RNA-seq quality control and pre-processing

AllBio workshop

January 8, 2014

Mikael Huss, SciLifeLab, Sweden

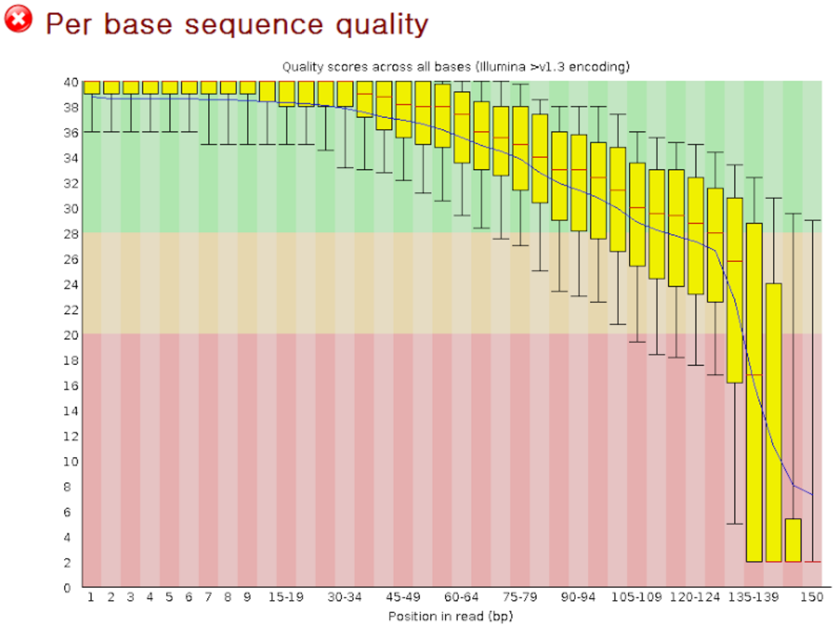
Enabler for Life Sciences

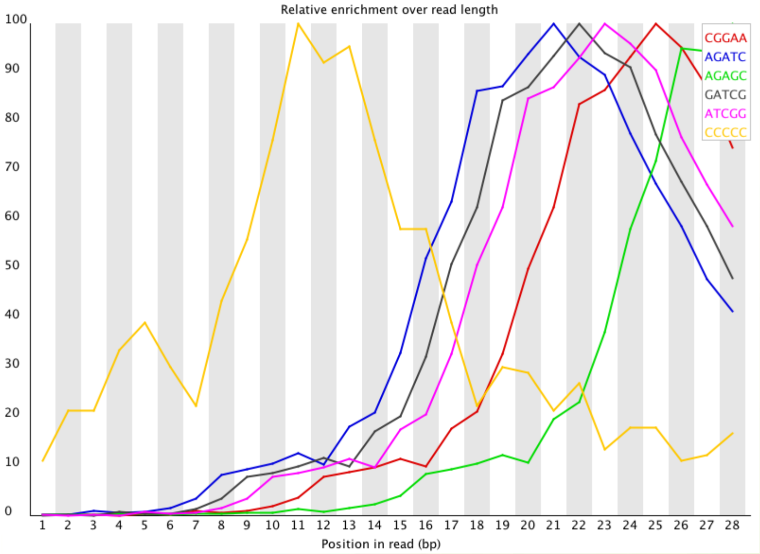
RNA-seq quality control and pre-processing

- Generic high-throughput sequencing QC tools (e g FastQC, PRINSEQ)
- RNA-seq specific QC tools (e g RSeQC, RNASeQC)

- Pre-mapping QC (sequence qualities, sequence overrepresentation)
- Pre-processing (trimming etc)
- Post-mapping QC (distribution of mapped regions, contamination etc)

FastQC

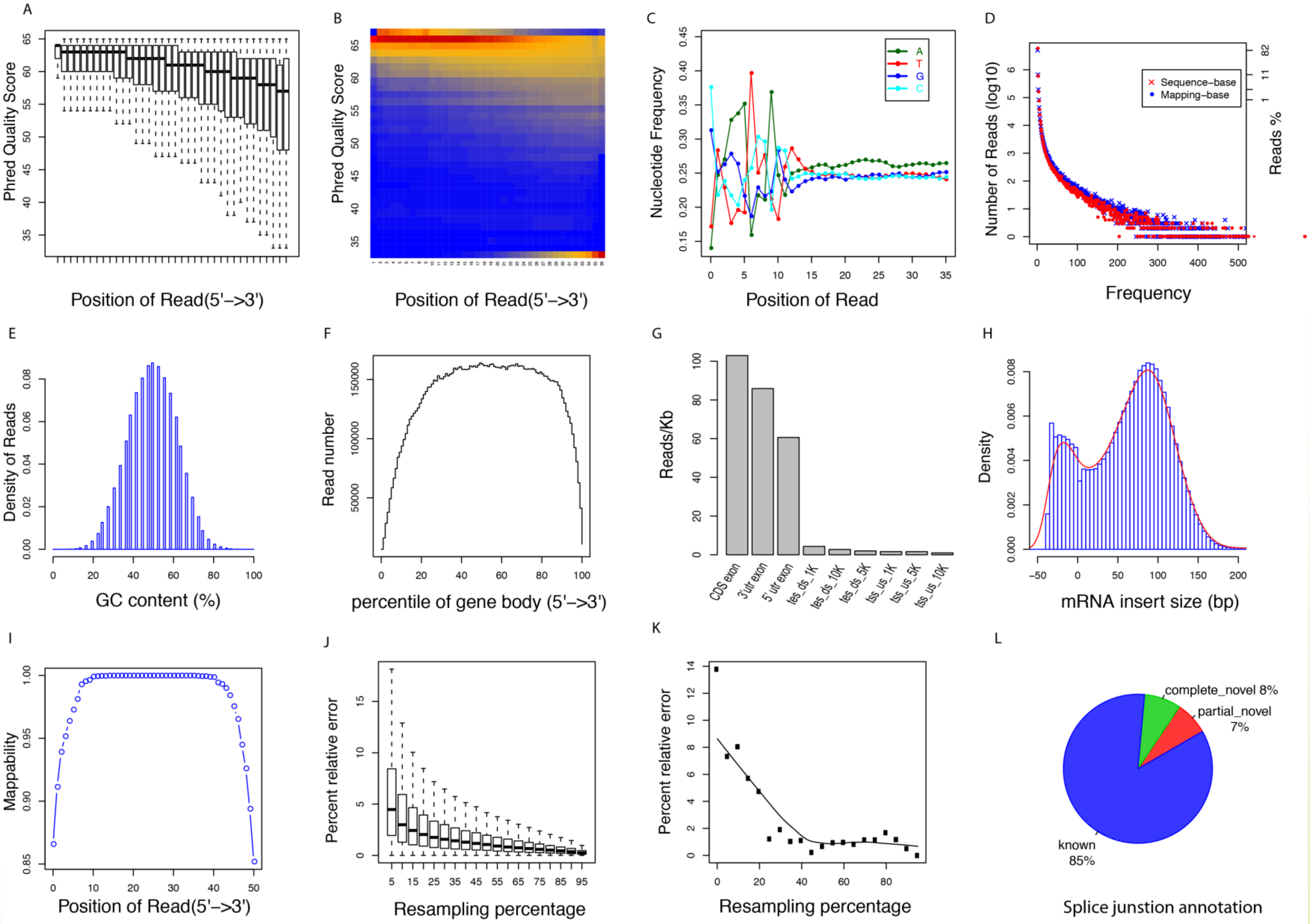




Sequence quality score plots

Sequence overrepresentation plots

FastQC (*http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*)
Also check out PRINSEQ (http://prinseq.sourceforge.net)



RSeQC (*https://code.google.com/p/rseqc/*)

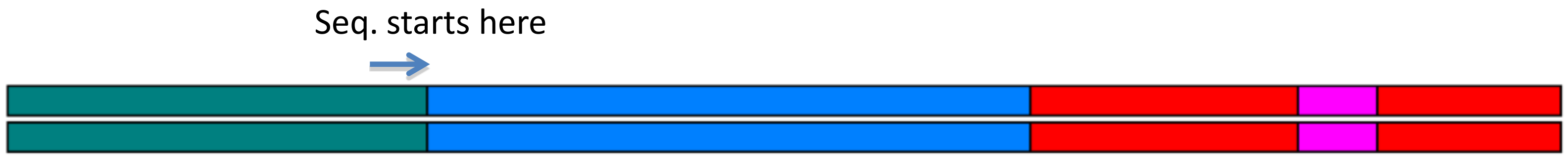
Trimming

- Adapter trimming
  - May increase mapping rates
  - Absolutely essential for small RNA
  - Probably improves *de novo* assemblies

- Quality trimming
  - May increase mapping rates
  - May also lead to loss of information

Lots of software doing either of these or both. E g Cutadapt, Trim Galore!, PRINSEQ, Trimmomatic, Sickle/Scythe, FASTX Toolkit, etc.

Adapter trimming



Seq. starts here

🟩 Universal Adapter

🟦 DNA Fragment of Interest

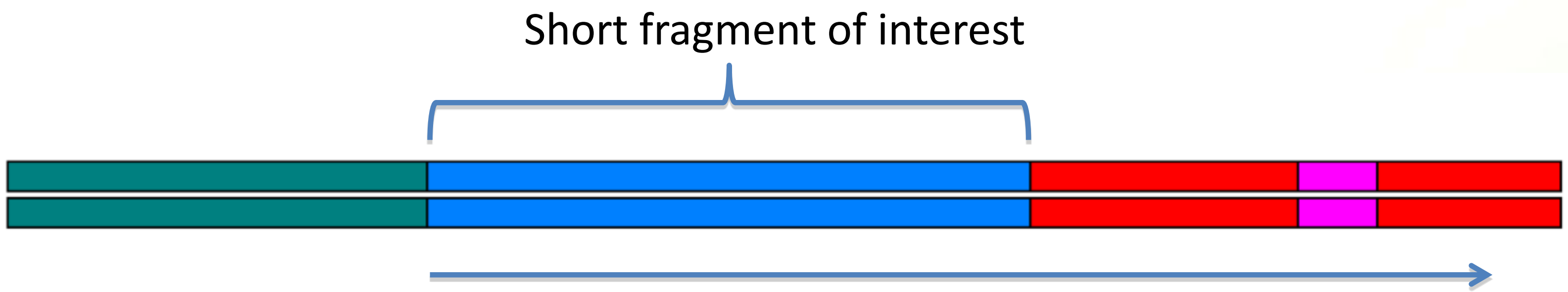
🟥 Indexed Adapter

🟪 6 Base Index Region

Illumina TruSeq DNA Adapters De-Mystified by James Schiemer
http://tucf-genomics.tufts.edu/documents/protocols/TUCF_Understanding_Illumina_TruSeq_Adapters.pdf

Most common case

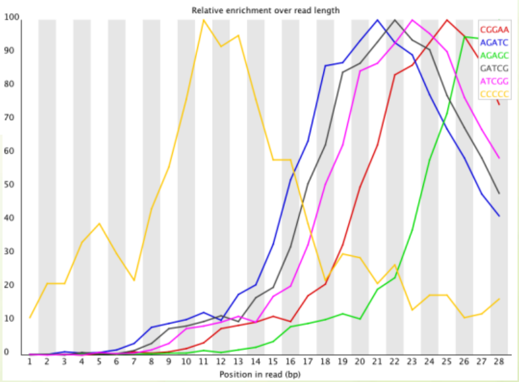
DNA fragment of interest shorter than read length



Short fragment of interest

100-bp read

- 🟩 Universal Adapter
- 🟦 DNA Fragment of Interest
- 🟥 Indexed Adapter
- 🟪 6 Base Index Region



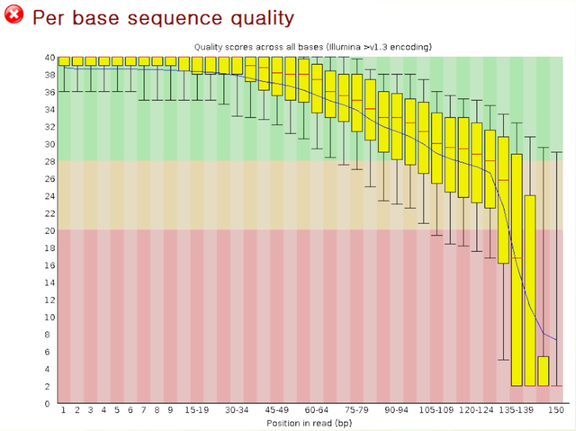
Will always happen for e g miRNA

Quality trimming

Rationale:

Erroneous base calls (often towards the ends of reads but also in the beginning) can have a detrimental effect on

- *de novo* assembly (spurious paths and bubbles in the assembly graph → increased memory consumption and complexity)
- mapping rates for reference based analysis
- variant calling



Assume that the reported quality values (QVs) for these erroneous base calls will be low. Therefore you want to trim away regions with average QVs below some threshold.

One way to quality trim

BWA, CutAdapt, CLC Bio and many others use slightly different versions of “PHRED trimming”, or the so-called “modified Mott algorithm”.

The basic idea is to trim from either the 3’ end, or both the 3’ and 5’ end, and keep track of a running sum of deviations from the threshold (negative if the base has lower quality than the cutoff, positive if higher). The read is trimmed where this sum is minimal.

If the trimmed sequence is too short (e.g. <30 bp), it is discarded.

So, (at least) 2 user defined parameters:

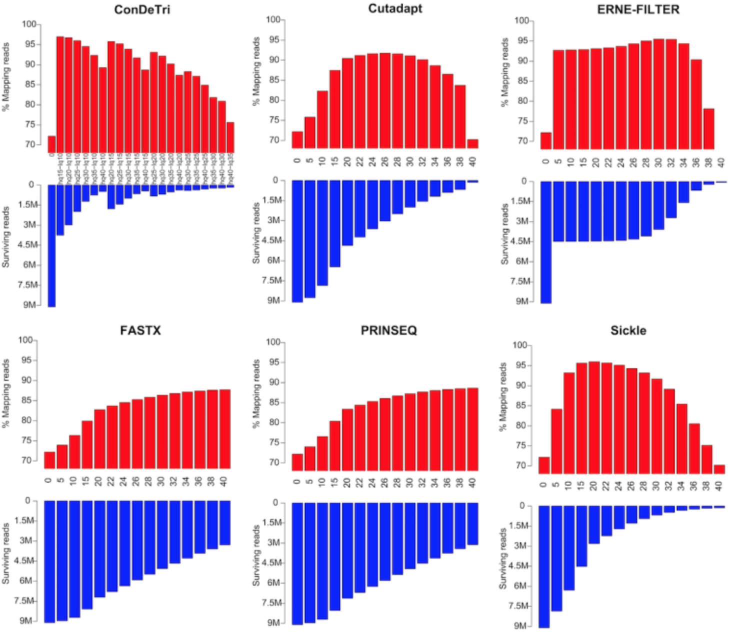
- quality score cutoff
- min length of sequence to keep

Details of the Mott algorithm plus several other trimming methods are given in
http://research.bioinformatics.udel.edu/genomics/ngsShoRT/download/advanced_user_guide.pdf

Is trimming beneficial?

Two recent papers + a blog post: *http://genomebio.org/is-trimming-is-beneficial-in-rna-seq/*

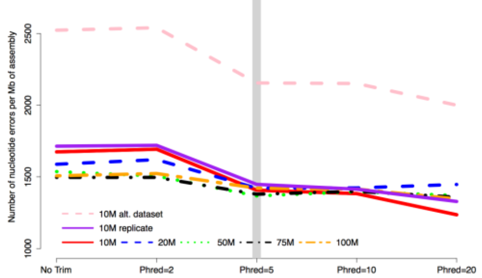
Software comparison, RNA/DNA-Seq



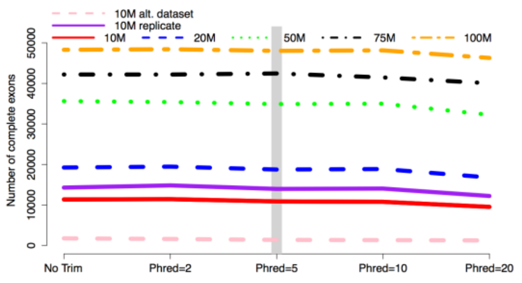
"trimming is beneficial in RNA-Seq, SNP identification and genome assembly procedures, with the best effects evident for intermediate quality thresholds (Q between 20 and 30)"

Del Fabbro C et al (2013) **An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis**. PLoS ONE 8(12): e85024. doi:10.1371/journal.pone.0085024

Assembly-oriented, RNA-seq only



Erroneous bases in assembly



# complete exons

"Although very aggressive quality trimming is common, this study suggests that a more gentle trimming, specifically of those nucleotides whose Phred score < 2 or < 5, is optimal for most studies across a wide variety of metrics."

MacManes MD (2013)
On the optimal trimming of high-throughput mRNAseq data doi: 10.1101/000422

Some comments on software

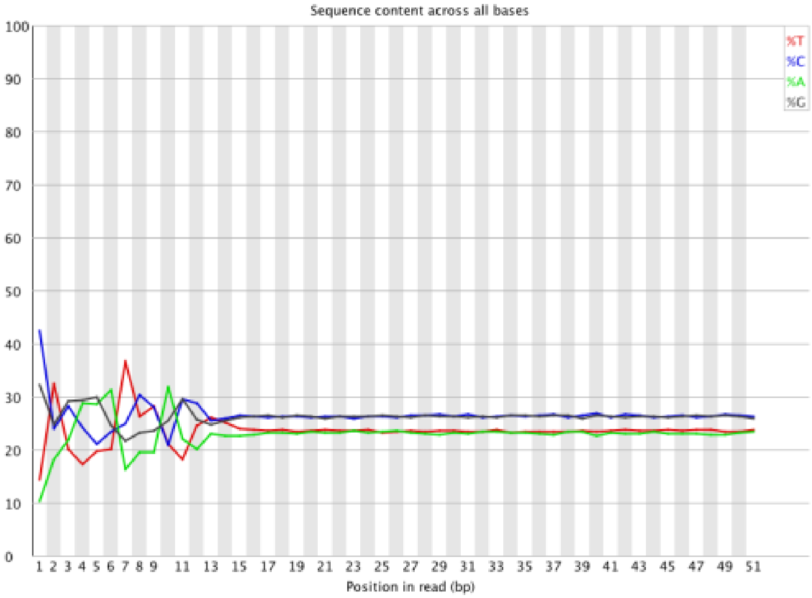
I do not always trim – just in cases where it appears to improve results

TrimGalore - wrapper to CutAdapt with both quality and adapter trimming
(my own choice as it works smoothly with paired-end reads as well)

Trimmomatic – quality and adapter trimming, lots of options

Scythe (adapter) – Sickle (quality) trimming combo

Beginnings of reads



Bias in sequence composition is often (always?) seen in the first 12-15 bp in Illumina RNA-seq data sets

Thought to be due to issues with “random” hexamer priming

Hansen et al. (2010) **Biases in Illumina transcriptome sequencing caused by random hexamer priming**
Nucleic Acids Res. 2010 July; 38(12): e131. doi: 10.1093/nar/gkq224

Not clear if trimming the 5’ helps here.
According to an authoritative source you should always remove the first base and preferably a couple of more bases afterwards ☺ (I have not personally done this so far)

Poly-A tails

Seldom captured in Illumina HiSeq runs

Could complicate mapping & lead to false positive hits in sequence databases

PRINSEQ low-complexity filter
EMBOSS TrimEST *http://emboss.sourceforge.net/apps/cvs/emboss/apps/trimest.html*
(etc).

GC bias

(disclaimer – I have never adjusted for this!)

"We […] demonstrate the existence of strong ***sample-specific*** *GC-content effects on RNA-Seq read counts, which can substantially bias differential expression analysis"*

Risso D et al. (2011) **GC-Content Normalization for RNA-Seq Data**. BMC Bioinformatics, 12:480 doi:10.1186/1471-2105-12-480

+ several other papers

CQN package for R (BioConductor)
http://www.bioconductor.org/packages/2.13/bioc/html/cqn.html

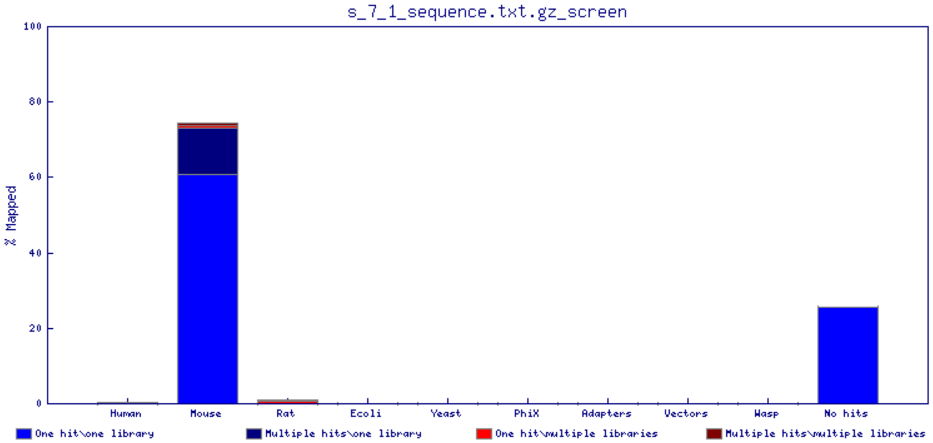
"The biochemistry of RNA-Seq library preparation results in cDNA fragments that are not uniformly distributed within the transcripts they represent. This non-uniformity must be accounted for when estimating expression levels …"

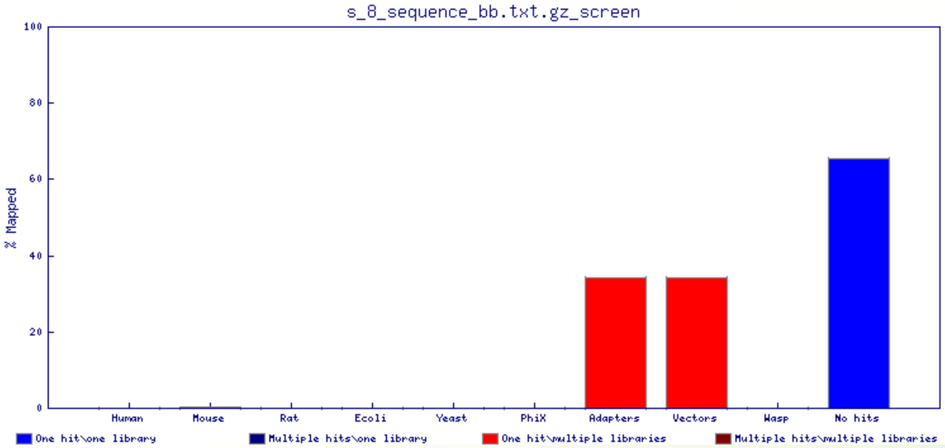
Roberts A et al. (2011) **Improving RNA-Seq expression estimates by correcting for fragment bias**. Genome Biology, 12:R22 doi:10.1186/gb-2011-12-3-r22

Post-mapping QC

- Contamination
- Duplicates
- Genomic features covered

What’s lurking in your data?



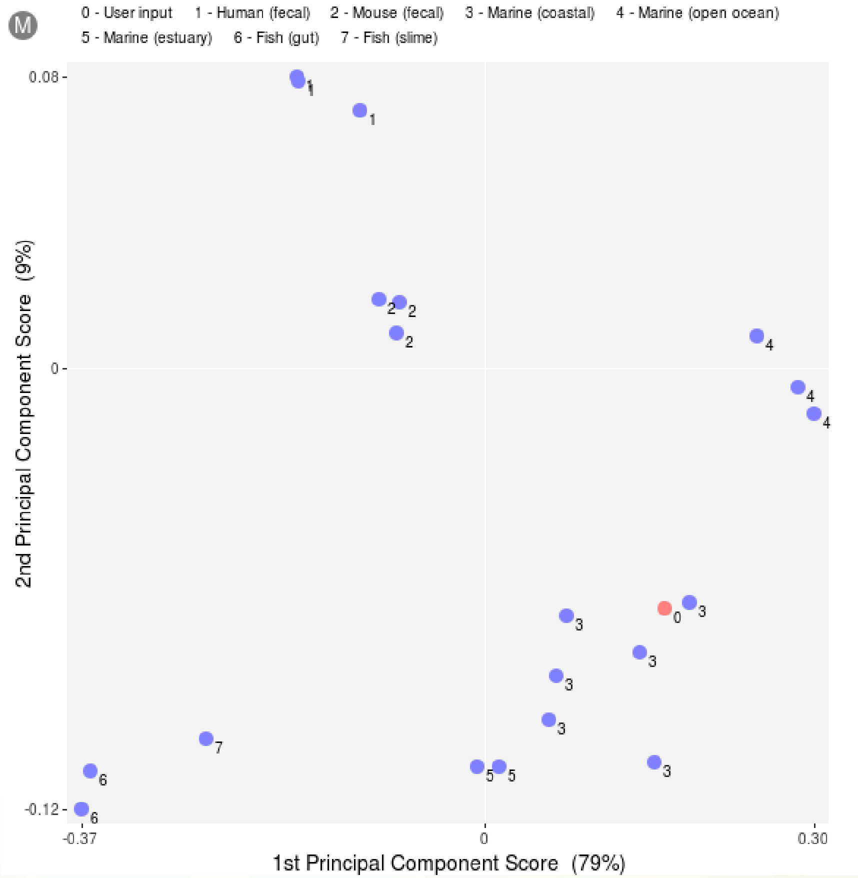


Screen for contaminating genomes, vectors, adapter sequences

FastQ Screen: http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/

Poor man’s version: simply BLAST (e.g.) 1000 random sequences against nt

What's lurking in your data?



Could also be done pre-mapping

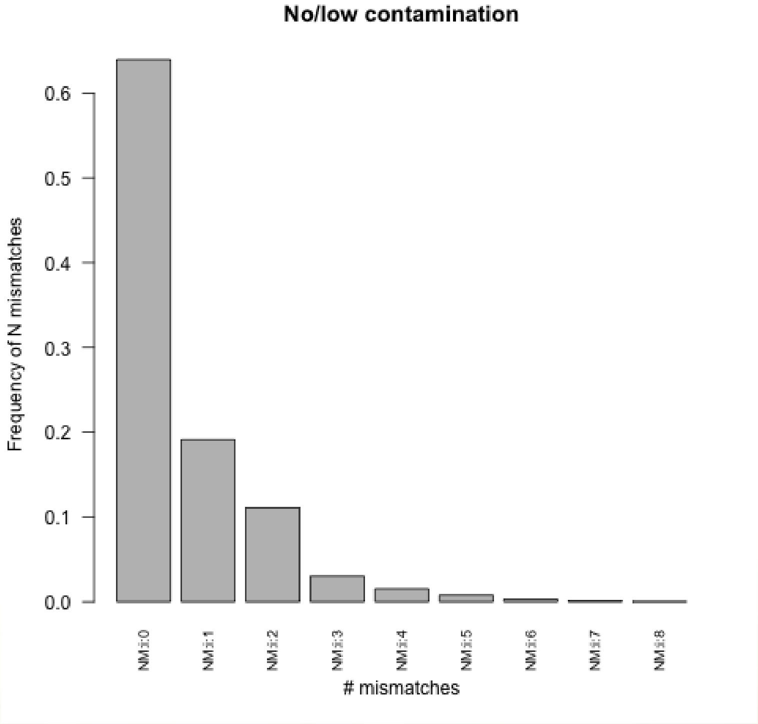
PRINSEQ

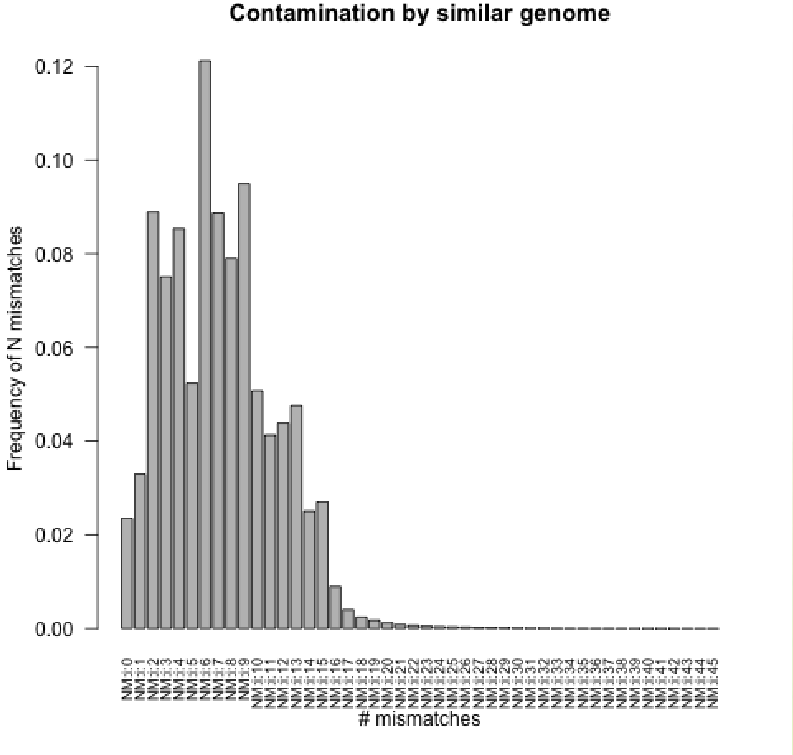
Dinucleotide frequencies

Comparing metagenomes

Contamination by similar genomes

Need to look at the distribution of the number of mismatches per alignment (e g NM:i: attribute in the BAM/SAM file)





Duplicate sequences

Observing identical sequences in a sequencing run could result from

- Genuine, multiple observations of the same sequence from different source molecules
- Amplification from PCR steps in library preparation or sequencing
- Optical duplicates
- Exhausting the library; sequencing the same molecule several times

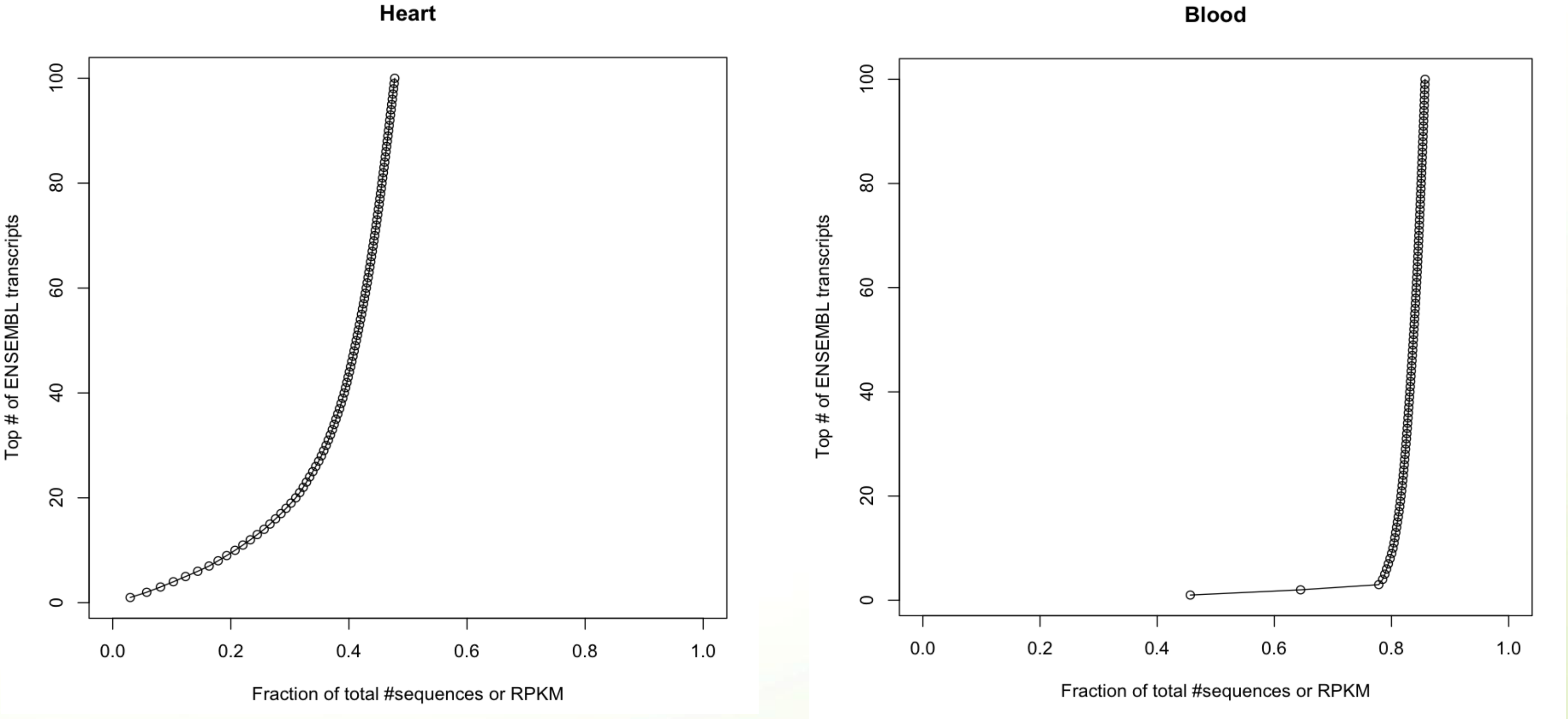
Note:

For resequencing applications (whole-genome, exome sequencing) it is standard practice to remove duplicate sequences. For RNA-seq, things are more complicated.

Duplicates are usually removed after mapping because it is simple. E g look for paired-end reads where both mates map to the same coordinates.

Duplicates and RNA-seq

# of sequences taking up X% of the sequences

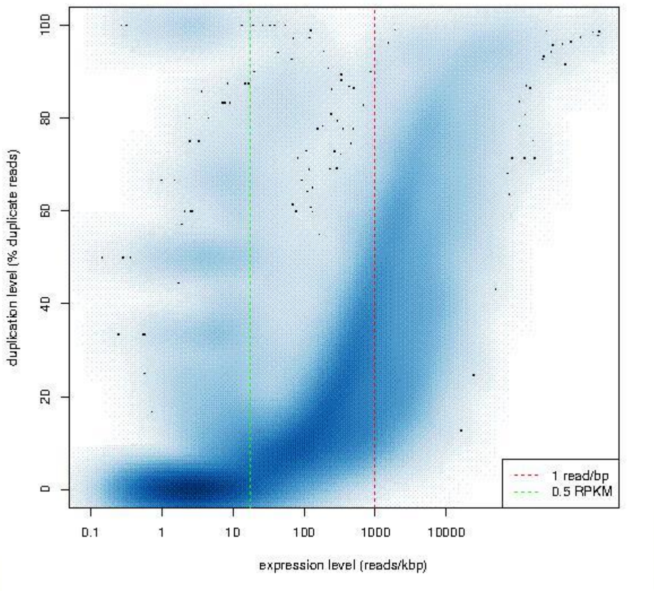


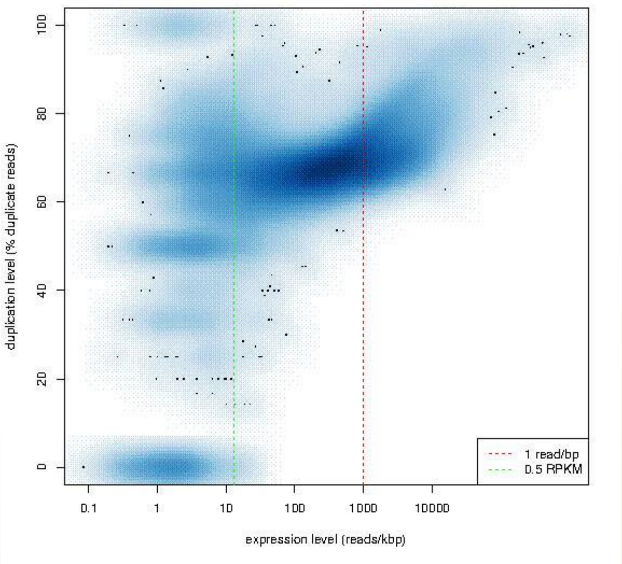
HBA1 HBB HBA2

Duplicates and RNA-seq

Millions of reads mapping to a single short transcript → will look like a LOT of duplicates! (this also happens with rRNA)

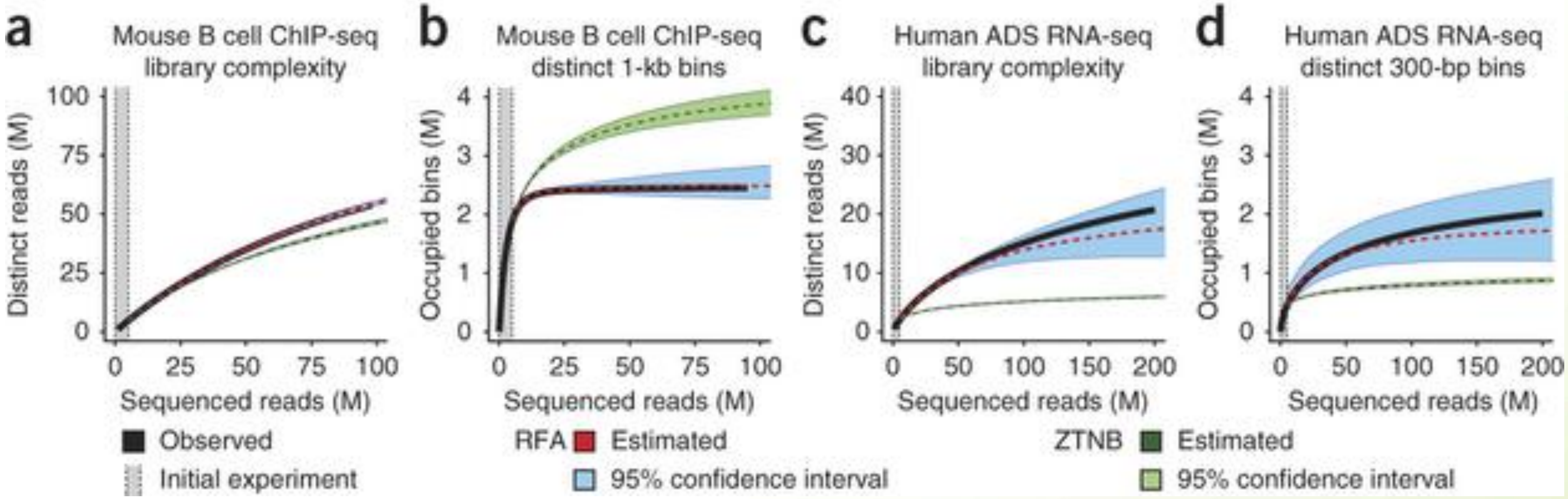
Thus, highly expressed transcripts having lots of “duplicate” reads is normal!



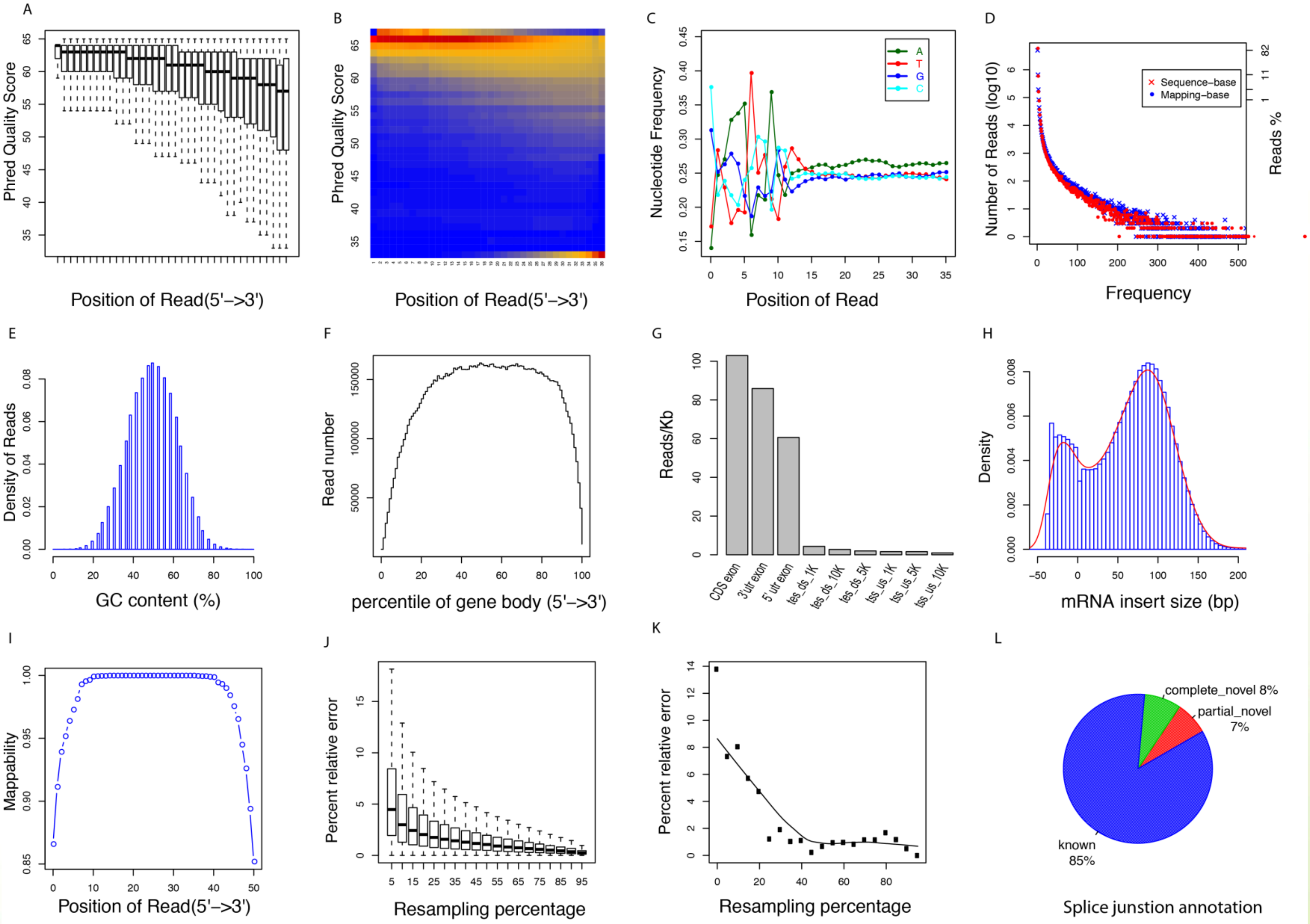


dupRadar (H. Klein et al)
http://sourceforge.net/projects/dupradar/

Predicting library complexity

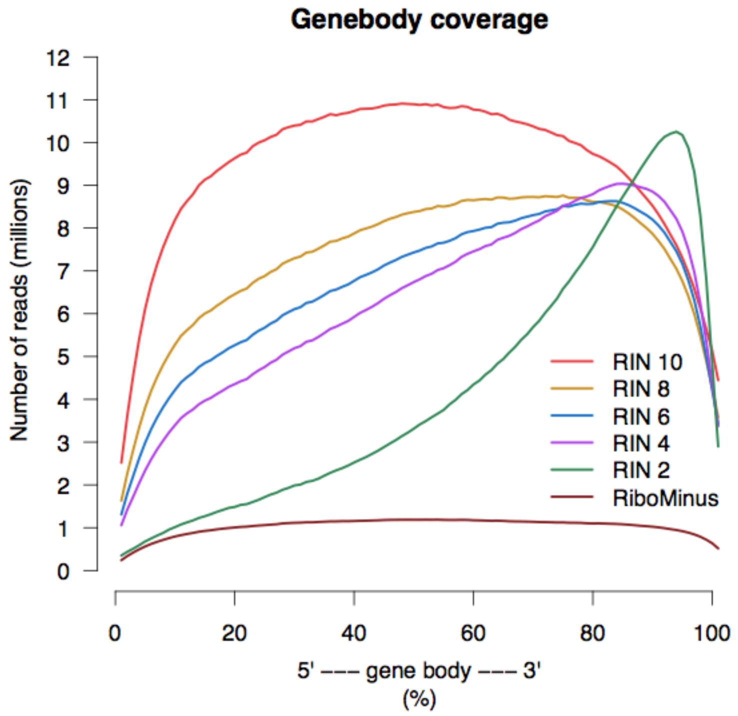


Daley T and Smith AD. ***Predicting the molecular complexity of sequencing libraries****.*
Nature Methods 10, 325–327 (2013) doi:10.1038/nmeth.2375



RSeQC (*https://code.google.com/p/rseqc/*)

Gene body coverage

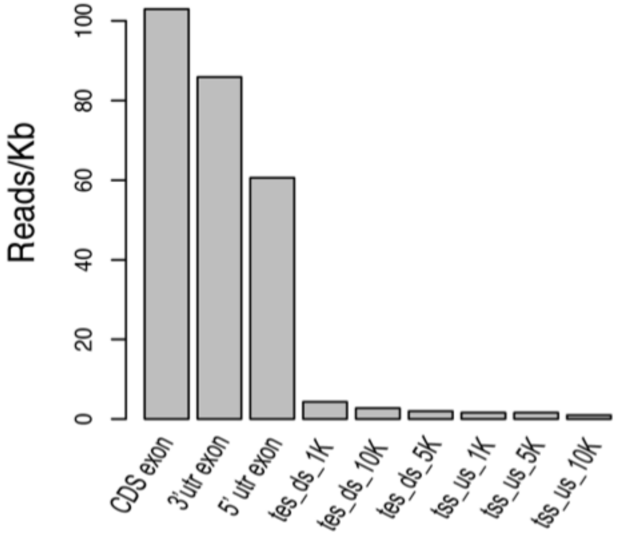


Benjamin Sigurgeirsson

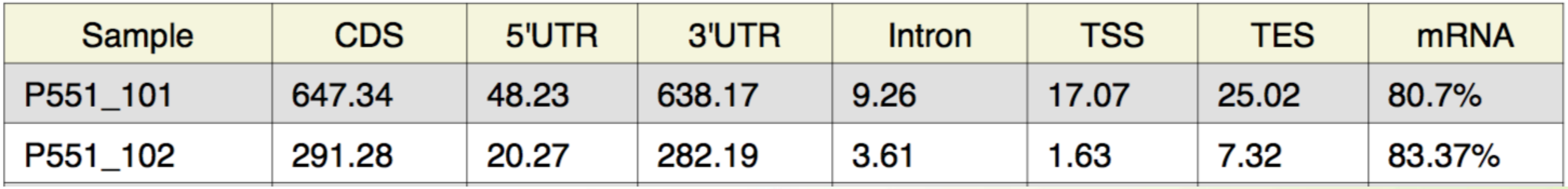
RNA quality affects the shape

If this profile has strange spikes, there may be extreme overrepresentation of sequences

Mapping to genomic features

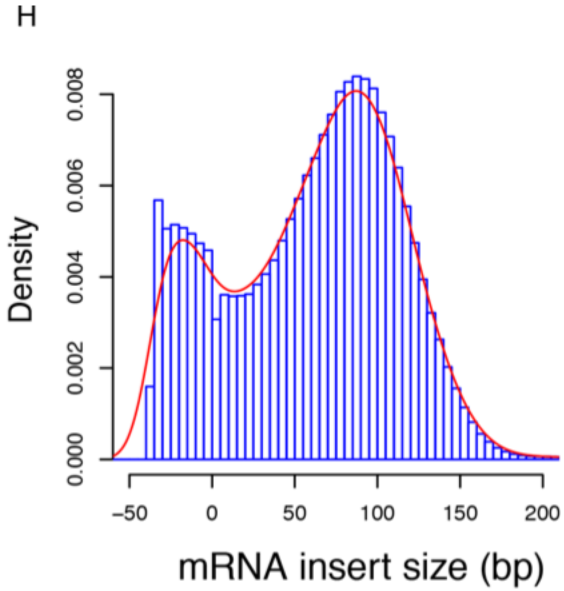


RSeQC plot



Sample	CDS	5'UTR	3'UTR	Intron	TSS	TES	mRNA
P551_101	647.34	48.23	638.17	9.26	17.07	25.02	80.7%
P551_102	291.28	20.27	282.19	3.61	1.63	7.32	83.37%

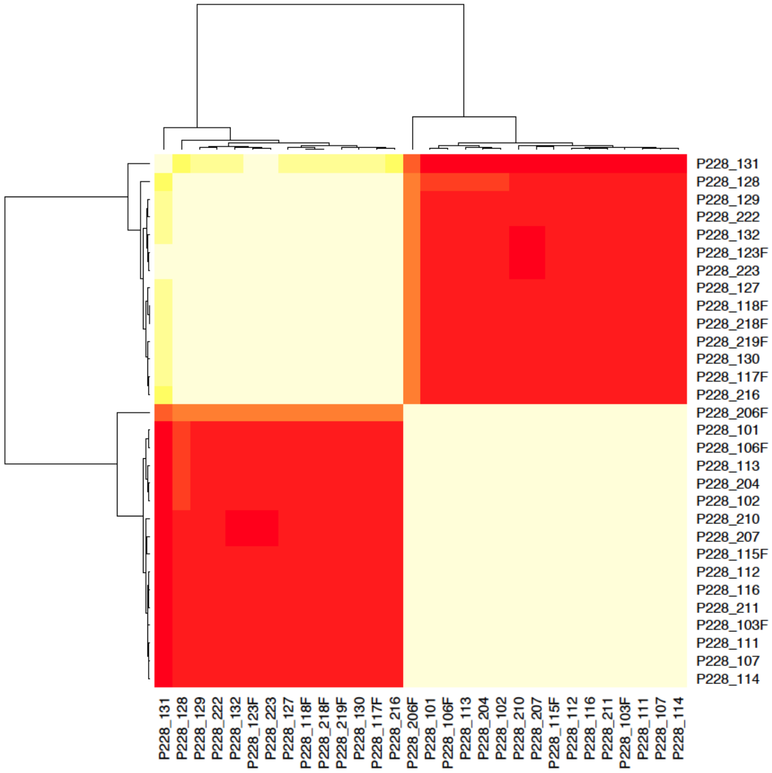
Insert size distribution



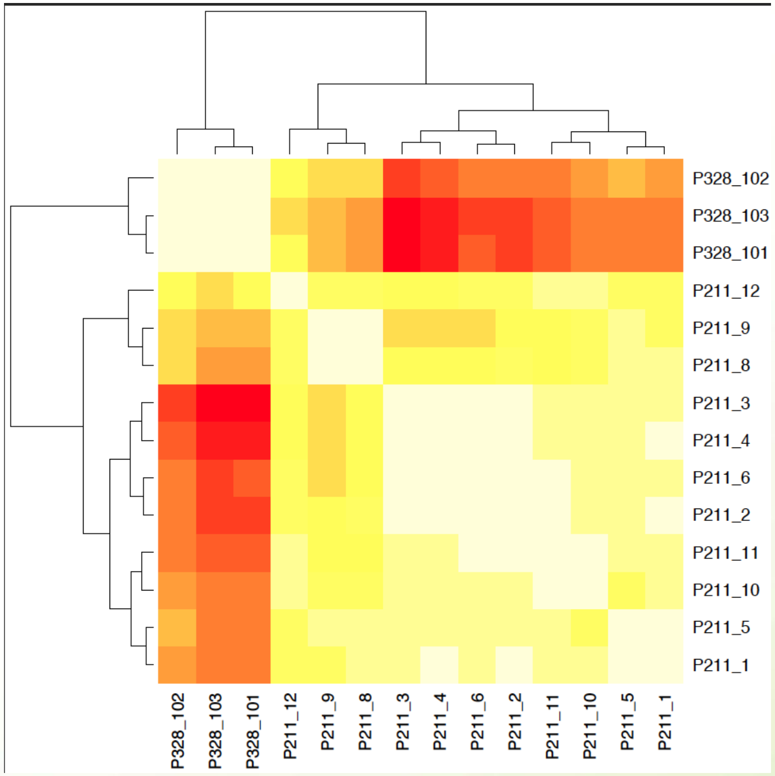
Negative insert size implies overlapping mate reads

For assembly, might want to join overlapping mates into “pseudo-single-end reads”
I have used FLASH; other tools mentioned here
http://thegenomefactory.blogspot.se/2012/11/tools-to-merge-overlapping-paired-end.html

Clustering to check for outliers and batch effects



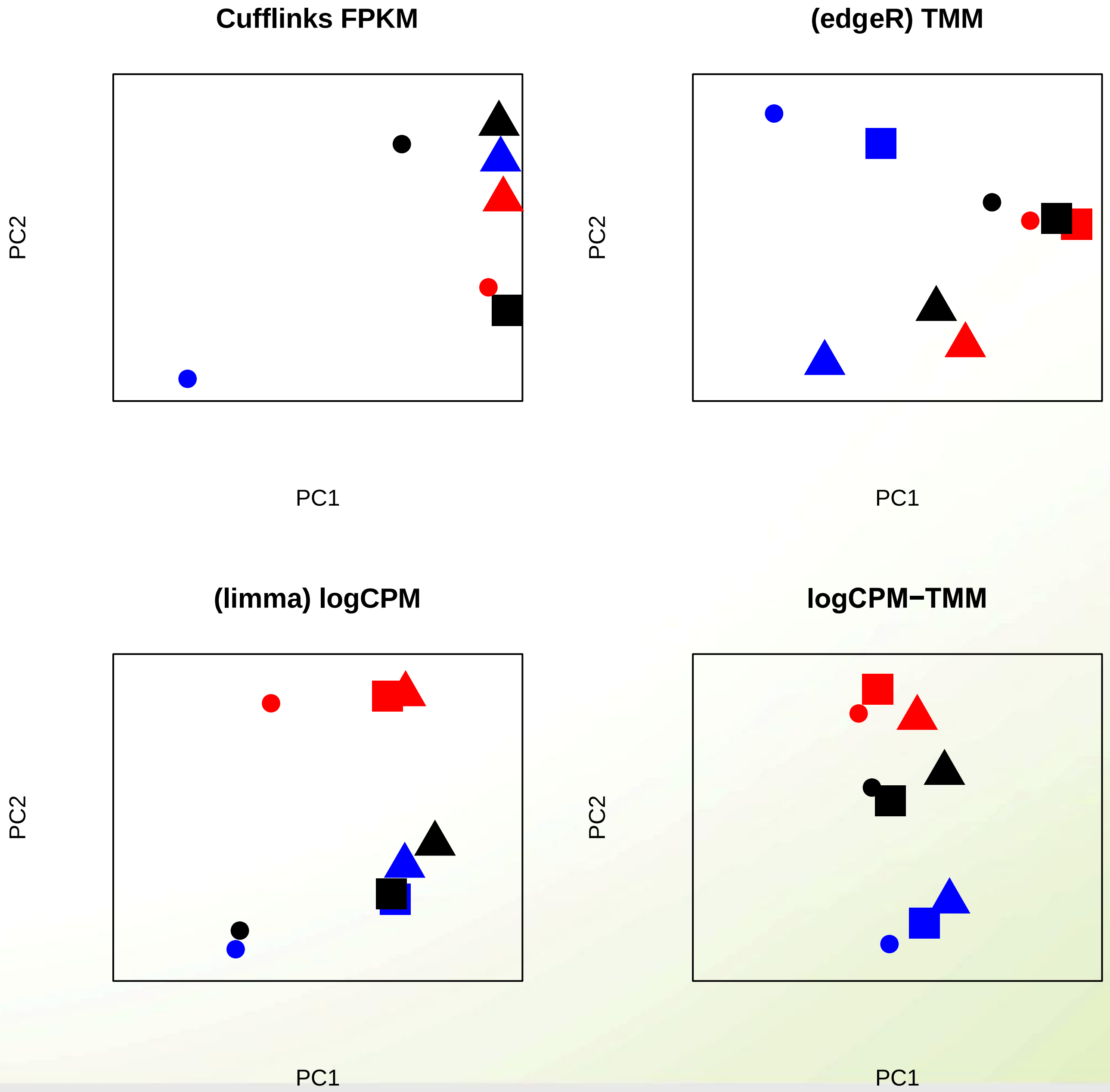
Cluster according to tissue



Cluster according to prep or sequencing batch

Red – brain
Blue – heart
Black – kidney

Circles – Study 1
Triangles – Study 2
Squares – Study 3



… or PCA plots
But it can be tricky because a lot depends on the normalization

Normalization: different goals

- **R/FPKM:** (Mortazavi et al. 2008)
  - Correct for: differences in sequencing depth and transcript length
  - Aiming to: compare a gene across samples and diff genes within sample

- **TMM**: (Robinson and Oshlack 2010)
  - Correct for: differences in transcript pool composition; extreme outliers
  - Aiming to: provide better across-sample comparability

- **TPM**: (Li et al 2010, Wagner et al 2012)
  - Correct for: transcript length distribution in RNA pool
  - Aiming to: provide better across-sample comparability

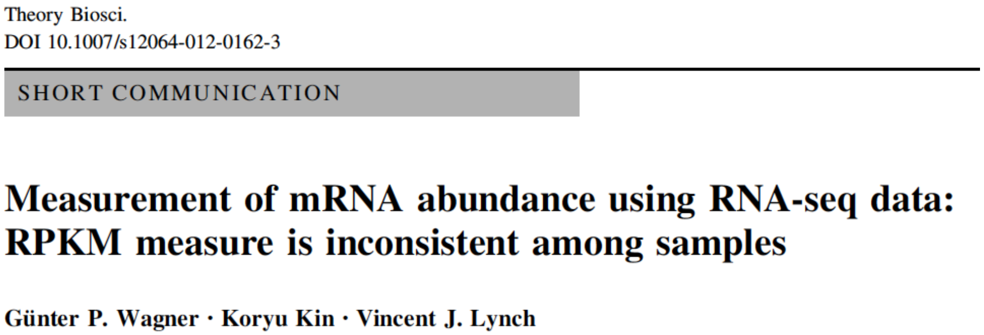
- **Limma voom (logCPM)**: (Lawet al 2013)
  - Aiming to: stabilize variance; remove dependence of variance on the mean

Optimal Scaling of Digital Transcriptomes

Gustavo Glusman ✉, Juan Caballero, Max Robinson, Burak Kutlu, Leroy Hood

Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

TPM – Transcripts Per Million



Theory Biosci.
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch

A slightly modified RPKM measure that accounts for differences in gene length distribution in the transcript population

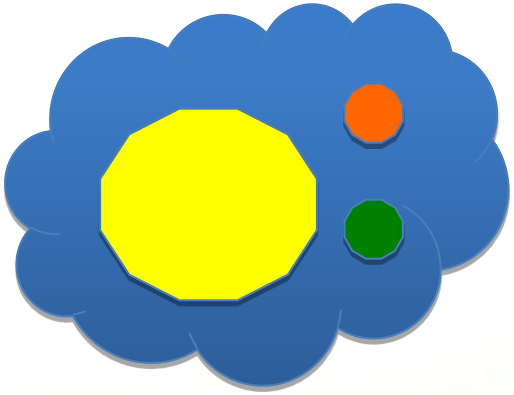
Blog post that explains how it works
http://blog.nextgenetics.net/?e=51

TMM – Trimmed Mean of M values

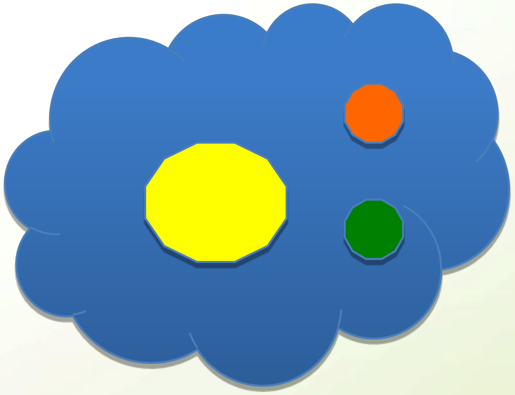
Attempts to correct for differences in RNA *composition* between samples

E g if certain genes are very highly expressed in one tissue but not another, there will be less "sequencing real estate" left for the less expressed genes in that tissue and RPKM normalization (or similar) will give biased expression values for them compared to the other sample

RNA population 1



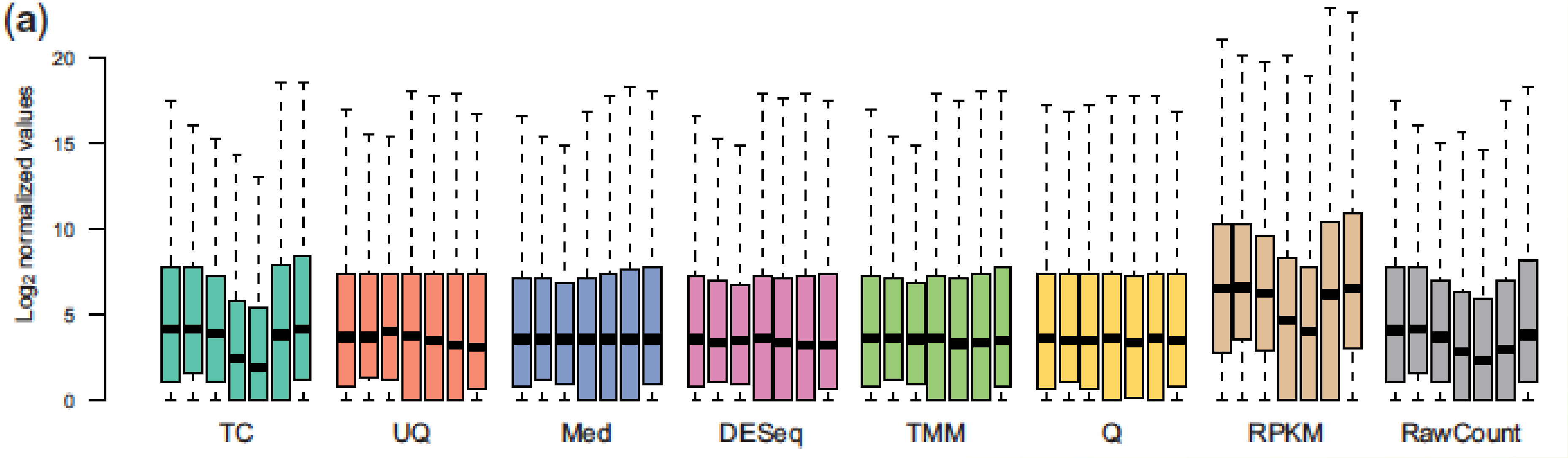
RNA population 2



Equal sequencing depth -> orange and red will get lower RPKM in RNA population 1 although the expression levels are actually the same in populations 1 and 2

Robinson and Oshlack Genome Biology 2010, 11:R25, http://genomebiology.com/2010/11/3/R25

Across-sample comparability



Dillies et al., Briefings in Bioinformatics, doi:10.1093/bib/bbs046

Comments on normalization

Constantly evolving area. My current recommendations:

For reporting gene expression estimates: Use TPM if possible (RSEM, Sailfish, eXpress)

For differential expression analysis: Use TMM, DESeq normalization or similar

For clustering and visualization: I prefer TMM + a log transform (limma-voom)

Questions?

Thanks to:

Thomas Svensson + the whole WABI group at SciLifeLab in Stockholm & Uppsala
Benjamin Sigurgeirsson (SciLifeLab/KTH)
Gary Schroth (Illumina)