Gene/transcript quantification

Mikhail Dozmorov

Spring 2018

RNA-seq statistical problems

- Summarization
- Normalization
- Differential expression testing
- Isoform expression estimation

Summarization of read counts

- From RNA-seq, the alignment result gives the chromosome/position of each aligned read.
- For a gene, there are reads aligned to the gene body. How to summarized them into a number for the expression?

Counts of reads

- Easiest: The relative expression of a transcript is proportional to the number of cDNA fragments that originate from it $\sim$ number of aligned reads.
- Disadvantages: longer gene produce more reads, library depth (total counts) influence counts of individual transcripts

Expression estimation for known genes and transcripts

- **HTSeq** - set of tools for analysing high-throughput sequencing data with Python
- `htseq-count` command line tool for counting reads in features

```
htseq-count --mode intersection-strict --stranded no
  --minaqual 1 --type exon --idattr transcript_id
  accepted_hits.sam chr22.gff > transcript_counts.tsv
```

https://htseq.readthedocs.io/en/release_0.9.1/count.html

Issues with `htseq-count`: http://seqanswers.com/forums/showthread.php?t=18068

Expression estimation for known genes and transcripts

- **featureCounts**, Summarize multiple datasets at the same time

```
featureCounts -t exon -g gene_id -a annotation.gtf
  -o counts.txt library1.bam library2.bam library3.bam
```

http://bioinf.wehi.edu.au/featureCounts/

Why is simple counting for transcript quantification not sufficient?

Each gene has multiple exonsStraightforward approaches

- **Union** - treat a gene as the union of its exons
- **Intersection** - treat a gene as the intersection of its exons

Problems

- Cannot correct for positional biases / insert length distributions since they don't model which transcript reads come from
- Intersection may throw away many reads
- Many more sophisticated approaches: Cufflinks (Trapnell, 2010), RSEM (Li, 2010), TIGAR (Nariai, 2014), eXpress (Roberts, 2013), Sailfish (Patro, 2014), Kallisto (Bray, 2015), more...

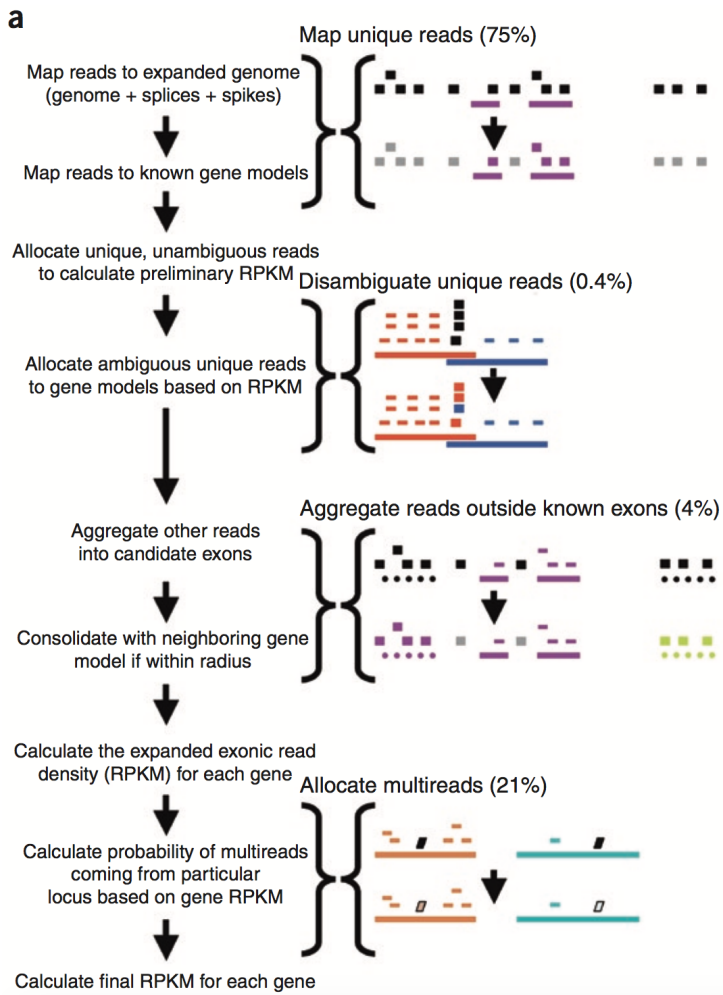
Trapnell et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." Nature Biotechnology 31 (2013): 46-53.

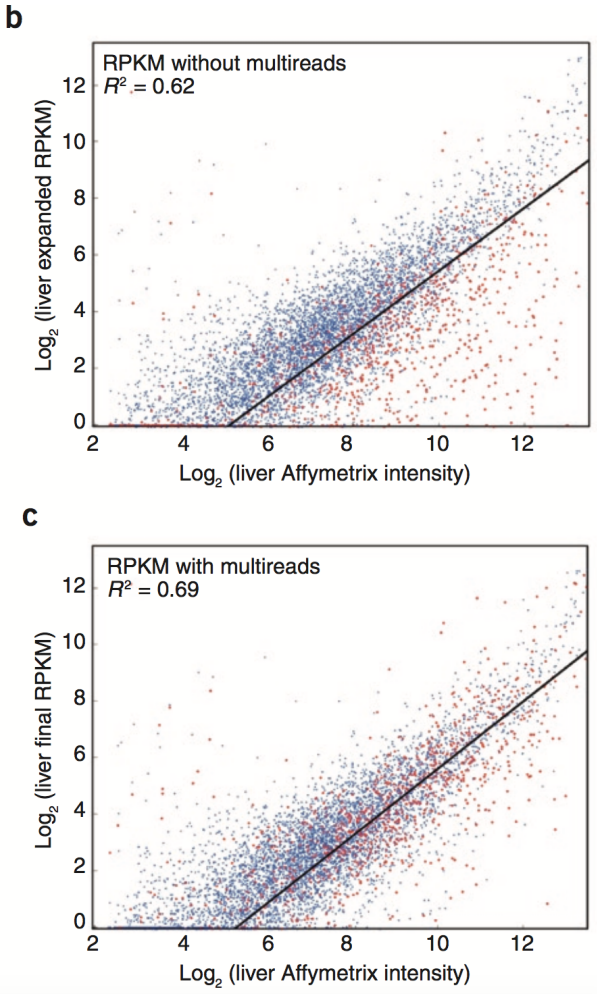
Multimapped reads

- A significant percentage of reads (up to 30% from the total mappable reads) are mapped to multiple locations (multireads) due to gene homology or low complexity.
- If the multireads are discarded, the expression levels of genes with homologous sequences will be artificially deflated
- If the multireads are split randomly amongst their possible loci, differences in estimates of expression levels for these genes between conditions will also be diminished leading to lower power to detect differential gene expression

Multimapped reads

- A heuristic solution
- divide the multireads amongst their mapped regions according to the distribution of the uniquely mapped reads in those regions.





Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." Nature Methods 5, no. 7 (July 2008): 621–28. https://doi.org/10.1038/nmeth.1226.

Multimapped reads

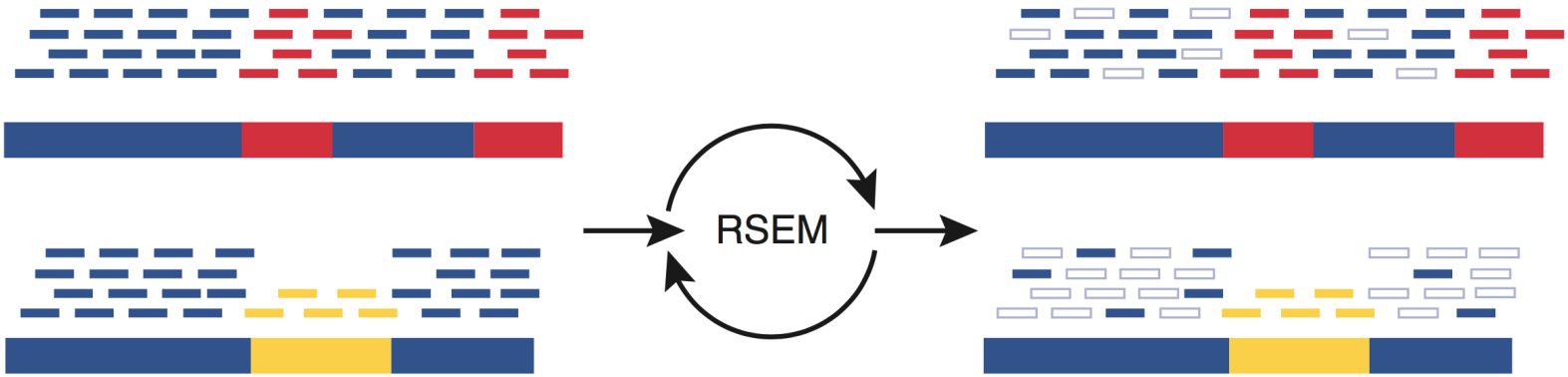
- More sophisticated approaches, e.g., considering genetic variation, have been developed

Hashimoto, Takehiro, Michiel J. L. de Hoon, Sean M. Grimmond, Carsten O. Daub, Yoshihide Hayashizaki, and Geoffrey J. Faulkner. "Probabilistic Resolution of Multi-Mapping Reads in Massively Parallel Sequencing Data Using MuMRescueLite." Bioinformatics (Oxford, England) 25, no. 19 (October 1, 2009): 2613–14. https://doi.org/10.1093/bioinformatics/btp438.

Paşaniuc, Bogdan, Noah Zaitlen, and Eran Halperin. "Accurate Estimation of Expression Levels of Homologous Genes in RNA-Seq Experiments." Journal of Computational Biology 18, no. 3 (March 2011): 459–68. https://doi.org/10.1089/cmb.2010.0259.

RSEM - RNA-Seq by Expectation-Maximization

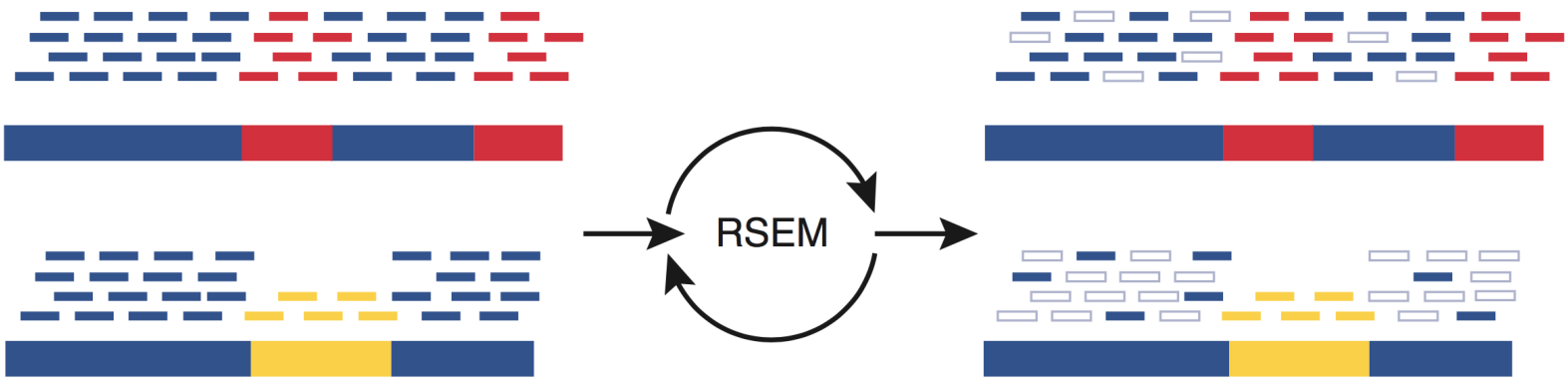
- Abundance estimation for two transcripts (long bars) with shared (blue) and unique (red, yellow) sequences
- RNA-seq reads (short bars) are first aligned to the transcript sequences (long bars, bottom)
- Unique regions of isoforms will capture uniquely mapping RNA-seq reads (red and yellow short bars), and shared sequences between isoforms will capture multimapped reads (blue short bars).



Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, et al. "De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." Nature Protocols 8, no. 8 (August 2013): 1494–1512. https://doi.org/10.1038/nprot.2013.084.

RSEM - RNA-Seq by Expectation-Maximization

- An expectation maximization algorithm **estimates the most likely relative abundances of the transcripts**
- Then, it **fractionally assigns reads to the isoforms based on these abundances.**
- The assignments of reads to isoforms resulting from iterations of expectation maximization are illustrated as filled short bars (right), and eliminated assignments are shown as hollow bars.



Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, et al. "De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." Nature Protocols 8, no. 8 (August 2013): 1494–1512. https://doi.org/10.1038/nprot.2013.084.

Data normalization

- Data from different samples need to be normalized so that they are comparable.
- Most important – sequencing depth: sample with more total counts will have more counts in each gene on average.
- Easiest method: divide by the total number of counts

Expression estimation for known genes and transcripts

- **Counts per million**: counts scaled by the library depth in million units. $CPM = C * 10^6 / N$
- **RPKM**: Reads Per Kilobase of transcript per Million mapped reads. Introduced by Mortazavi, 2008
- **FPKM**: Fragments Per Kilobase of transcript per Million mapped reads. Introduced by Salzberg, Pachter, 2010

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628 (2008).

Expression estimation for known genes and transcripts

- **FPKM** (or **RPKM**) attempt to normalize for gene size and library depth

$$RPKM\ (or\ FPKM_i) = (10^9 * C_i)/(N * L_i)$$

- C_i - number of mappable reads/fragments for a i gene/transcript/exon/etc.
- N - total number of mappable reads/fragments in the library
- L_i - number of base pairs in the i gene/transcript/exon/etc.

https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

TPM: Transcript per Kilobase Million

- **TPM**: Transcripts per million. Introduced by Li, 2011. Normalized by total transcript count instead of read count in addition to average read length.

If you were to sequence one million full length transcripts, TPM is the number of transcripts you would have seen for transcript i.

$$TPM_i = 10^6 * Z * \frac{C_i}{N * L_i}$$

- Z - sum of all length normalized transcript counts

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26, 493–500 (2010).

TPM: Transcript per Kilobase Million

FPKM is calculated as

1. Sum sample/library fragments per million
2. Divide gene/transcript fragment counts by #1 – fragments per million, FPM
3. Divide FPM by length of gene in kilobases (FPKM)

TPM reverses the order - length first, library size second

1. Divide fragment count by length of transcript – fragments per kilobase, FPK
2. Sum all FPK for sample/library per million
3. Divide #1 by #2 (TPM)

https://youtu.be/TTUrtCY2k-w?t=23

https://www.ncbi.nlm.nih.gov/pubmed/22872506

Alignment-free methods: kallisto

- Use transcriptome to estimate probability of a read being generated by a transcript
- Hashing technique and pseudoalignment via the transcript-specific Target de Bruijn Graphs
- 500-1,000x faster than previous approaches. RNA-seq analysis of 30 million reads takes ~2.5 minutes
- Speed allows for bootstrapping to obtain uncertainty estimates, thus leading to new methods for differential analysis

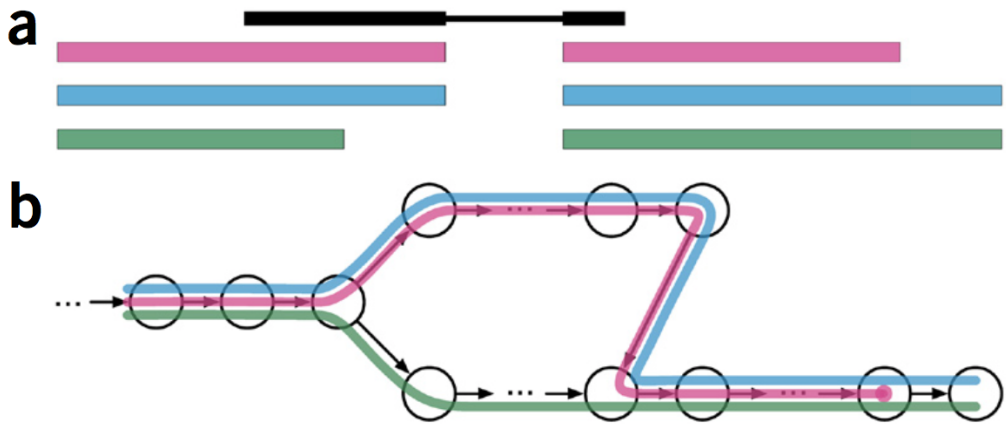
Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. "Near-Optimal Probabilistic RNA-Seq Quantification." Nature Biotechnology 34, no. 5 (May 2016): 525–27. https://doi.org/10.1038/nbt.3519.

https://pachterlab.github.io/kallisto/

https://liorpachter.wordpress.com/2015/05/10/near-optimal-rna-seq-quantification-with-kallisto/

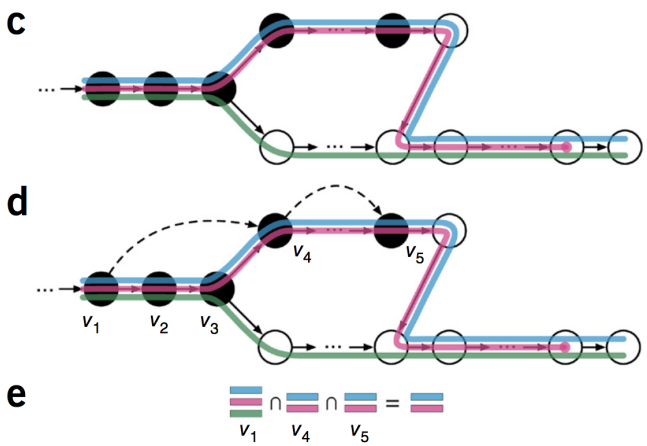
kallisto: Target de Bruijn Graph (T-DBG)

- Create every k-mer in the transcriptome (k=31), build de Bruijn Graph and color each k-mer
- Preprocess the transcriptome to create the T_DBG
- Index (fast)



kallisto: Target de Bruijn Graph (T-DBG)

- Use k-mers in read to find which transcript it came from
- Want to find pseudoalignment - which transcripts the read (pair) is compatible with (**not** an alignment of the nucleotide sequences)
- Can jump over k-mers that provide the same information - ~8x speedup over checking all k-mers
- Each k-mer appears in a set of transcripts
- The intersection of all sets is our pseudoalignment



sleuth: Differential analysis of RNA-seq incorporating quantification uncertainty

- Uses `kallisto` for transcripts quantification
- Separates the between-sample variability into two components:
  - **'biological variance'** that arises from differences in expression between samples as well as from variability due to library preparation
  - **'inferential variance'** which includes differences arising from computational inference procedures in addition to measurement 'shot noise' arising from random sequencing of fragments.
- Differential expression using an extension of the general linear model where the total error has two additive components.

https://pachterlab.github.io/sleuth/, https://liorpachter.wordpress.com/2015/08/17/a-sleuth-for-rna-seq/

Pimentel, Harold J, Nicolas Bray, Suzette Puente, Páll Melsted, and Lior Pachter. "Differential Analysis of RNA-Seq Incorporating Quantification Uncertainty." BioRxiv, 2016, 058164. https://www.nature.com/articles/nmeth.4324.

Sailfish: Ultrafast Gene Expression Quantification

- Fast expectation maximization algorithm
- Uses small data atoms rather than long sequences
- More tolerant of genetic variation between individuals
- Extremely parallelized

Patro, Mount, Kingsford, Nature Biotech, 2014. https://www.nature.com/articles/nbt.2862

https://www.cs.cmu.edu/~ckingsf/software/sailfish/

Salmon: fast & accurate method for RNA-seq-based quantification

- Pseudo-alignment, or using precomputed alignment to transcriptome
- Dual-phase statistical inference procedure
- Uses sample-specific bias models that account for sequence-specific, fragment, GC content, and positional biases
- Includes its own aligner RapMap, or can take transcriptome-mapped BAM files

Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." Nature Methods 14, no. 4 (March 6, 2017): 417–19. https://doi.org/10.1038/nmeth.4197.

https://github.com/COMBINE-lab/Salmon

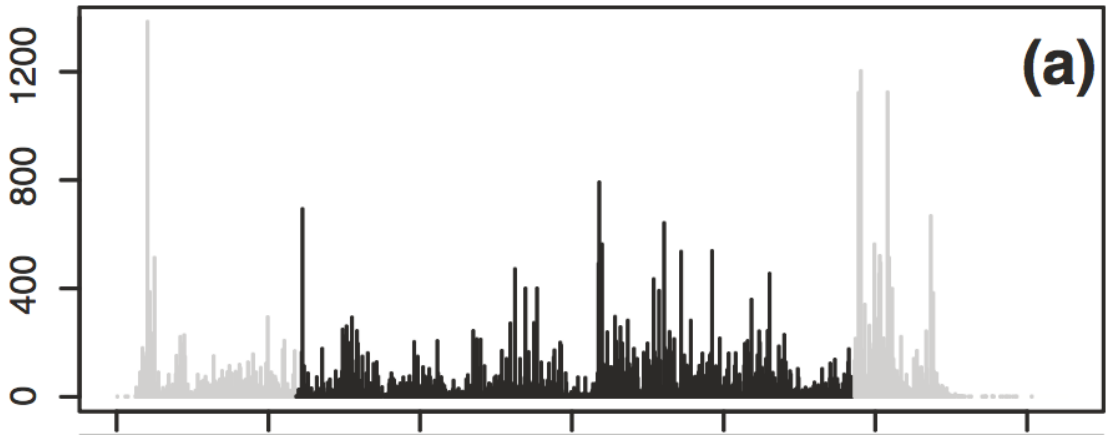
How to get the estimated values into R?

- tximport - Import and summarize transcript-level estimates for transcript- and gene-level analysis
- Imports transcript-level abundance, estimated counts and transcript lengths, and summarizes into matrices for use with downstream gene-level analysis packages.
- Average transcript length, weighted by sample-specific transcript abundance estimates, is provided as a matrix which can be used as an offset for different expression of gene-level counts.

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

Artifacts in the reads distribution

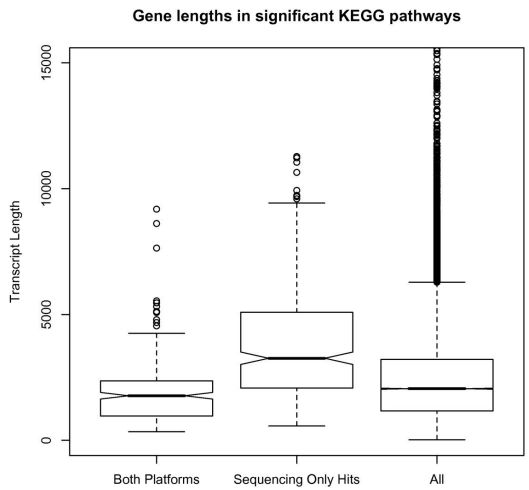
- The reads are NOT uniformly distributed within gene bodies. Think affinity of probes on a microarray.
- Need to account for read position to quantify gene expression from read counts
- Example: Counts of reads along gene *Apoe*



Li et al. Genome Biology 2010, https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-5-r50

Preferential sequencing for longer reads

- Longer genes produce more reads, have more chance to be sequenced
- The ability to call differentially expressed genes is strongly associated with the length of the transcript
- Example: Length of differentially expressed genes detected by microarray and RNA-seq (first box), by RNA-seq only (second box), and for all genes common for both platforms (third box)

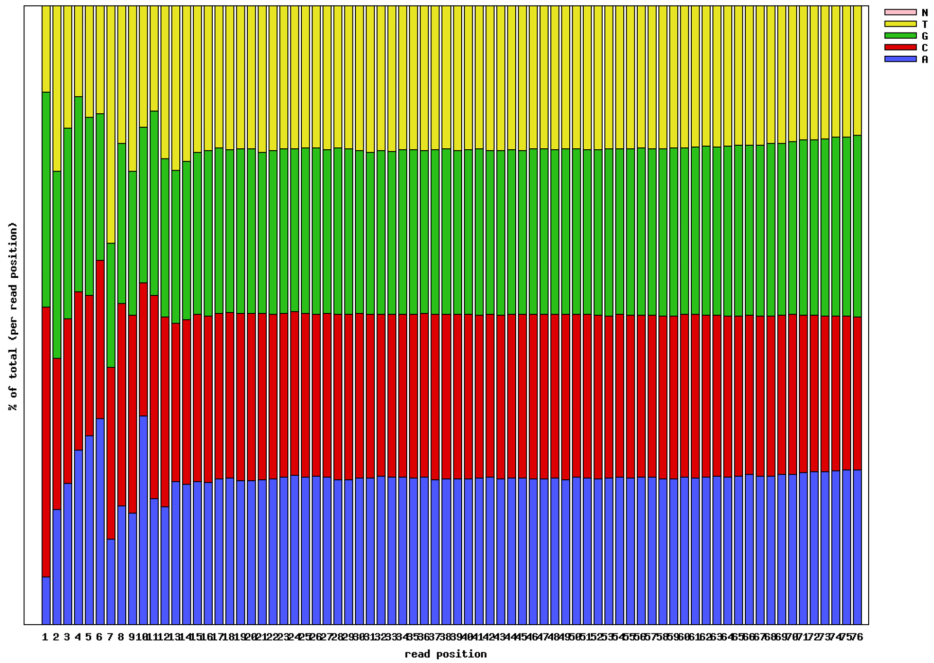


Olshak et.al. 2009 (Biology Direct) https://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-4-14

Need to weight counts by gene length in differential analysis settings. Bullard et.al. 2010 (BMC Bioinformatics) https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-94#Bib1

Random hexamer priming

- Single-stranded RNA needs to be converted to cDNA for sequencing
- Need a small double-stranded segment for reverse transcriptase to achieve the conversion
- Random hexamers - small single-stranded DNA stretches six nucleotides long - which bind RNA at random places
- Random priming is not so random



Downweighting hexamer effect

- Discovered that reads from Illumina have a 7bp motif at beginning: there are more reads started with certain 7-bp due to technical artifacts (the random priming bias).
- The effect stretches up to position 12-13.
- Downweight the reads started with the motif

$$w(h) = \frac{\frac{1}{6}\sum_{i=24}^{29} \hat{p}_{hep:i}(h)}{\frac{1}{2}(\hat{p}_{hep:1}(h) + \hat{p}_{hep:2}(h))}$$

- Read length is at least 35bp
- $w(h)$ - weights for reads starting with heptamer h
- $\hat{p_{hep:i}}$ - observed proportion of heptamers (7 bases) starting at position i.

Hansen et al. 2010 NAR, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896536/

Example: Downweighting hexamer effect

Table 1. Data from a small genomic region in the sense strand of the YOL086C gene in *S. cerevisiae*

Strand	Location l	Heptamer $h(l)$	Count $c(l)$	Weight $w(h(l))$	Reweighted count $c_w(l)$
...					
−1	159792	TTGGTCG	17	1.39	23.6
−1	159793	TTTGGTC	17	0.25	4.3
−1	159794	TTTTGGT	65	0.31	20.4
−1	159795	GTTTTGG	72	0.32	23.3
−1	159796	CGTTTTG	10	1.66	16.6
...					

$c(l)$ denotes the number of mapped reads starting at a particular (stranded) location l and $h(l)$ is the unique heptamer associated with this location. $w(h(l))$ are weights such as in Equation (1) and $c_w(l) = c(l)w(h(l))$ are the location-specific reweighted counts. For this particular small genomic region, reweighting makes the counts more comparable between different locations. Data from the WT experiment.

Downweighting hexamer effect

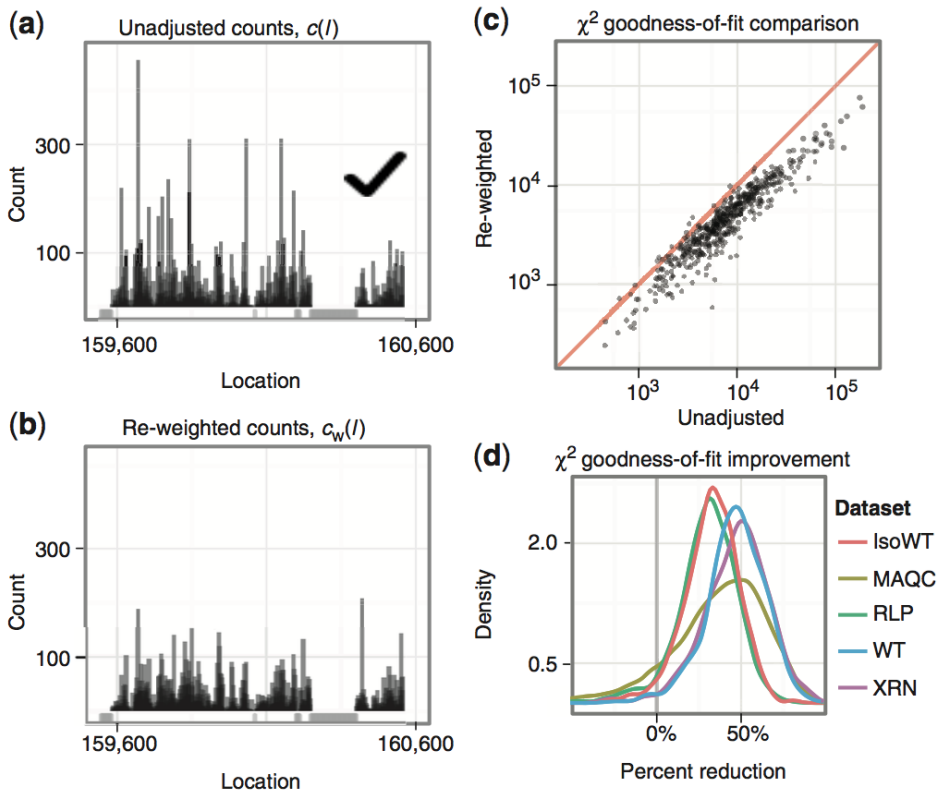


Figure 4. Evaluation of the reweighting scheme. (**a** and **b**) Unadjusted and re-weighted base-level counts for reads from the WT experiment mapped to the sense strand of a 1-kb coding region in *S. cerevisiae* (YOL086C). The graey bars near the *x*-axis indicate unmappable genomic locations. (**c**) The χ^2 goodness-of-fit statistics based on unadjusted and reweighted counts for 552 highly expressed regions of constant expression. (**d**) Smoothed histograms of the reduction in χ^2 goodness-of-fit statistics when using the re-weighting scheme, evaluated in five different experiments. Values greater than zero indicate that the re-weighting scheme improves the uniformity of the read distribution.