Introduction to Differential Expression Analysis with Next generation Sequencing

Lieven Clement

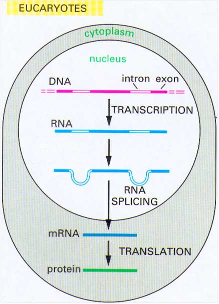
Transcriptomics

Outline

1. Intro
   - Technology
   - Data
   - II.1. Normalization
2. II.2. Statistical Model
   - Poisson
   - GLM
   - Normalization
   - Overdispersion
3. Statistical inference

Introduction

Central Dogma:



mRNA concentration ↑↓

↓

mRNA templates ↑↓

↓

protein production ↑↓

↓

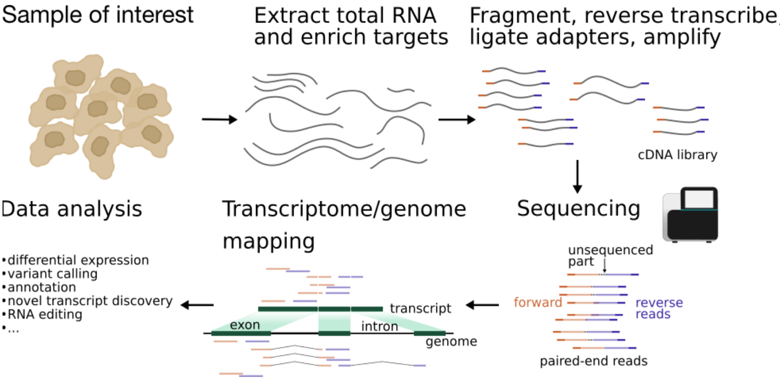
protein concentration ↑↓

↓

protein activity ↑↓

↓

phenotype changes



Sample of interest

Extract total RNA and enrich targets

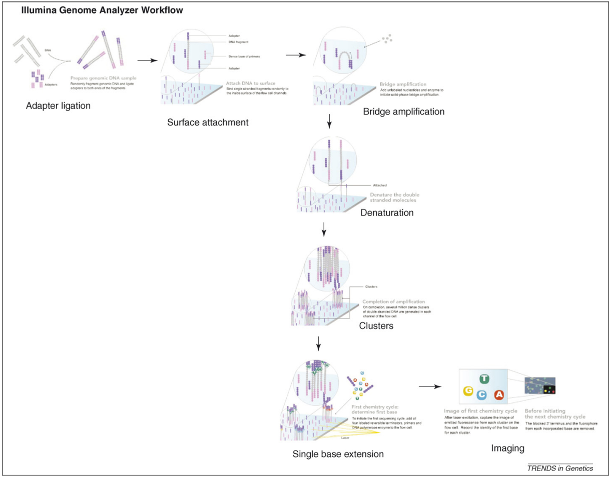
Fragment, reverse transcribe, ligate adapters, amplify

Sequencing

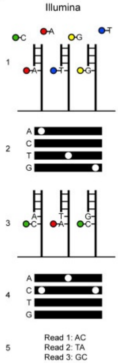
Transcriptome/genome mapping

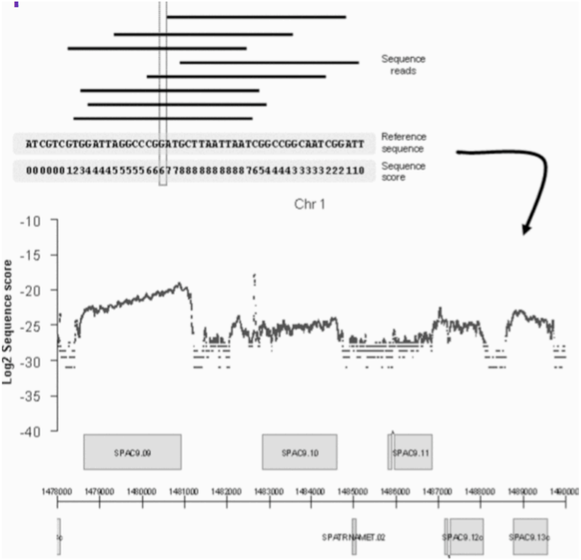
Data analysis

- differential expression
- variant calling
- annotation
- novel transcript discovery
- RNA editing
- ...


Illumina Genome Analyzer Workflow
DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments
Adapter ligation
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single stranded fragments randomly to the inside surface of the flow cell channels.
Surface attachment
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Bridge amplification
Attached
Denature the double stranded molecules
Denaturation
Clusters
Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.
Clusters
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
Laser
Single base extension
T
G
C
A
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.
Imaging

TRENDS in Genetics




Sequence reads
ATCGTCGTGGATTAGGCCCGGATGCTTAATTAATCGGCCGGCAATCGGATT
Reference sequence
000000123444455555666778888888888876544443333322110
Sequence score
Chr 1
Log2 Sequence score
-10
-15
-20
-25
-30
-35
-40
SPAC9.09
SPAC9.10
SPAC9.11
1478000
1479000
1480000
1481000
1482000
1483000
1484000
1485000
1486000
1487000
1488000
1489000
1490000
SPATRNAMET.02
SPAC9.12c
SPAC9.13c

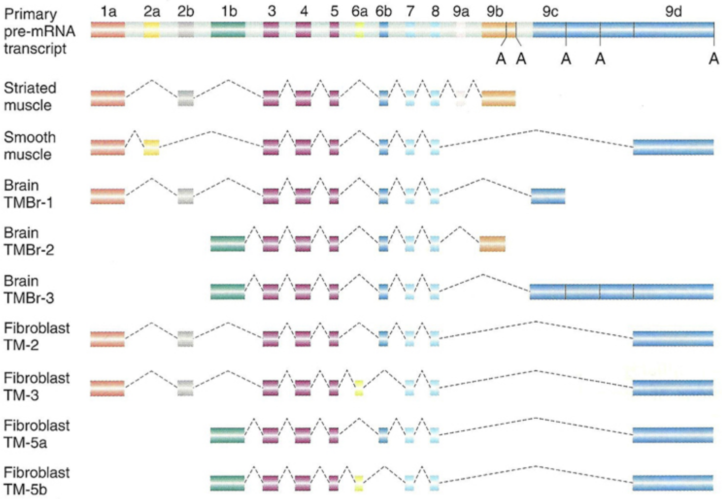
	Potential Problems
read counts	
↑	mapping, lane, flow cell, run bias
cDNA library	
↑	RNA extraction, rRNA, DNA conversion,...
mRNA levels	
↓	post transcriptional regulation, translation speed
protein levels	
↓	post translational regulation, modification, activity regulation...
phenotype	

	Potential Problems
read counts	
↑	mapping, lane, flow cell, run bias
cDNA library	
↑	RNA extraction, rRNA, DNA conversion,...
mRNA levels	
↓	post transcriptional regulation, translation speed
protein levels	
↓	post translational regulation, modification, activity regulation...
phenotype	

- Number of reads depends on many factors
- expression level, total number of reads per library, transcript length etc.
- Here we focus on differences at gene level: transcripts have the same length.
- Systematic differences in read counts → systematic differences in mRNA levels

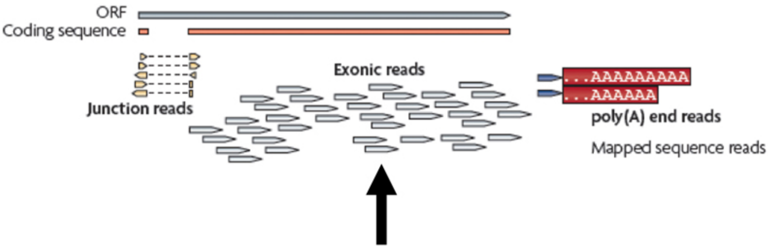
Alternative Splicing

Alternative splicing in tropomyosin



Alternative Splicing

Data from RNA-seq

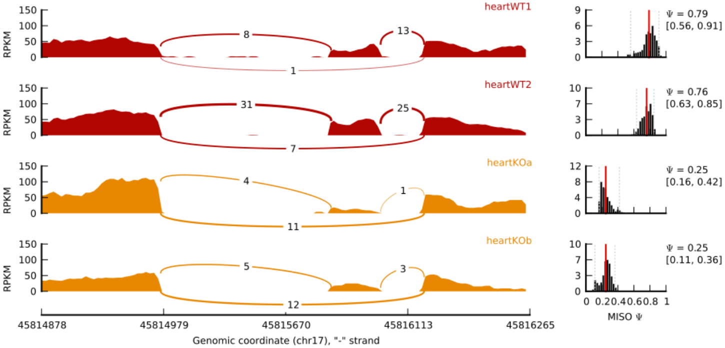


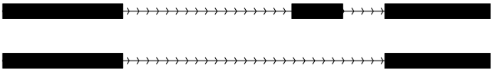
Determine mRNA abundance

Infer relative level of expression

Alternative Splicing

chr17:45816186:45816265:-@chr17:45815912:45815950:-@chr17:45814875:45814965:-





Single end vs paired end

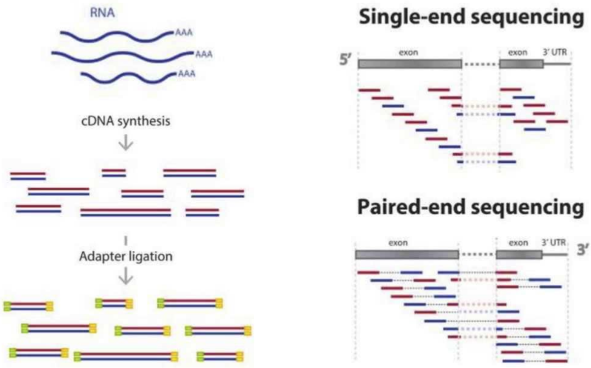
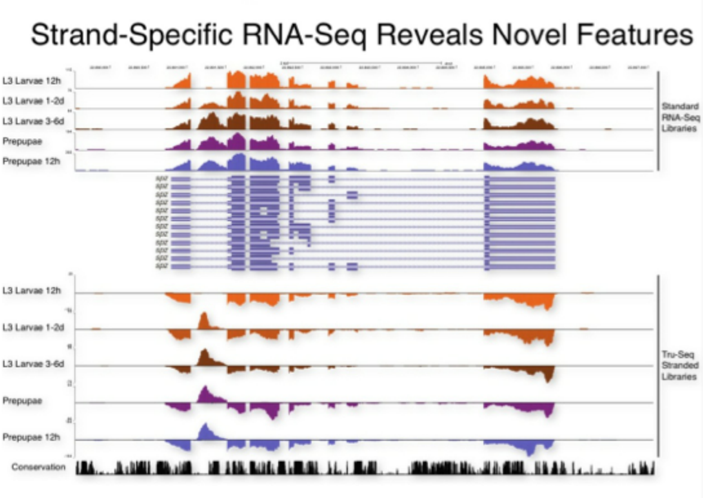


Image adapted from Zhernakova, et al., PLoS Genet. 2013 June; 9(6): e1003594.

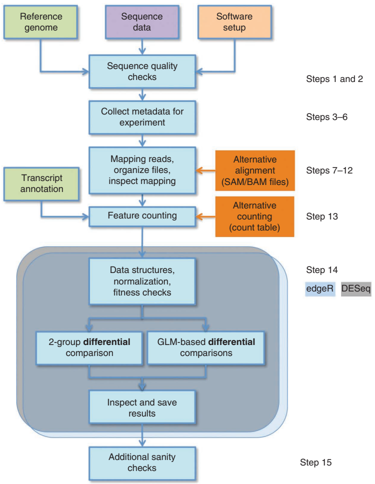
Naive vs Strand specific



Steps in a RNA-seq Experiment

1. Experimental design
2. Experiment
3. Sampling
4. Library Prep
5. Sequencing/basecalling
6. Quality assessment of Reads
7. Read alignment to reference genome
8. Quality assessment of alignment
9. Summarization: read counts per feature (gene, exon, ...)
10. Gene Prioritization: Analysis of differential expression
11. Downstream Analysis

RNA-seq Data Analysis work flow



Reference genome
Sequence data
Software setup
Sequence quality checks
Steps 1 and 2
Collect metadata for experiment
Steps 3–6
Mapping reads, organize files, inspect mapping
Alternative alignment (SAM/BAM files)
Steps 7–12
Transcript annotation
Feature counting
Alternative counting (count table)
Step 13
Data structures, normalization, fitness checks
Step 14
edgeR
DESeq
2-group **differential** comparison
GLM-based **differential** comparisons
Inspect and save results
Additional sanity checks
Step 15

Anders et al. (2013) Nature Protocols.

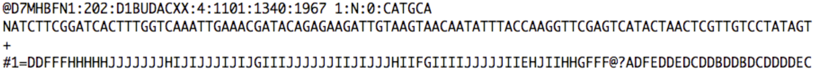
Part I. Basecalling, Alignment and Summarization

I.1. Basecalling

Most researchers use standard base caller:
Illumina → Cassava → fastq files
`http://en.wikipedia.org/wiki/FASTQ_format`

FASTQ format

- Combines sequence and base quality information
- Four lines per sequence (read)
- ID line (starting with @) sequence
- another ID line (starting with +) base qualities
- For paired-end sequencing: one file for "first" reads and one for "second" reads of the read pairs



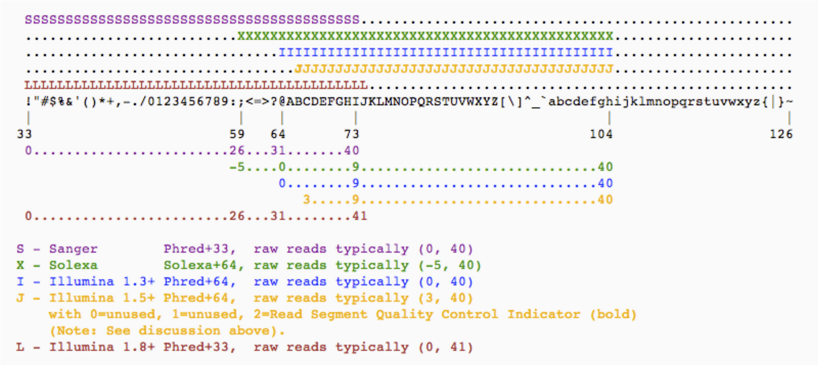
```
@D7MHBFN1:202:D1BUDACXX:4:1101:1340:1967 1:N:0:CATGCA
NATCTTCGGATCACTTTGGTCAAATTGAAACGATACAGAGAAGATTGTAAGTAACAATATTTACCAAGGTTCGAGTCATACTAACTCGTTGTCCTATAGT
+
#1=DDFFFHHHHHJJJJJJHIJIJJJIJIJGIIIJJJJJJIIJIJJJHIIFGIIIIJJJJJIIEHJIIHHGFFF@?ADFEDDEDCDDBDDBDCDDDDEC
```

- Line 1
  - D7MHBFNI: unique instrument name
  - 202: run ID
  - D1BUDACXX: flowcell ID
  - 4: flowcell lane
  - 1101: tile number within lane
  - 1340: x-coordinate of cluster within tile
  - 1967: y-coordinate of cluster within tile
  - 1: member of pair (1 or 2). Older versions: /1 and /2
  - Y/N: did quality control of read failed (Y: bad)
  - 0: none of the control bits are on
  - CATGCA: index sequence (barcode)
- Line 2: read sequence
- Line 4: Base quality

FASTQ format - base qualities

Quality score: $Q = -10 \log_{10} p$

Phred Score	Probability on incorrect base call	base call accuracy
10	1/10	90%
20	1/100	99%
30	1/1000	99.9%
40	1/10000	99.99%
50	1/100000	99.999%



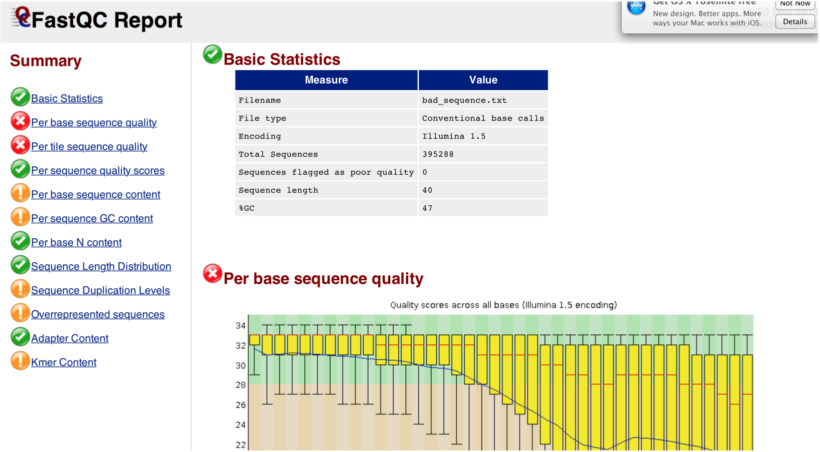
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
  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........................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)
```

https://en.wikipedia.org/wiki/FASTQ_format

Raw base quality QC

fastQC: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`



Summary

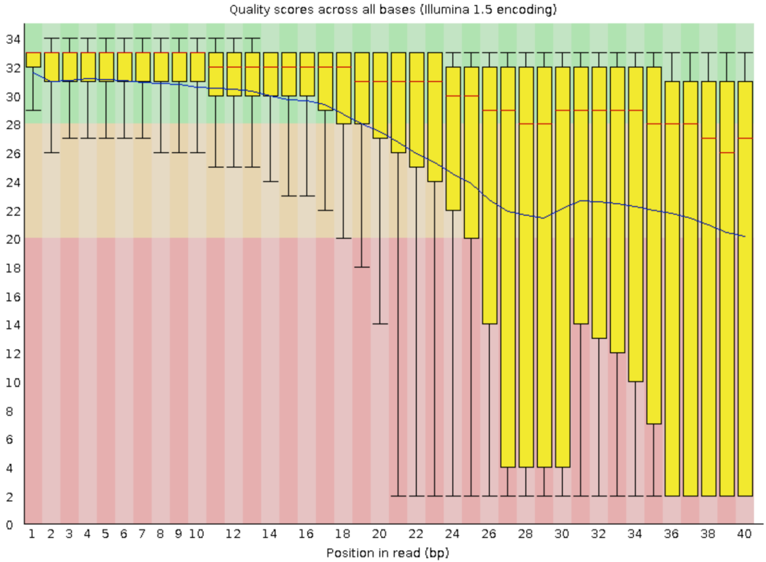
- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✓ Basic Statistics

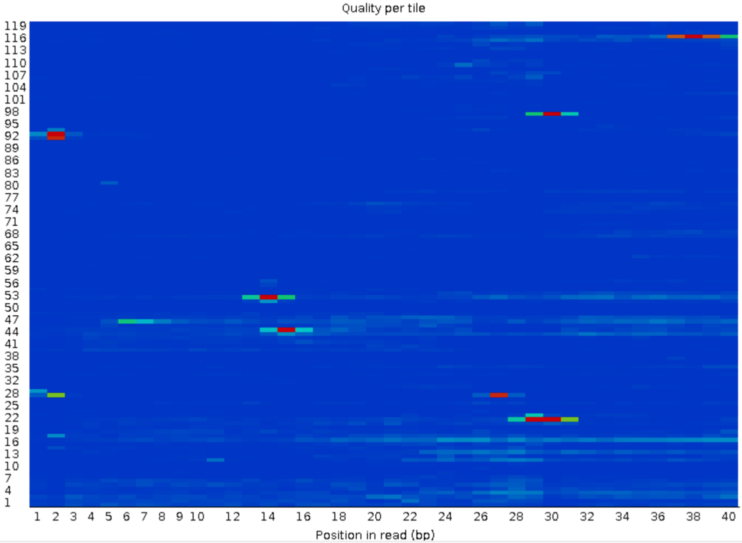
Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

✗ Per base sequence quality

Raw base quality QC

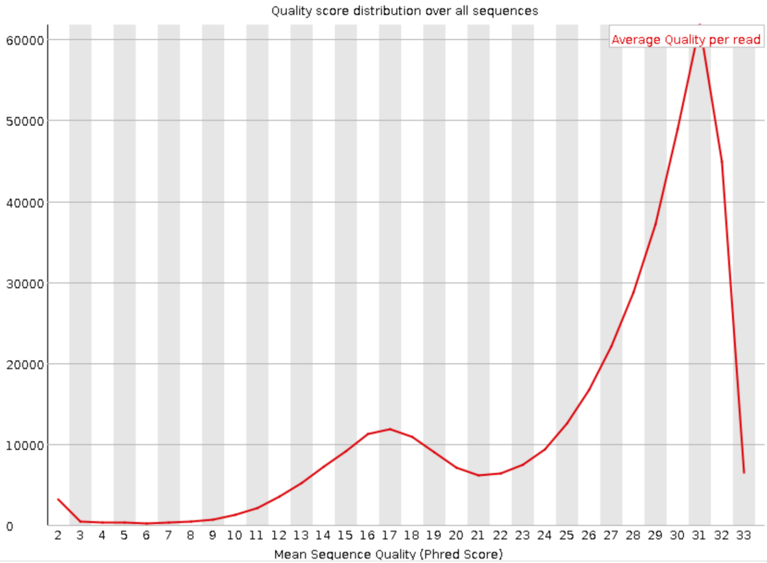


Raw base quality QC



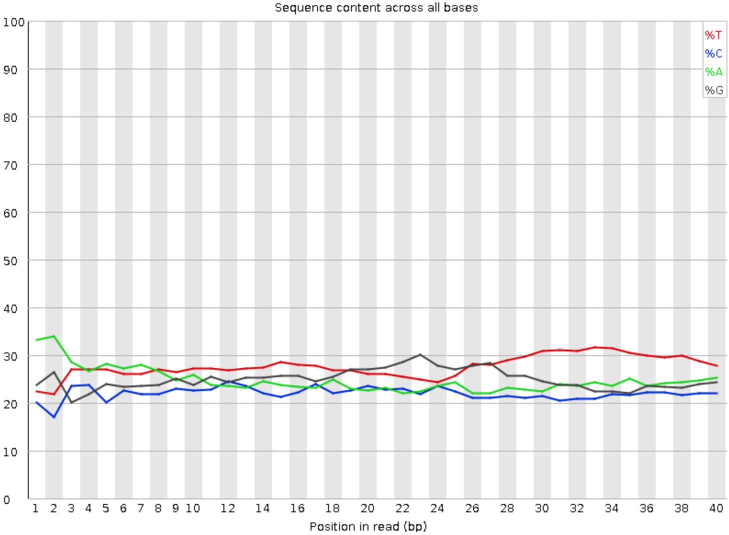
Raw base quality QC

✓ Per sequence quality scores



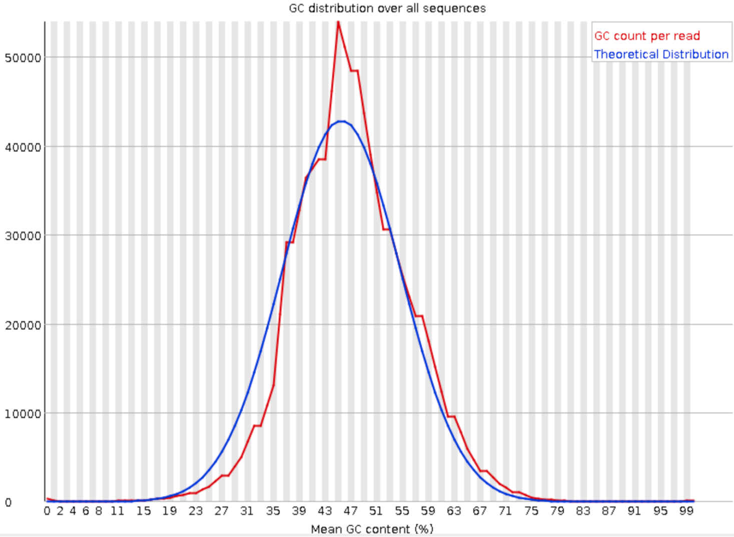
Raw base quality QC

Per base sequence content



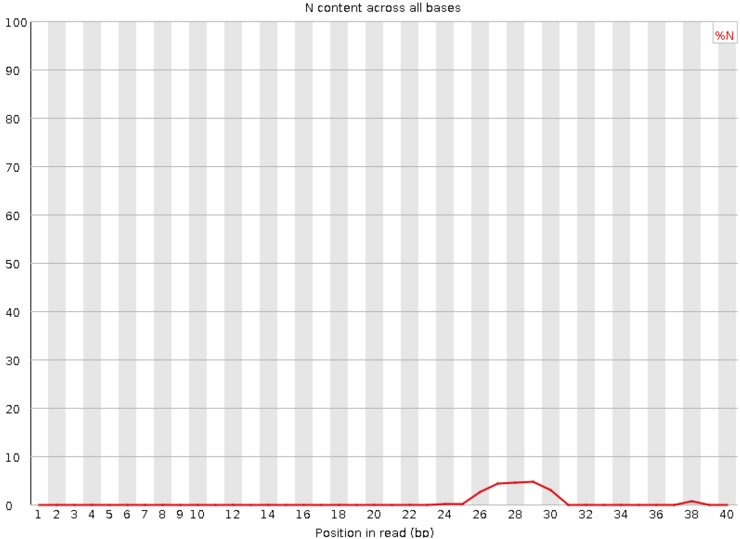
Raw base quality QC

⚠ **Per sequence GC content**



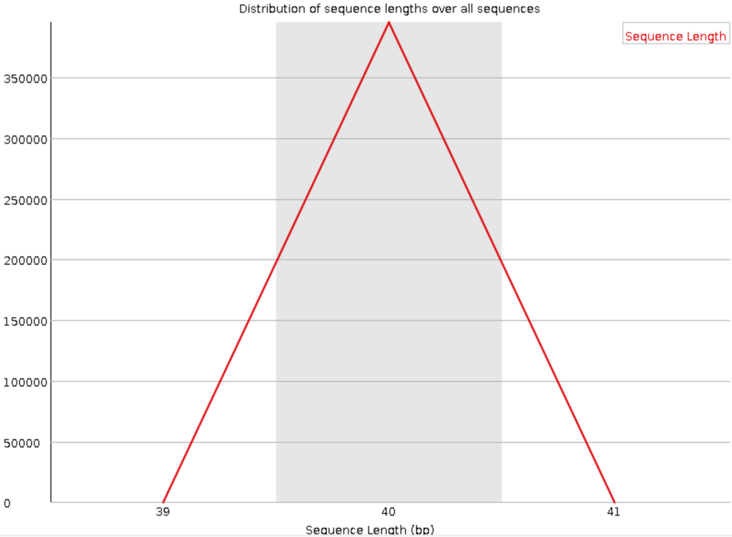
Raw base quality QC

✓ **Per base N content**



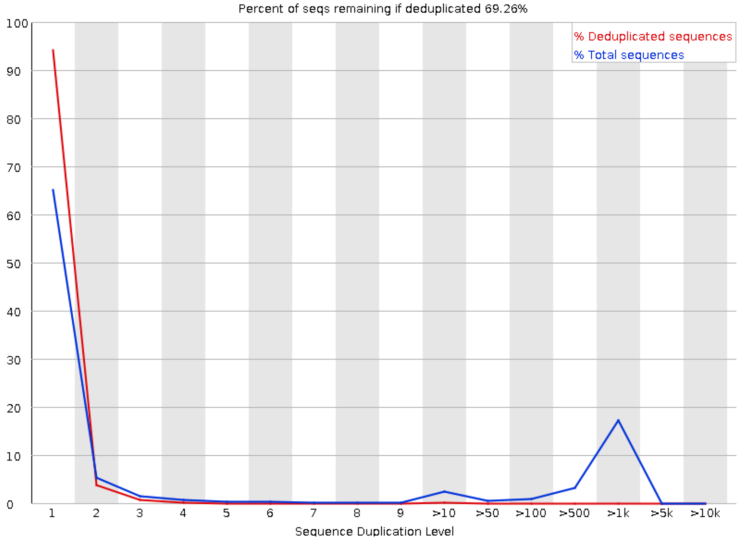
Raw base quality QC

✓ Sequence Length Distribution

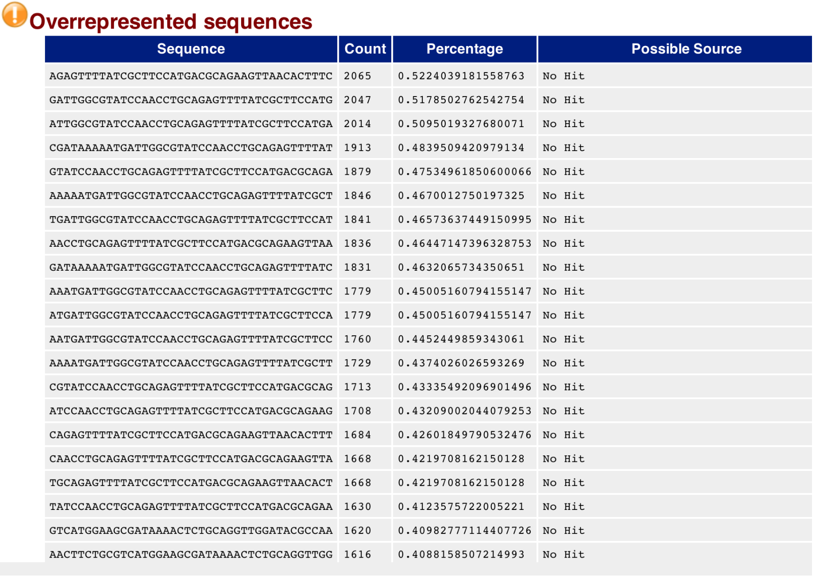


Raw base quality QC

Sequence Duplication Levels



Raw base quality QC

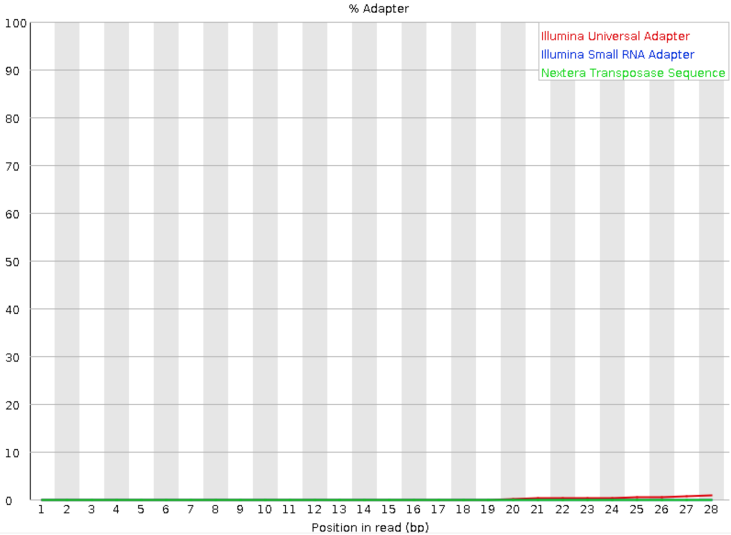


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.517850276254275 4	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.43209002044079253	No Hit
CAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTT	1684	0.42601849790532476	No Hit
CAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTA	1668	0.4219708162150128	No Hit
TGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT	1668	0.4219708162150128	No Hit
TATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAA	1630	0.4123575722005221	No Hit
GTCATGGAAGCGATAAAACTCTGCAGGTTGGATACGCCAA	1620	0.40982777114407726	No Hit
AACTTCTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGG	1616	0.4088158507214993	No Hit

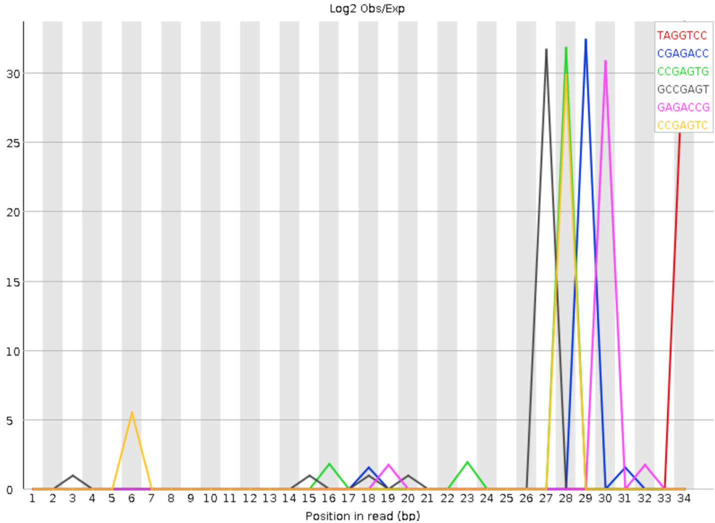
Raw base quality QC

✓ Adapter Content



Raw base quality QC

Kmer Content



Preprocessing

- Read Trimming
  - Adaptor sequence
  - Bar code
  - (deteriorating bases at the end of reads)
  - often already done by the sequencing provider.
  - remaining polyA tails
- Read filtering
  - low quality reads
  - PhiX reads (should be removed already by sequence provider)
  - in RNA-seq never remove duplicates because they can occur for highly expressed transcripts
- Perform fastQC again

I.2 Alignment

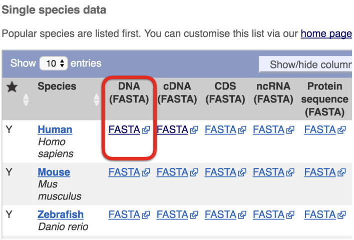
- DNA
  - bowtie2, BWA,...
  - Needs: genomic reference sequence + cleaned reads
- RNAseq
  - Aligning to transcriptome: annotation-bias, you throw away some data: very fast: Salmon and Kallisto.
  - Genome: problem Gaps
    Star, tophat2, Rsubread, ...
  - Needs: genomic reference sequence + genomic annotation+ cleaned reads

`http://wwwdev.ebi.ac.uk/fg/hts_mappers/`

The human reference genome

www.ensembl.org/info/data/ftp/index.html

https://www.gencodegenes.org/human/



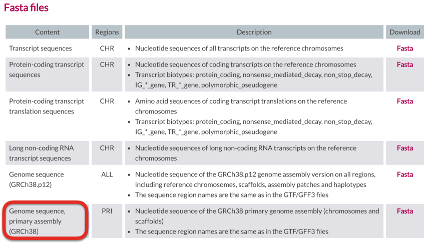
Single species data

Popular species are listed first. You can customise this list via our home page

Show 10 entries

Show/hide column

★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)
Y	**Human** *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA
Y	**Mouse** *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA
Y	**Zebrafish** *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA



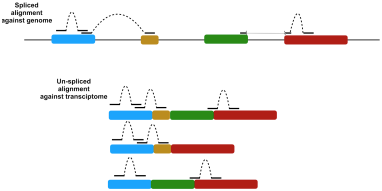
Fasta files

Content	Regions	Description	Download
Transcript sequences	CHR	• Nucleotide sequences of all transcripts on the reference chromosomes	Fasta
Protein-coding transcript sequences	CHR	• Nucleotide sequences of coding transcripts on the reference chromosomes • Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene	Fasta
Protein-coding transcript translation sequences	CHR	• Amino acid sequences of coding transcript translations on the reference chromosomes • Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene	Fasta
Long non-coding RNA transcript sequences	CHR	• Nucleotide sequences of long non-coding RNA transcripts on the reference chromosomes	Fasta
Genome sequence (GRCh38.p12)	ALL	• Nucleotide sequence of the GRCh38.p12 genome assembly version on all regions, including reference chromosomes, scaffolds, assembly patches and haplotypes • The sequence region names are the same as in the GTF/GFF3 files	Fasta
Genome sequence, primary assembly (GRCh38)	PRI	• Nucleotide sequence of the GRCh38 primary genome assembly (chromosomes and scaffolds) • The sequence region names are the same as in the GTF/GFF3 files	Fasta

Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz 840 MB

slide courtesy Charlotte Soneson

To the genome: gap aware!



STAR: Spliced Transcripts Alignment to a Reference

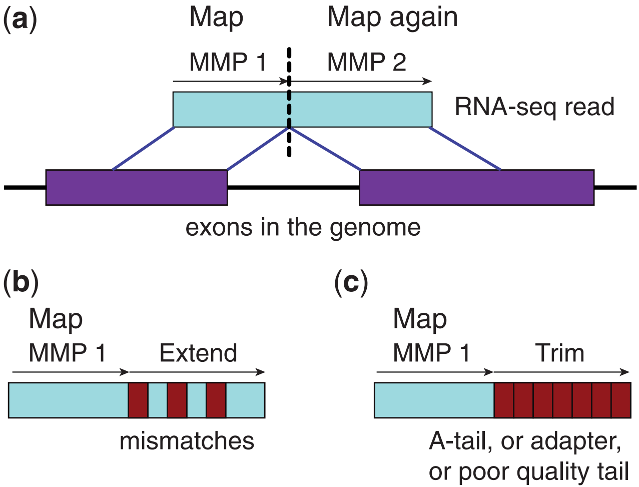
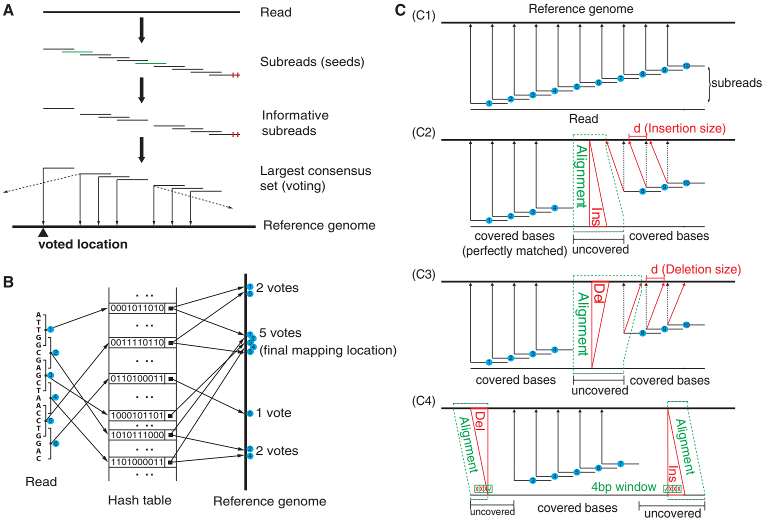


Fig. 1. Schematic representation of the Maximum Mappable Prefix search in the STAR algorithm for detecting (**a**) splice junctions, (**b**) mismatches and (**c**) tails

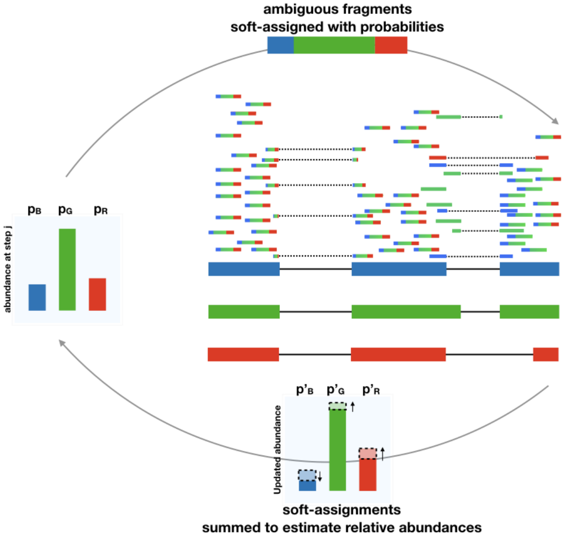
Dobin et al. (2013) Bioinformatics 29(1), 15-21

Rsubread: integration read alignment into R



Liao et al. (2013) Nucleic Acids Research, 41(10):e108, 2013

Salmon: fast and bias-aware quantification of transcript expression (Mapping to the transcriptome)



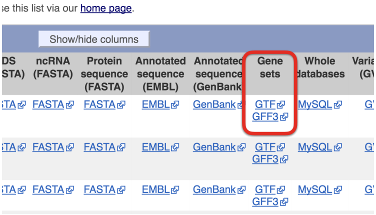
Post alignment QC

- fastQC
- Coverage plots
- Removal of biological contamination if not of interest mRNA (only small fraction of RNA pool):
  rRNA, ncRNA, mitochondrial RNA

→ Normally removed with kits prior to sequencing

1.3. Summarization

www.ensembl.org/info/data/ftp/index.html

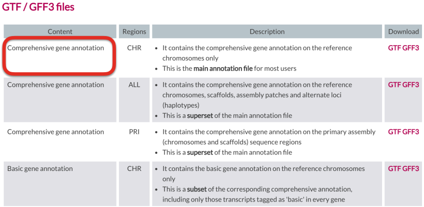


se this list via our home page.

Show/hide columns

DS STA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank	Gene sets	Whole databases	Varia (GV
TA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	G
TA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	G
TA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	G

https://www.gencodegenes.org/human/



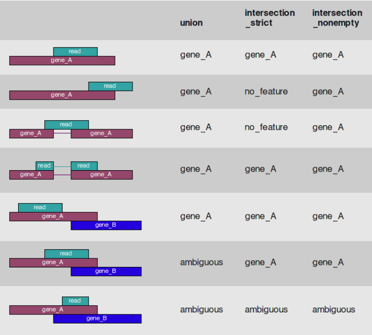
GTF / GFF3 files

Content	Regions	Description	Download
Comprehensive gene annotation	CHR	• It contains the comprehensive gene annotation on the reference chromosomes only • This is the **main annotation file** for most users	GTF GFF3
Comprehensive gene annotation	ALL	• It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a **superset** of the main annotation file	GTF GFF3
Comprehensive gene annotation	PRI	• It contains the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions • This is a **superset** of the main annotation file	GTF GFF3
Basic gene annotation	CHR	• It contains the basic gene annotation on the reference chromosomes only • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3

slide courtesy Charlotte Soneson

I.3a Summarization upon mapping to the genome

- Most applications summarize reads based upon known annotation: bias
- Generate counts for genes, transcript or exons
- Count read instead of nt
- Count each read only ones
- Discard reads that
  - do not map uniquely
  - overlap with several genes
  - with a bad quality score



	union	intersection _strict	intersection _nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

Figure 4.1: Overlap modes; Image from the HTSeq package developed by Simon Anders.

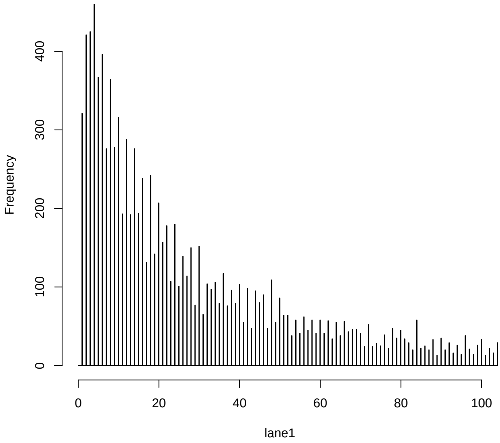
I.3b Summarization upon mapping to the transcriptome

- Results in counts at the transcript level.
- Sum of transcript level counts to obtain gene-level count
- Account for potential difference in transcript usage between samples : via average transcript length (see normalisation, why?)
- Has been shown to be more accurate: e.g. Soneson et al. (2015). F1000Research, 4. doi: 10.12688/f1000research.7563.1

Part II: Analysis of count data

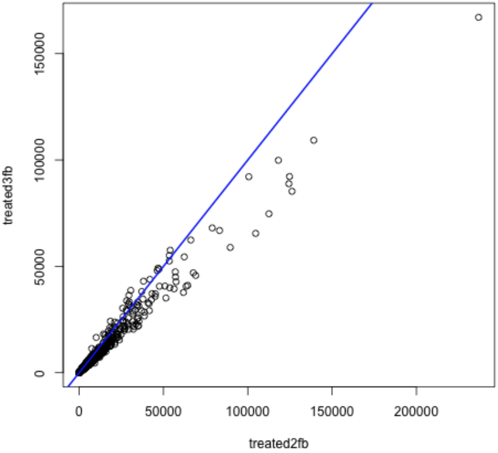
Classical Approach: Gene level data

- gene x sample matrix
- Differential expression well studied by statisticians
- Count data:
  - many zeroes
  - very large range
  - biological variability?

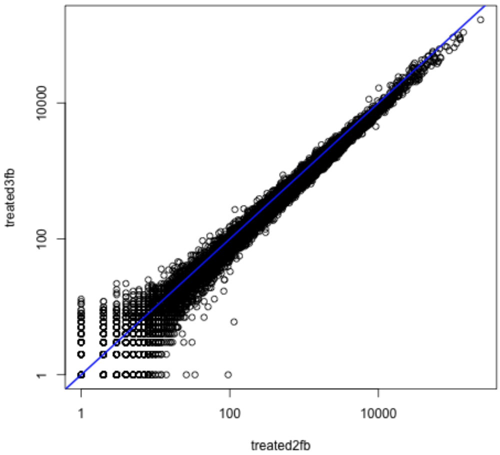


- Discrete data
- Skewed distribution

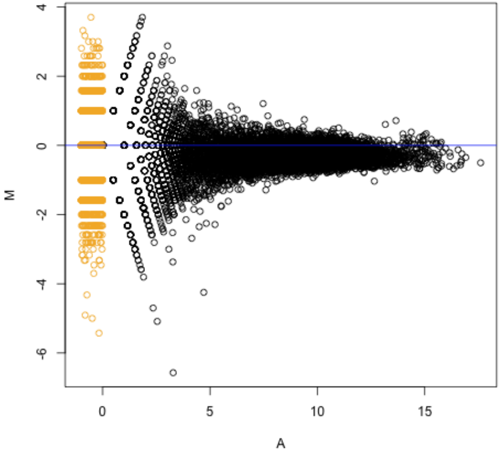
Normalization



Normalization



Normalization



- $M = \log_2(Y_2) - \log_2(Y_1)$
- $A = \frac{[\log_2(Y_2)+\log_2(Y_2)]}{2}$

Normalization

- Sequencing depth

	group	lib.size	norm.factors
treated2fb	treated	15620018.00	1.00
treated3fb	treated	12733865.00	1.00
untreated3fb	untreated	10283129.00	1.00
untreated4fb	untreated	11653031.00	1.00

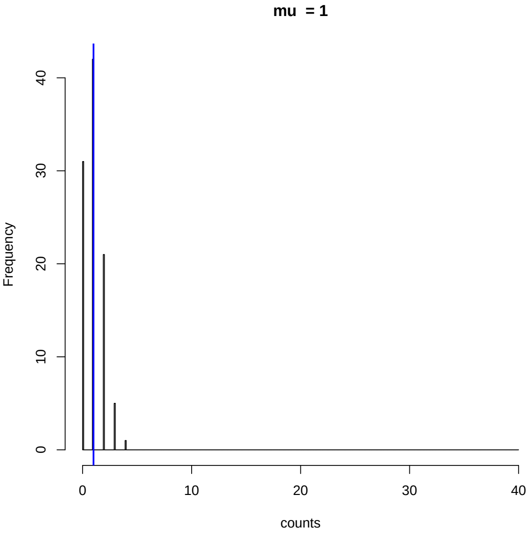
- use lib.size to normalize?

→ Convert reads in counts per million

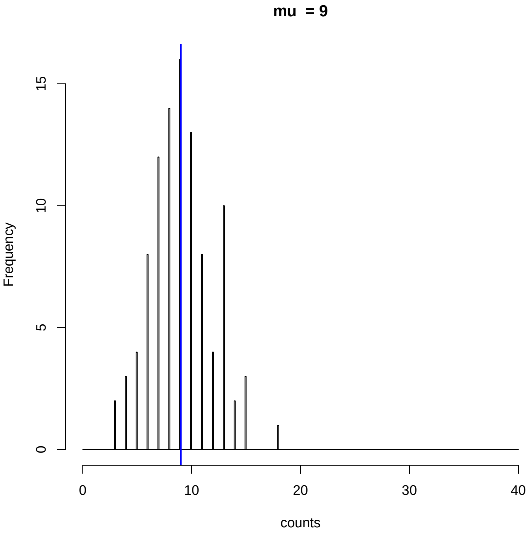
Modeling Counts

- Marioni (2008) Genome Research showed that technical replicates are *Poisson*(μ)
- Properties: μ=mean=variance

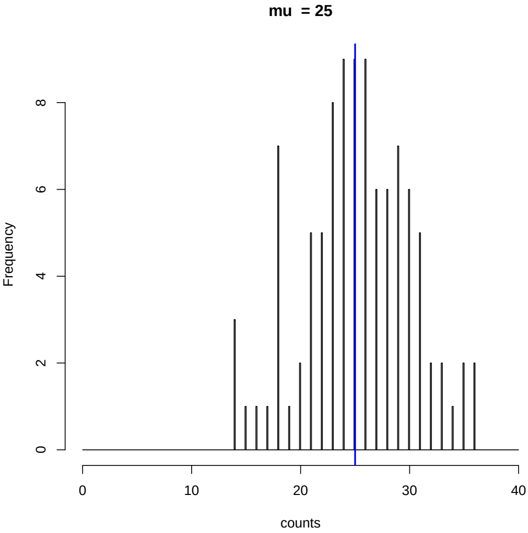
Modeling Counts



Modeling Counts



Modeling Counts



Modeling Counts

- The Poisson distribution is commonly used $Poisson(\mu)$
- Properties: μ=mean=variance
- Relative error decreases with increasing mean
- CV=standard deviation/mean=$\sqrt{\mu}/\mu = 1/\sqrt{\mu}$

Mean	CV
1	1
9	1/3
25	1/5
100	1/10

Generalized linear model for seq data

$$\begin{cases} y_{ig} & \sim & \text{Poisson}(\mu_{ig}) \\ \log(\mu_{ig}) & = & \eta_{ig} \\ \eta_{ig} & = & \sum_{k=1}^{N} x_{ik}\beta_{gk} \end{cases}$$

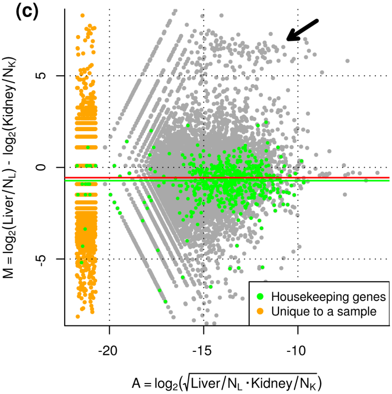
- y_{gi}:count for gene g of subject i
- x_{ik}: predictor variabele k evaluated for subject i
- η: linear predictor
- β_{gk}: effect for predictor variable k and gene g

GLM with normalization

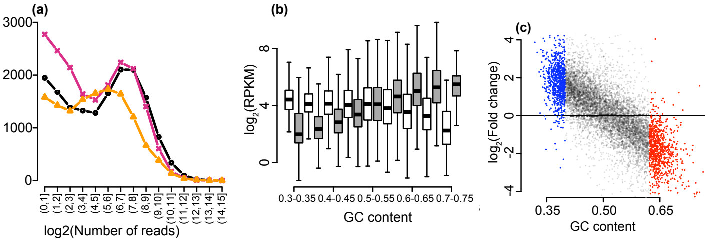
$$\begin{cases} y_{ig} & \sim & \text{Poisson}(\mu_{ig}) \\ \mu_{ig} & = & \lambda_{ig} S_{ig} \\ \log(\mu_{ig}) & = & \eta_{ig} \\ \eta_{ig} & = & \sum_{k=1}^{N} x_{ik}\beta_{gk} + \log S_{ig} \end{cases}$$

- y_{gi}:count for gene g of subject i
- x_{ik}: predictor variabele k evaluated for subject i
- η: linear predictor
- β_{gk}: effect for predictor variable k and gene g
- S_{ig}: effective library size for gene g of subject i

Normalization with lib.size??



Robinson and Oshlack (2010). Genome Biology.

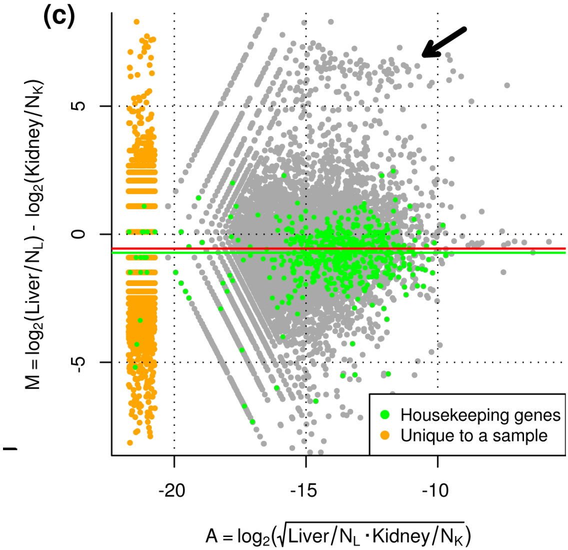


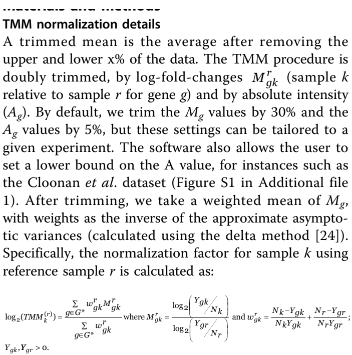
Hansen, Irizarry and Wu (2012). Biostatistics.

Normalization: S_{ij}

- Upper quartile Bullard et al. (2010) BMC Bioinformatics.
- scale normalization: edgeR package, Robinson and Oshlack (2010) Genome biology
- Geometric mean: DESeq, Anders and Huber (2010) Genome Biology
- Gene specific normalization: S_{gij}
  - GC content
  - gene length
  - cqn package, Hansen, Irizarry and Wu (2012) Biostatistics.

Normalization EdgeR





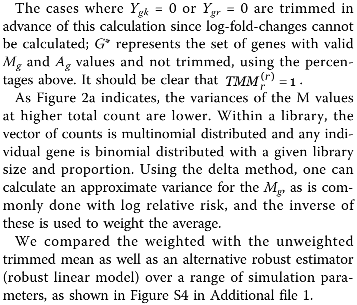
Materials and methods

TMM normalization details

A trimmed mean is the average after removing the upper and lower x% of the data. The TMM procedure is doubly trimmed, by log-fold-changes M^r_{gk} (sample *k* relative to sample *r* for gene *g*) and by absolute intensity (A_g). By default, we trim the M_g values by 30% and the A_g values by 5%, but these settings can be tailored to a given experiment. The software also allows the user to set a lower bound on the A value, for instances such as the Cloonan *et al*. dataset (Figure S1 in Additional file 1). After trimming, we take a weighted mean of M_g, with weights as the inverse of the approximate asymptotic variances (calculated using the delta method [24]). Specifically, the normalization factor for sample *k* using reference sample *r* is calculated as:

$$\log_2(TMM_k^{(r)}) = \frac{\sum_{g \in G^*} w^r_{gk} M^r_{gk}}{\sum_{g \in G^*} w^r_{gk}} \text{ where } M^r_{gk} = \frac{\log_2\left(Y_{gk} / N_k\right)}{\log_2\left(Y_{gr} / N_r\right)} \text{ and } w^r_{gk} = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}};$$

$$Y_{gk}, Y_{gr} > 0.$$

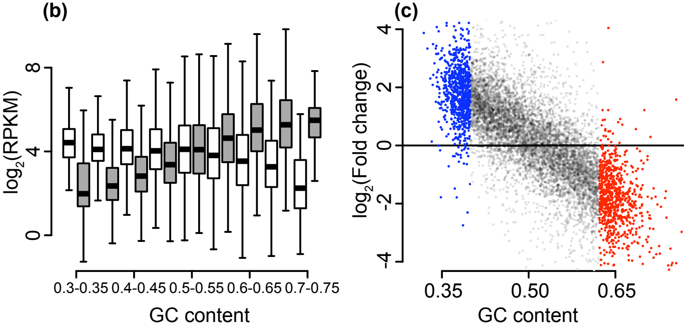


The cases where $Y_{gk} = 0$ or $Y_{gr} = 0$ are trimmed in advance of this calculation since log-fold-changes cannot be calculated; G^* represents the set of genes with valid M_g and A_g values and not trimmed, using the percentages above. It should be clear that $TMM_r^{(r)} = 1$.

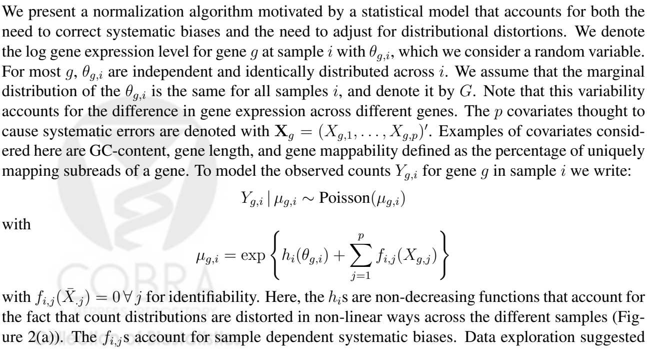
As Figure 2a indicates, the variances of the M values at higher total count are lower. Within a library, the vector of counts is multinomial distributed and any individual gene is binomial distributed with a given library size and proportion. Using the delta method, one can calculate an approximate variance for the M_g, as is commonly done with log relative risk, and the inverse of these is used to weight the average.

We compared the weighted with the unweighted trimmed mean as well as an alternative robust estimator (robust linear model) over a range of simulation parameters, as shown in Figure S4 in Additional file 1.

Normalization cqn (Hansen, Irizarry and Wu (2012). Biostatistics)



Normalization cqn (Hansen, Irizarry and Wu (2012). Biostatistics)



We present a normalization algorithm motivated by a statistical model that accounts for both the need to correct systematic biases and the need to adjust for distributional distortions. We denote the log gene expression level for gene g at sample i with $\theta_{g,i}$, which we consider a random variable. For most g, $\theta_{g,i}$ are independent and identically distributed across i. We assume that the marginal distribution of the $\theta_{g,i}$ is the same for all samples i, and denote it by G. Note that this variability accounts for the difference in gene expression across different genes. The p covariates thought to cause systematic errors are denoted with $\mathbf{X}_g = (X_{g,1}, \ldots, X_{g,p})'$. Examples of covariates considered here are GC-content, gene length, and gene mappability defined as the percentage of uniquely mapping subreads of a gene. To model the observed counts $Y_{g,i}$ for gene g in sample i we write:

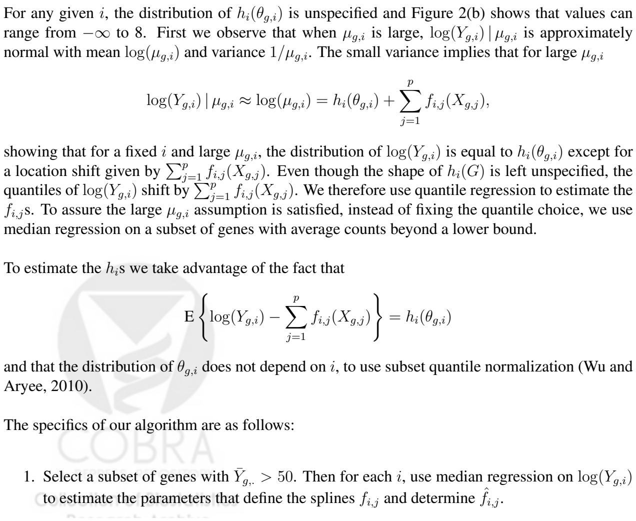
$$Y_{g,i} \,|\, \mu_{g,i} \sim \text{Poisson}(\mu_{g,i})$$

with

$$\mu_{g,i} = \exp\left\{ h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}) \right\}$$

with $f_{i,j}(\bar{X}_{\cdot j}) = 0 \,\forall\, j$ for identifiability. Here, the h_is are non-decreasing functions that account for the fact that count distributions are distorted in non-linear ways across the different samples (Figure 2(a)). The $f_{i,j}$s account for sample dependent systematic biases. Data exploration suggested

Normalization cqn (Hansen, Irizarry and Wu (2012). Biostatistics)



For any given i, the distribution of $h_i(\theta_{g,i})$ is unspecified and Figure 2(b) shows that values can range from $-\infty$ to 8. First we observe that when $\mu_{g,i}$ is large, $\log(Y_{g,i}) \mid \mu_{g,i}$ is approximately normal with mean $\log(\mu_{g,i})$ and variance $1/\mu_{g,i}$. The small variance implies that for large $\mu_{g,i}$

$$\log(Y_{g,i}) \mid \mu_{g,i} \approx \log(\mu_{g,i}) = h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}),$$

showing that for a fixed i and large $\mu_{g,i}$, the distribution of $\log(Y_{g,i})$ is equal to $h_i(\theta_{g,i})$ except for a location shift given by $\sum_{j=1}^{p} f_{i,j}(X_{g,j})$. Even though the shape of $h_i(G)$ is left unspecified, the quantiles of $\log(Y_{g,i})$ shift by $\sum_{j=1}^{p} f_{i,j}(X_{g,j})$. We therefore use quantile regression to estimate the $f_{i,j}$s. To assure the large $\mu_{g,i}$ assumption is satisfied, instead of fixing the quantile choice, we use median regression on a subset of genes with average counts beyond a lower bound.

To estimate the h_is we take advantage of the fact that

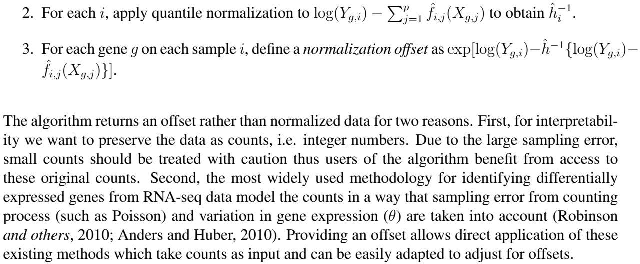
$$\mathrm{E}\left\{\log(Y_{g,i}) - \sum_{j=1}^{p} f_{i,j}(X_{g,j})\right\} = h_i(\theta_{g,i})$$

and that the distribution of $\theta_{g,i}$ does not depend on i, to use subset quantile normalization (Wu and Aryee, 2010).

The specifics of our algorithm are as follows:

1. Select a subset of genes with $\bar{Y}_{g,\cdot} > 50$. Then for each i, use median regression on $\log(Y_{g,i})$ to estimate the parameters that define the splines $f_{i,j}$ and determine $\hat{f}_{i,j}$.

Normalization cqn (Hansen, Irizarry and Wu (2012). Biostatistics)

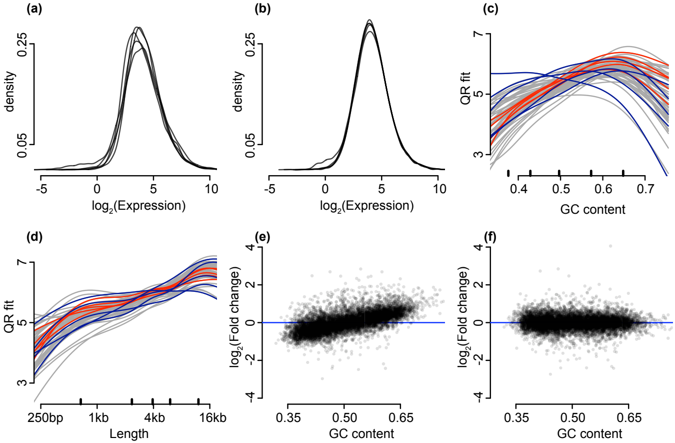


2. For each i, apply quantile normalization to $\log(Y_{g,i}) - \sum_{j=1}^{p} \hat{f}_{i,j}(X_{g,j})$ to obtain $\hat{h}_i^{-1}$.

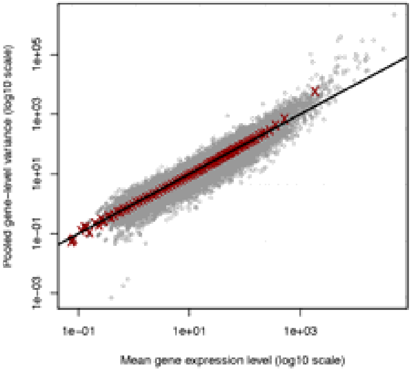
3. For each gene g on each sample i, define a *normalization offset* as $\exp[\log(Y_{g,i}) - \hat{h}^{-1}\{\log(Y_{g,i}) - \hat{f}_{i,j}(X_{g,j})\}]$.

The algorithm returns an offset rather than normalized data for two reasons. First, for interpretability we want to preserve the data as counts, i.e. integer numbers. Due to the large sampling error, small counts should be treated with caution thus users of the algorithm benefit from access to these original counts. Second, the most widely used methodology for identifying differentially expressed genes from RNA-seq data model the counts in a way that sampling error from counting process (such as Poisson) and variation in gene expression (θ) are taken into account (Robinson *and others*, 2010; Anders and Huber, 2010). Providing an offset allows direct application of these existing methods which take counts as input and can be easily adapted to adjust for offsets.

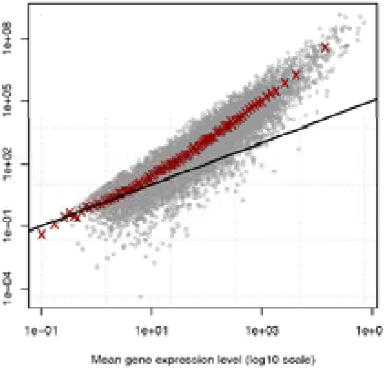
Normalization cqn (Hansen, Irizarry and Wu (2012). Biostatistics)



Mean Variance relationship



Technical replicates

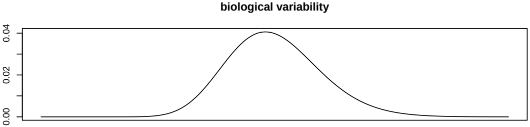


Biological replicates

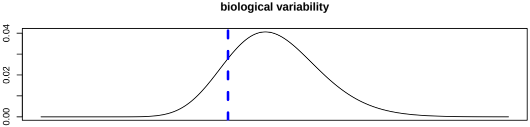
Mean variance relationship

		Seq technology		true expression
total variability	$=$	technical variability	$+$	biological variability
$\mathrm{Var}\left[y_{gi}\right]$	$=$	μ_{gi}	$+$	$\phi_g * \mu_{gi}^2$
Total CV^2	$=$	Technical CV^2	$+$	biological CV^2
Total CV^2	$=$	$\frac{1}{\mu_{gi}}$	$+$	ϕ_g

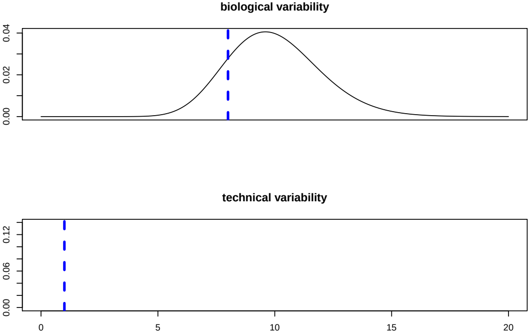
Sources of variability in a sequencing experiment



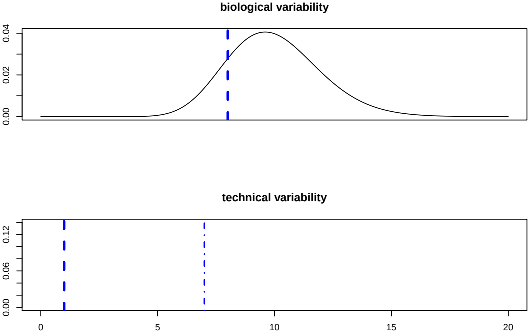
Sources of variability in a sequencing experiment



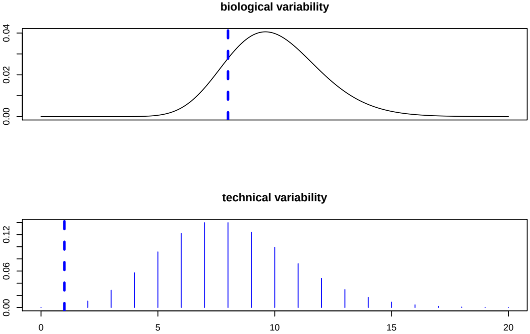
Sources of variability in a sequencing experiment



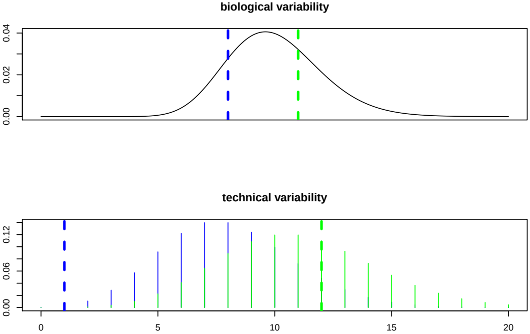
Sources of variability in a sequencing experiment



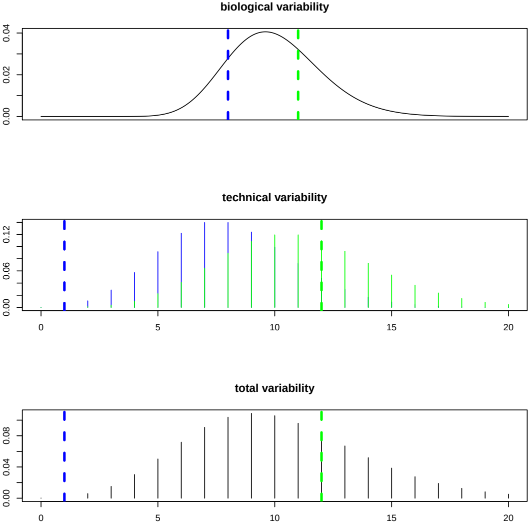
Sources of variability in a sequencing experiment



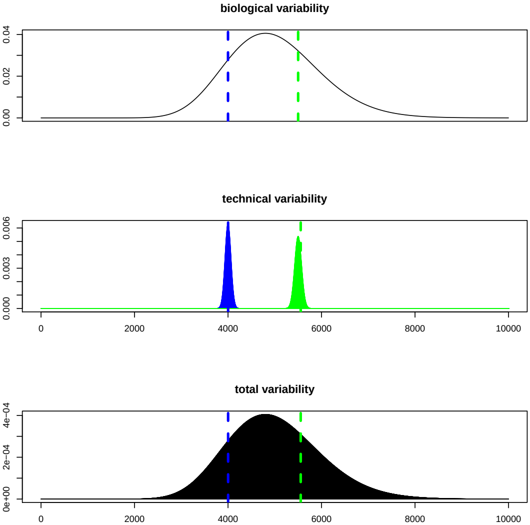
Sources of variability in a sequencing experiment



Sources of variability in a sequencing experiment



Sources of variability in a sequencing experiment



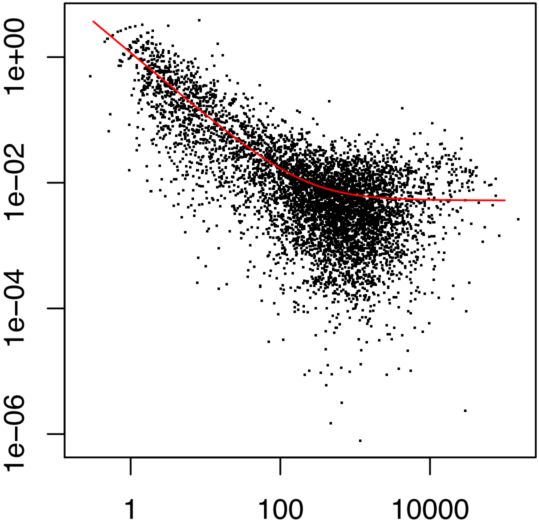
Improved model for RNA-seq data

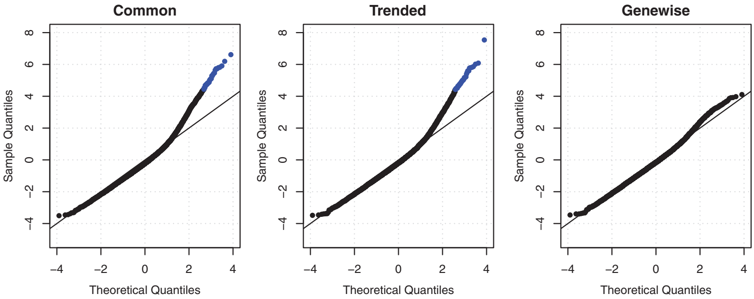
$$\begin{cases} y_{ig} & \sim & \mathrm{NB}(\mu_{ig}, \phi_{ig}) \\ \mu_{ig} & = & \lambda_{ig} S_{ig} \\ \log(\mu_{ig}) & = & \eta_{ig} \\ \eta_{ig} & = & \sum_{k=1}^{N} x_{ik}\beta_{gk} + \log S_{ig} \end{cases}$$

- y_{gi}:count for gene g of subject i
- x_{ik}: predictor variabele k evaluated for subject i
- β_{gk}: effect for predictor variable k and gene g
- S_{ig}: effective library size for gene g of subject i

Estimating overdispersion

- For every single gene: not enough data
- Common dispersion for all genes
- Trended dispersion
- Gene wise, EB shrinkage to a common (trended) dispersion: Borrow strength across genes (McCarthy & Smyth (2012). Nucleic Acid Research)


1e+00
1e−02
1e−04
1e−06
1
100
10000



McCarthy et al. (2012) NAR

II. 3. Statistical inference

- Asymptotic statistical tests exist to test if (contrasts of the) parameters of the GLM are different form zero.
- Implemented in edgeR and DESeq2.
- Again we have to correct for multiple testing !!!