Analysis of Metatranscriptomic Data from Defined Communities

Dr. Anna Sintsova
Sunagawa Lab

Part 1: From raw data to count tables

Part 2: Data analysis and normalization

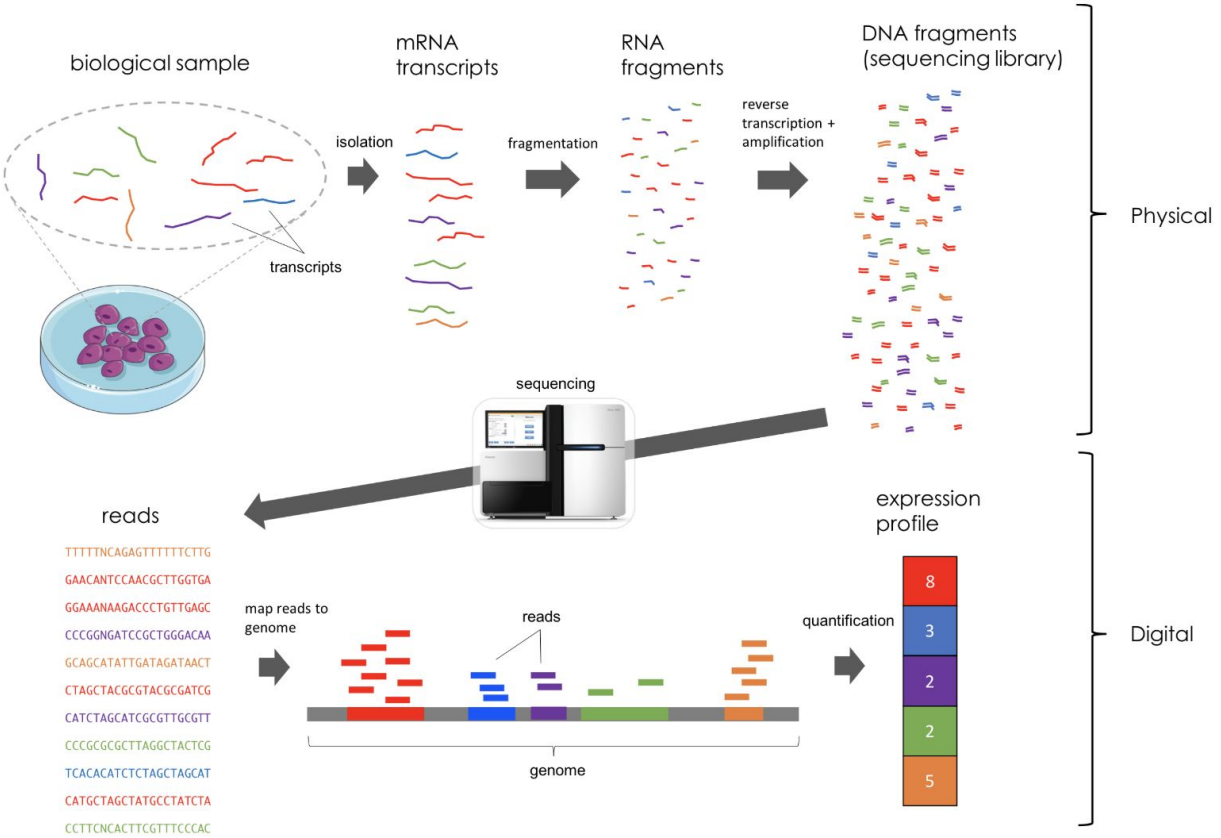
Part 3: Case study with 12-strain mouse community

Part 1: From raw data to count tables

Learning Objectives

- How metatranscriptomics differs from single-organism RNAseq
- Experimental design considerations for community studies
- Data preprocessing and QC specific to metatranscriptomics
- Different alignment/quantification approaches and their trade-offs
- Why competitive mapping matters in community data

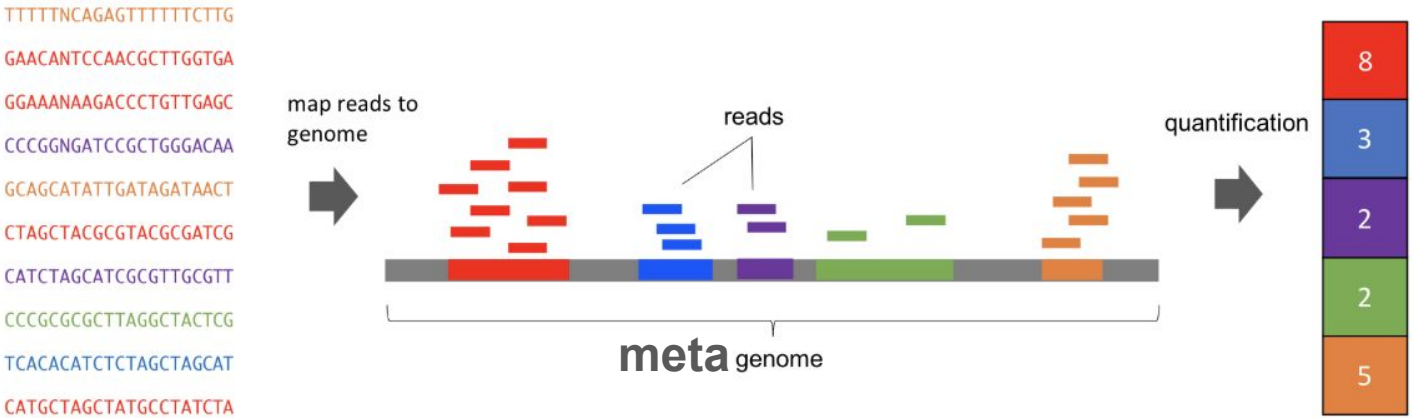
What is RNAseq?



Ref

Metatranscriptomics: who is active and what are they doing?

- RNAseq on the whole community
- Determines which genes and pathways are being expressed
- Can reveal with functions are active
- Can also reveal which organisms are active.

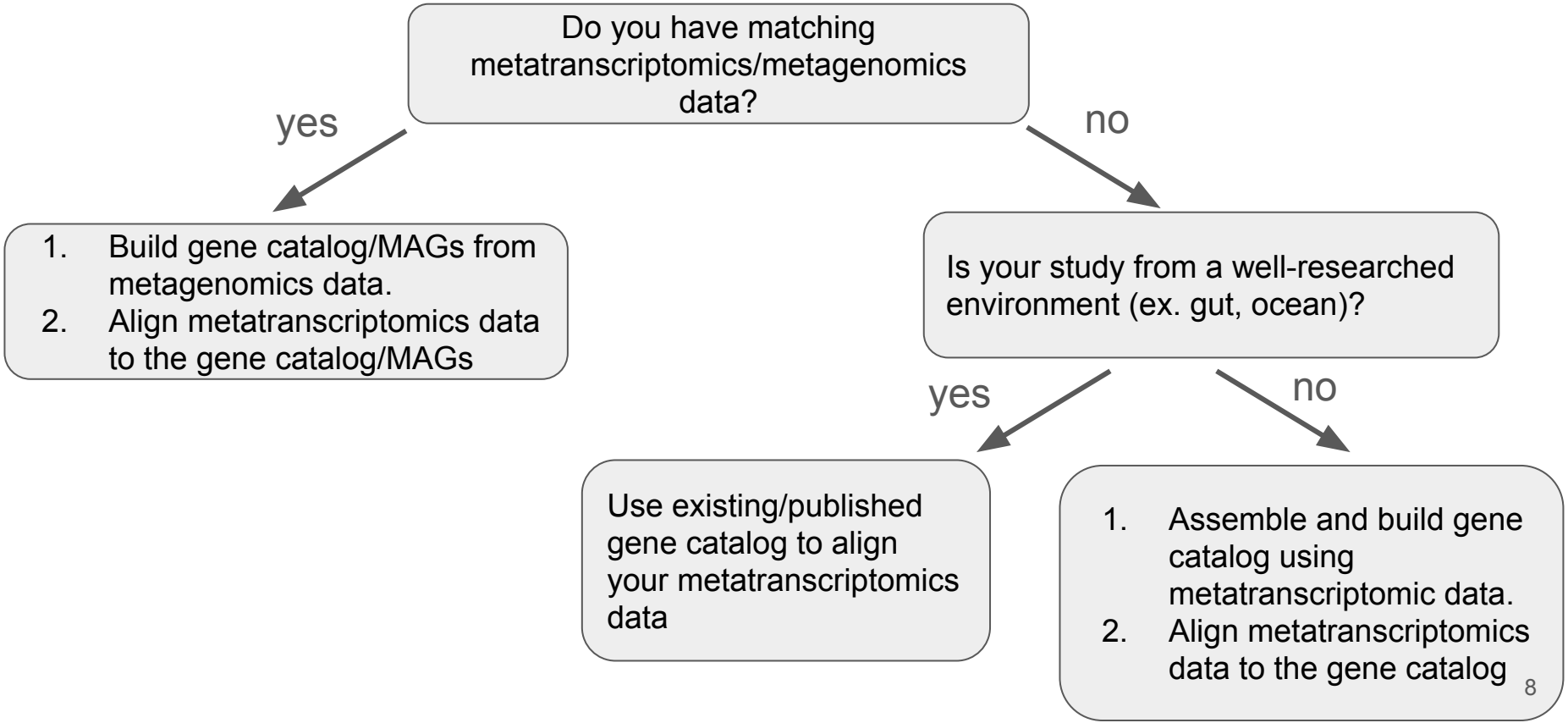


Ref

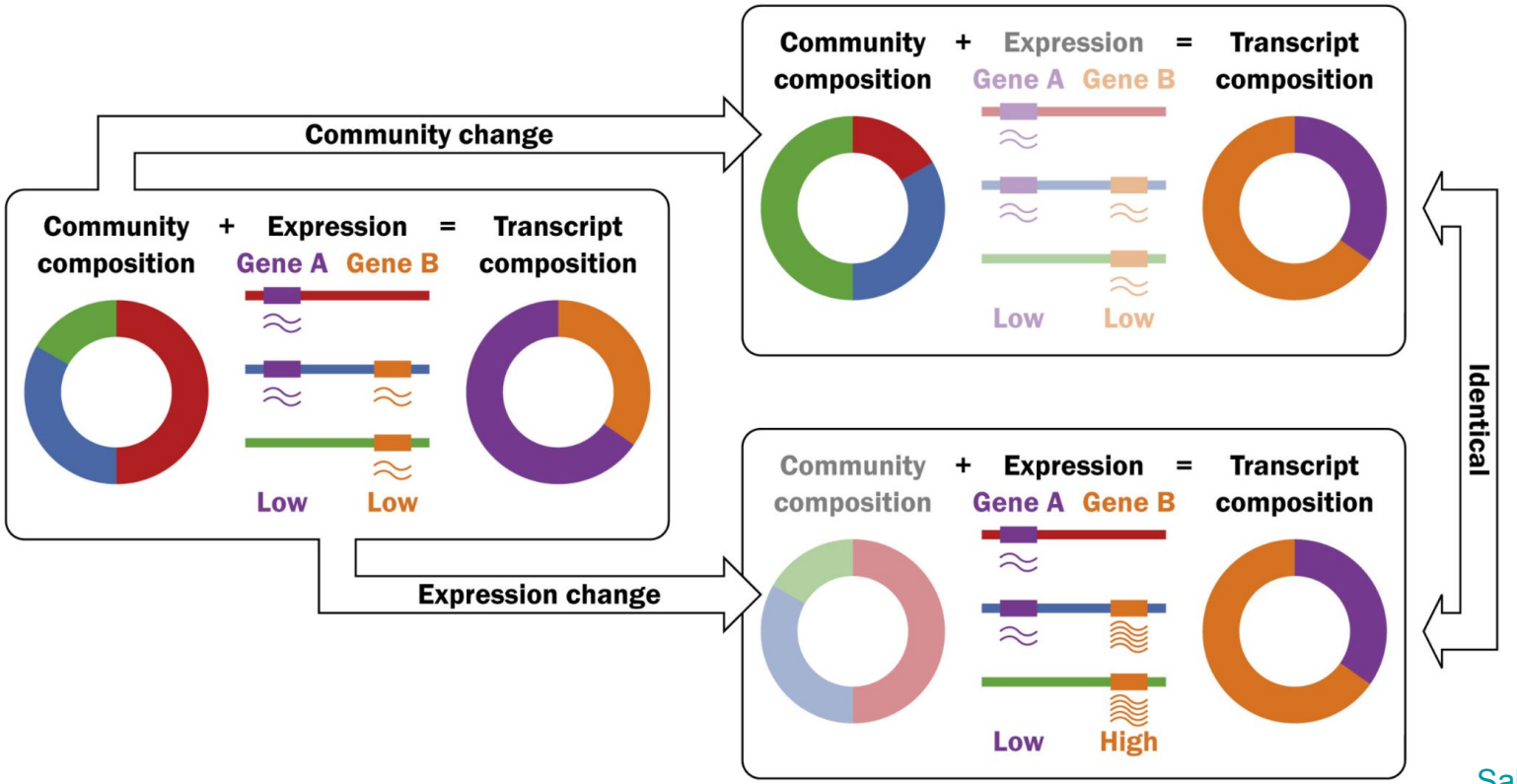
Differences from RNAseq:

- Sequencing multiple organisms with varying abundances
- **Changes in community composition confound expression analysis**
- Ideally needs matching metagenomic data
- Analysis methods are not standardized
- Will only get information for most abundant species

Metatranscriptomics analysis workflow:



Challenges: is it changes in gene expression or community composition?



Salazar et al 2019

Pros and cons of sequencing technologies

Technology	Questions answered	Pros	Cons
Amplicon sequencing			
Metagenomic sequencing			

Pros and cons of sequencing technologies

Technology	Questions answered	Pros	Cons
Amplicon sequencing	Who is there?	Cheap and (relatively) easy Well established analysis methods A lot of data from different environments allows global comparisons (MicrobeAtlas)	Limited in terms of resolution and questions it can answer Primer bias can skew the abundance results Many prokaryotes have multiple 16S rRNA copies, which are not accounted for in relative abundance calculations
Metagenomic sequencing	Who is there? What can they do? Strain transmission between hosts	**Get functional information** Higher resolution than 16S - important if want to trace transmission events	Financially and computationally expensive Bioinformatic expertise required Validation / culture of novel strains is a challenge

Pros and cons for metatranscriptomic sequencing

Technology	Questions answered	Pros	Cons
Metatranscriptomic sequencing	(Who is there?*) What are they actually doing?	**Get information on who is active and what functions are being expressed**	Financially and computationally expensive Bioinformatic expertise required Identifying who's transcribing is difficult

Metatranscriptomics of defined communities

Advantages:

- Known community members (genome sequences available)
- Easier to interpret biological relevance
- Reproducible experimental system
- **Bridge between single-organism and complex microbiome studies**

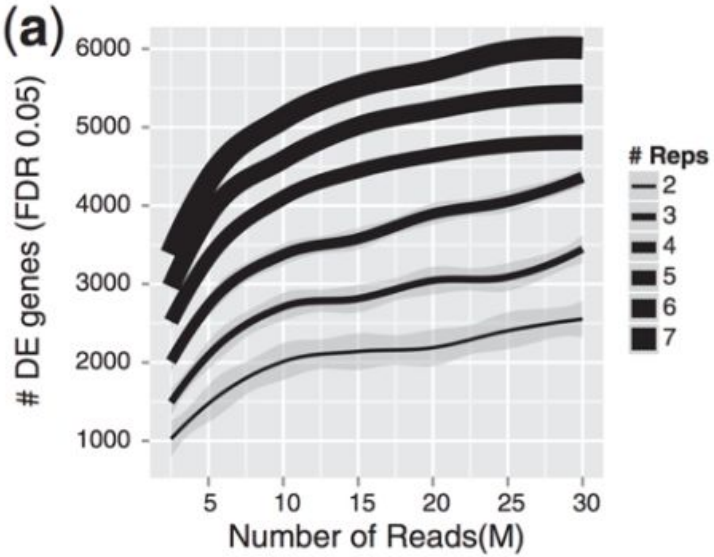
Example: Oligo-MM12 model

- 12 bacterial strains representing different gut phyla
- Genomes sequenced and well-annotated*****
- Colonizes germ-free mice stably
- Provides partial colonization resistance to Salmonella infection

Experimental Design Considerations

1. **Biological replicates > Sequencing depth**
    - Minimum 3-5 replicates per condition
    - Captures biological variation
2. **Sequencing depth:**
    - Single isolate: ~5M reads sufficient
    - Communities: need more for low-abundance members
3. **Technical replicates:**
    - Generally not necessary with modern protocols
    - Better to invest in biological replicates

Experimental design



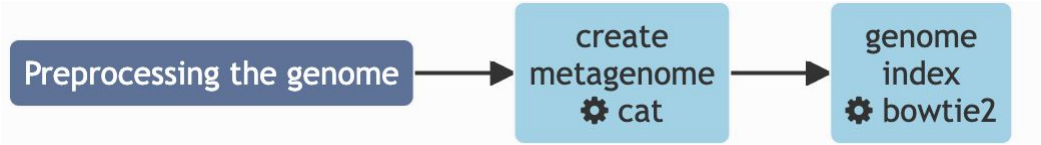
- Biological replicates are more important than depth of sequencing
- 5 M reads is usually sufficient for bacterial **isolate** RNAseq
- Will need to sequence deeper, if have low abundance members of the community
- More on experimental design considerations

Liu, Y., et al., Bioinformatics (2014)

(meta)RNAseq workflows

1. Preparing your (meta)genome
2. Quality Control & Preprocessing of raw sequencing data
3. Read Mapping/Quantification
4. Normalization & Statistical Analysis
5. Functional Interpretation

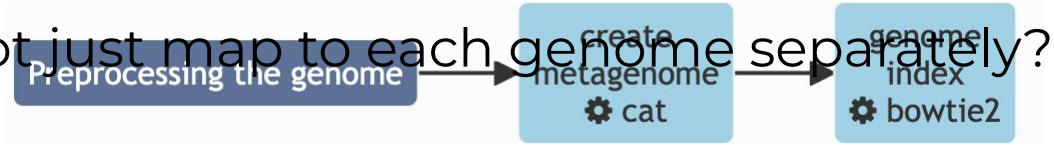
(meta)-transcriptomic workflow: metagenome



Why not just map to each genome separately?

(meta)-transcriptomic workflow: metagenome

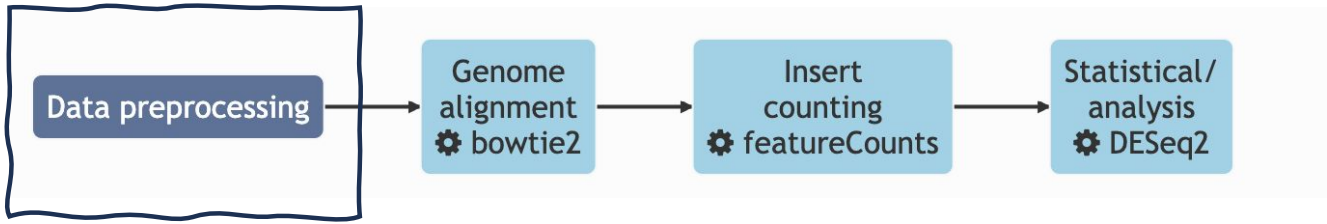
Why not just map to each genome separately?



Note

Why is competitive mapping important? It properly accounts for sequences that potentially map to multiple targets/species (multi-mappers, count as fractions). If the sample was mapped to each species individually, these reads will be counted towards each genome, overestimating the counts.

(meta)-transcriptomic workflow



Data QC and Preprocessing - Overview

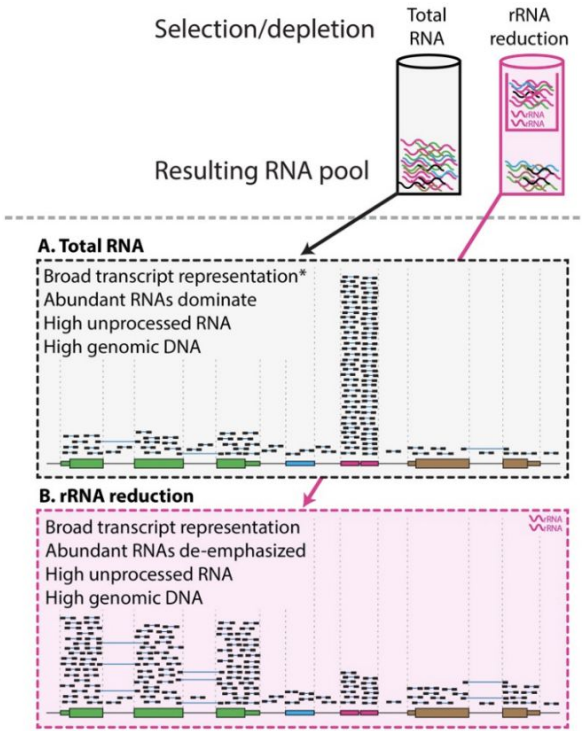
Quality control: FastQC, MultiQC

Adapter trimming: Cutadapt, Trimmomatic

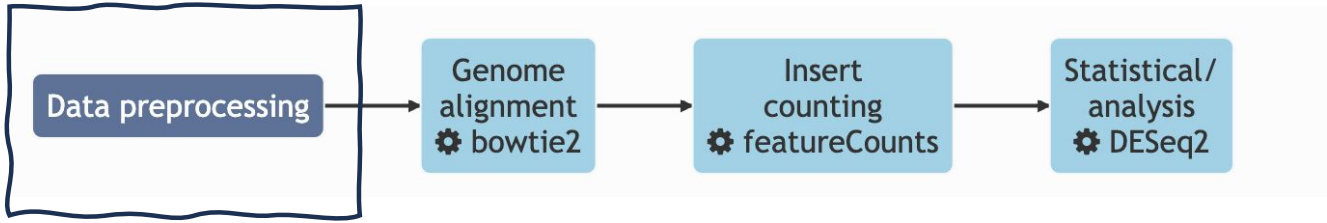
Quality filtering: Remove low-quality reads

rRNA assessment: fastq_screen

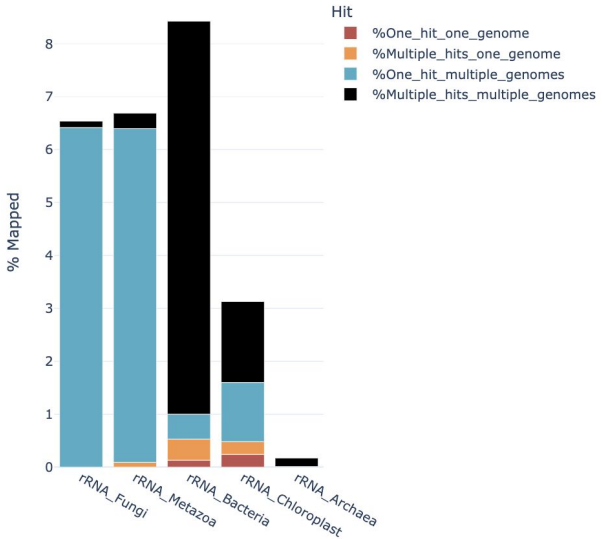
rRNA removal: SortMeRNA (if needed)



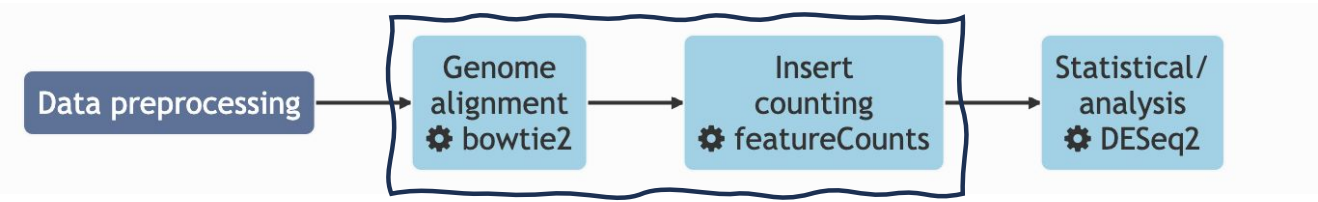
(meta)-transcriptomic workflow: QC



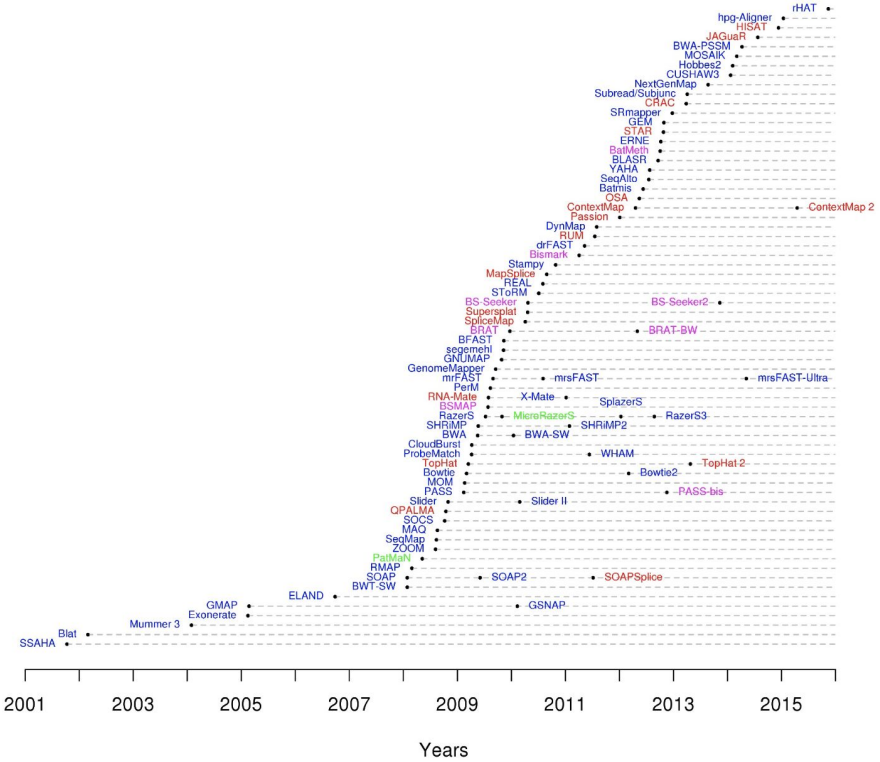
- Data preprocessing and QC:
  - Usual qc: trimming and filtering
  - Fastq_screen to estimate rRNA levels
  - SortMeRNA to remove the rRNA reads

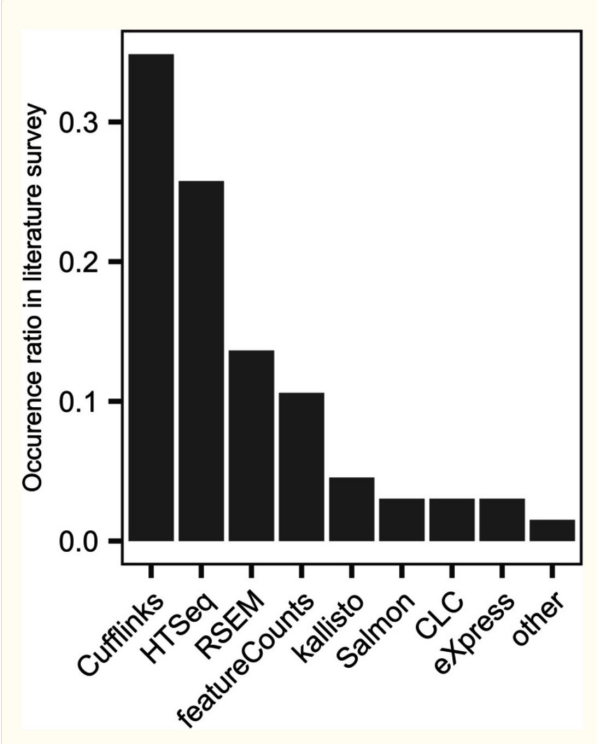


How do I measure transcript abundance?



How do I measure transcript abundance?





Mapping Approaches - Overview

Two main paradigms:

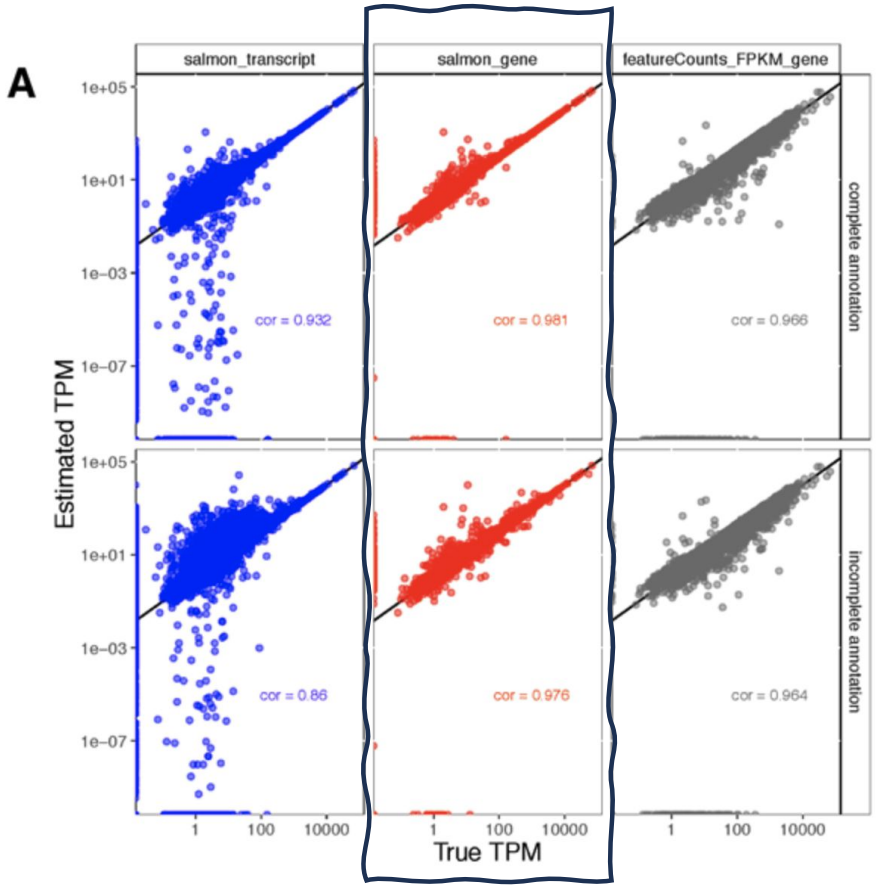
1. Alignment-based:
   - STAR, HISAT2, bowtie2
   - Base-by-base alignment
   - Outputs: BAM files with read positions
   - Slower but interpretable

2. Alignment-free (pseudo-mapping):
   - Salmon, kallisto
   - k-mer based matching
   - Faster, requires transcriptome
   - Outputs transcript quantification
   - Probabilistic assignment

Salmon

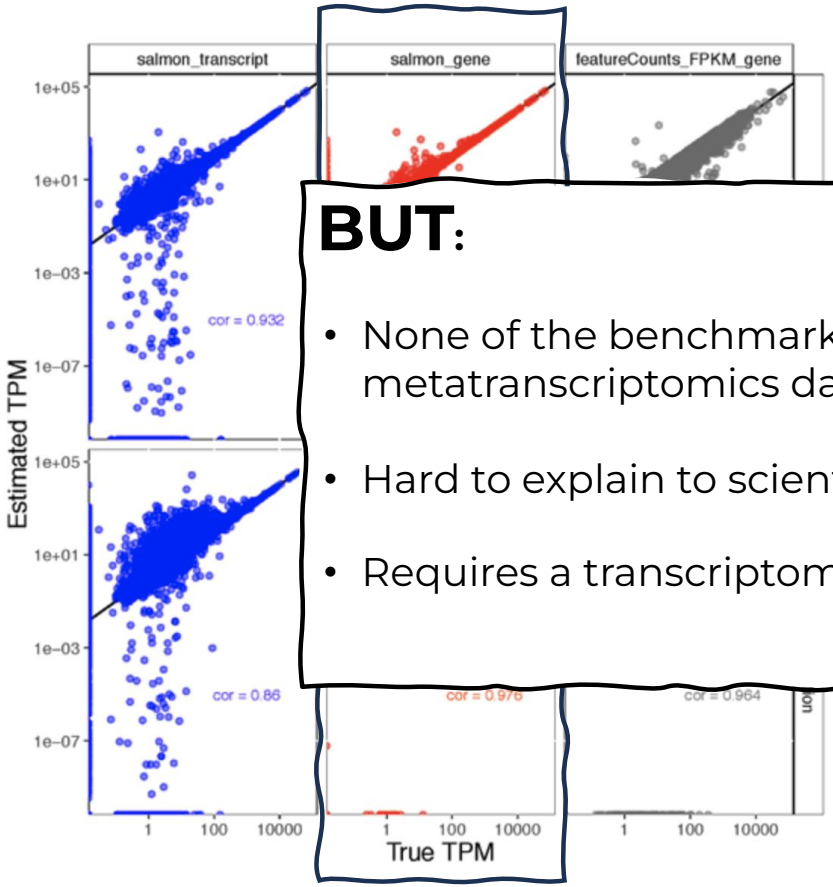
- Uses a reference **transcriptome** to perform both mapping and quantification of reads
- **Quasi-mapping:** perform a k-mer based search (instead of base-by-base alignment) to detect which transcripts the reads maps to
- **Quantification**:
  - Calculate abundances based on quasi-mapping results
  - For multimappers, the counts will be divided between transcripts
  - Uses modelling to estimate final transcript abundances
  - Tries to account for known biases:
    - GC bias
    - Positional coverage bias

How do I measure transcript abundance?



- Lot's of tools!
- Recently 'pseudo-aligners' have gained popularity
- From multiple benchmarks:
  - All of them (i.e. salmon, kallisto, RSEM) perform similarly
  - Under specific simulated conditions perform better than HISAT2/STAR + featureCounts

How do I measure transcript abundance?



- Lot's of tools!

BUT:

- None of the benchmarks are done using (simulated) metatranscriptomics data
- Hard to explain to scientists what it actually does
- Requires a transcriptome (not a genome)

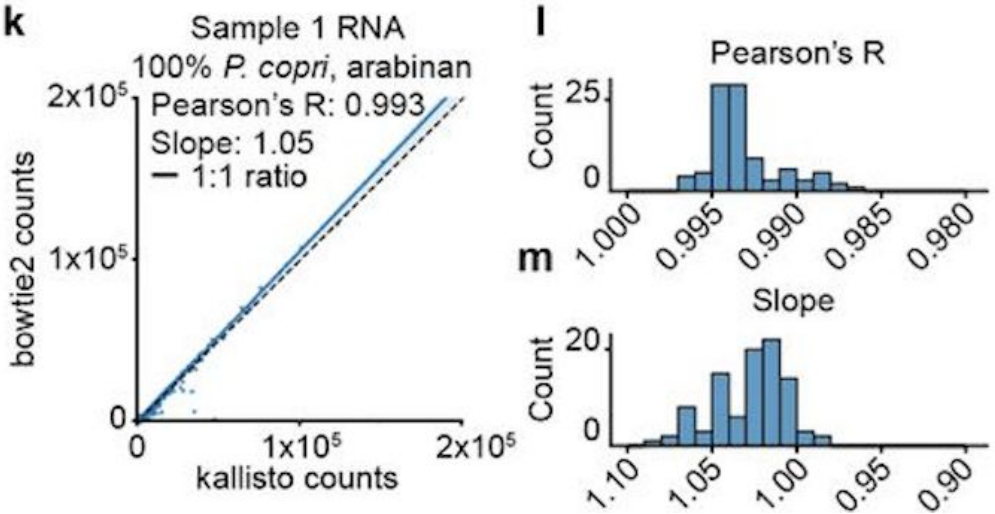
' have gained

ks:
, kallisto,
ly

ed conditions
atureCounts

- DTE is unreliable and should be avoided

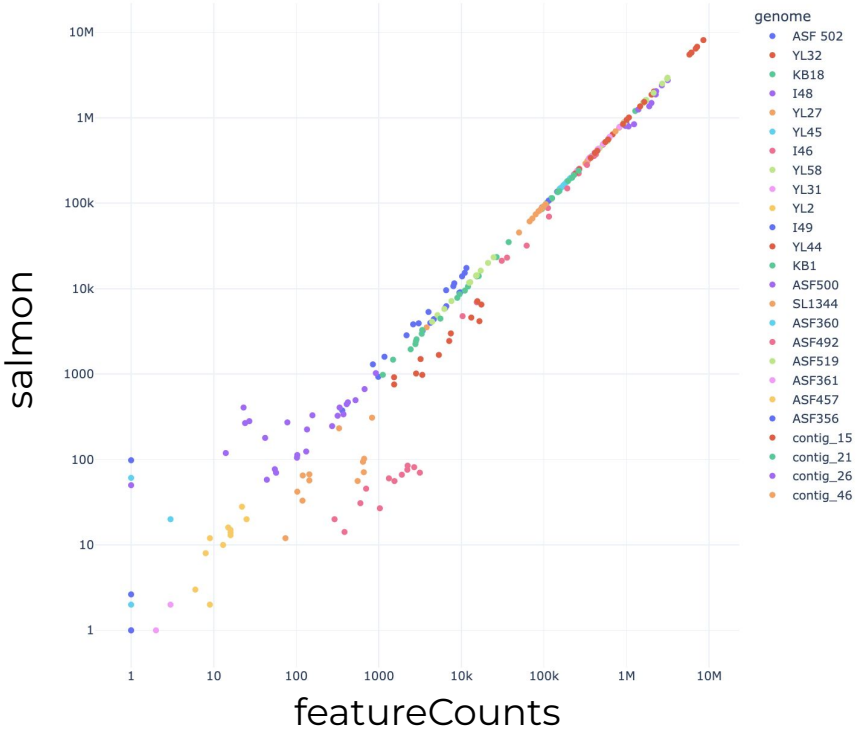
How do I measure transcript abundance?

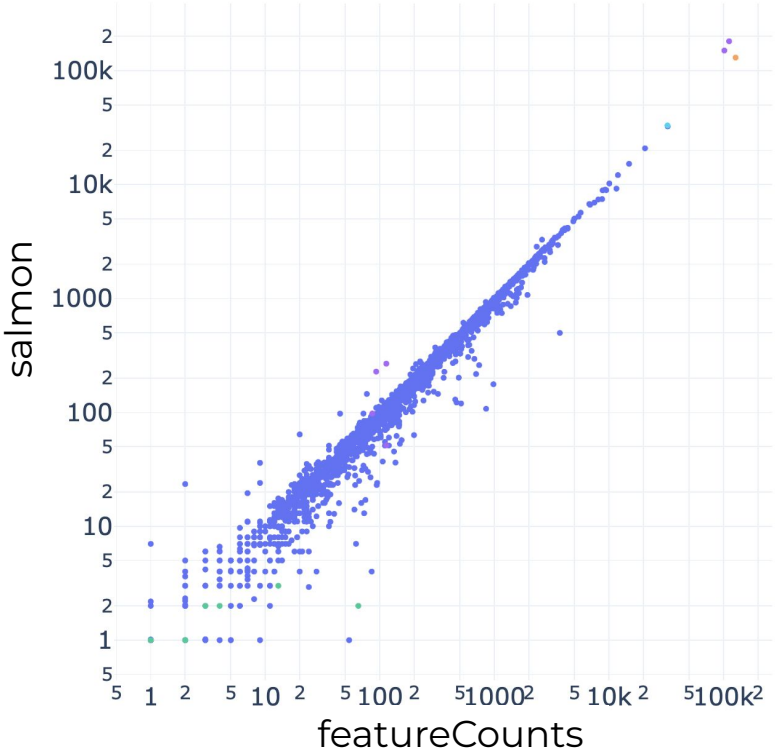


- Recent pre-print evaluates kallisto vs bowtie2/featureCounts in a metatranscriptomic benchmarking study

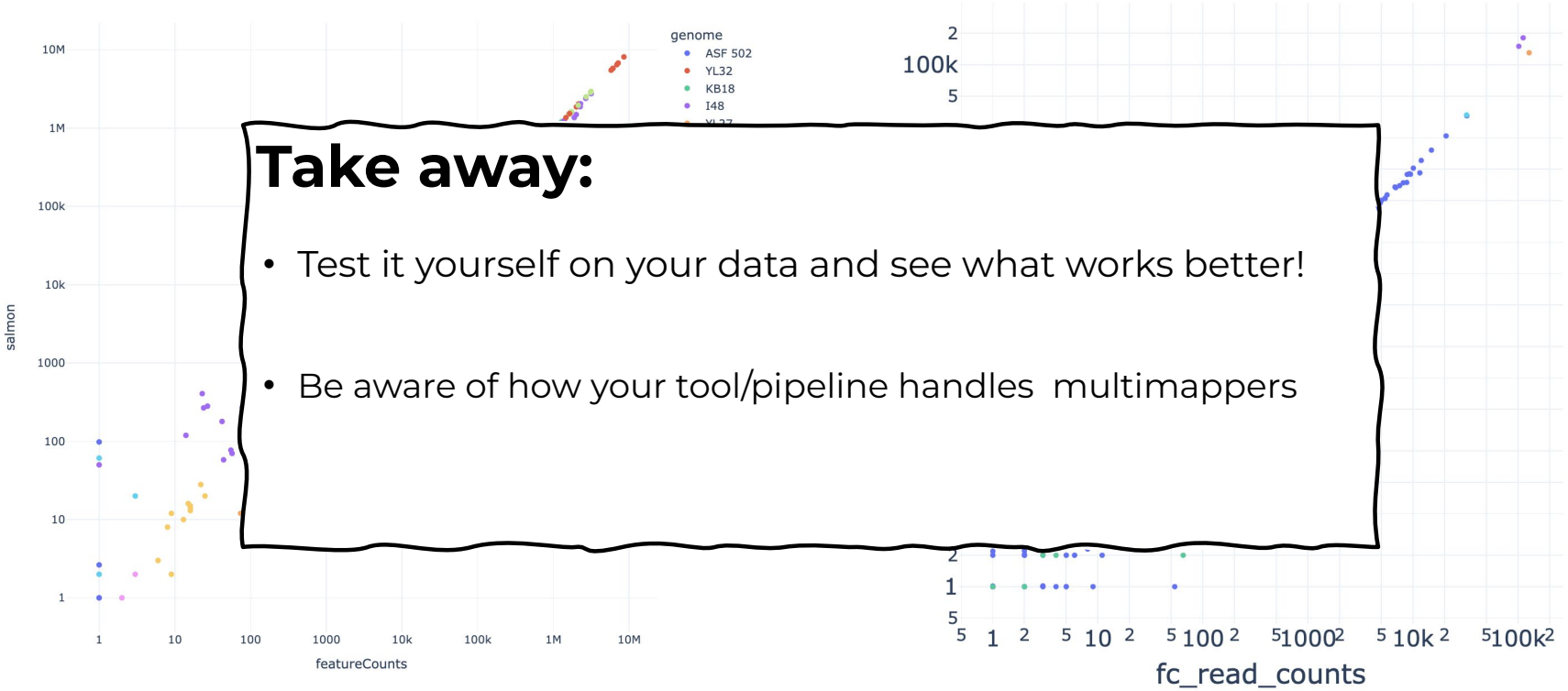
Ref

Different counting methods produce consistent results





Different counting methods produce consistent results



Take away:

- Test it yourself on your data and see what works better!
- Be aware of how your tool/pipeline handles multimappers

From Counts to Biology

	SAMPLE0001	SAMPLE0002	SAMPLE0003	SAMPLE0004	SAMPLE0005
BUG0001_GROUP000001	17	35	9	4	0
BUG0001_GROUP000003	1	27	1	7	8
BUG0001_GROUP000005	30	0	15	10	0
BUG0001_GROUP000006	0	20	6	18	16
BUG0001_GROUP000007	0	28	6	6	7
BUG0001_GROUP000008	7	14	0	0	6

But wait!

Can we directly compare counts?

NO! Because of:

- Sequencing depth differences
- Gene length differences
- **Taxonomic abundance differences** ← Unique to metatranscriptomics!

Part 1: From raw data to count tables

Part 2: Data analysis and normalization

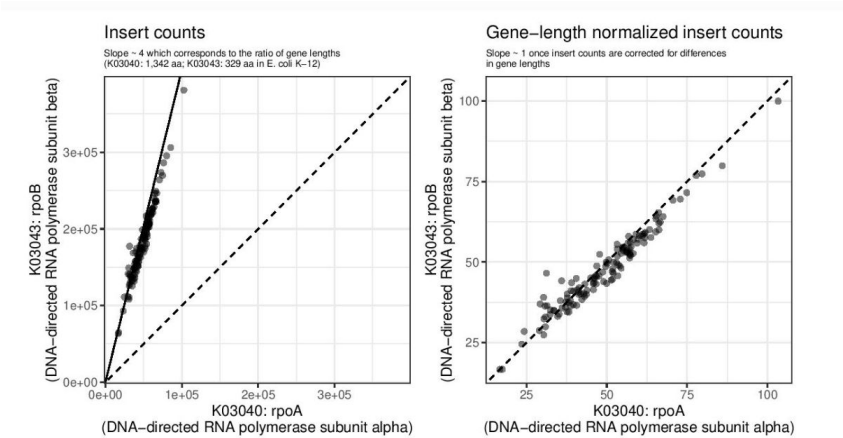
Learning Objectives

- Sources of bias in metatranscriptomic data (gene length, depth, composition)
- Normalization strategies
- DESeq2's median-of-ratios method and its limitations for metatranscriptomics
- Taxon-specific normalization for handling compositional effects
- Statistical testing with negative binomial models (DESeq2)

Sources of Bias in Metatranscriptomics

Gene Length

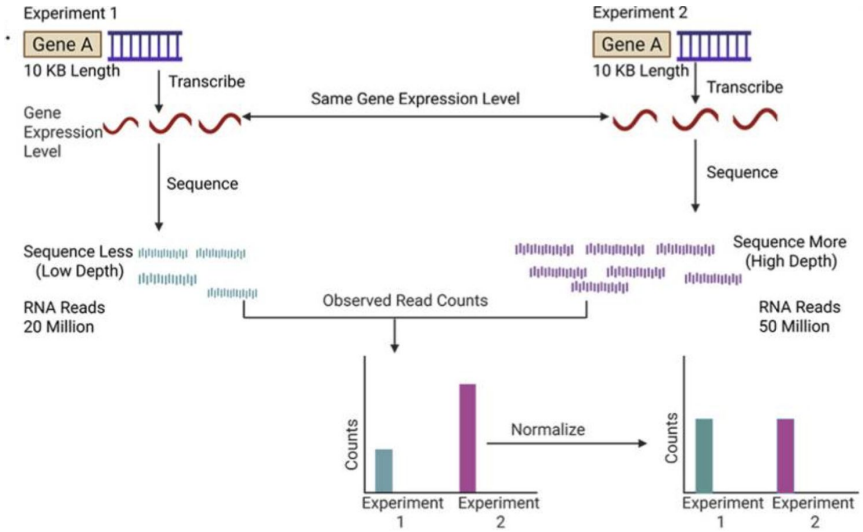
- Longer genes accumulate more reads
- Example: 3kb gene gets 3× more reads than 1kb gene at same expression level



Sources of Bias in Metatranscriptomics

Differences in sequencing depth between samples

- Example: 10M reads vs 5M reads - twice as many counts, but not twice the expression

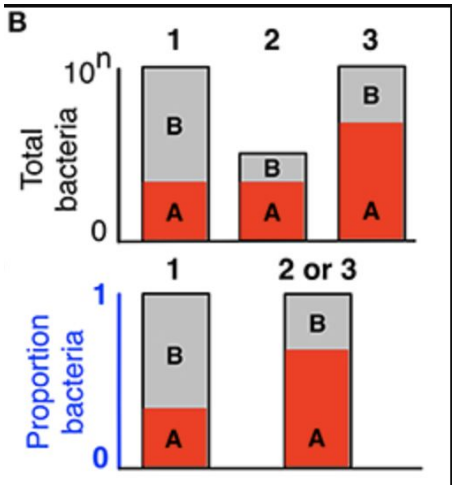


Ref

Sources of Bias in Metatranscriptomics

Compositionality

- The number of inserts for a given gene in a given sample is in itself arbitrary and can only be interpreted relative to the rest of the genes in the sample.

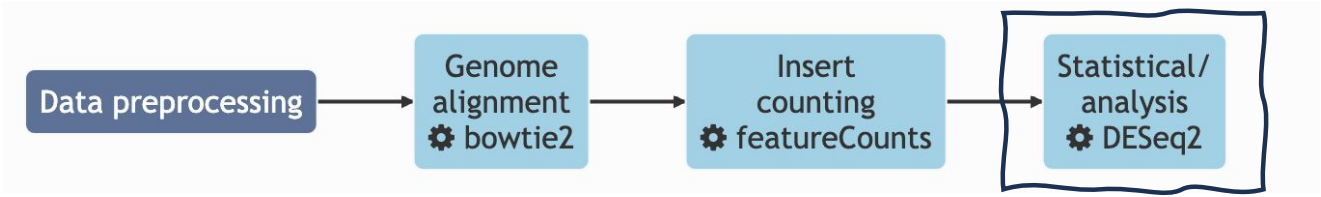


Common normalisation strategies

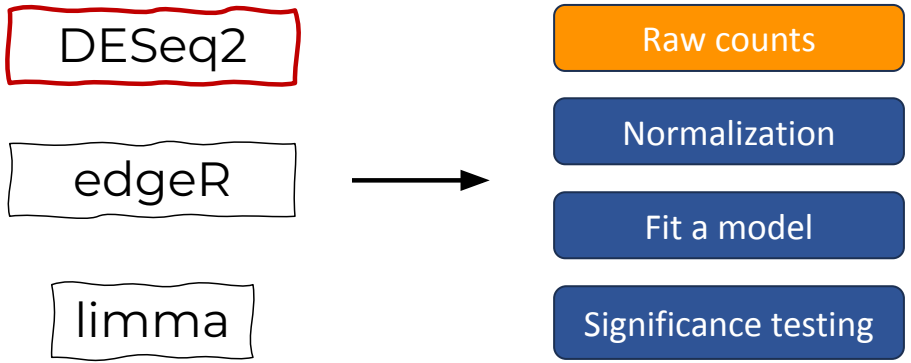
Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**

https://github.com/hbctraining/DGE_workshop/

How do we perform DE analysis?



How do we perform DE analysis?



Common sources of bias (review):

- Sequencing depth
- Gene length
- Compositionality

Samples

Genes/Features

	SAMPLE0001	SAMPLE0002	SAMPLE0003	SAMPLE0004	SAMPLE0005
BUG0001_GROUP000001	17	35	9	4	0
BUG0001_GROUP000003	1	27	1	7	8
BUG0001_GROUP000005	30	0	15	10	0
BUG0001_GROUP000006	0	20	6	18	16
BUG0001_GROUP000007	0	28	6	6	7
BUG0001_GROUP000008	7	14	0	0	6

DESeq2-normalized counts: median of ratios method

- Step 1: Create a ‘reference sample’: geometric mean of each gene

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**
...	...	...	...

DESeq2-normalized counts: median of ratios method

- Step 2: Calculate ratio of each sample to the reference sample

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

https://github.com/hbctraining/DGE_workshop/

DESeq2-normalized counts: median of ratios method

- Step 3: Calculate sample specific **size factor** (median of ratios)

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

Size factor for sampleB is median of 0.78, 0.77, 0.72, 0.74, 1.35 ...

Size factor for sampleA is median of 1.28, 1.30, 1.39, 1.35, 0.59 ...

DESeq2-normalized counts: median of ratios method

- Step 4: Divide counts in each sample by the samples **size factor**

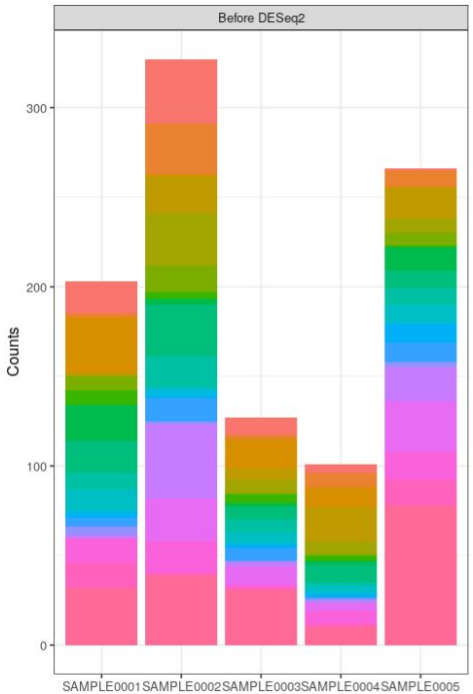
gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

Size factor for sampleB is median of 0.78, **0.77**, 0.72, 0.74, 1.35 ...

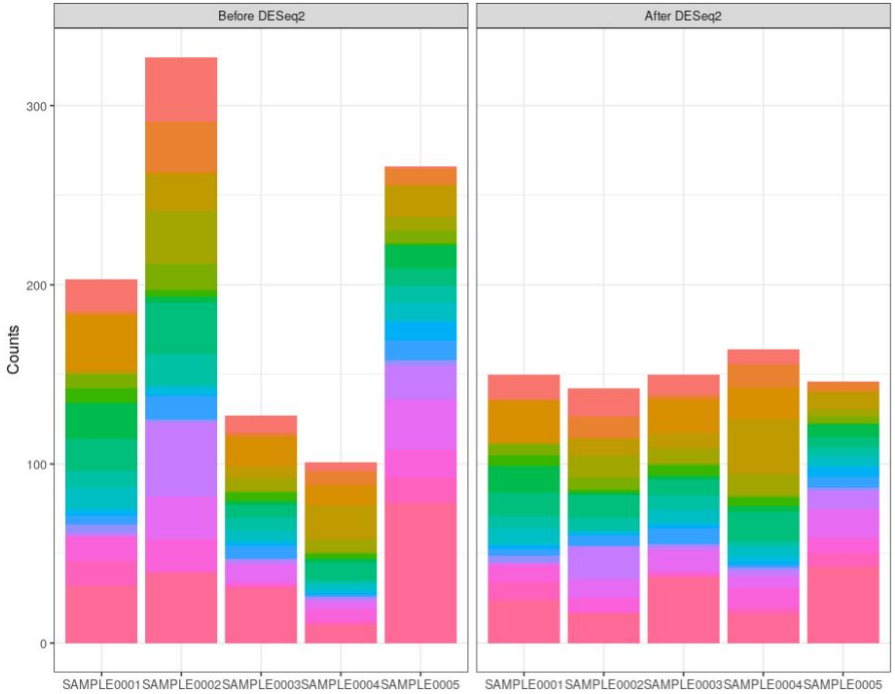
Size factor for sampleA is median of 1.28, 1.30, 1.39, **1.35**, 0.59 ...

https://github.com/hbctraining/DGE_workshop/

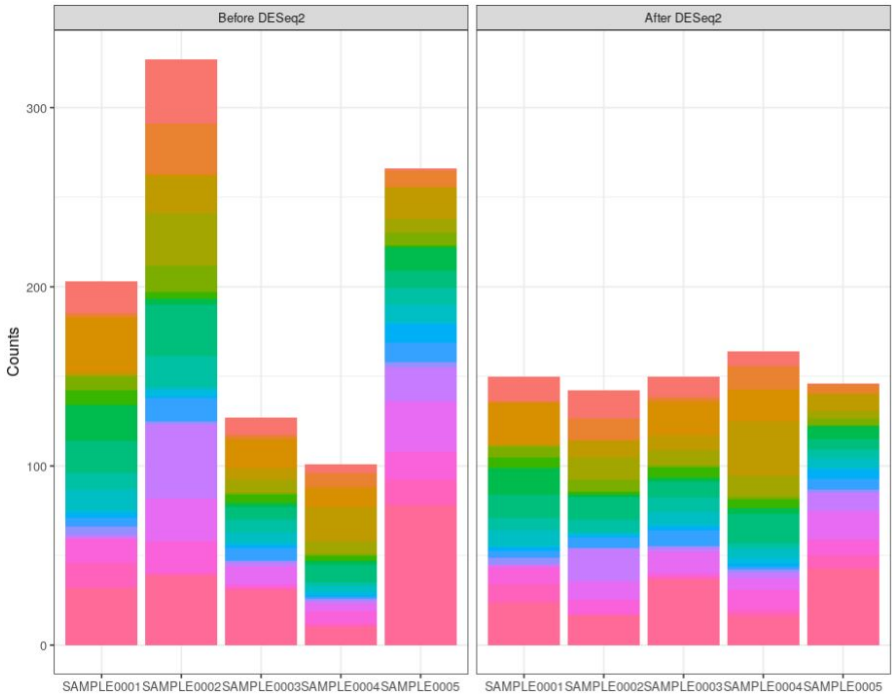
DESeq2-normalized counts: effects of normalisation

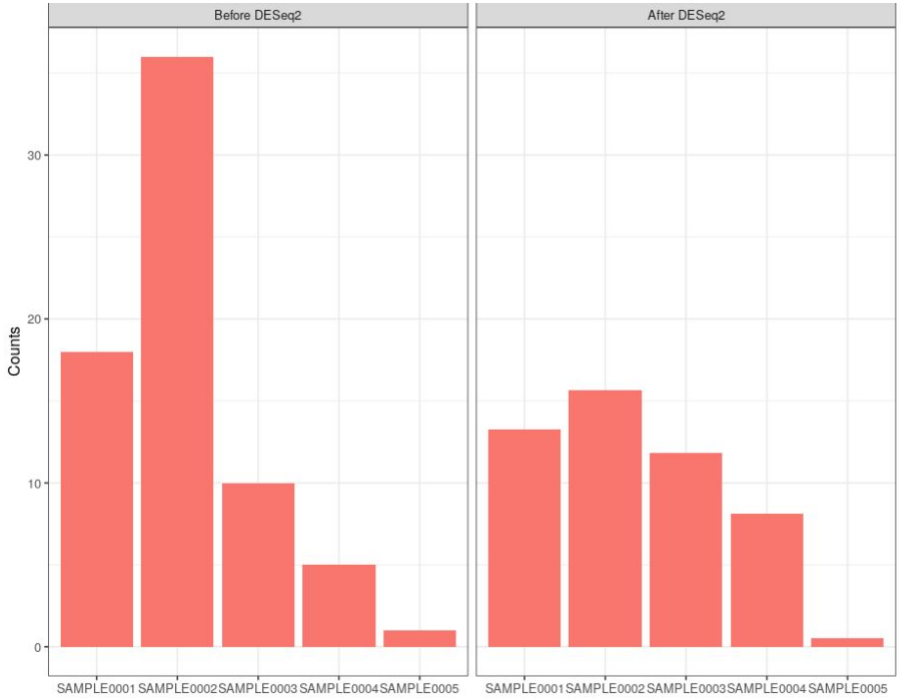


DESeq2-normalized counts: effects of normalisation



DESeq2-normalized counts: effects of normalisation





DESeq2-normalized counts: important considerations

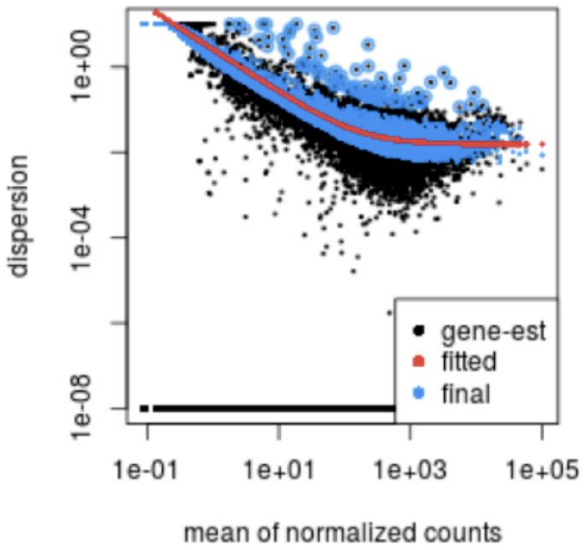
- **This method assumes that the majority of genes are not differentially expressed (i.e. majority of genes should have similar counts between samples)**

- Accounts for sequencing depth and composition (median values are robust against extreme outliers)

- Most of the time size factors should be around 1. Take note if there's a large variation between samples

DESeq2 estimation of variance

- Estimating within-group variance for each gene is not straight forward
- DESeq2 does this by estimating dispersion for each gene
- More about varia



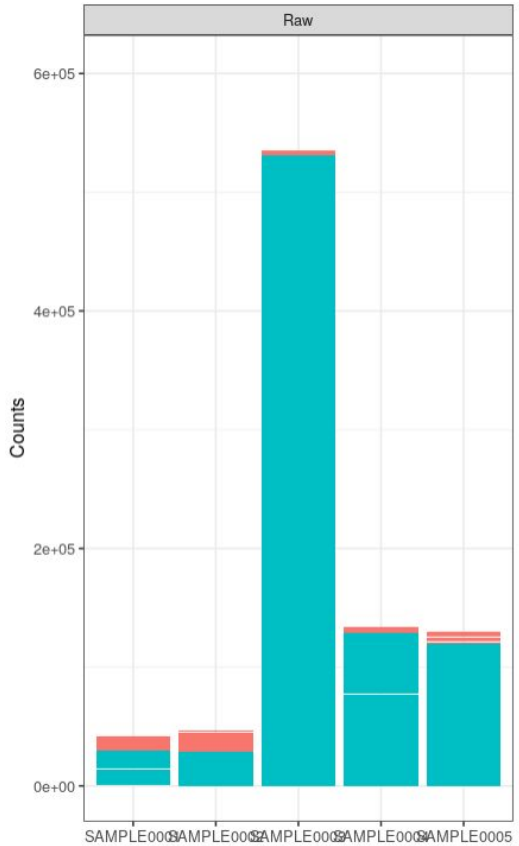
Fitting a model and testing

- DESeq2 uses **negative binomial distribution** to model the data
- For each gene, fit a GLM model

Gene expression ~ Treatment
Gene expression ~ genotype + treatment

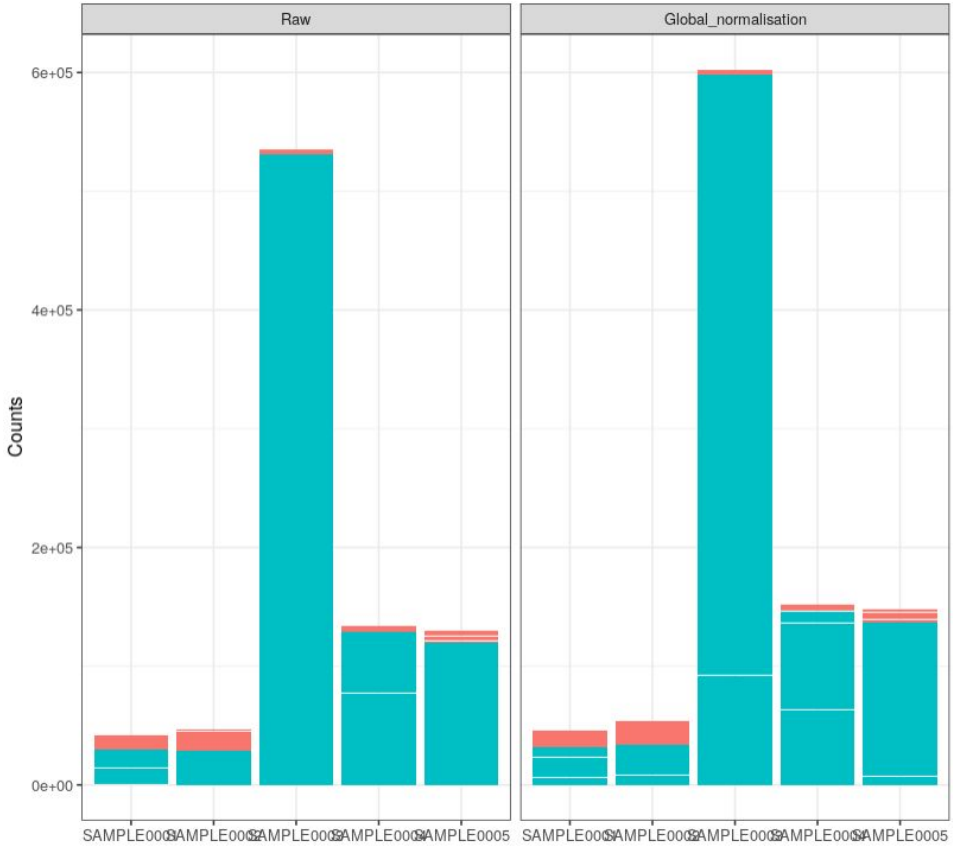
- By default runs Wald test to test for significance (Null hypothesis is no change between groups (LFC == 0)

Does DESeq2 work for metatranscriptomic data?



Changes in taxonomic abundance can have a confounding effect

Does DESeq2 work for metatranscriptomic data?



- **Want to investigate expression of each community member independently of changes in community composition**
- Standard RNA-seq normalisation fails in metatranscriptomics
- When one taxon blooms, all others appear to decrease in relative abundance, even if their absolute expression is unchanged. This zero-sum constraint makes it impossible to distinguish true functional changes from taxonomic shifts.

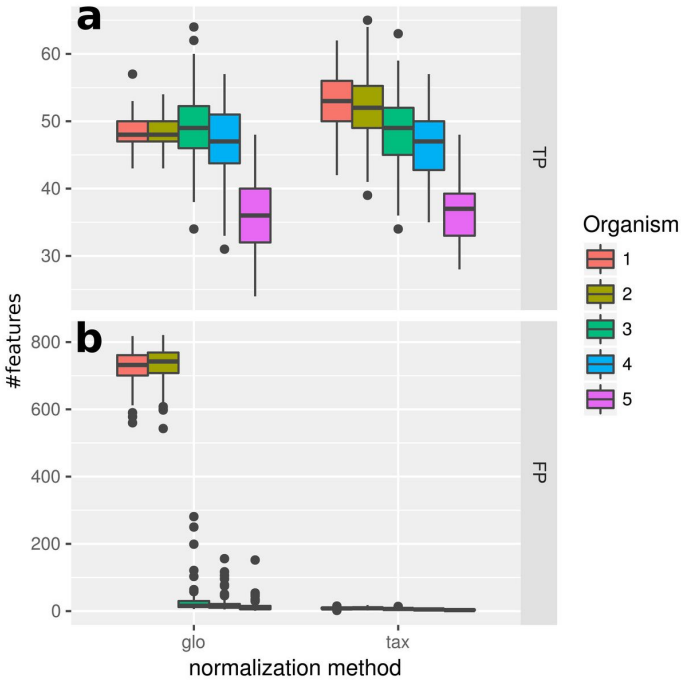
JOURNAL ARTICLE

Statistical approaches for differential expression analysis in metatranscriptomics

Yancong Zhang, Kelsey N Thompson, Curtis Huttenhower, Eric A Franzosa

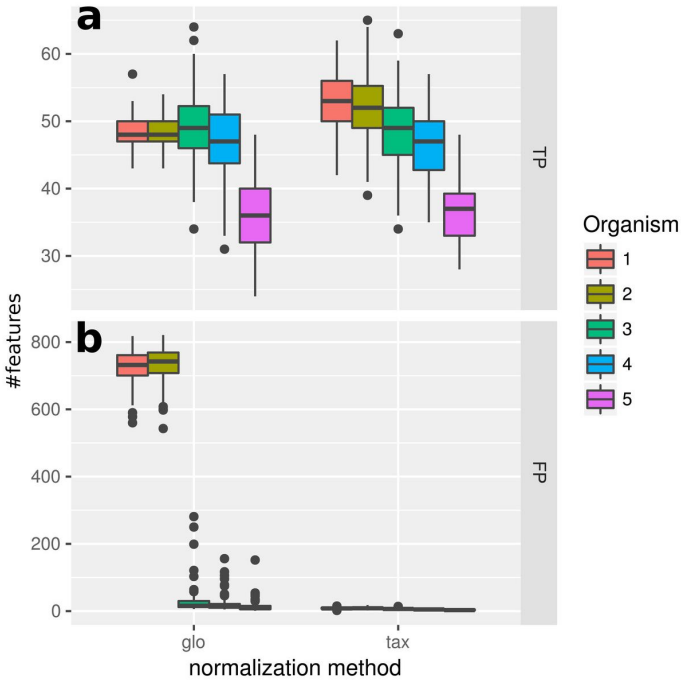
Bioinformatics, Volume 37, Issue Supplement_1, July 2021, Pages i34–i41,
https://doi.org/10.1093/bioinformatics/btab327
Published: 12 July 2021

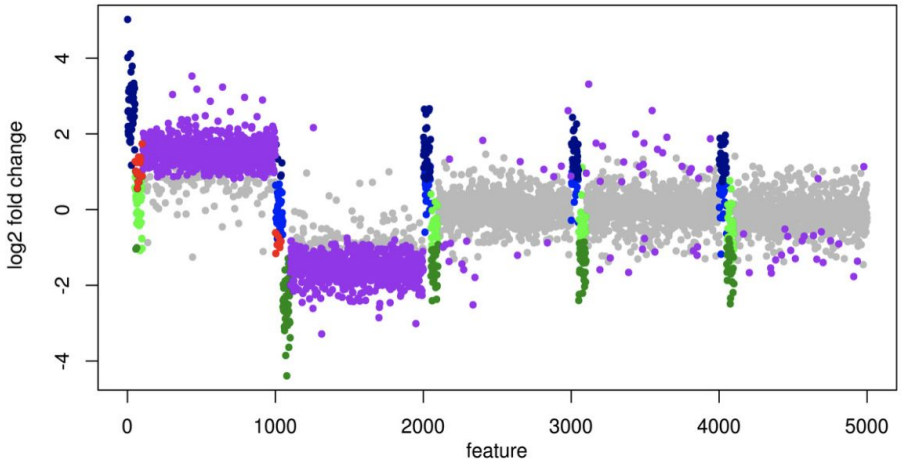
Why taxon-specific scaling?



Klingenberg H, Meinicke P. 2017

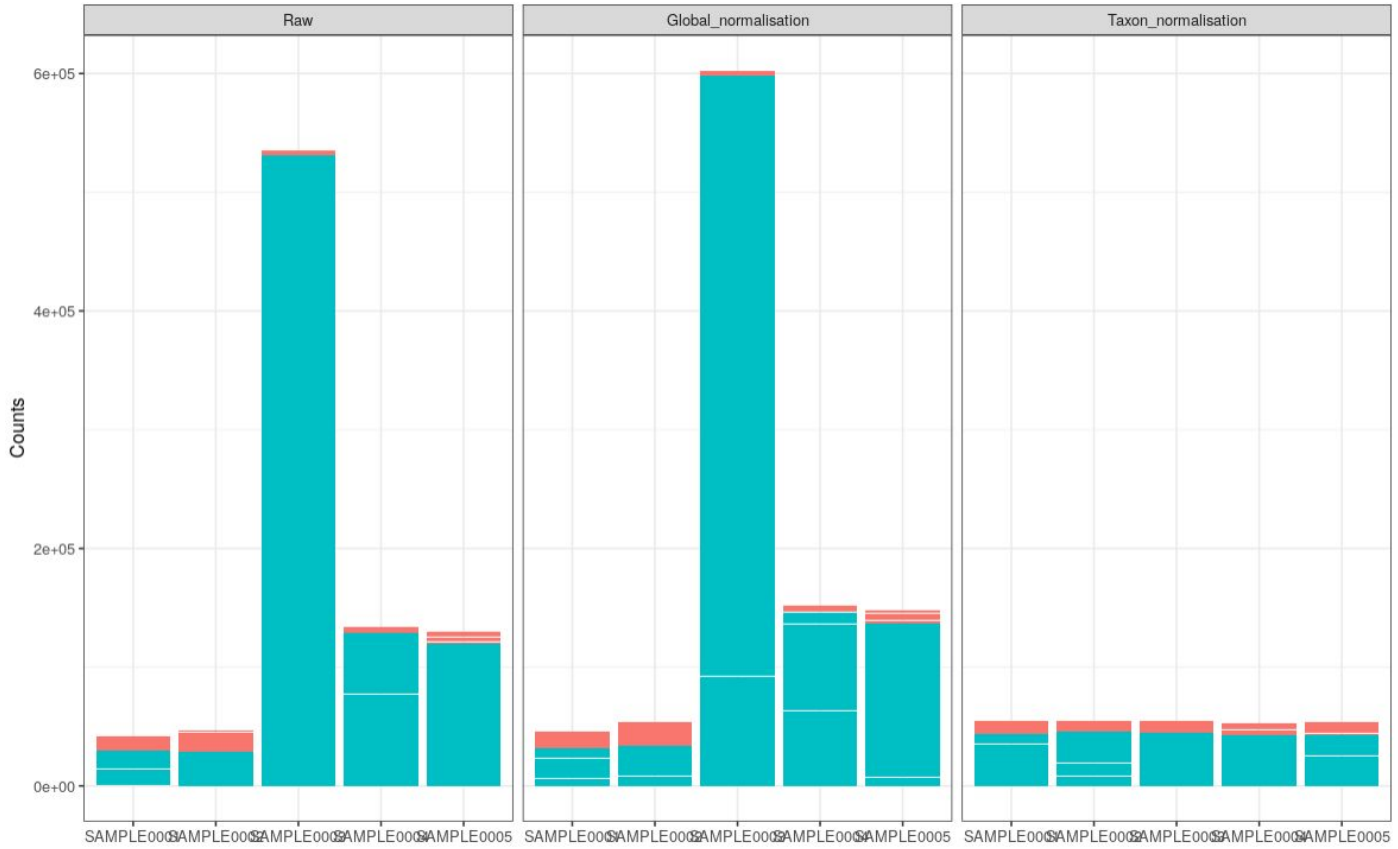
Why taxon specific scaling?





Klingenberg H, Meinicke P. 2017

Does DESeq2 works with within-taxon normalisation?



Take away

- Taxon specific scaling allows to account for variation in taxonomic abundances
- DE genes reflect actual change in the gene expression/behaviour
- Practically equivalent to analyzing N different RNAseq experiments, where N is the number of species in the community
- The authors provide R code to adapt DESeq2 for metatranscriptomic studies

From Reads to Counts - The Challenge

Goal: Quantify expression of each gene in each organism

Challenges specific to metatranscriptomics:

1. Reads can map to multiple organisms (homologous genes)
2. Changes in organism abundance affect total counts
3. Need to attribute reads to correct taxon
4. Computational efficiency with large references

Solution: Competitive mapping to concatenated genomes

Resources

- Methods in microbiomics now has pipelines for
  - Data preprocessing
  - Gene catalog building
  - Matching METAT/METAG data
- Harvard Chan Bioinformatics Core has some good resources for general RNAseq data analysis & other bioinformatic workflows

Case Study 1: The "Ambiguous Cousin"

Scenario: A researcher is studying a defined community containing *E. coli* K-12 with and without Salmonella Typhimurium. To get their count tables, they decide to align their metatranscriptomic reads first to the K-12 genome reference, and then take all unmapped reads and align them to the Salmonella genome.

Questions:

- What is the major flaw in this sequential mapping approach?
- How would the counts for the pathogenic strain be biased?
- **Bonus:** What is the standard "best practice" pipeline to fix this?

Case Study 2: The "Blooming" Problem

Scenario: You are analyzing a 4-species community in a gut model. You treat the community with a prebiotic that causes Species A to grow explosively (10x increase in biomass). Species B, C, and D are unaffected physiologically—they do not grow or change their gene expression. You sequence the total RNA of the community and normalize the data using standard TPM (Transcripts Per Million) or standard DESeq2 (calculating size factors across the whole community).

Questions:

- What will the differential expression results look like for Species B, C, and D?
- Why does this happen?

Case Study 3: The Budget Dilemma

Scenario: A student wants to compare the metatranscriptome of a mouse gut community before and after a dietary switch. They have a limited budget. They propose sequencing only 2 biological replicates per group but sequencing them very deeply (100 million reads each) to ensure they capture low-abundance transcripts.

Questions:

- Critique this experimental design. Is sequencing depth the limiting factor here?
- If they find a gene is 2-fold upregulated, can they trust it?

Case Study 4: Distinguishing Growth from Regulation

Scenario: In a defined community experiment, you see that *Salmonella* transcript counts of the virulence factor increase 4-fold in the treatment group compared to control (assume same sequencing depth across samples). You also know from 16S data that the abundance of *Salmonella* cells increased 4-fold.

Questions:

- Is the transcript abundance for the virulence factor different between treatment and control?
- Is the gene expression of the virulence factor different between treatment and control?
- How should you normalize the data to find genes that are truly being turned on specifically to handle the treatment?

Case Study 1: The "Ambiguous Cousin"

Scenario: A researcher is studying a defined community containing *E. coli* K-12 with and without Salmonella Typhimurium. To get their count tables, they decide to align their metatranscriptomic reads first to the K-12 genome reference, and then take all unmapped reads and align them to the Salmonella genome.

Questions:

- What is the major flaw in this sequential mapping approach?
- How would the counts for the pathogenic strain be biased?
- **Bonus:** What is the standard "best practice" pipeline to fix this?

Case Study 2: The "Blooming" Problem

Scenario: You are analyzing a 4-species community in a gut model. You treat the community with a prebiotic that causes Species A to grow explosively (10x increase in biomass). Species B, C, and D are unaffected physiologically—they do not grow or change their gene expression. You sequence the total RNA of the community and normalize the data using standard TPM (Transcripts Per Million) or standard DESeq2 (calculating size factors across the whole community).

Questions:

- What will the differential expression results look like for Species B, C, and D?
- Why does this happen?

Case Study 3: The Budget Dilemma

Scenario: A student wants to compare the metatranscriptome of a mouse gut community before and after a dietary switch. They have a limited budget. They propose sequencing only 2 biological replicates per group but sequencing them very deeply (100 million reads each) to ensure they capture low-abundance transcripts.

Questions:

- Critique this experimental design. Is sequencing depth the limiting factor here?
- If they find a gene is 2-fold upregulated, can they trust it?

Case Study 4: Distinguishing Growth from Regulation

Scenario: In a defined community experiment, you see that *Salmonella* transcript counts of the virulence factor increase 4-fold in the treatment group compared to control (assume same sequencing depth across samples). You also know from 16S data that the abundance of *Salmonella* cells increased 4-fold.

Questions:

- Is the transcript abundance for the virulence factor different between treatment and control?
- Is the gene expression of the virulence factor different between treatment and control?
- How should you normalize the data to find genes that are truly being turned on specifically to handle the treatment?

Part 1: From raw data to count tables

Part 2: Data analysis and normalization

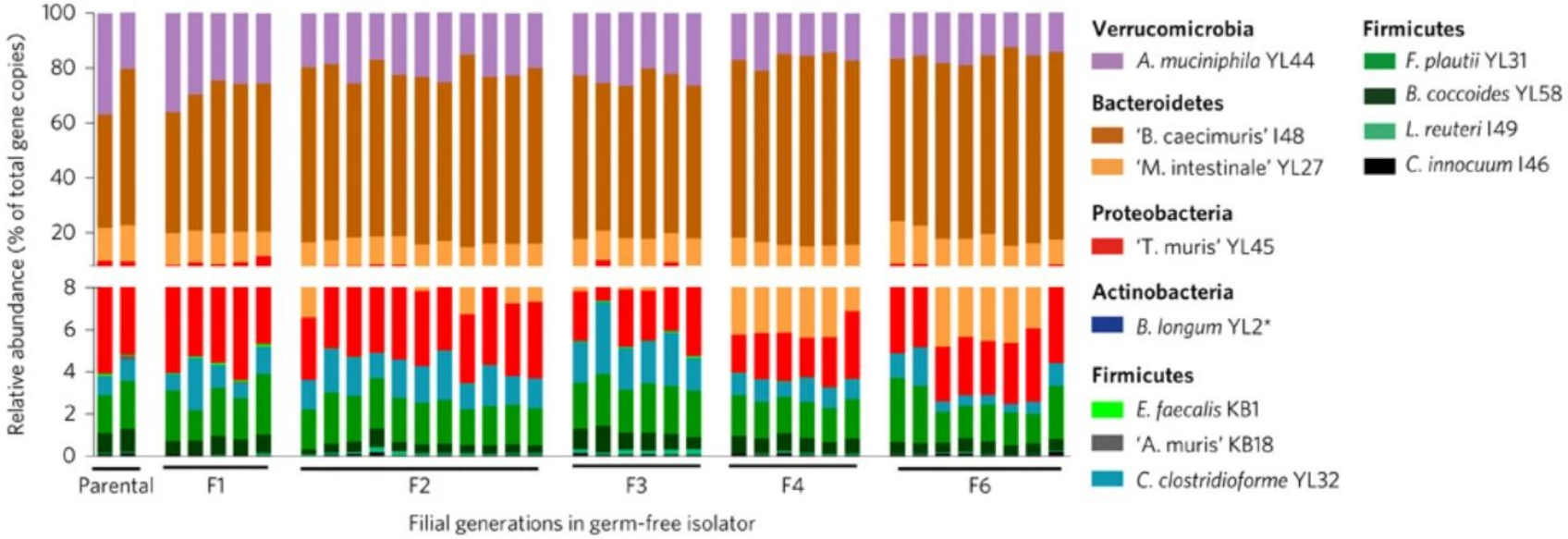
Part 3: Case study with 12-strain mouse community

OligoMM12 community

- 12 bacterial strains representing the five most prevalent and abundant phyla of mouse gut microbiome
- Genomes sequenced and well-annotated*
- Colonizes germ-free mice stably
- Provides partial colonization resistance to Salmonella infection

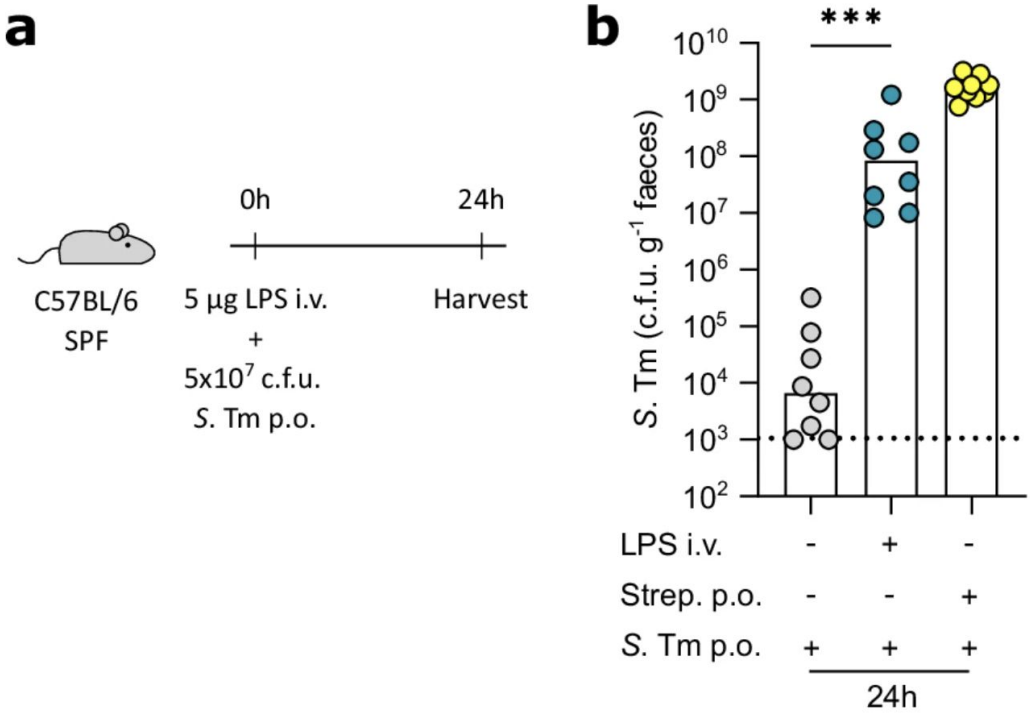
Ref

OligoMM12 community



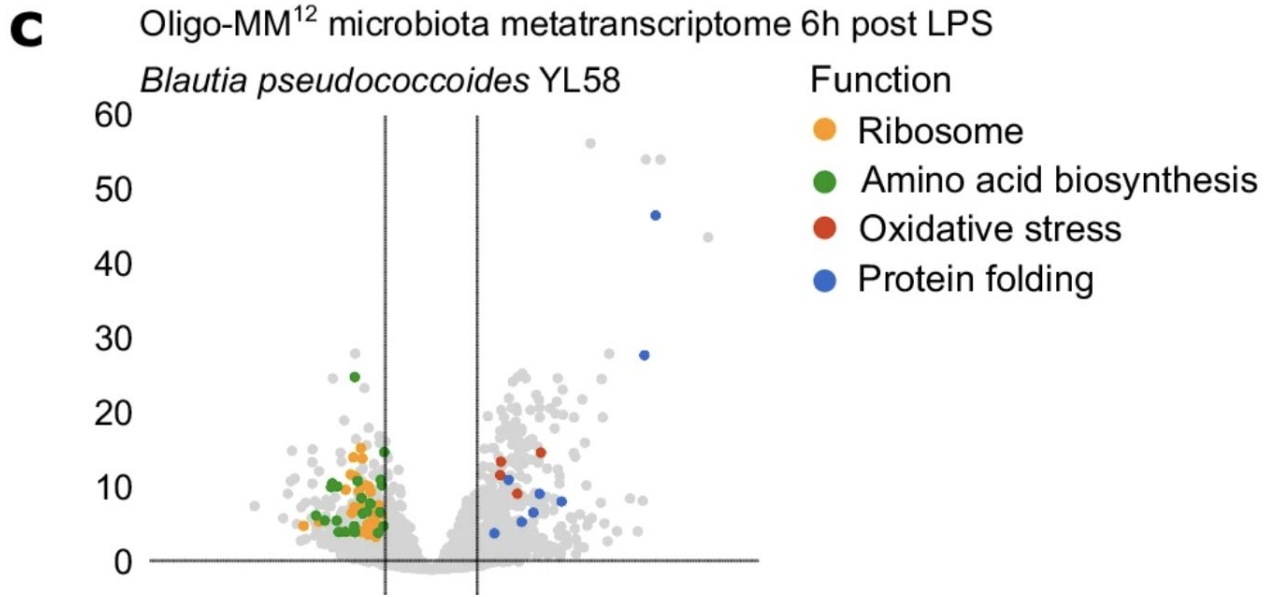
Ref

Sublethal systemic LPS exposure promotes gut-luminal pathogens to bloom in a TLR4-dependent manner



Ref

How does microbiome responds to LPS treatment?



Before we begin...

- Does everyone know how to login to the server?

  `https://motus-sib-course.ethz.ch/`

- Can everyone copy `/course_data/nccr_course_jan26/metatranscriptomics/metat_notebooks` to their home folder?