Alimentary Pharmabiotic Centre

Interfacing Food & Medicine

Understanding the Microbiome: *Metatranscriptomics*

Marcus Claesson

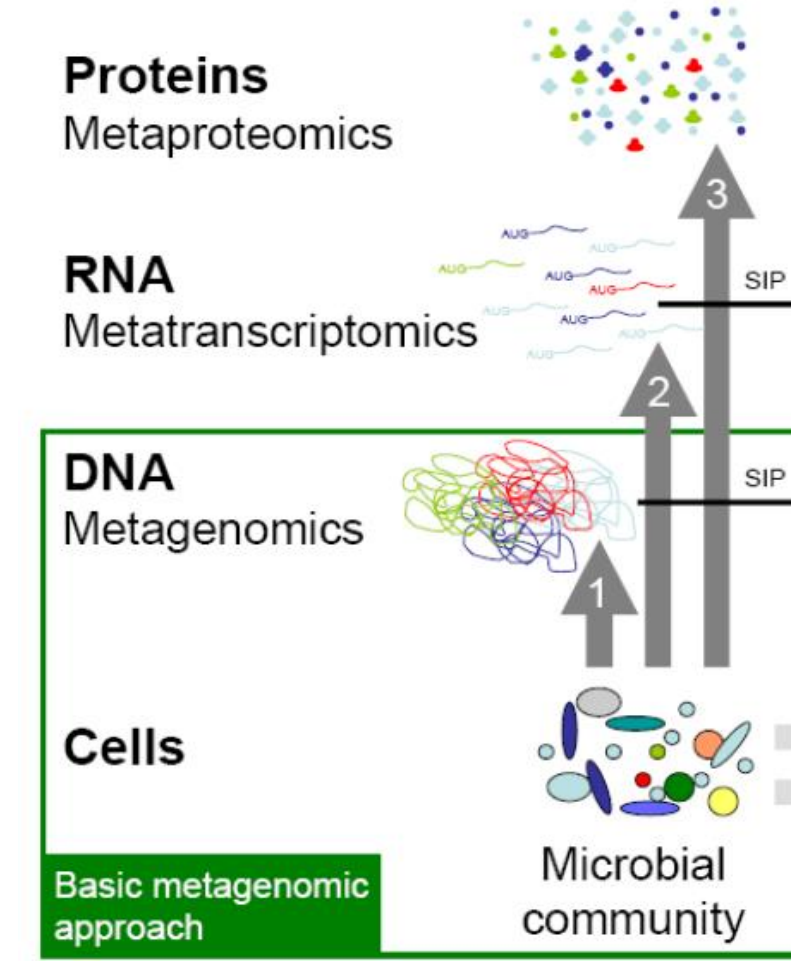
APC Microbiome Symposium 2015

Metatranscriptomics

Definition

(genetics, ecology)
A branch of transcriptomics that studies and correlates, the transcriptomes of a group of interacting organisms or species

--Wiktionary--



Warnecke & Hugenholtz (2007) Genome Biology

Sequence-based 'omics' technologies

Metatranscriptome:

- Protein-coding RNA (mRNA)
- Non-coding RNA (rRNA, tRNA, regulatory RNA, etc)

Metatranscriptomics studies:

- Community functions
- Response to different environments / treatments; differential gene expression
- Regulation of gene expression

Metatranscriptomics cDNA/mRNA sequencing

- Microbial gene expression at certain times and/or locations
- *What ARE they doing?*

Metagenomics shotgun sequencing

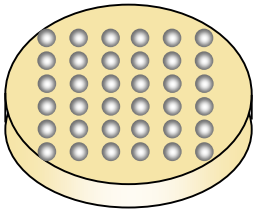
- Encoded potential functions of the microbiota
- *What CAN they do?*

Microbiota compositional analysis

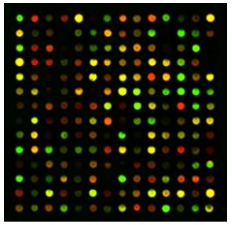
- Quantification of the ubiquitous 16S rRNA gene
- *WHAT organisms are there?*

Measuring the transcriptome

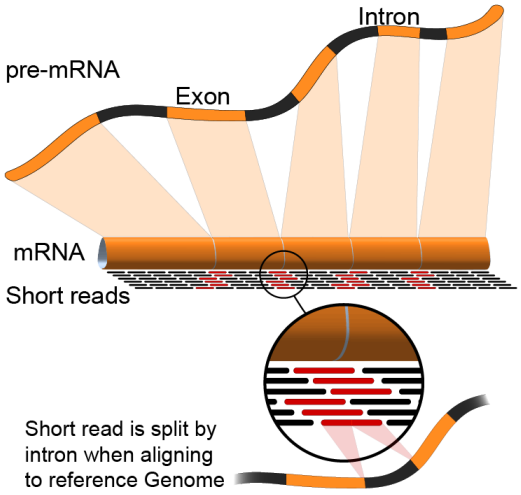
- cDNA clone libraries + Sanger sequencing



- Microarrays: hybridizing mRNA onto cDNA/oligonucleotide proves on glass slide

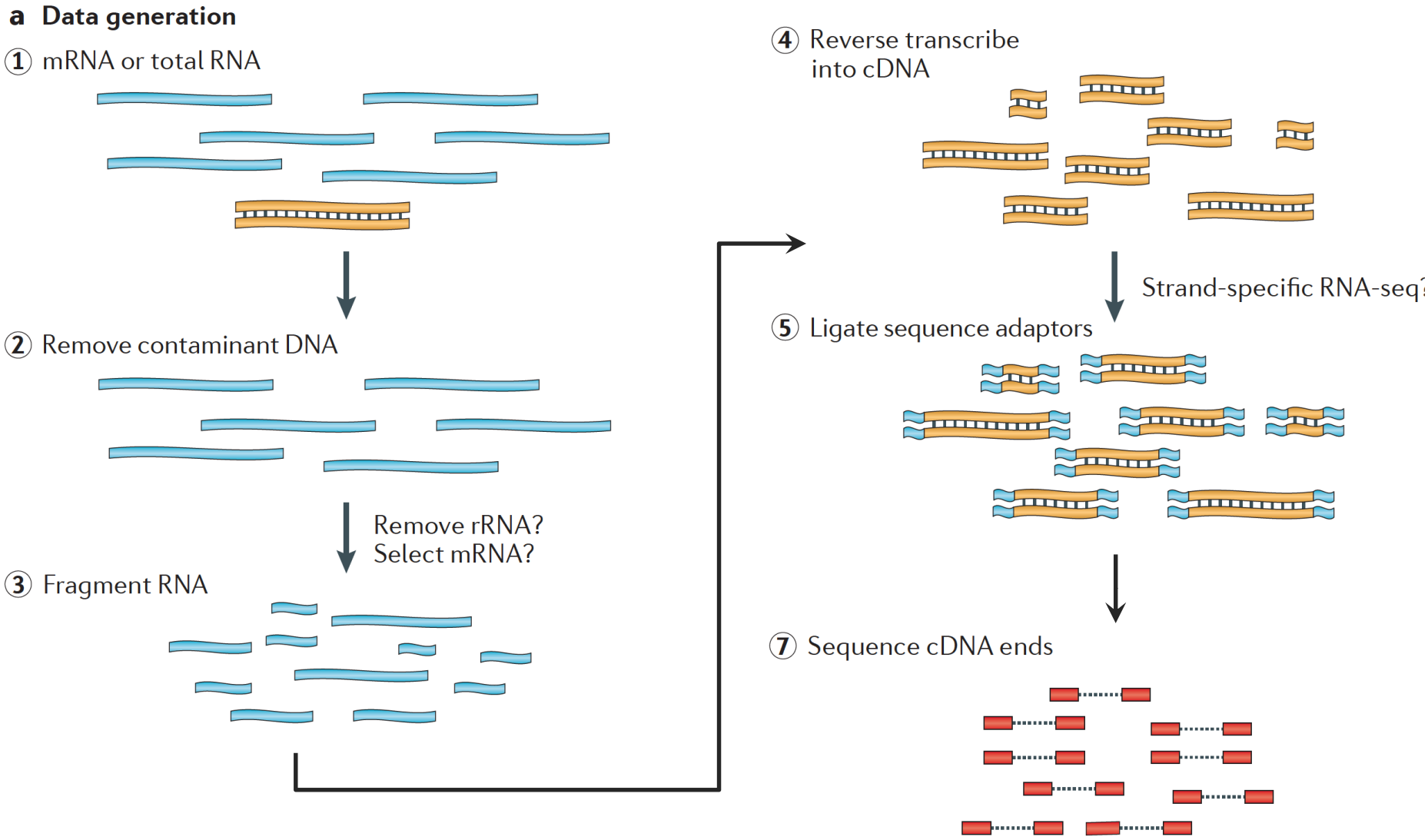


- RNA-seq enabled by next-generation sequencing technologies – usually Illumina HiSeq



RNA-seq superior to microarrays for gene expression analysis of microbial communities

From RNA to sequence data



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Challenges and considerations

Wet lab

- Deplete host RNA or not? Dual RNA-Seq an option
- Instability of RNA (half-lives of minutes)
- High rRNA content in total RNA (mRNA<5% of total RNA)
- Single-end or Paired-end?
- Stranded cDNA transcription or not?
- How much sequencing per sample?
- Technical replicates not necessary any more, but biological are
- Avoid batch effect -> randomize samples across runs!

http://www.nwfsc.noaa.gov/index.cfm

http://cybernetnews.com/vista-recovery-disc/

Bioinformatics

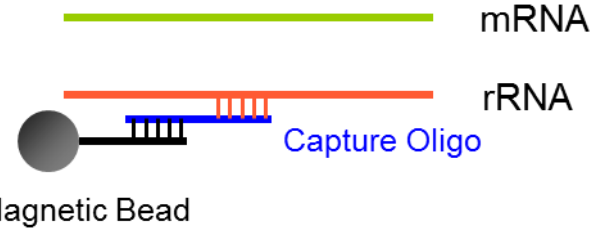
- General challenges with short reads and large data size
- Lack of metagenome/genome reference
- Statistical considerations
- Assemble or map reads, or both?

rRNA removal methods

Majority of bacterial mRNA is not polyadenylated, can't be isolated using oligo-dT selection

Subtractive Hybridization

Ribo-Zero / MICROBExpress
Bacterial mRNA Enrichment



Exonuclease Digestion

mRNA-ONLY Prokaryotic mRNA Isolation

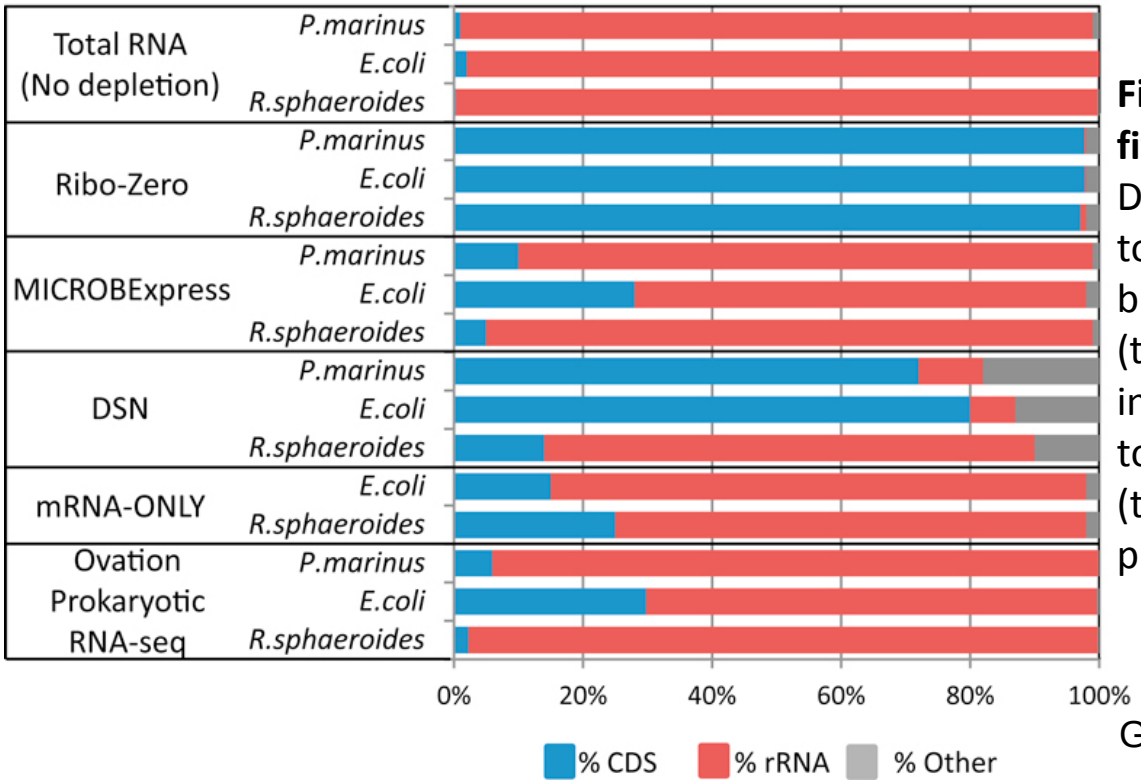


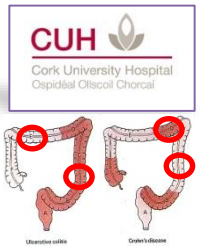
Figure 1. Performance evaluation of five rRNA depletion methods. Distribution of RNA-seq reads aligning to protein-coding sequences (CDS; blue), rRNA (red), and other regions (tRNA, non-coding RNA, small RNA, and intergenic regions; gray) for undepleted total RNA (top) and five rRNA depletion protocols.

Giannoukos et al. Genome Biology 2012,

Experimental pipeline

- Inflamed/uninflamed colonic biopsies
  - **Pilot project**: 12 CD & 6 UC
  - **Main project**: 60 CD, 86 UC & 30 HC

PostDoc
Emilio Laserna



Total RNA extracted: *RNAlater*

~~Eukaryotic RNA depleted: *Microbe Enrich*~~

Ribosomal RNA depleted: *RiboZero*

cDNA synthesized: *ScriptSeq transcriptase*

~~cDNA amplification: *GenomiPhi*~~

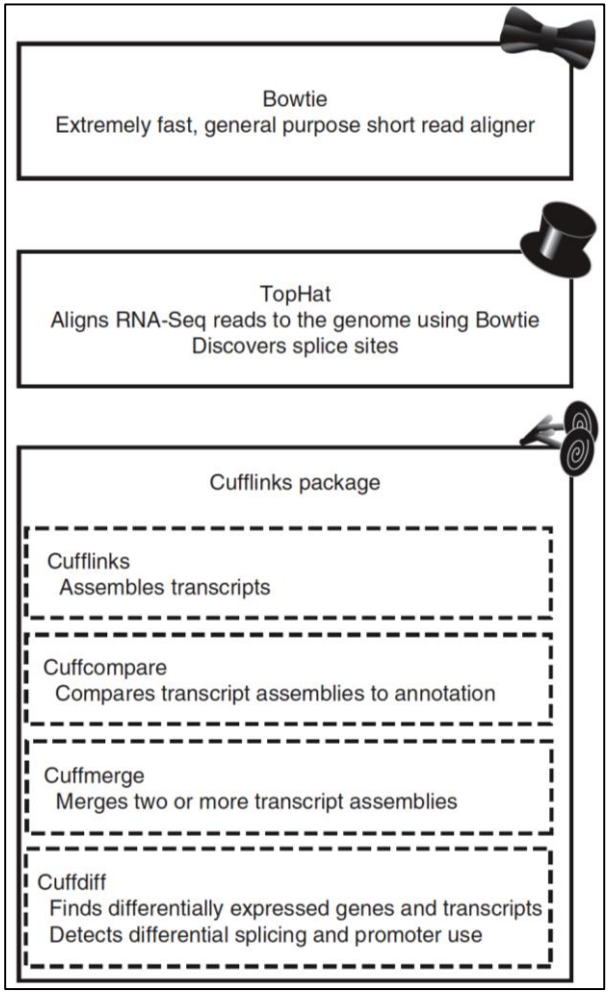
RNA-Seq library prep: *Illumina TruSeq*

RNA-Seq: *Illumina HiSeq 2500*

15 million 125 bp PE reads per sample

Differential Gene Expression

Options for DGE analysis (tuxedo suite)



Trapnell *et al.*, Nature Protocols, March 2012

Bowtie2 and Bowtie use Burrows-Wheeler indexing for aligning reads. With bowtie2 there is no upper limit on the read length

Tophat2 uses Bowtie2 to align reads in a splice-aware manner and aids the discovery of new splice junctions

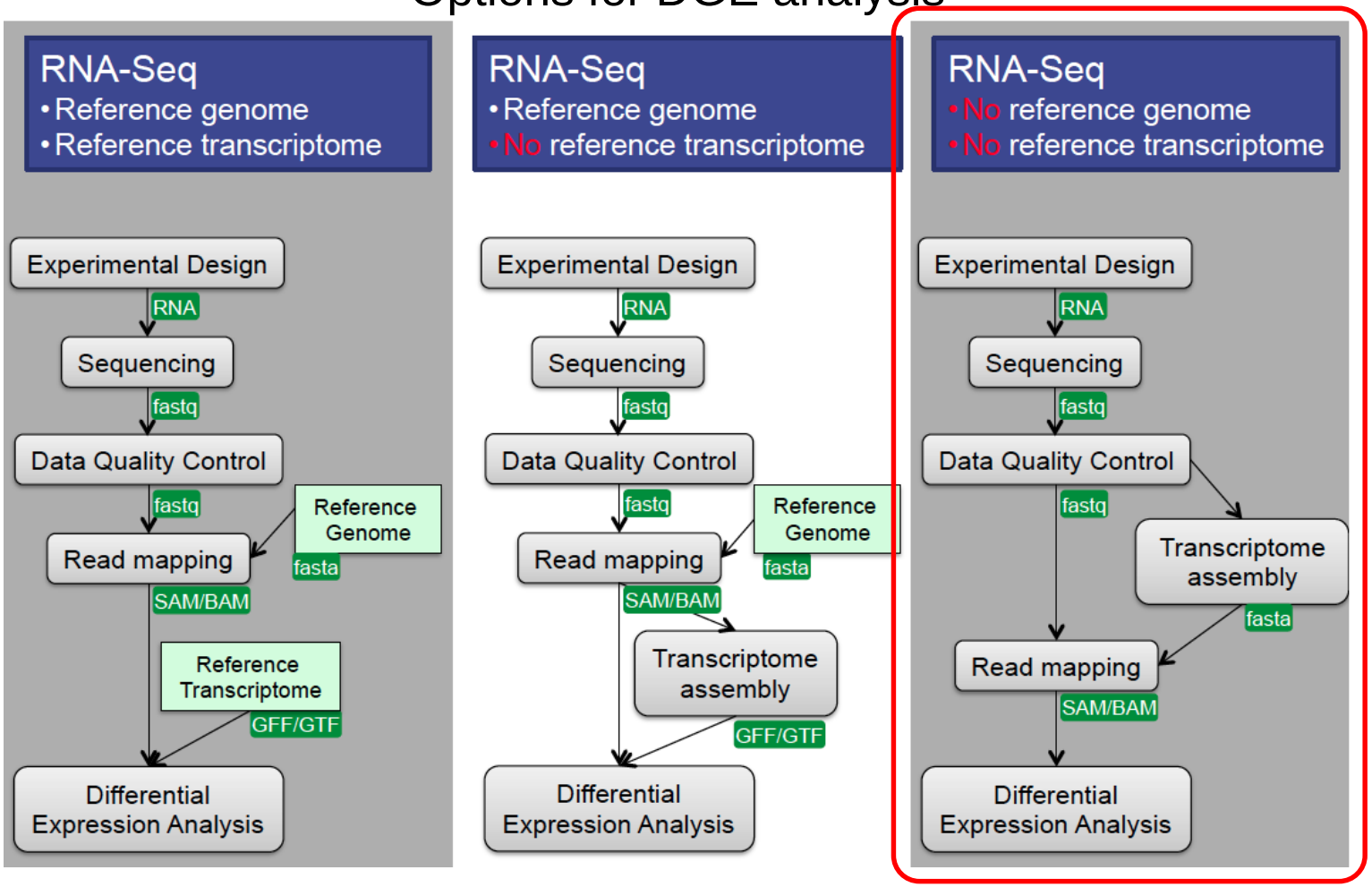
The Cufflinks2 package has 4 components, the 2 major ones are listed below -

Cufflinks2 does both *de novo* and reference-based transcriptome assembly

Cuffdiff2 does statistical analysis and identifies differentially expressed transcripts in a simple pairwise comparison, and a series of pairwise comparisons in a time-course experiment

Differential Gene Expression

Options for DGE analysis



Want to learn more about the formats?
https://genome.ucsc.edu/FAQ/FAQformat.html

Bioinformatics pipeline

PhD student
Feargal Ryan

Check MD5 of downloaded FASTQ files

FastQC

Quality filtering/trimming using *Trimmomatic*

FastQC

sortmeRNA to remove rRNA

Align vs *hg20* using *STAR*

Align each sample against all known microbial genes (IGC) with *Bowtie2*

Make count table using *HTseq*

Visualise normalized counts using *DEXSeq* for Differential exon usage *DESeq2* for Differential Expression

Combine samples and assemble with *Trinity* or *IDBA-MT/IDBA-UD*

Assembles and annotate transcripts

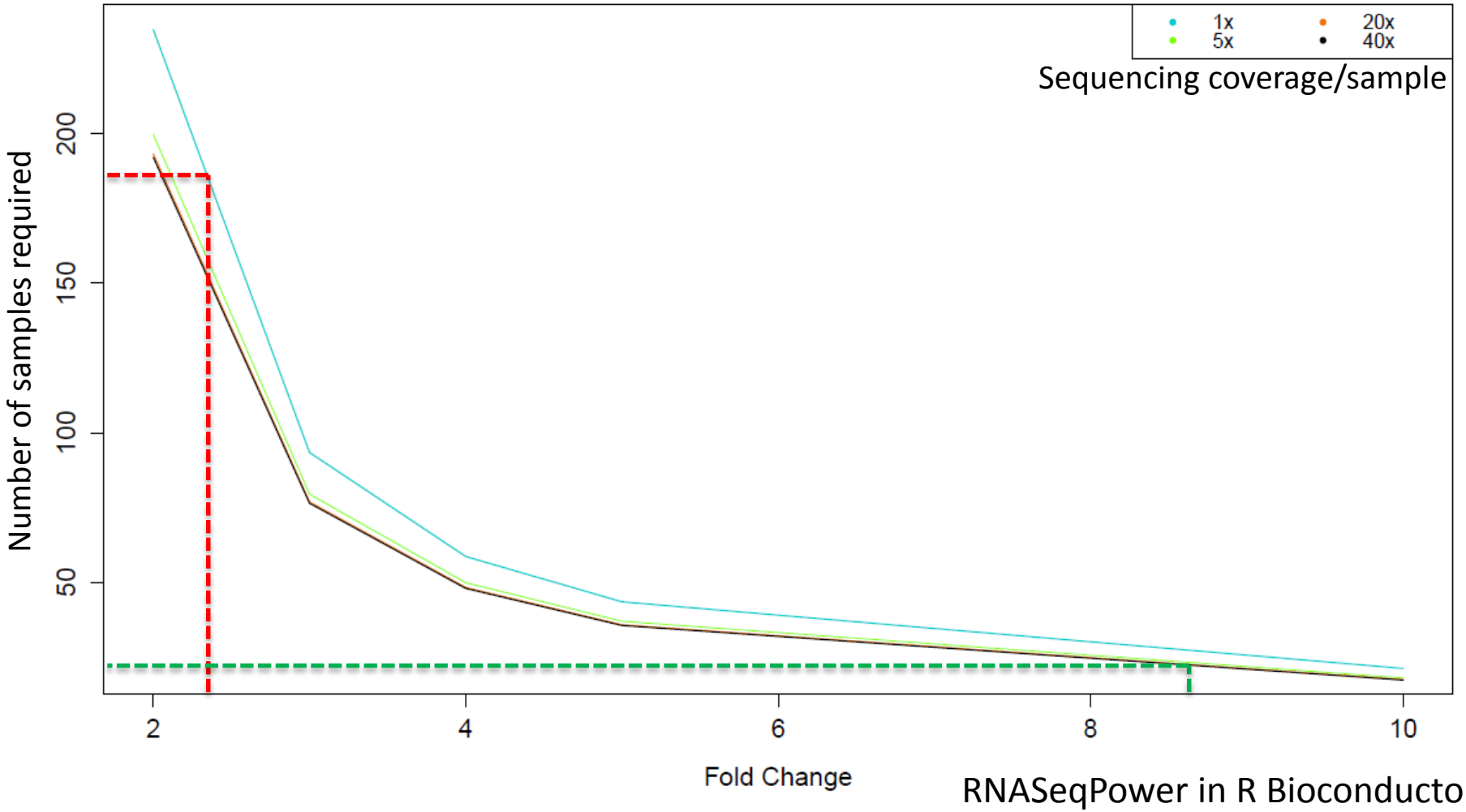
Visualise sample composition using annotation information

DESeq2 – Normalize counts and perform differential expression analysis

Visualize normalized counts. Any batch effects? Any other confounding factors separating out the data to control for in *DESEeq2*

Power calculation:

Read coverage vs. sample size



What the Pilot Study taught us

- Differential gene expression for inflamed and non-inflamed in both UC and CD
  => higher *n* required to confirm
- Use a more effective rRNA depletion method
- Don't deplete host mRNA if also interested in it
- No extra cDNA amplification needed
- High *n* more important than sample coverage
- Experimental & bioinformatics pipelines
- SIRG & Second Genome study:
  - 200 subjects
  - Genotype, disease activity, diet, medication