Lesson 13: RNASeq and qRT-PCR

Spring 2015

RNASeq

- Short pieces of mRNA transcripts called reads.
- Count how many reads in each sample align with the genome.
- Compare # of counts in differential expression.

RNAseq- Next Generation Sequencing (NGS)

- First generation sequencing- Sanger Sequencing: Sequence of long pieces of DNA or mRNA over 1 Kb long (slow –accurate)
- Next generation sequencing – sequence short fragments reads 50-100 bp long. Map to genome or assemble (fast-many replicates for accuracy). Illumina is the main platform.
- Third generation sequencing- (Fast, long, and accurate – but not cheap yet).

Advantages of RNASeq

- Not limited to genes on chip
- More sensitive
- Better Dynamic Range- a larger range of concentrations reliably
- Does not suffer from cross-hybridization
- Can detect new splice variants

Caveat 1: Not always reliably!

Caveat 2: Relative concentration of splice variants NOT reliable.

Other Uses of NGS

Genome assembly:

Genomes of new species.

Precision medicine.

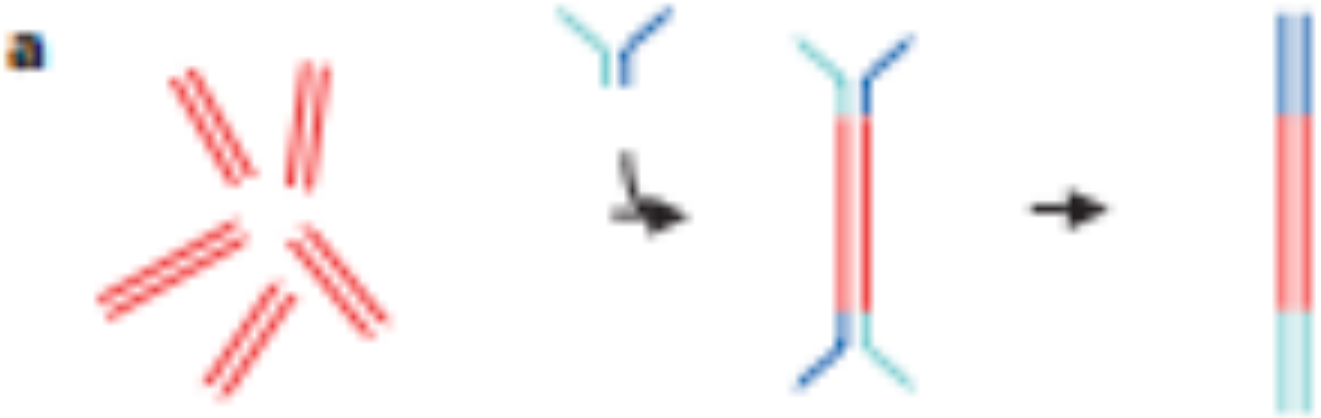
Bacterial, viral, and immunological evolution.

DNAseq: Open Chromatin.

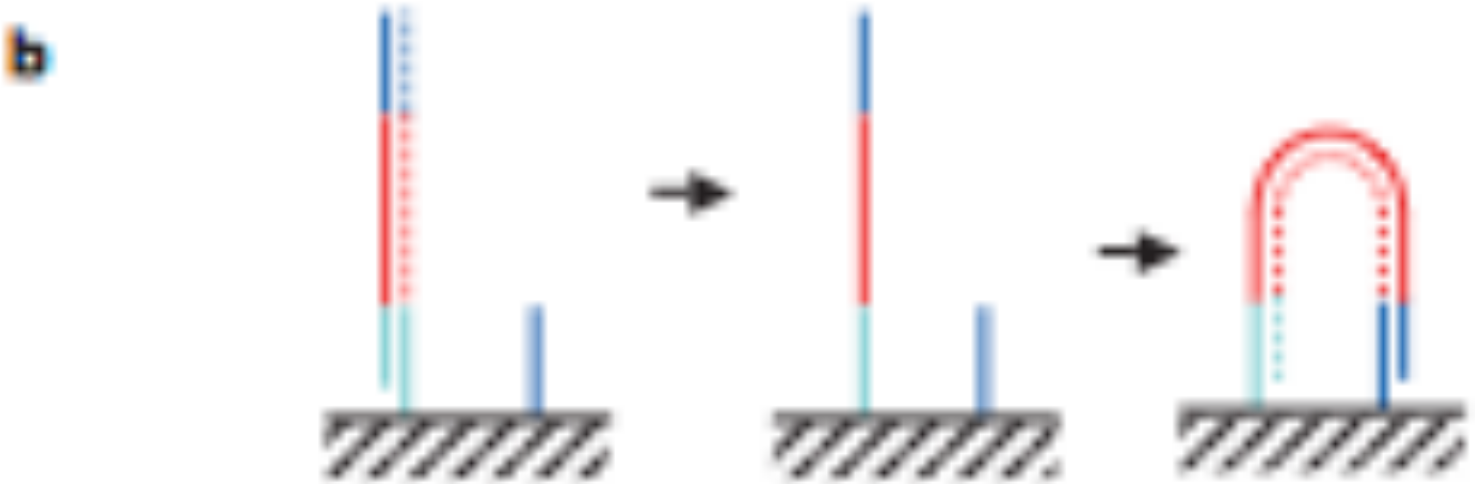
ChipSeq: Transcription-Factor and histone binding.

MethSeq: DNA methylation.

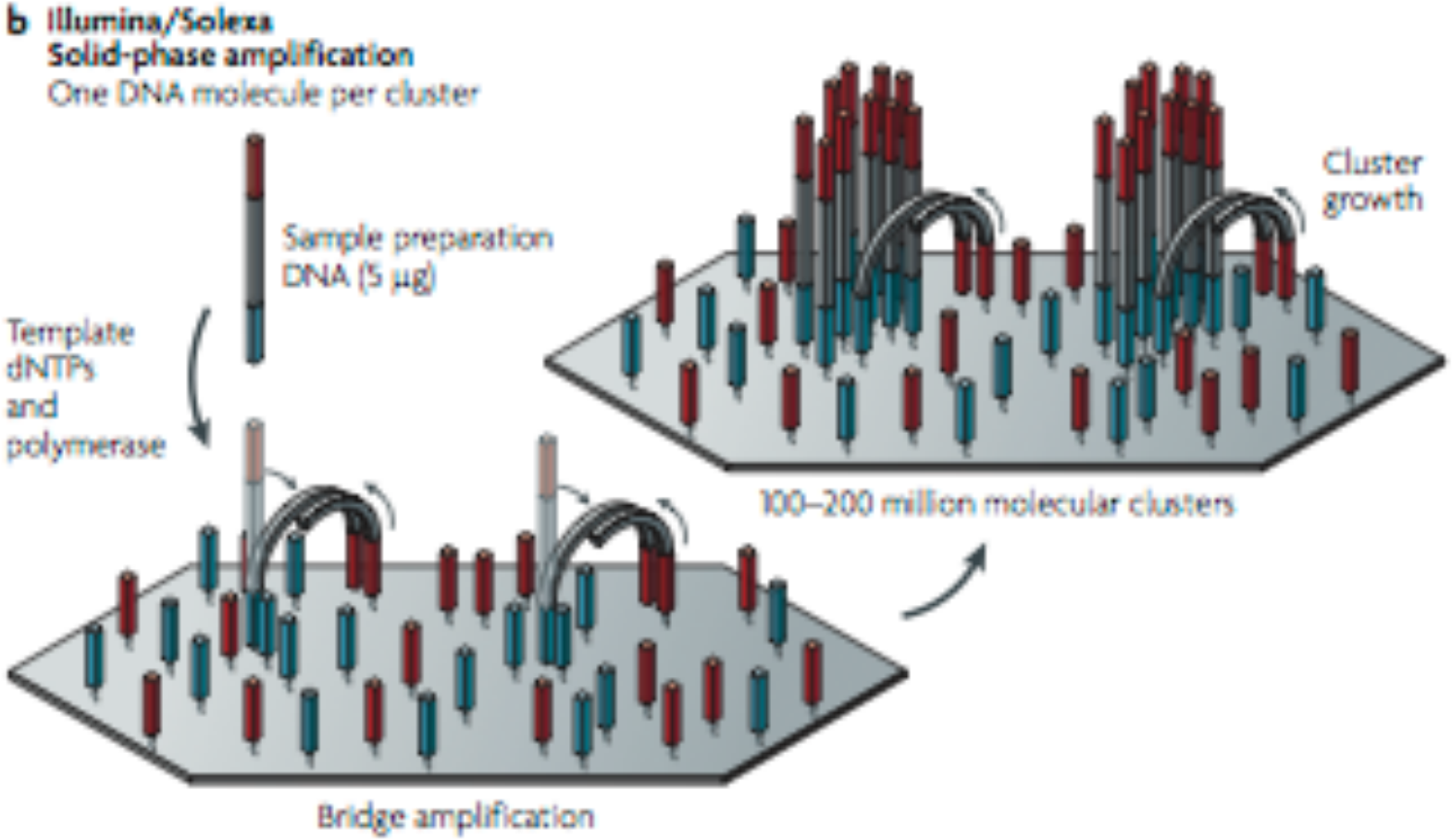
Illumina Protocol 1- Preparing Reads



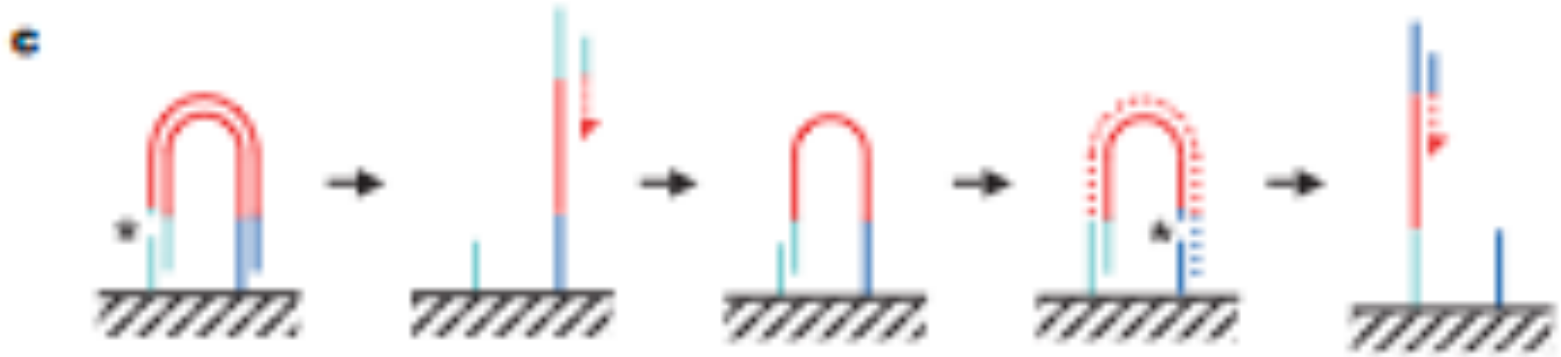
Illumina Protocol 2- Synthesis of clones of single strands



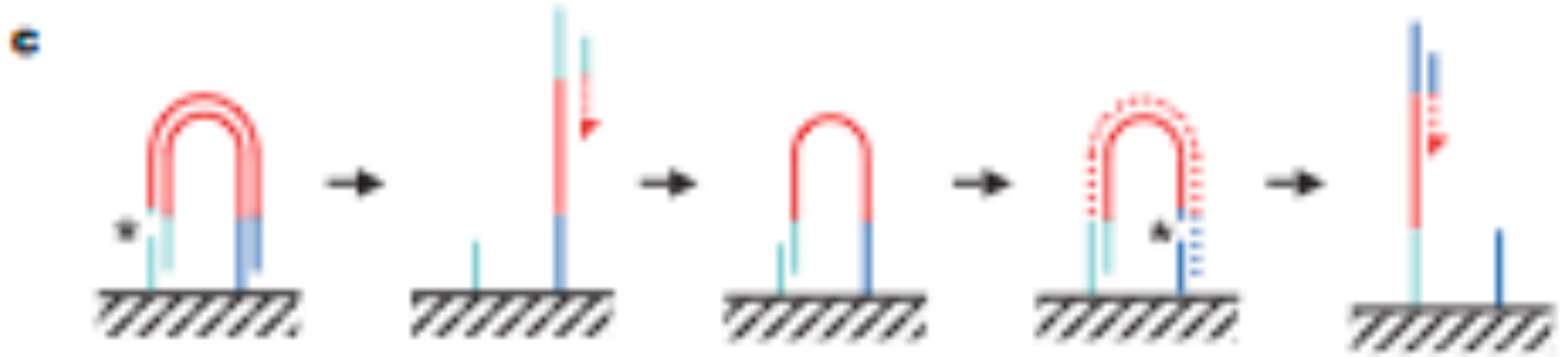
Illumina Protocol 3- Growth of single molecule cluster



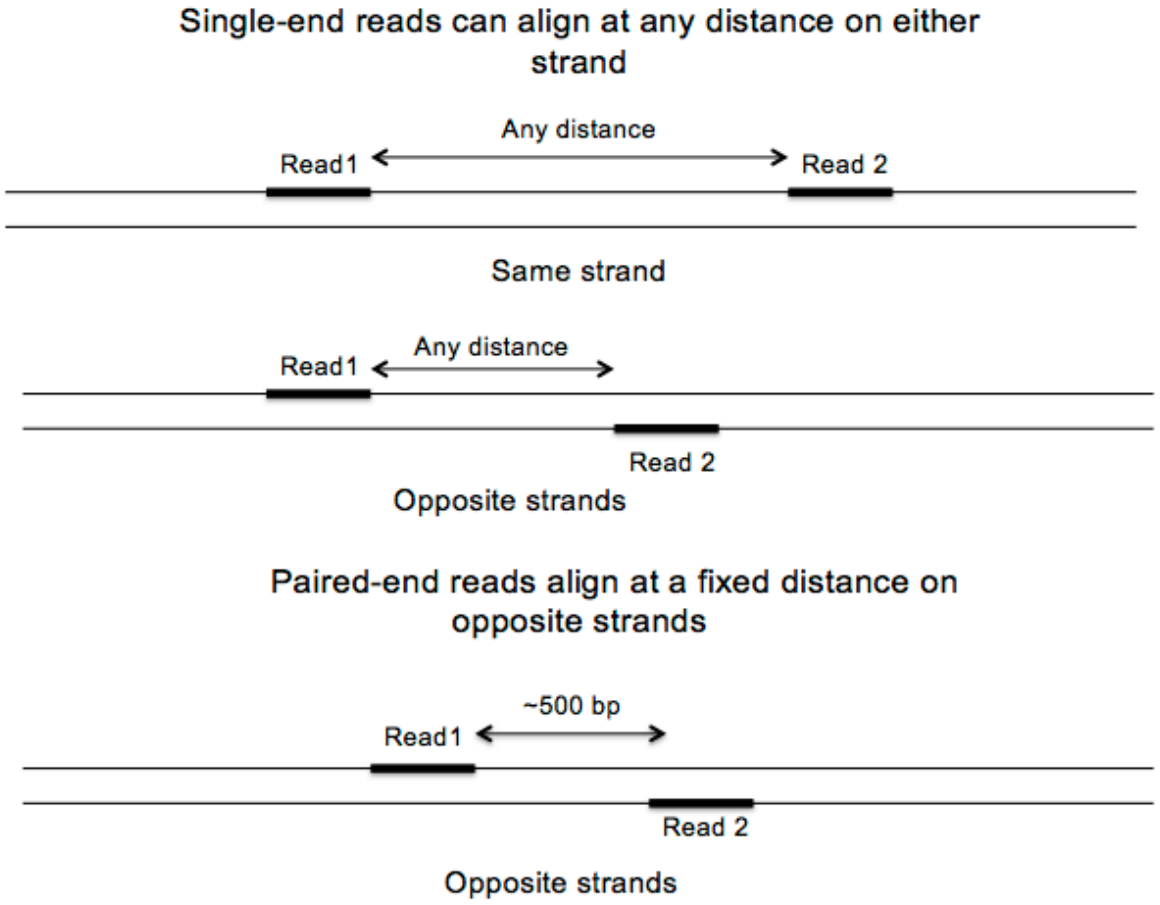
Illumina Protocol 4- Sequencing by Synthesis



Illumina Protocol 5- Paired-end Sequencing



Alignment of single-end vs paired-end reads



Phred Score

P= Probability that base is in error.

Phred score:

$Q=-10\log_{10}P$

The higher the reliability, the lower P.

The lower P, the higher Phred score Q.

Therefore: The greater the reliability, the higher Q.

Recommended: Q>22 optimizes alignment reliability.

FastQ files

Short read information is in FastQ files.

FastQ files consist of

1. Format of sequence in Fasta format.
2. Phred quality score in Ascii code.

Quality checking of FastQ files FastQC and PRINSEQ

- Base quality
- Read length
- K-mer distribution
- Ambiguous reads
- Contamination by vectors and adaptors
- Poly A/T tails

Filtering and Trimming

Trimmomatic:

Filters vector and adapter contaminants and low average quality.

Trims reads so that ALL BASES to ≥ a given quality.

Recommended: Q≥22.

Alignment to Genome

Alignment program:

Input:

1. FastQ file (short reads).

2. GTF file (sequence range of genes on genome)

Output: SAM (Sequence Alignment Mapped) file

Or BAM (Binary Alignment Mapped) file

Alignment Program

FastQ file + GTF file → SAM or BAM file

Ungapped alignment

Ungapped Alignment – Burrough-Wheeler Transform.

Indexes genome sequence for rapid searching.

Free Burroughs Wheeler Transform program:

Bowtie.

The Tuxedo Suite

Bowtie- Ungapped alignment.

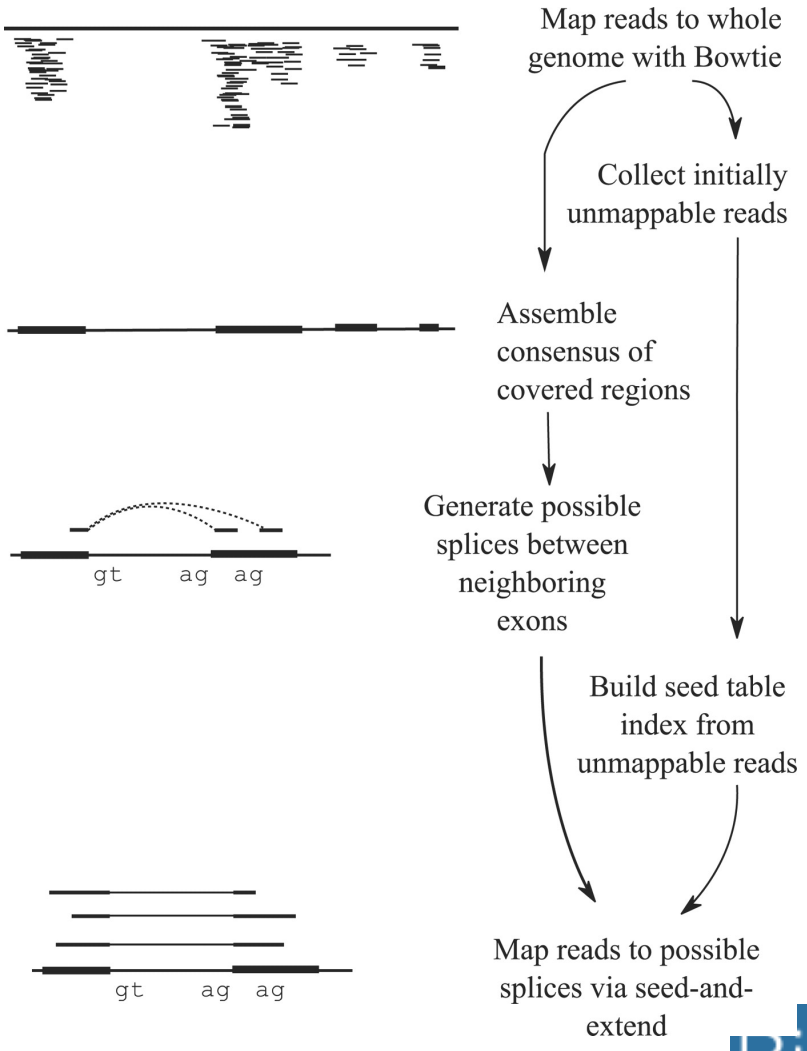
Tophat- Gapped alignment.

Cufflinks- Identification of transcripts.

Cummerbund- Visualization.

Monocle – Single cell.

TopHat Aligner



Cole Trapnell et al. Bioinformatics 2009;25:1105-1111

Bioinformatics

STAR: Spliced Transcript Alignment to a Reference

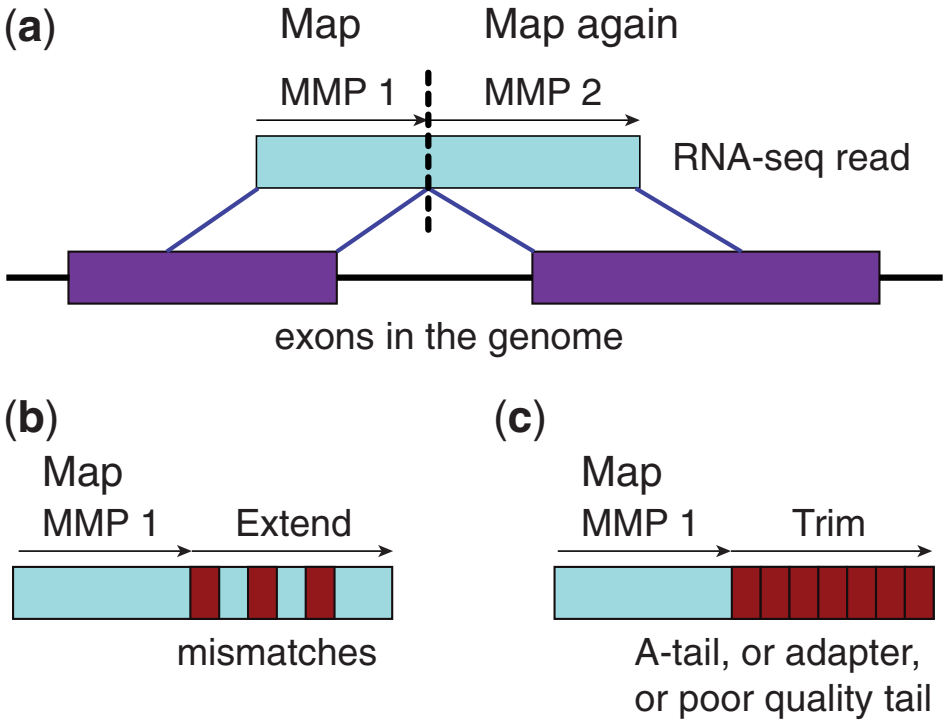
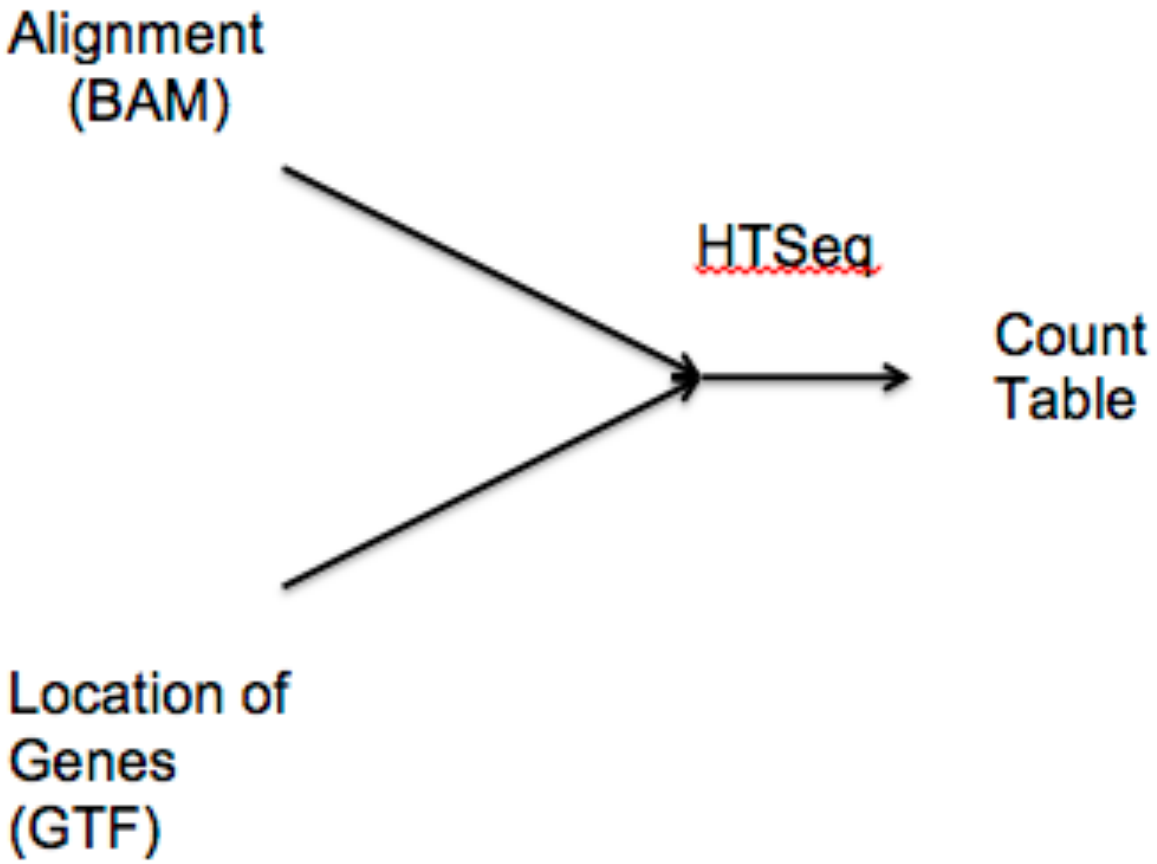


Fig. 1. Schematic representation of the Maximum Mappable Prefix search in the STAR algorithm for detecting (**a**) splice junctions, (**b**) mismatches and (**c**) tails

HTseq- Counting the reads



Meaning of counts

Raw counts of each genes in each sample

proportional to expression of that gene in each sample

NOT

Reads (or Fragments) per million per kilobase

NOT

Counts per transcript.

Need for normalization

Experiment- 100 Million total counts.

Control – 30 Million total counts.

Cause of difference in counts of a gene between the 2 samples:

1. Biological.
2. Size of library.

TMM: Trimmed Mean Method

$$\mathrm{TMM(BvsA)} = \sum_{\text{genes}} \log \left(\frac{\dfrac{\#\text{Reads of gene i in sample B}}{\text{Total number of reads in sample B}}}{\dfrac{\#\text{Reads of gene i in sample A}}{\text{Total number of reads in sample A}}} \right)$$

Complications in trimmed mean

- Sum is actually weighted
- More extreme terms (5-10% highest and lowest) not included in the sum.

Use TMM not FPKM.

Get raw counts from the core.

Use TMM in program.

Probability of RNASeq read for a gene

Picking out transcripts belonging to

50,000

different genes

from 10,000,000,000 transcripts

in 70,000,000

reads

Poisson distribution

Mean-variance properties of statistical distributions

Normal distribution (continuous variables):

var uncorrelated with μ.

Poisson distribution (counts from a single population):

$$\text{var} = \mu$$

Negative binomial distribution (counts from a population with variable members- e.g. Biological variation):

$$\text{var} = \mu + \phi\mu^2$$

Programs for RNASeq differential expression

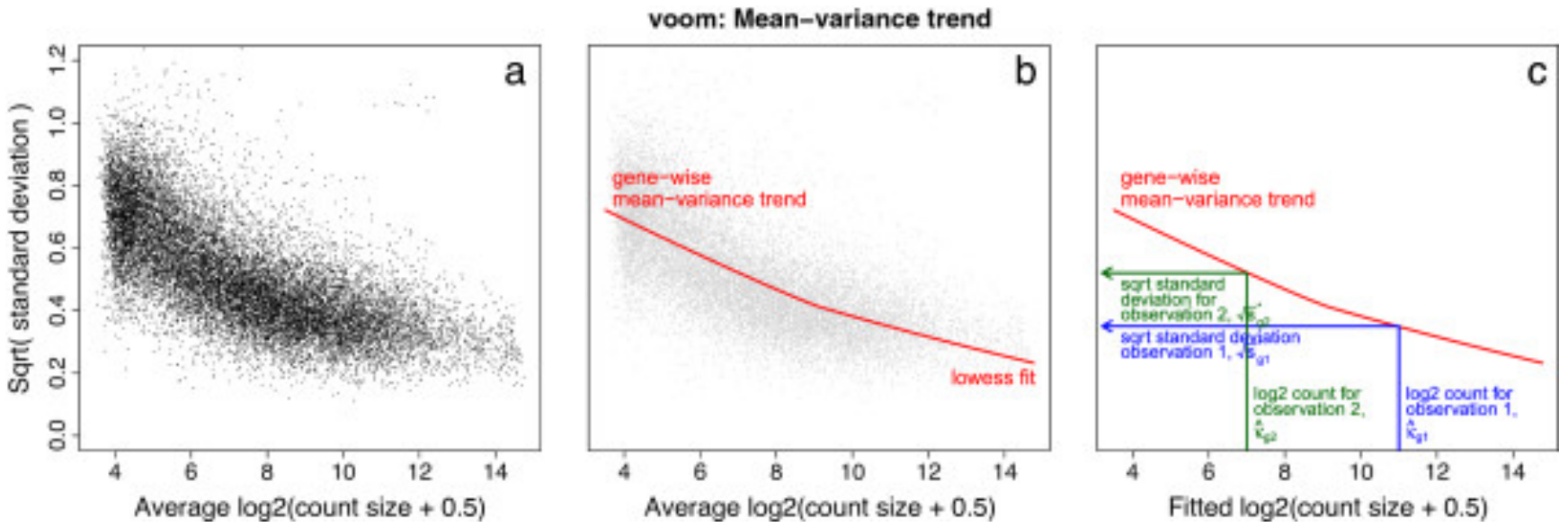
DeSeq –Uses the negative binomial distribution.

Estimates a single ϕ based on the means and variances of all genes

EdgeR,DeSeq2 – Both use the negative binomial distribution and estimates a ϕ for each gene based upon its own mean and variance and the means and variance of all of the genes.

Limma-Voom Uses a normal distribution with a variance for each gene based on its own variance and the variance of all genes as a function of μ.

Voom- Global contribution to the variance



Limma-Voom works best!

Limma-Voom works better than Deseq and EdgeR

(Not compared to Deseq2 rigorously yet).

How many samples?
How many reads?

USUAL: 30 M TOTAL reads

OPTIMAL: 70M mapped ~(90M total) will saturate before going to more samples.

USUAL: 1-3 Biological replicates

OPTIMAL: At least 5 Biological replicates.

Validation by qRT-PCR

1. PCR reaction doubles concentration of RNA until a concentration passes threshold. Cycles of doubling are counted.
2. Ct=Cycles to threshold inversely related to concentration:

Less concentration: more cycles.

More concentration: less cycles.

Cycles and concentration

3. Ct α $-\log_2$[Concentration]

4. Ct must be corrected for house keep gene control (actin, 18S RNA, etc)

$-(Ct_{exp}-Ct_{ctrl}) \alpha -\log_2$[Concentration]

PCR-Validation

- Don’t: Take wells as replicates (Pseudoreplication).
- Do: Average wells for a biological replicate over technical replicates.

PCR-Validation

Don't average and do your statistics on ratios ($2^{-\Delta\Delta Ct}$)

Do average and do your statistics on cycles.

Null Hypothesis:

$-(Ct_{exp}-Ct_{ctrl})= -(Ct_{ref}-Ct_{ctrl})$

$\Delta\Delta Ct=\Delta Ct_{ref}-\Delta Ct_{exp}=\log_2[Exp]-\log_2[Ref]$

$=\log_2[Exp/Ref]=\log_2 FC$

Multiple Comparisons

- t-test only for 2 groups
- More than 2 groups
- ANOVA- ANalysis Of VAriance

(whether at least 1 group not the same as others).

Multiple comparison pairwise tests if more than 2 groups

- Each group agains every other group:
  Tukey test
- Each group against a reference:
  Dunnett test.
  (Both implemented in PRISM)
- You pick the comparisons:

R Multcomp package

(also more powerful comparisons)