Outline

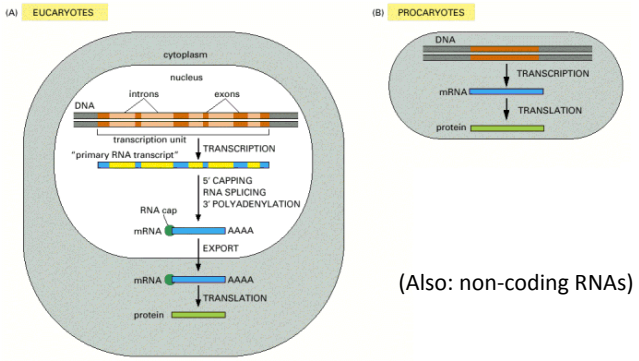
- Lecture 1
  - Measuring gene expression – RNA-seq
  - RNA-seq preprocessing & QC
  - Quantification & Normalization
- Lecture 2
  - Statistics of counts
  - Differential expression analysis with RNA-seq

RNA – seq
Lecture 1

STAT/GSAT/BIOF 540
2015

Paul Pavlidis

Refresher: gene expression



(Also: non-coding RNAs)

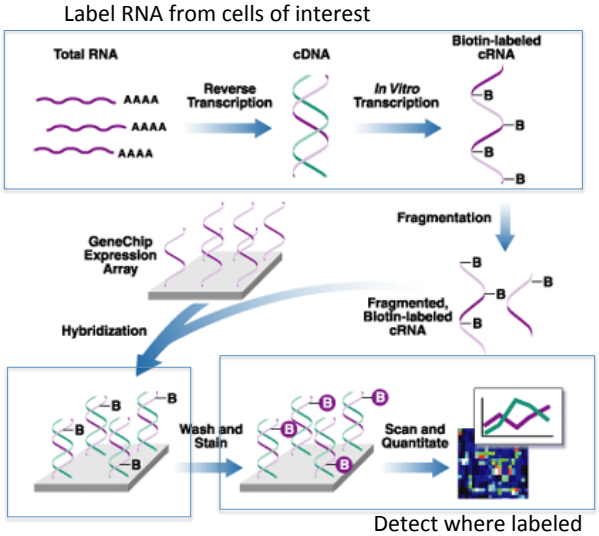
At RNA level, regulatory points include:

- Rate of transcription
- Rate of modification and export to the cytoplasm
- Rate of degradation

Alberts, Molecular Biology of the Cell

Refresher: Measuring RNA expression with Hybridization (Affymetrix)

Label RNA from cells of interest



Hybridize to chip

Detect where labeled RNA has bound.

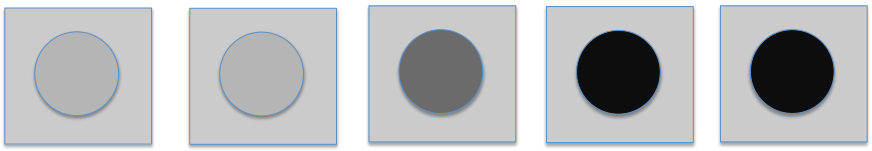
Affymetrix, Inc.

Pros and cons of hybridization-based assays

Mature, inexpensive, and accessible

But:

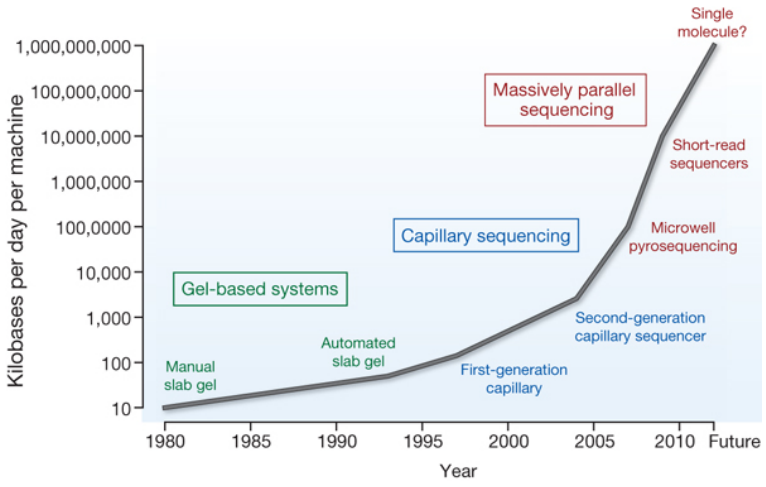
- Only detect what you probe
- Cross hybridization → background + ambiguity
- Difficult to assess alternative splicing etc.
- Sensitivity limited by background & noise in image
- Dynamic range further limited by saturation



Replacing hybridization with sequencing

- Directly analyze the nucleotides in the sample using sequence analysis: count occurrences
- Background and saturation are not an issue in the same way
- Until recently, this was too expensive to do because sequencing was rate limiting (old method: SAGE)
- Now: RNA-seq or Whole Transcriptome Shotgun Sequencing, WTSS

Improvements in the rate of DNA sequencing



MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

Steps for RNA-seq
(focus on quantitation)

Data generation

- RNA isolation and library preparation
- Sequencing

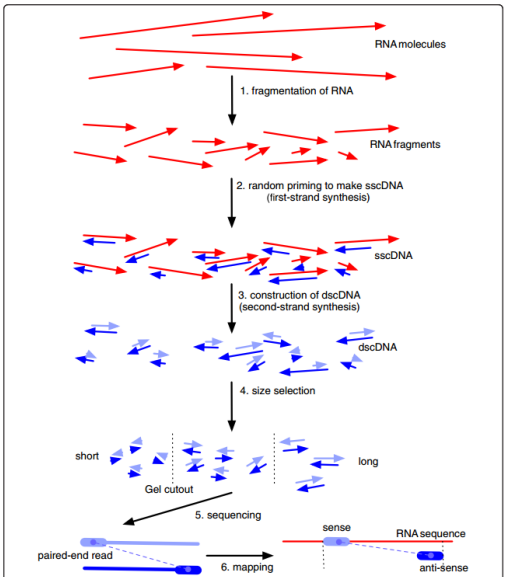
Analysis

- Sequence QC
- Sequence alignment
- Gene/transcript quantification

RNA-seq library preparation

- Randomly sample molecules from your RNA sample
- Sequence (pieces of) them to find out what they are
- Count up how many times you see sequences for a given gene.

(Figure shows a generic simplified protocol; details vary)

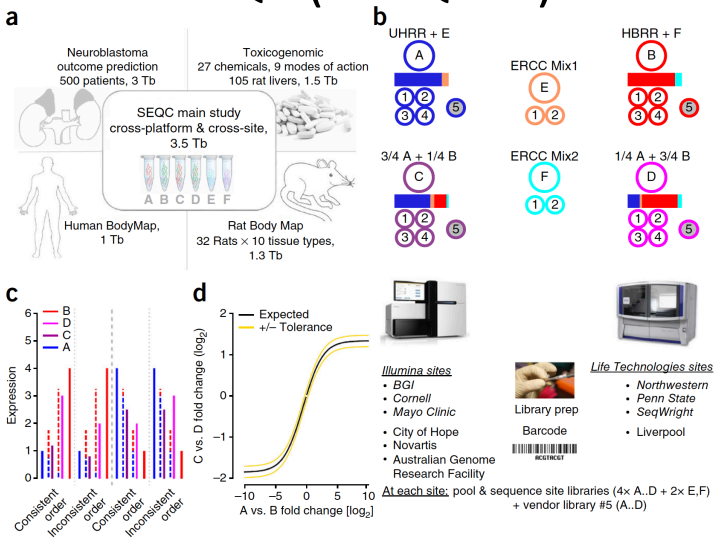


Roberts et al. Genome Biology 2011, 12:R22
http://genomebiology.com/2011/12/3/R22

Advantages of RNA-seq

- Doesn't require prior genome annotation
- Increased dynamic range and sensitivity (in theory)
- Able to detect splicing variation, gene fusions
- Allows other applications such as mutation detection or RNA editing
- Allele-specific expression can be assessed
- Adjustable depth – pay more to sequence more → increase sensitivity, precision

SEQC (MAQC III)



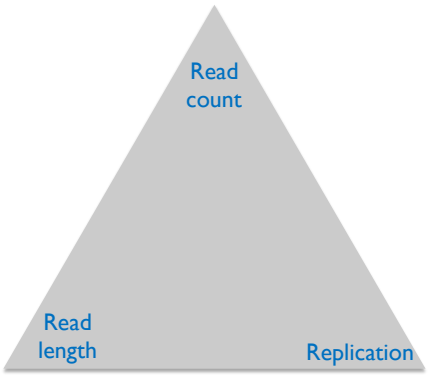
Nature Biotechnology 32, 903–914 (2014)

Some SEQC conclusions

- "RNA-seq can be used as a versatile tool for relative expression profiling, with comparable or superior performance to microarrays in many applications given sufficient read depth and appropriate choice of analysis pipeline"
- Low expression very difficult to measure accurately; ~1/3 of genes removed to get reasonable results (similar to microarray)
- "Substantial improvements in short-read RNA-seq analysis are still required"
- "In line with our observations, earlier comparisons of RNA-Seq and microarrays have identified the techniques as complementary, advising applications to combine their respective strengths, especially in large clinical studies"

RNA-seq cost trade-offs

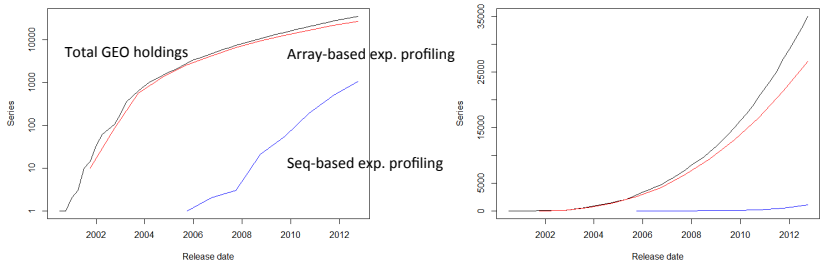
Abundance precision
Sensitivity



Mapping accuracy
Transcript structure

Diff. ex. Power

Still a lot more microarray data in public database, but RNA-seq is growing fast



Update Feb 2014: 2295 RNA-seq, 33256 microarray studies
Feb 2015: 4113 RNA-seq, 38622 microarray studies
Caveat: RNA-seq studies tend to have fewer replicates

My analysis of counts from http://www.ncbi.nlm.nih.gov/geo/

Sample preparation: don't sequence what you don't care about

- Transcriptome composed of large number of RNA species: mRNA and non-coding RNA (rRNA, tRNA, microRNA, other ncRNAs,…)
- Can start with:
  - total RNA
  - polyA selection
  - small RNA
- Ribosomal RNA constitutes >90% of transcripts, so have to deplete them
  - e.g. "RiboMinus" (Invitrogen), "Ribo-Zero" (Epicentre)

RNA-seq variants

- **Paired-end sequencing (pretty standard)**
  - RNA fragments are size-selected
  - Both ends of fragment sequenced
  - Allows for better alignment across repetitive regions
  - Easier identification of structural rearrangements
- **Strand-specific sequencing (common)**
  - By default, we don't get strand info
  - Strand specific variants are more complicated but useful for assessing overlapping transcripts, antisense transcription, non-coding RNAs
- **Molecule-specific barcodes (newer)**
  - Each RNA molecule is tracked; reduces effect of library preparation
  - Especially interesting when applied to single cell sequencing

Sequencing platforms

Illumina HiSeq

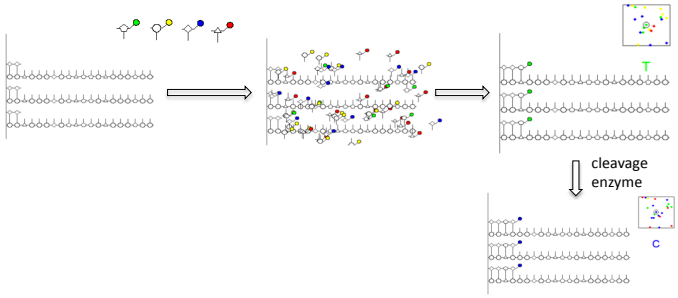
Applied Biosystems SOLiD

Roche 454

Illumina; BGI; http://cage.unl.edu

Illumina/Solexa reversible terminator sequencing

- Introduced in 2006
- Sequences up to ~250bp as of 2014 (x 2 for paired ends)
- Hundreds of millions of fragments at once
- Per-base sequence error rate ~1%

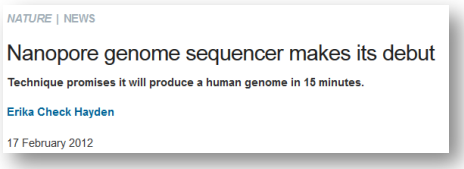


Third generation sequencing

Single molecule sequencing

- (e.g. Pac. Bio., Oxford Nanopore)
- Potential for very long reads (>1kb)
- Higher base-call error rates (20% or more)

Not quite ready for prime time


NATURE | NEWS

Nanopore genome sequencer makes its debut

Technique promises it will produce a human genome in 15 minutes.

Erika Check Hayden

17 February 2012


NATURE | NEWS

Data from pocket-sized genome sequencer unveiled

Results from Oxford Nanopore's MinION are promising, but fall short of high expectations.

Erika Check Hayden

14 February 2014

http://www.nature.com/news/data-from-pocket-sized-genome-sequencer-unveiled-1.14724

Processing RNA-seq data

- Sequence file format; quality assessment
- Alignment/assembly

Raw data: FASTQ format

Text file; Sequences with quality information; ~60GB file for 500M paired-end reads, 100bp

Data for one read:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

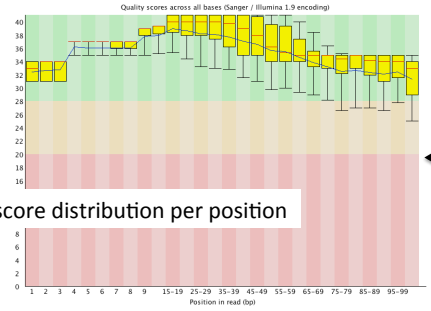
```
@HWI-ST909_0091:8:2207:10507:60108#0/1
```

example of Illumina sequence ID

- unique instrument name
- flow cell lane
- tile number
- x and y coordinates of the cluster
- index number
- member of a pair

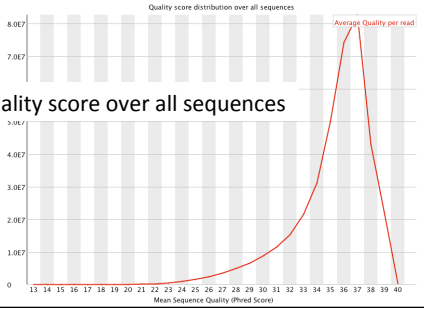
http://en.wikipedia.org/wiki/FASTQ_format

Sequence quality: FASTQC



Quality score distribution per position

Phred quality Q20 = 1 in 100 chance of being incorrect



Mean quality score over all sequences

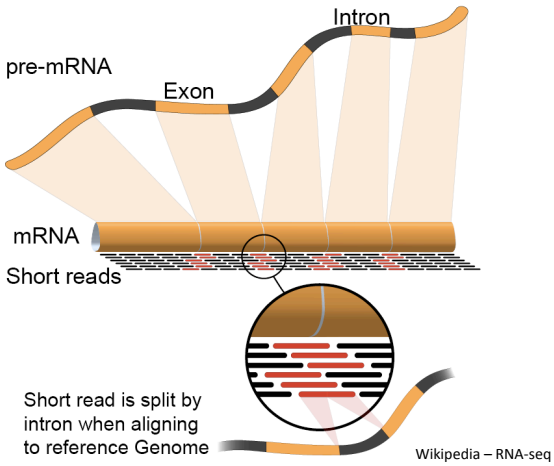
Also:

- GC content
- sequence length distribution
- sequence duplication level

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

Aligning raw reads

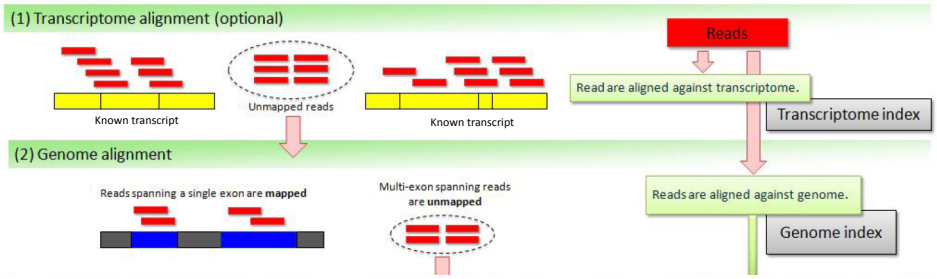
Challenges: junction-spanning reads, pseudogenes, genome variation, sequencing errors



Spliced sequence aligners

- TopHat, GSNAP, STAR, GSTRUCT…
- Different ways to deal with splice junctions
  - Align to the annotated transcriptome – good for well-annotated genomes, can be augmented by:
  - *de novo* splice junction discovery by alignment to the genome
  - Or without any reference: "*de novo* assembly" e.g. Trans-Abyss
- Challenge: speed and memory usage
  - Some trade-off between speed and sensitivity
  - Parallelization important
- No consensus yet on best method; see http://www.nature.com/nmeth/journal/v10/n12/full/nmeth.2714.html

Spliced alignment steps pt. 1

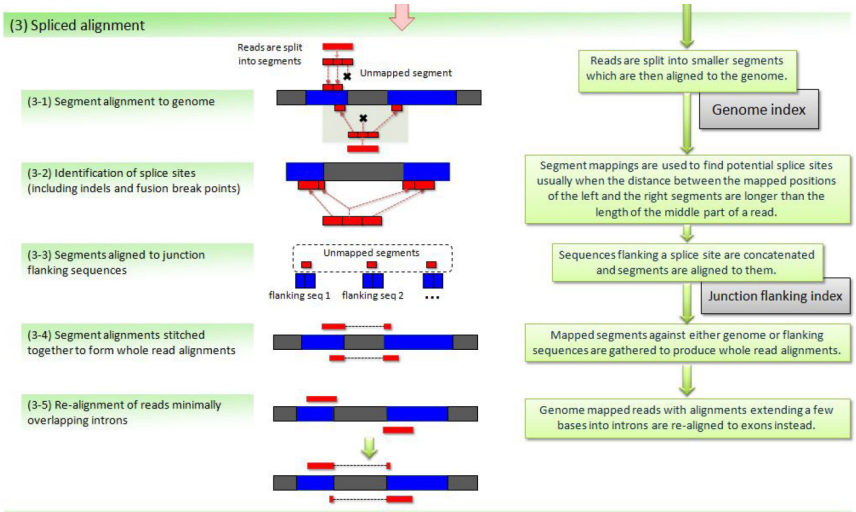


Genome Biology
METHOD Open Access

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

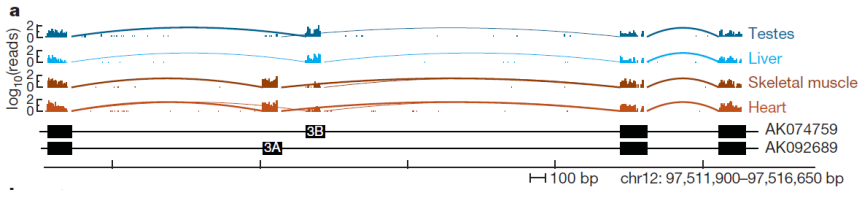
Daehwan Kim[1,2,3*], Geo Pertea[3], Cole Trapnell[4,5], Harold Pimentel[2], Ryan Kelley[6] and Steven L Salzberg[3,4]

Spliced alignment steps pt. 2



Gene/transcript quantification

- Gene quantification is relatively simple
  - Basically, look at what gene is at the point the sequence aligned. Assign read to that gene.
- Transcript quantification is much harder
  - Have to guess which transcript a read came from
  - In the face of uncertainty of not knowing the transcript structures present



http://www.nature.com/nature/journal/v456/n7221/abs/nature07509.html

Transcript quantification

Recent evaluation: Steijger et al. Nature Methods 10, 1177–1184 (2013): Cufflinks, Exonerate, GSTRUCT

- Correlation with a gold standard was 0.34-0.68
- "Expression-level estimates can vary considerably [across methods]" and "each [method] failed to report a subset of exons and junctions despite the availability of adequate RNA-seq alignments"

"The complexity of higher eukaryotic genomes imposes severe limitations on transcript recall and splice product discrimination that are likely to remain limiting factors for the analysis of current-generation RNA-seq data."

Aligned data – SAM/BAM format

- SAM - tab-delimited text file with alignment info
- BAM – binary version

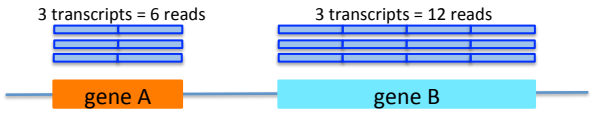
- Indexing for quick data retrieval by genome coordinate
- *Relatively* compact, but BAM files can still be 30Gb or more per RNA-seq sample

Data analysis

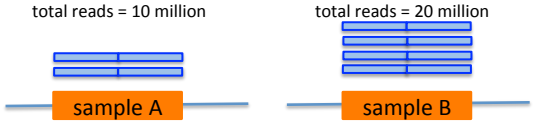
- Normalization
- Differential expression (next lecture)

Normalization

- Enables comparison of expression between and within samples
- **Within-sample comparison** – at the same expression level, longer transcripts have more read counts



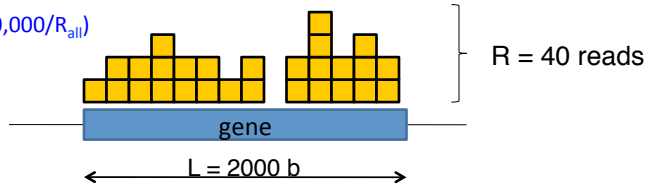
- **Between samples** – higher counts at higher sequencing depths



RPKM – normalizing across genes (and libraries)

- Number of reads mapping to a gene will depend on total number of reads sequenced and length of a gene
- RPKM - reads per kb of exon model per million mapped reads
- A.K.A. **FPKM** ("fragment" so "read" includes "paired read")
- Gene length correction doesn't matter for inter-sample comparisons

$RPKM=(R/L)*(1{,}000{,}000/R_{all})$

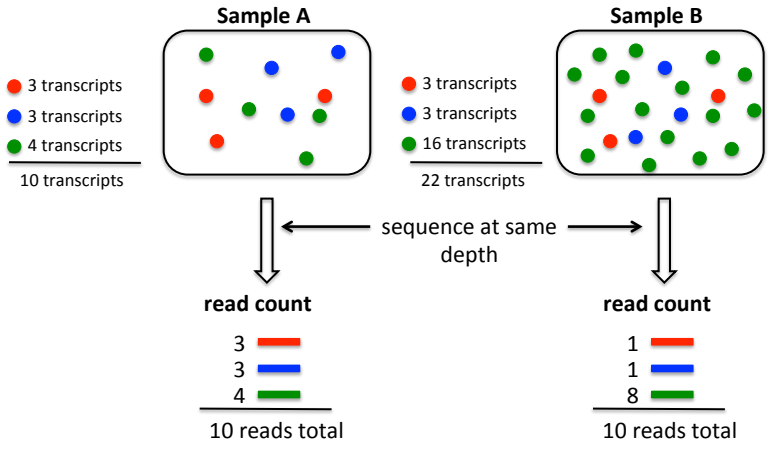


- If there were R_{all}= 1 million reads mapped to the genome then this exon would have 20 RPKM
- If there are 20 millions mapped reads = 1 RPKM

Inter-sample normalization

- Important for comparisons across samples
  - In contrast to estimating relative abundance within a sample.
- Key concept: "Sequencing space"
- Finite # of reads implies that measuring one gene always decreases ability to measure another

Effect of RNA population composition



⇨ red and blue gene appear to be down-regulated in sample B

Normalization in DEseq

for samples *k* and *k'*, estimate normalization factor f_K

$$f_k = \underset{g}{median} \frac{Y_{gk}}{Y_{gk'}}$$

Y_g –count of gene *g*

for *n* samples:

$$f_k = \underset{g}{median} \frac{Y_{gk}}{\left(\prod_{i=1}^{n} Y_{gi}\right)^{1/n}}$$

← pseudo-reference sample

Also implemented in edgeR

Anders and Huber Genome Biol. 2010;11(10):R106

"Trimmed Mean of M component" (TMM) normalization

- Same idea as DESeq normalization, a little more complicated:

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

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

Y_{gk} – count of gene *g* in sample k
N_k – total # of reads in sample k
G* - set of genes after trimming

- Mean instead of median; weighted by est. var.
- Trimming: exclude genes for which expression level or M is too high or low
- Default in edgeR (calcNormFactors)

Robinson and Oshlack Genome Biol. 2010;11(3):R25.