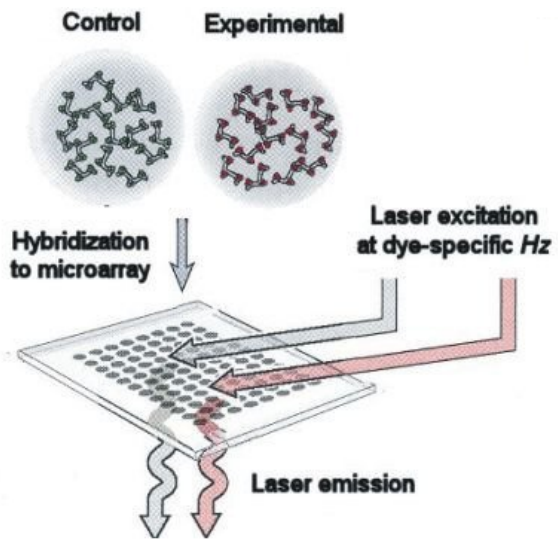
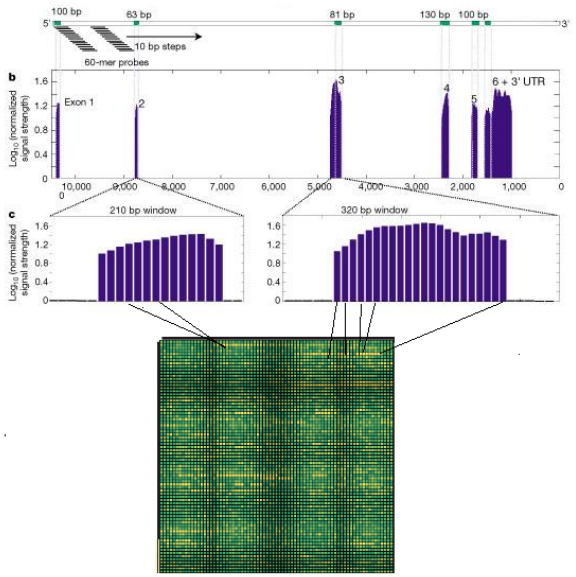
RNA-seq and Gene Expression

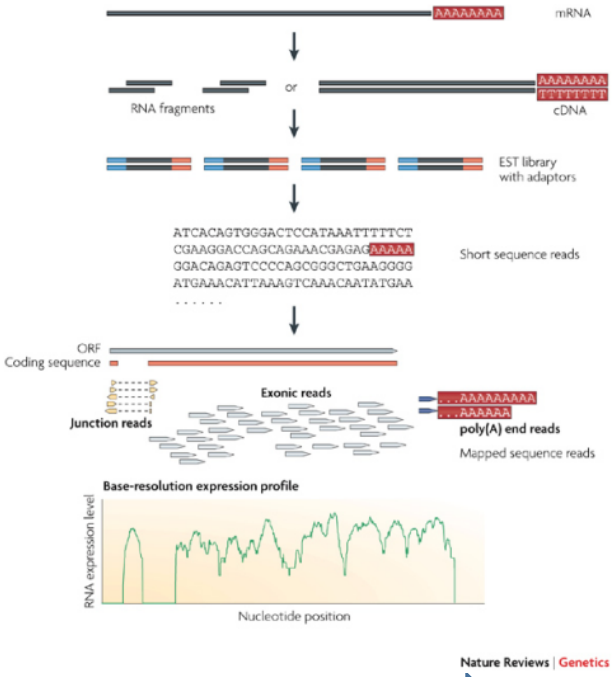
Gail Rosen

The evolution of transcriptomics

Hybridization-based





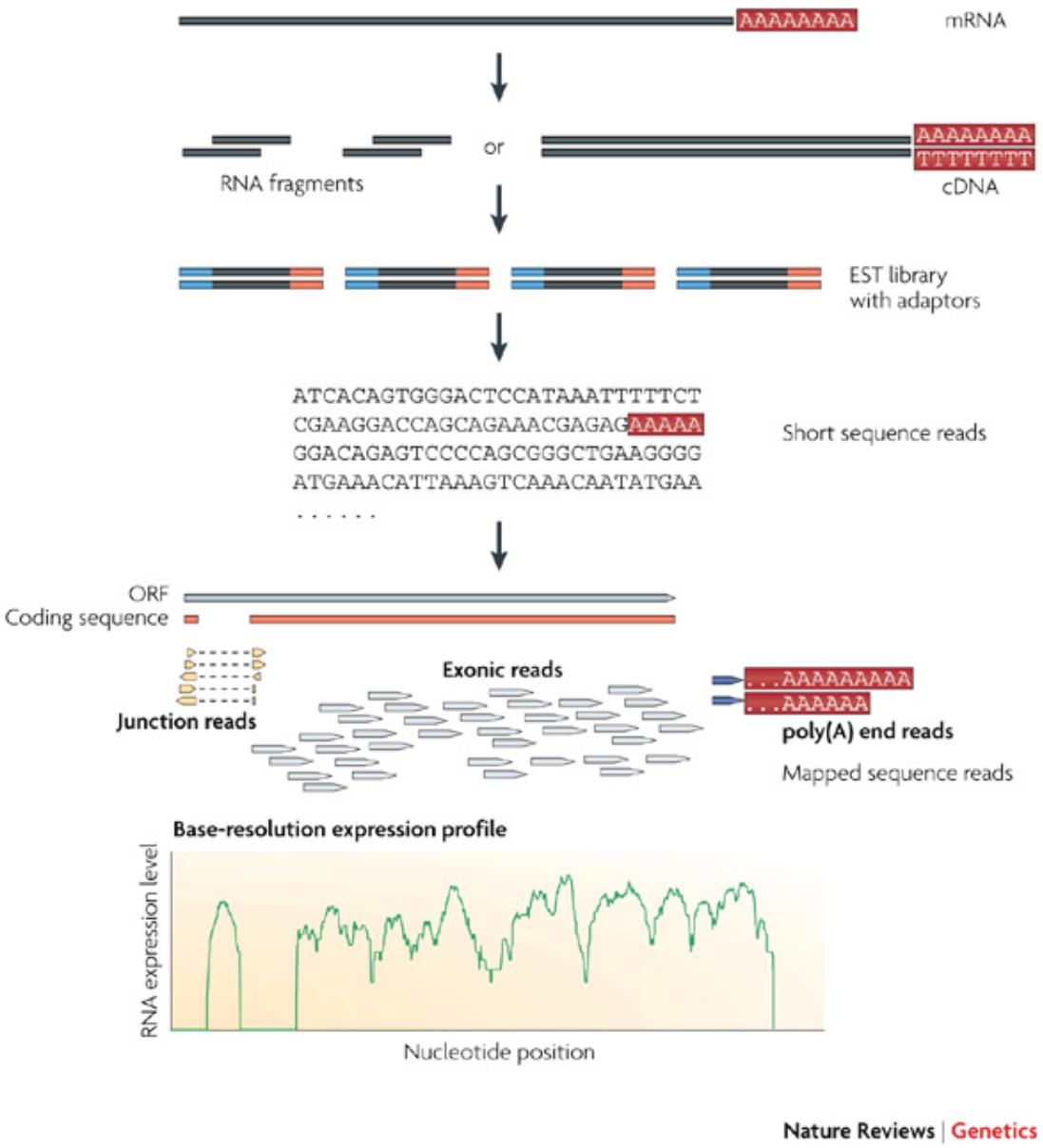


RNA-seq is still a technology under active development

1995 P. Brown, et. al. Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)



Sample preparation

Next generation sequencing (NGS)

Data analysis:
- ✓Mapping reads
- ✓Visualization (Gbrowser)
- ✓De novo assembly
- ✓Quantification

Figure from Wang et. al, **RNA-Seq: a revolutionary tool for transcriptomics,** Nat. Rev. Genetics 10, 57-63, 2009)**.**

RNA-seq vs. microarray

- RNA-seq can be used to characterize novel transcripts and splicing variants as well as to profile the expression levels of known transcripts (but hybridization-based techniques are limited to detect transcripts corresponding to known genomic sequences)
- RNA-seq has **higher resolution** than whole genome tiling array analysis
  - In principle, mRNA can achieve single-base resolution, where the resolution of tiling array depends on the density of probes
- RNA-seq can apply the same experimental protocol to various purposes, whereas specialized arrays need to be designed in these cases
  - Detecting single nucleotide polymorphisms (needs SNP array otherwise)
  - Mapping exon junctions (needs junction array otherwise)
  - Detecting gene fusions (needs gene fusion array otherwise)

– Next-generation sequencing (NGS) technologies are now challenging microarrays as the tool of choice for genome analysis.

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Data Analysis of RNA-seq for metagenomics

- Mapping reads to the *reference* genome
  - Read mapping of 454 sequencers can be done by conventional sequence aligners (BLAST, BLAT, etc)
  - Short read aligner needed for Illumina or SOLiD reads
- Quantifying the known genes
- Prediction of novel transcripts
  - Assembly of short reads: comparative vs. *de novo*

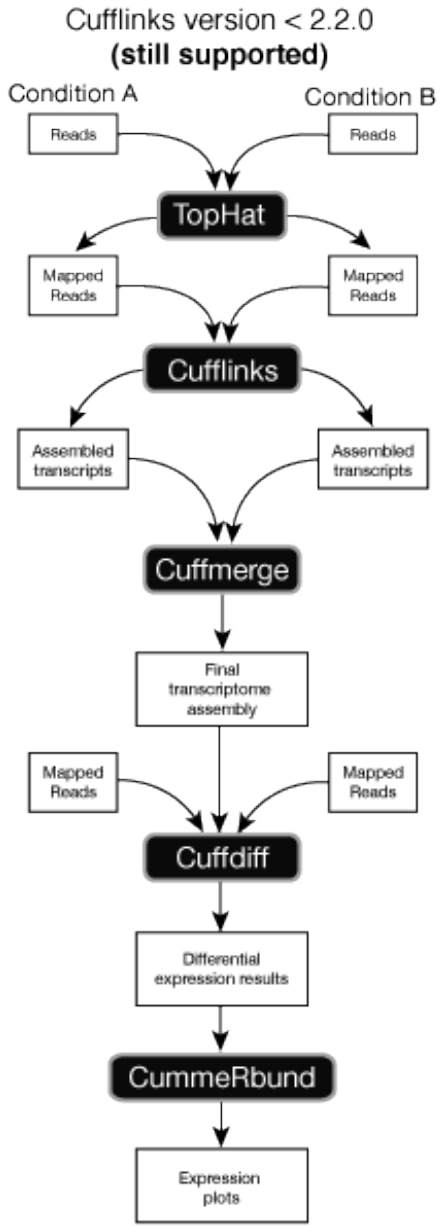
Prediction of Novel Transcripts: Assembling Short Reads

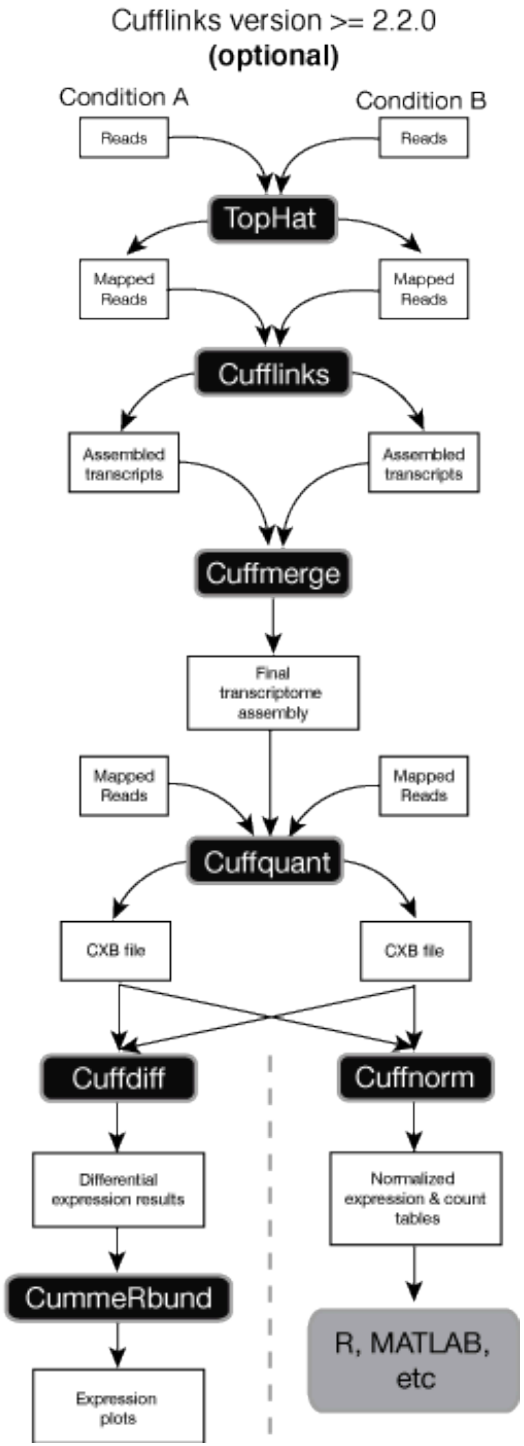
- *De novo* assembly: without using a reference genome
  - Splicing graph approach: Heber, et. al. ISMB 2002 for EST assembly
  - K-mer based approach, working efficiently for short reads (e.g. Velvet, ALLPATH and EULER-ESR)

Challenges: mapping reads to reference genome

- Sequencing errors and polymorphisms
- Repetitive sequences: a significant portion of sequence reads match multiple locations in the genome
  - Obtaining longer sequence reads, or paired-end sequencing strategy, should help alleviate the multi-matching problem.

Cufflinks


Cufflinks version < 2.2.0
(still supported)
Condition A
Condition B
Reads
Reads
TopHat
Mapped Reads
Mapped Reads
Cufflinks
Assembled transcripts
Assembled transcripts
Cuffmerge
Final transcriptome assembly
Mapped Reads
Mapped Reads
Cuffdiff
Differential expression results
CummeRbund
Expression plots


Cufflinks version >= 2.2.0
(optional)
Condition A
Condition B
Reads
Reads
TopHat
Mapped Reads
Mapped Reads
Cufflinks
Assembled transcripts
Assembled transcripts
Cuffmerge
Final transcriptome assembly
Mapped Reads
Mapped Reads
Cuffquant
CXB file
CXB file
Cuffdiff
Cuffnorm
Differential expression results
Normalized expression & count tables
CummeRbund
R, MATLAB, etc
Expression plots

FPKM

- Fragments Per Kilobase of sequence per Million fragments mapped

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

Xi – number of counts of reads that align to a genomic region i

~li – effective length of genomic region i, calculated by $\tilde{l}_i = l_i - \mu_{FLD} + 1$,

Mean fragment length distribution (mean of aligned reads)

N – number of fragments that were sequenced

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

Counts

Rely on:

- The amount of fragments you sequenced (related to relative abundances)
- (Effective) Length of the feature

Effective length is the number of possible start sites a feature could have generated a fragment of that particular length

Counts Per Million

$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

Within Sample Normalization

- The number of fragments you see from a feature depends on its length
- In order to compare features of different length, you should normalize counts by the length of the feature. Doing so, allows the summation of expression across features to get the expression of a group of features (think a set of transcripts which make up a gene).

No units introduced so far are comparable across experiments.

Advocating use of TPM

$$\mathrm{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}} \right) \cdot 10^6.$$

Normalized to sum of abundance of each transcript

While FPKM is normalized to number of reads sequenced

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j} \right) \cdot 10^6.$$

Comparison of software packages for detecting differential expression in RNA-seq studies

Table 1:
Software packages for detecting differential expression

Method	Version	Reference	Normalization[a]	Read count distribution assumption	Differential expression test
edgeR	3.0.8	[4]	TMM/Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	[5]	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	[6]	Scaling factors (quantile/TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	[7]	RPKM/TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	[8]	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	[9]	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff 2 (Cufflinks)	2.0.2-beta	[10]	Geometric (DESeq-like)/quartile/classic-fpkm	Beta negative binomial distribution	*t*-test
EBSeq	1.1.7	[11]	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods

edgeR and DeSeq

- Use Negative Binomial model to estimate dispersion in the counts…..