

Phred Quality Score	Probability of incorrect call	Base accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10 000	99.99%
50	1 in 100 000	99.999%
60	1 in 1 000 000	99.9999%





SAM/BAM

- **S**equence **A**lignment/**M**ap Format
- TAB-delimited text format
- header section (optional) and alignment section
- header lines start with '@', alignment lines not
- each alignment line has 11 mandatory fields (mapping position, ...) and variable number of optional fields for flexible aligner specific information

SAM/BAM – Example <https://samtools.github.io/hts-specs/SAMv1.pdf>

```
Coord      12345678901234 5678901234567890123456789012345
ref        AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1    TTAGATAAAGGATA*CTG
+r002      aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004      ATAGCT.....TCAGC
-r003      ttagctTAGGC
-r001/2    CAGCGGCAT
```

The corresponding SAM format is:¹

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

¹The values in the FLAG column correspond to bitwise flags as follows: 99 = 0x63: first/next is reverse-complemented/ properly aligned/multiple segments; 0: no flags set, thus a mapped single segment; 2064 = 0x810: supplementary/reverse-complemented; 147 = 0x93: last (second of a pair)/reverse-complemented/properly aligned/multiple segments.

SAM/BAM – Example <https://samtools.github.io/hts-specs/SAMv1.pdf>

- mandatory fields

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2 ¹⁶ -1]	bitwise FLAG
3	RNAME	String	* [!-()+-<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 ³¹ -1]	1-based leftmost mapping POSITION
5	MAPQ	Int	[0,2 ⁸ -1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	* = [!-()+-<>-~] [!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2 ³¹ -1]	Position of the mate/next read
9	TLEN	Int	[-2 ³¹ +1,2 ³¹ -1]	observed Template LENgth
10	SEQ	String	* [A-Za-z=.]+	segment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

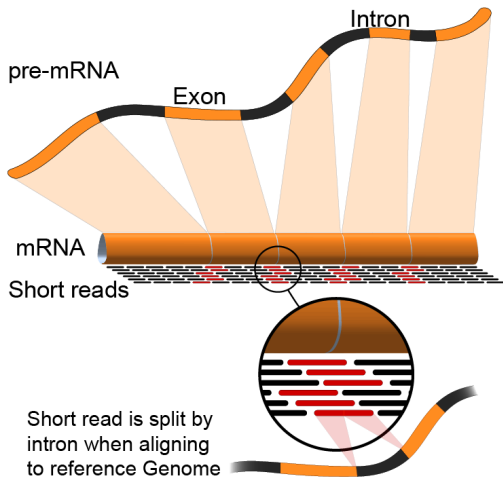
SAM/BAM – Example <https://samtools.github.io/hts-specs/SAMv1.pdf>

- **FLAGS**

Bit	Description
1	0x1 template having multiple segments in sequencing
2	0x2 each segment properly aligned according to the aligner
4	0x4 segment unmapped
8	0x8 next segment in the template unmapped
16	0x10 SEQ being reverse complemented
32	0x20 SEQ of the next segment in the template being reverse complemented
64	0x40 the first segment in the template
128	0x80 the last segment in the template
256	0x100 secondary alignment
512	0x200 not passing filters, such as platform/vendor quality controls
1024	0x400 PCR or optical duplicate
2048	0x800 supplementary alignment

<http://broadinstitute.github.io/picard/explain-flags.html>

Measure the amount of reads mapping to a certain feature



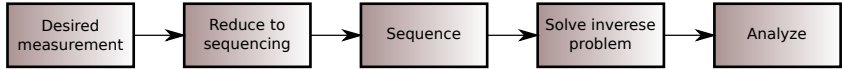
Measure the amount of reads mapping to a certain feature

“**Stories from the Supplement**”, Genome Informatics Meeting, Dr. Lior Pachter, UC Berkeley, 11/1/2013¹

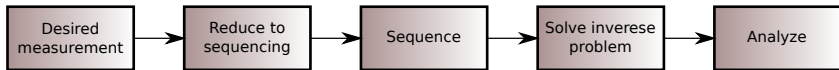


¹ https://www.youtube.com/watch?feature=player_detailpage&v=5NiFibnbE8o#t=1831

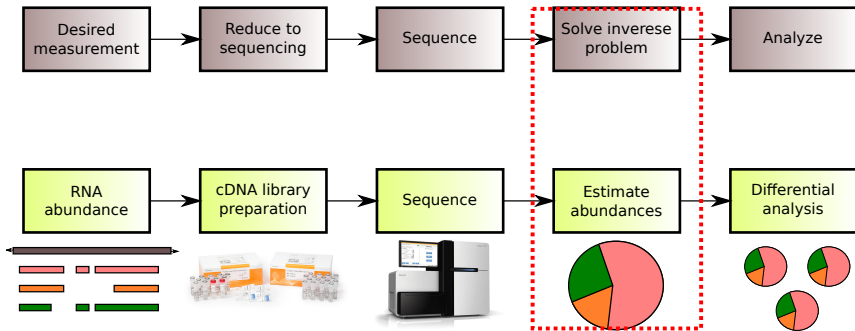
RNA-Seq assay



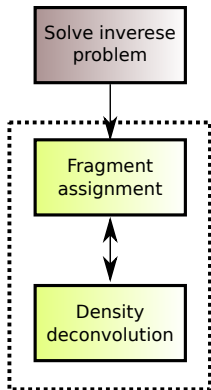
RNA-Seq assay



RNA-Seq assay

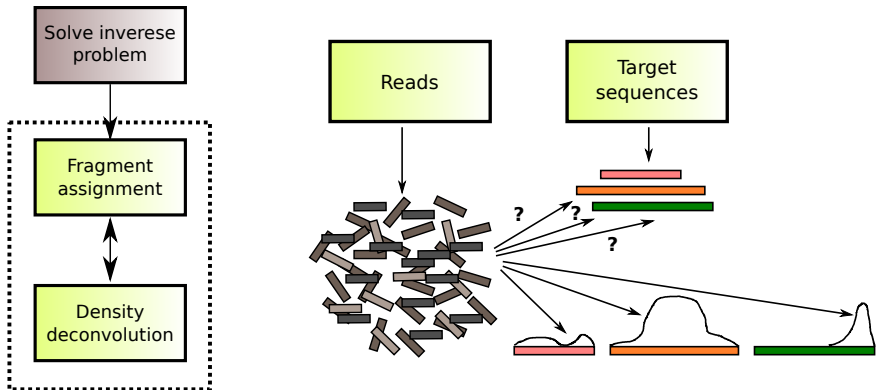


Analysis of the sequencing output



- **For every read**, find the target sequence and position in that target that it originated from
- After normalizing for non-uniform generation of reads from target sequences, **for every position in every target sequence** estimate the number of reads that originated there

Analysis of the sequencing output



The fragment assignment problem



target sequences



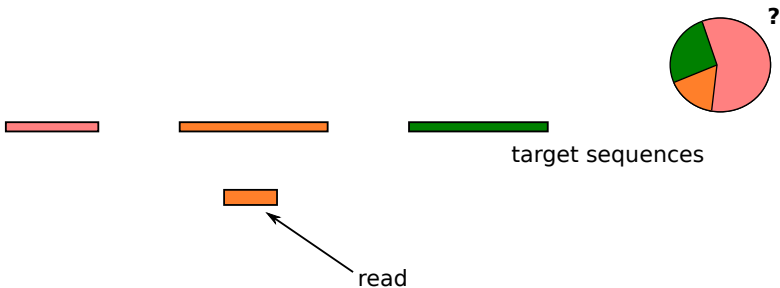
The fragment assignment problem



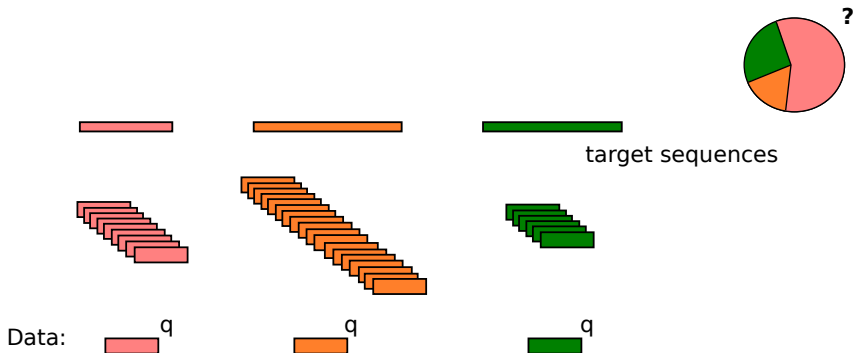
target sequences



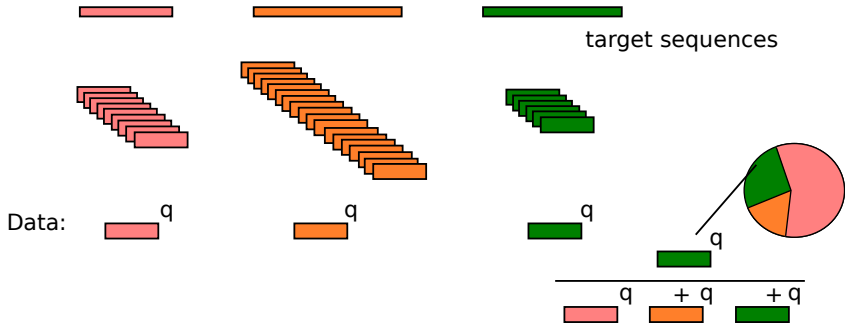
The fragment assignment problem



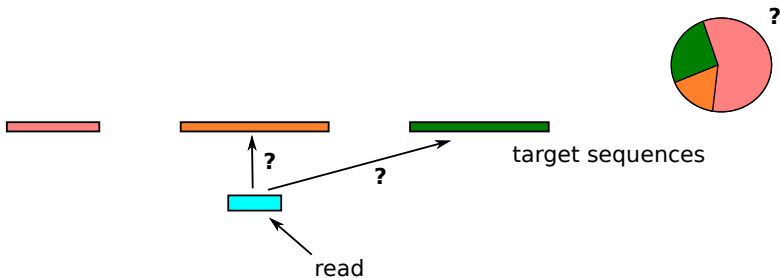
The fragment assignment problem



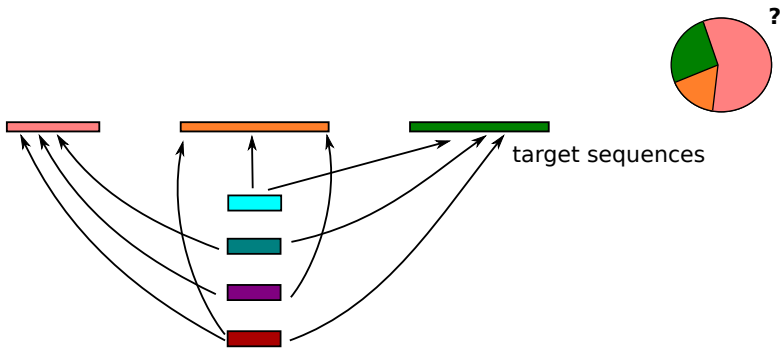
The fragment assignment problem



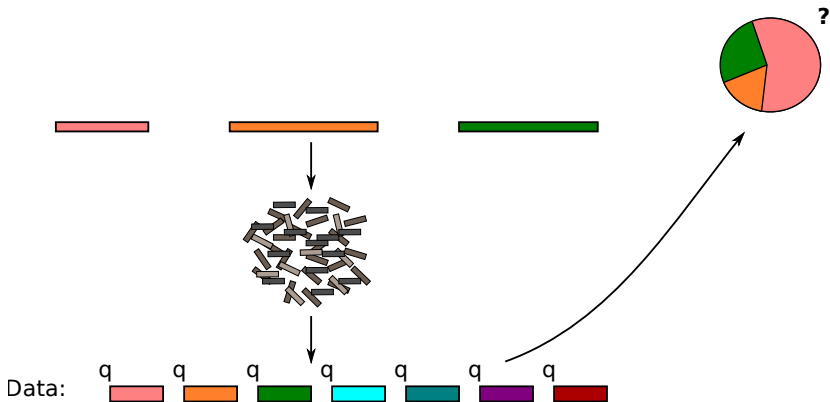
The fragment assignment problem



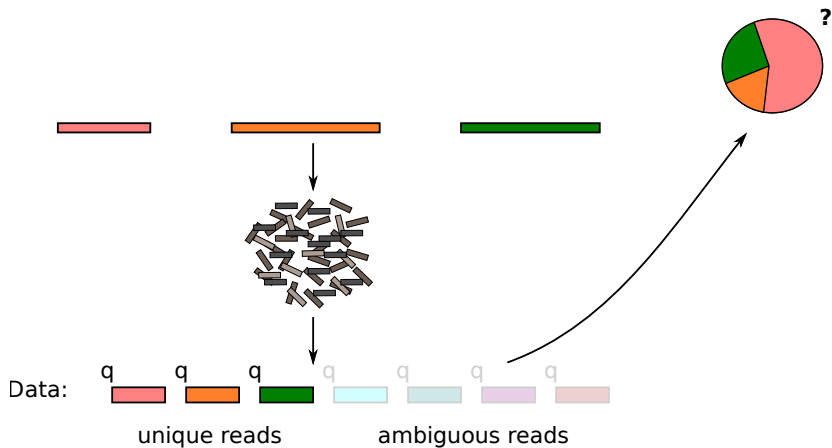
The fragment assignment problem



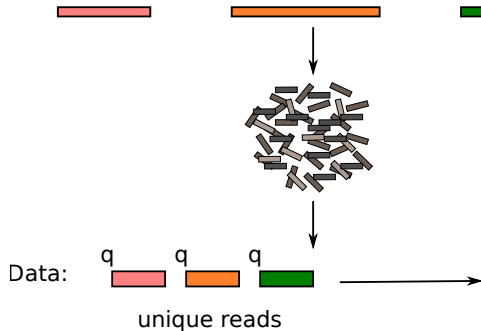
The fragment assignment problem



The fragment assignment problem



The fragment assignment problem



The read numbers here are random!
Repeat the experiment and get other numbers!
The pie chart is fuzzy!

RNA-Seq expression units

Units for quantification of mapped RNA-Seq reads.

- CPM
- RPKM
- FPKM
- TPM

RNA-Seq expression units

Units for quantification of mapped RNA-Seq reads.

- CPM
- RPKM
- FPKM
- TPM

NONE of these units are comparable across experiments. This is a result of RNA-Seq being a relative measurement, not an absolute one.

Preliminaries

- a **read** refers to both single-end or paired-end reads
 - concept of counting is the same with either type of read, as each read represents a fragment that was sequenced
- a **feature** refers to a genomic region containing a sequence that can normally appear in an RNA-Seq experiment
 - e.g. gene, isoform, exon
- we use X_i to denote the counts we observe from a feature of interest i
 - note: with alternative splicing you do not directly observe X_i , so often $\mathbb{E}[X_i]$ is used instead
 - (estimated using EM² algorithm by tools like eXpress, RSEM, Sailfish, Cufflinks, ...)

²expectation-maximization

Preliminaries

- a **read** refers to both single-end or paired-end reads
 - concept of counting is the same with either type of read, as each read represents a fragment that was sequenced
- a **feature** refers to a genomic region containing a sequence that can normally appear in an RNA-Seq experiment
 - e.g. gene, isoform, exon
- we use X_i to denote the **counts** we observe from a **feature** of interest i
 - note: with alternative splicing you do not directly observe X_i , so often $\mathbb{E}[X_i]$ is used instead
 - (estimated using EM² algorithm by tools like eXpress, RSEM, Sailfish, Cufflinks, ...)

²expectation-maximization

Counts

- “Counts” usually refer to the number of reads that align to a particular feature (X_i)
- these numbers are heavily dependent on two things:
 - (1) the amount of fragments sequenced (this is related to relative abundances)
 - (2) the length of the feature, or more appropriately, the effective length
- the effective length refers to the number of possible start sites a feature could have generated a fragment of that particular length:

$$\tilde{l}_i = l_i - \mu_{FLD} + 1$$

- where μ_{FLD} is the mean of the fragment length distribution

Since counts are NOT scaled by the length of the feature, all units in this category are not comparable within a sample without adjusting for the feature length.

Counts are often used by differential expression methods since they are naturally represented by a counting model, such as a negative binomial

Effective counts

$$effCounts_i = X_i \cdot \frac{l_i}{\bar{l}_i}$$

- idea: if the effective length is much shorter than the actual length, then in an experiment with no bias you would expect to see more counts
- the effective counts are scaling the observed counts up, e.g.

$$\mu_{FLD} = 200:$$

$$effCounts_{i_1} = 100 \cdot \frac{2000}{1801} = 111.05$$

$$effCounts_{i_2} = 100 \cdot \frac{300}{101} = 297.03$$

Counts per million (CPM)

- CPM mapped reads are counts scaled by the number of fragments we sequenced N times one million

$$CPM_i = \frac{X_i}{\frac{N}{10^6}}$$

Counts per million (CPM), within sample normalization

- CPM mapped reads are counts scaled by the number of fragments we sequenced N times one million
- as noted in the counts section, the number of fragments you see from a feature depends on its length

$$CPM_i = \frac{X_i}{\frac{N}{10^6}}$$

Counts per million (CPM), within sample normalization

- CPM mapped reads are counts scaled by the number of fragments we sequenced N times one million
- as noted in the counts section, the number of fragments you see from a feature depends on its length
- therefore: to compare features of different length you should normalize counts by the length of the feature.

$$CPM_i = \frac{X_i}{\frac{N}{10^6}}$$

Counts per million (CPM), within sample normalization

- CPM mapped reads are counts scaled by the number of fragments we sequenced N times one million
- as noted in the counts section, the number of fragments you see from a feature depends on its length
- therefore: to compare features of different length you should normalize counts by the length of the feature.

$$CPM_i = \frac{X_i}{\frac{N}{10^6}}$$

- CPM is related to RPKM/FPKM without length normalization and a factor of 10^3

RPKM

- Reads per **k**ilobase per **m**illion mapped reads (RPKM)
- RPKM is NOT a method, it is simply a unit of expression
- RPKM takes the same rate we will discuss in the TPM section and instead of dividing it by the sum of rates, divides it by the total number of reads sequenced (N) and multiplies by a big number (10^9):

$$RPKM_i = \frac{X_i}{\left(\frac{\tilde{l}_i}{10^3}\right) \left(\frac{N}{10^6}\right)} = \frac{X_i}{\tilde{l}_i N} \cdot 10^9$$

RPKM/FPKM

- **F**ragments per **k**ilobase per **m**illion mapped reads (**FPKM**)
- **FPKM** is NOT a method, it is simply a unit of expression
- **FPKM** takes the same rate we will discuss in the TPM section and instead of dividing it by the sum of rates, divides it by the total number of reads sequenced (N) and multiplies by a big number (10^9):

$$FPKM_i = \frac{X_i}{\left(\frac{\tilde{l}_i}{10^3}\right) \left(\frac{N}{10^6}\right)} = \frac{X_i}{\tilde{l}_i N} \cdot 10^9$$

Which N to use? “The total number”?

- number of sequenced reads
- number of mapped reads
- number of assigned reads



Which N to use? “The total number”?

- number of sequenced reads
- number of mapped reads
- **number of assigned reads**

- within a sample everything will be normalized by the same N
- between samples it is probably more important to just stay consistent

The problem with FPKM

- FPKM values are *experiment specific*! FPKM values are not comparable between experiments!
- go one step back and use a *universal* proportionality constant like 10^6
- **please use TPM** and not FPKM like proposed by the Cufflinks paper

“In the case of RPKM, the problem originates from the fact that there is no biological interpretation of the *denominator* [N], the total number of reads. It is a variable that characterizes a particular sequencing run, but does not correspond in any direct way to a biological variable, ...”

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch. “Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.” *Theory in Biosciences* 131.4 (2012): 281-285.

⇒ Nice explanation why RPKM is inconsistent among samples: <http://blog.nextgenetics.net/?e=51>

Transcripts per million (TPM)

$$TPM_i = \frac{X_i}{\tilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) \cdot 10^6$$

Transcripts per million (TPM)

$$TPM_i = \frac{X_i}{\tilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) \cdot 10^6$$

Instead of normalizing the reads to library size by total tag count division (**N**), we should divide by the sum of all length normalized tag counts

Transcripts per million (TPM)

$$TPM_i = \frac{X_i}{\tilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) \cdot 10^6$$

- taking the length into consideration: $\frac{X_i}{\tilde{l}_i}$ (counts per base)
- this number is also dependent on the total number of fragments sequenced
- to adjust this, simply divide by the sum of all rates and this gives the proportion of transcripts i in the sample
- scale by one million because the proportion is often very small

Transcripts per million (TPM)

$$TPM_i = \frac{X_i}{\tilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) \cdot 10^6$$

- the interpretation is that if you were to sequence one million full length transcripts, TPM is the number of transcripts you would have seen of type i , given the abundances of the other transcripts in your sample

Relationship between TPM and FPKM

$$TPM_i = \left(\frac{FPKM_i}{\sum_j FPKM_j} \right) \cdot 10^6$$

R example of CPM, TPM, FPKM



www.rna.uni-jena.de/teaching.php

Comparative analysis of RNA-Seq with DESeq2

Love et al. *Genome Biology* (2014) 15:550
DOI 10.1186/s13059-014-0550-8



METHOD

Open Access

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love^{1,2,3}, Wolfgang Huber² and Simon Anders^{2*}

Abstract

In comparative high-throughput sequencing assays, a fundamental task is the analysis of count data, such as read counts per gene in RNA-seq, for evidence of systematic changes across experimental conditions. Small replicate numbers, discreteness, large dynamic range and the presence of outliers require a suitable statistical approach. We present *DESeq2*, a method for differential analysis of count data, using shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates. This enables a more quantitative analysis focused on the strength rather than the mere presence of differential expression. The *DESeq2* package is available at <http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html>.

Two applications of RNA-Seq data

Discovery

- find new transcripts
- find transcript boundaries
- find splice junctions

Comparison

- given samples from different experimental conditions, find effects of the treatment on
 - gene expression strengths
 - isoform abundance ratios, splice patterns, transcript boundaries, ...

Sequencing count data

ID	shSCR_iPSc_1	shSCR_iPSc_2	shSCR_iPSc_3	shSCR_noniPSc_1	shSCR_noniPSc_2	shSCR_noniPSc_3
ENSMUSG000000000001	6000	5754	6116	4865	4615	6246
ENSMUSG000000000003	0	0	0	0	0	0
ENSMUSG000000000028	3282	3026	3147	761	710	837
ENSMUSG000000000031	24789	21466	23701	38593	40969	49883
ENSMUSG000000000037	868	881	844	206	192	260
ENSMUSG000000000049	23	17	26	6	1	4
ENSMUSG000000000056	1875	1729	1832	1304	1333	1685
ENSMUSG000000000058	155	155	144	1476	1315	1918
ENSMUSG000000000078	2635	2527	2609	4506	4495	5701

Sequencing count data

ID	shSCR_iPSc_1	shSCR_iPSc_2	shSCR_iPSc_3	shSCR_noniPSc_1	shSCR_noniPSc_2	shSCR_noniPSc_3
ENSMUSG000000000001	6000	5754	6116	4865	4615	6246
ENSMUSG000000000003	0	0	0	0	0	0
ENSMUSG000000000028	3282	3026	3147	761	710	837
ENSMUSG000000000031	24789	21466	23701	38593	40969	49883
ENSMUSG000000000037	868	881	844	206	192	260
ENSMUSG000000000049	23	17	26	6	1	4
ENSMUSG000000000056	1875	1729	1832	1304	1333	1685
ENSMUSG000000000058	155	155	144	1476	1315	1918
ENSMUSG000000000078	2635	2527	2609	4506	4495	5701

- count reads, not base-pairs
- count each read at most once
- discard a read if
 - it cannot be uniquely mapped
 - its alignment overlaps with several genes
 - the alignment score is bad
 - the mates (paired-end reads) do not map to the same gene

Sequencing count data

ID	shSCR_1PSc_1	shSCR_1PSc_2	shSCR_1PSc_3	shSCR_noniPSc_1	shSCR_noniPSc_2	shSCR_noniPSc_3
ENSMUSG000000000001	6000	5754	6116	4865	4615	6246
ENSMUSG000000000003	0	0	0	0	0	0
ENSMUSG000000000028	3282	3026	3147	761	710	837
ENSMUSG000000000031	24789	21466	23701	38593	40969	49883
ENSMUSG000000000037	868	881	844	206	192	260
ENSMUSG000000000049	23	17	26	6	1	4
ENSMUSG000000000056	1875	1729	1832	1304	1333	1685
ENSMUSG000000000058	155	155	144	1476	1315	1918
ENSMUSG000000000078	2635	2527	2609	4506	4495	5701

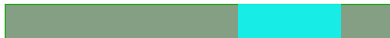
We need a good annotation for counting!

Why discard non-unique alignments?

transcript A



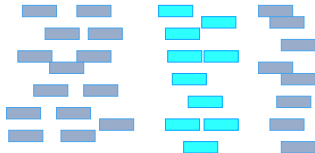
transcript B



control condition



treatment condition



What to do with these counts?

What to do with these counts?

- in principal: we want to calculate a **fold change** between the expression level of two genes and how significant this number is among replicates

$$\#_reads_gene_A_1 = 3267; \#_reads_gene_A_2 = 866$$

What to do with these counts?

- in principal: we want to calculate a **fold change** between the expression level of two genes and how significant this number is among replicates
- a fold change is simply a ratio between two numbers

$$\frac{\#_reads_gene_A_1}{\#_reads_gene_A_2} = \frac{3267}{866} = 3.7725$$

What to do with these counts?

- in principal: we want to calculate a **fold change** between the expression level of two genes and how significant this number is among replicates
- a fold change is simply a ratio between two numbers
- in general, the 2-based-logarithmic fold change is used for differential gene expression

$$\frac{\#_reads_gene_A_1}{\#_reads_gene_A_2} = \log_2\left(\frac{3267}{866}\right) = 1.9155$$

What to do with these counts?

- in principal: we want to calculate a **fold change** between the expression level of two genes and how significant this number is among replicates
- a fold change is simply a ratio between two numbers
- in general, the 2-based-logarithmic fold change is used for differential gene expression
- another advantage: directly see up- and down-regulated genes

$$\frac{\#_reads_gene_X_1}{\#_reads_gene_X_2} = \frac{13}{231} = 0.05627$$

What to do with these counts?

- in principal: we want to calculate a **fold change** between the expression level of two genes and how significant this number is among replicates
- a fold change is simply a ratio between two numbers
- in general, the 2-based-logarithmic fold change is used for differential gene expression
- another advantage: directly see up- and down-regulated genes

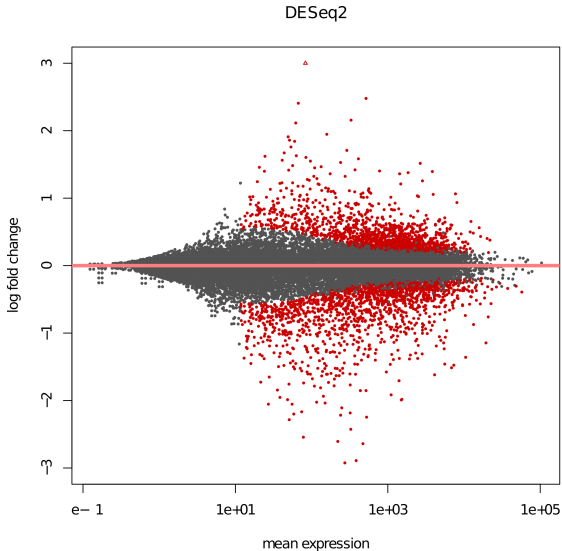
$$\frac{\#_reads_gene_X_1}{\#_reads_gene_X_2} = \log_2\left(\frac{13}{231}\right) = -4.1513$$

Down- and up-regulated genes

$$\frac{\#_reads_gene_X_1}{\#_reads_gene_X_2} = \log_2\left(\frac{13}{231}\right) = -4.1513$$

$$\frac{\#_reads_gene_Z_1}{\#_reads_gene_Z_2} = \log_2\left(\frac{201}{17}\right) = 3.5635$$

Significant differential expressed genes (DEGs)



Normalization for library size

- if sample A has been sequenced deeper than sample B, we expect counts to be higher for A
- Naive Approach: divide by the total number of reads per sample
- Problem: genes that are strongly and differentially expressed may distort the ratio of total reads

Poisson distribution

- if length of the genome is N , then the probability of the event that a single read position starts at a single position in the genome is $\frac{1}{N}$ (very small)

Poisson distribution

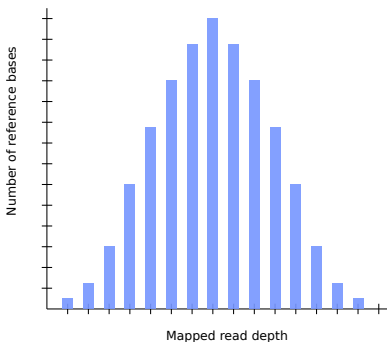
- if length of the genome is N , then the probability of the event that a single read position starts at a single position in the genome is $\frac{1}{N}$ (very small)
- if the number of reads is K , the total number of read positions that start at a single genome position is the number of times that an event with probability $\frac{1}{N}$ happens out of K trials

Poisson distribution

- if length of the genome is N , then the probability of the event that a single read position starts at a single position in the genome is $\frac{1}{N}$ (very small)
- if the number of reads is K , the total number of read positions that start at a single genome position is the number of times that an event with probability $\frac{1}{N}$ happens out of K trials

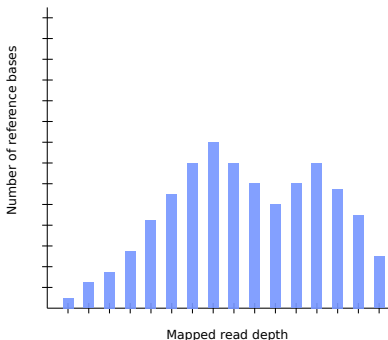
Take the genome and choose a location at random to produce a read. This is a Poisson process.

Poisson distribution



If you plot the depth of sequence along this theoretical genome, it will be a poisson distribution.

Poisson distribution



Data from real genomes aren't exactly poisson, due to biases related to the composition of the genome, chemistry of sequencing, assembly errors in the reference genome, and inability to map to repetitive regions

RNA-Seq and Poisson distribution

- same is true for RNA-Seq, only instead of a genome, we are choosing a read from the transcriptome

RNA-Seq and Poisson distribution

- same is true for RNA-Seq, only instead of a genome, we are choosing a read from the transcriptome
- more complex: some transcripts are present at higher abundance than others

RNA-Seq and Poisson distribution

- same is true for RNA-Seq, only instead of a genome, we are choosing a read from the transcriptome
- more complex: some transcripts are present at higher abundance than others
- but at the core: still poisson process

RNA-Seq and Poisson distribution

- same is true for RNA-Seq, only instead of a genome, we are choosing a read from the transcriptome
- more complex: some transcripts are present at higher abundance than others
- but at the core: still poisson process
- BUT: in the case of RNA-Seq and DEGs we model the distribution of counts for a given gene **across** biological replicates

RNA-Seq and Poisson distribution

- same is true for RNA-Seq, only instead of a genome, we are choosing a read from the transcriptome
- more complex: some transcripts are present at higher abundance than others
- but at the core: still poisson process

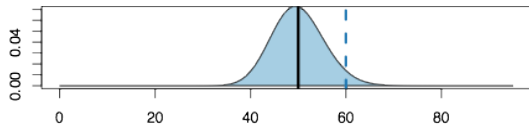
- BUT: in the case of RNA-Seq and DEGs we model the distribution of counts for a given gene **across** biological replicates
- in real biological applications, there is simply more variability than Poisson can explain

RNA-Seq and Poisson distribution

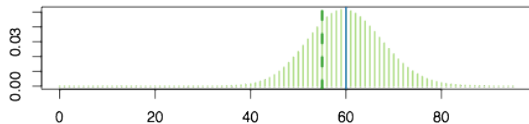
- same is true for RNA-Seq, only instead of a genome, we are choosing a read from the transcriptome
- more complex: some transcripts are present at higher abundance than others
- but at the core: still poisson process

- BUT: in the case of RNA-Seq and DEGs we model the distribution of counts for a given gene **across** biological replicates
- in real biological applications, there is simply more variability than Poisson can explain
- more fitting: Negative Binomial Distribution

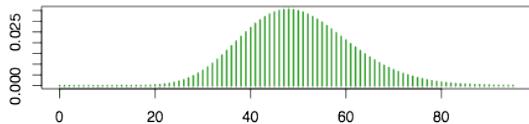
Negative Binomial Distribution



Biological sample with mean μ and variance v



Poisson distribution with mean q and variance q .



Negative binomial with mean μ and variance $q+v$.

Negative Binomial Distribution

- (1) the observed data at gene level is inherently counts or estimated counts of fragments for each feature
- (2) the spread of values among biological replicates is more than given by a simpler, one parameter distribution, the Poisson; and it seems to be captured by the NB sufficiently well

How does DESeq2 normalization work?

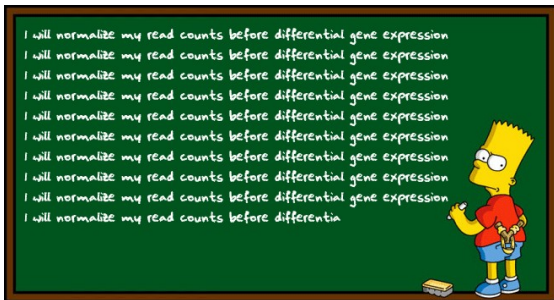
- DESeq2 calculates a pseudo-reference sample to normalize each single sample against
- uses geometric mean:

$$\sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdot X_4 \cdot \dots \cdot X_n}$$

How does DESeq2 normalization work?

- DESeq2 calculates a pseudo-reference sample to normalize each single sample against
- uses geometric mean:

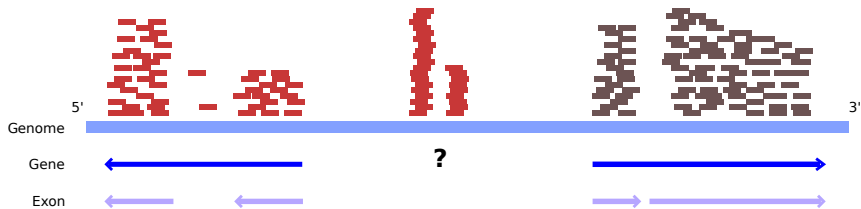
$$\sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdot X_4 \cdot \dots \cdot X_n}$$



De novo transcriptome assembly/quantification



De novo transcriptome assembly/quantification



De novo transcriptome assembly/quantification

nature
biotechnology

LETTERS

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell¹⁻³, Brian A Williams⁴, Geo Pertea², Ali Mortazavi⁴, Gordon Kwan⁴, Marijke J van Baren⁵, Steven L Salzberg^{1,2}, Barbara J Wold⁴ & Lior Pachter^{3,6,7}

High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation¹⁻³. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over

(75 bp in this work versus 25 bp in our previous work) and pairs of reads from both ends of each RNA fragment can reduce uncertainty in assigning reads to alternative splice variants¹². To produce useful transcript-level abundance estimates from paired-end RNA-Seq data, we developed a new algorithm that can identify complete novel transcripts and probabilistically assign reads to isoforms.

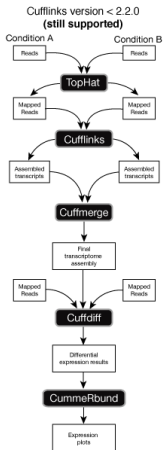
For our initial demonstration of Cufflinks, we performed a time course of paired-end 75-bp RNA-Seq on a well-studied model of

.. All rights reserved.

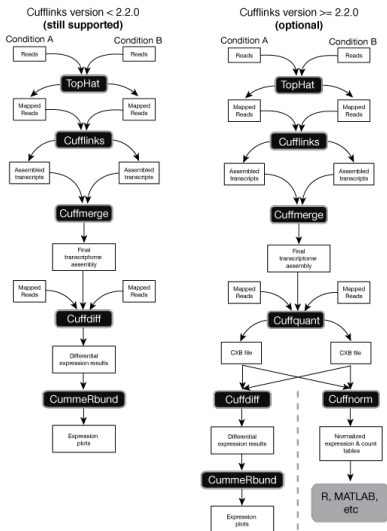
Cufflinks

- Transcriptome assembly and differential expression analysis for RNA-Seq
- *Cufflinks* is both the name of a suite of tools and a program within that suite. *Cufflinks* the program **assembles transcriptomes** from RNA-Seq data and **quantifies their expression**

Cufflinks RNA-Seq workflow



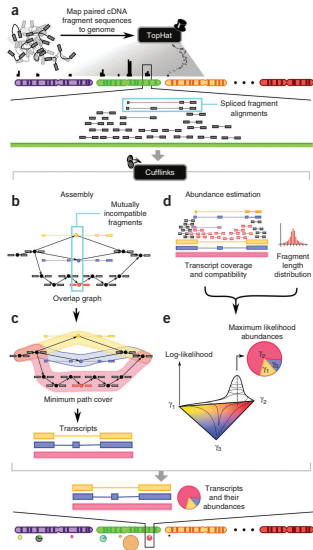
Cufflinks RNA-Seq workflow



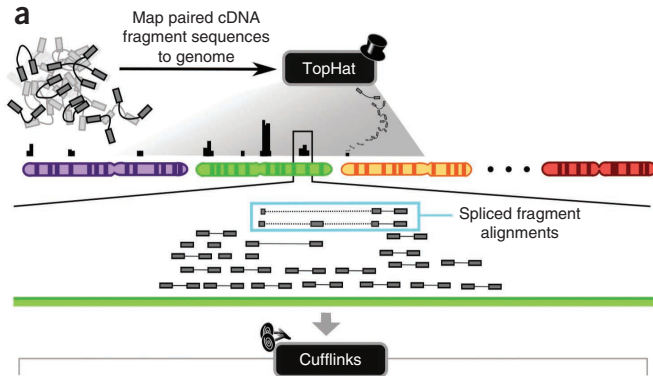
Cufflinks de novo and with annotation

- Cufflinks can be used with a supplied reference annotation in GFF format to guide the assembly
 - `-g/--GTF-guide <reference_annotation.(gtf/gff)>`
- Output will include all reference transcripts as well as any novel genes and isoforms that are assembled

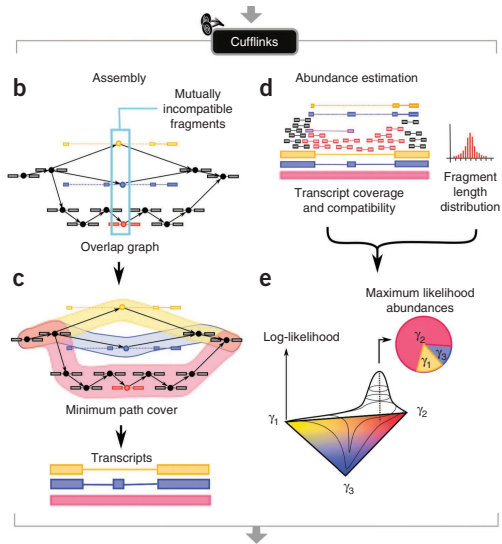
Overview of Cufflinks



Overview of Cufflinks



Overview of Cufflinks



Overview of Cufflinks

