Data Workshop 3

BABS4 (66I) – Gene expression and biochemical interactions strand

Haemophilus influenzae (Hi) gene expression analysis

Remember they key questions of this module:

1) Is there a **functional gam homologue** in *Haemophilus influenzae* (HiGam), and, if so,
2) Is HiGam part of a **non-canonical** competence response?

"not the usual"

In these data practicals (data workshops 3 and 4) you will ask:

a) How does gene expression change during **canonical** competence in Hi?
b) What is the role (if any) of HiGam in this response?

Haemophilus influenzae (Hi) gene expression analysis

You will analyse RNAseq counts data which will support your lab practicals

The data is from this publication:

PLOS ONE

RESEARCH ARTICLE

A competence-regulated toxin-antitoxin system in *Haemophilus influenzae*

Hailey Findlay Black[1¤a], Scott Mastromatteo[1¤b], Sunita Sinha[2], Rachel L. Ehrlich[3], Corey Nislow[4], Joshua Chang Mell[3], Rosemary J. Redfield[1*]

Citation: Findlay Black H, Mastromatteo S, Sinha S, Ehrlich RL, Nislow C, Chang Mell J, et al. (2020) A competence-regulated toxin-antitoxin system in *Haemophilus influenzae*. PLoS ONE 15(1): e0217255. https://doi.org/10.1371/journal.pone.0217255

What are counts?

In RNAseq, counts refer to the number of reads which came from a gene

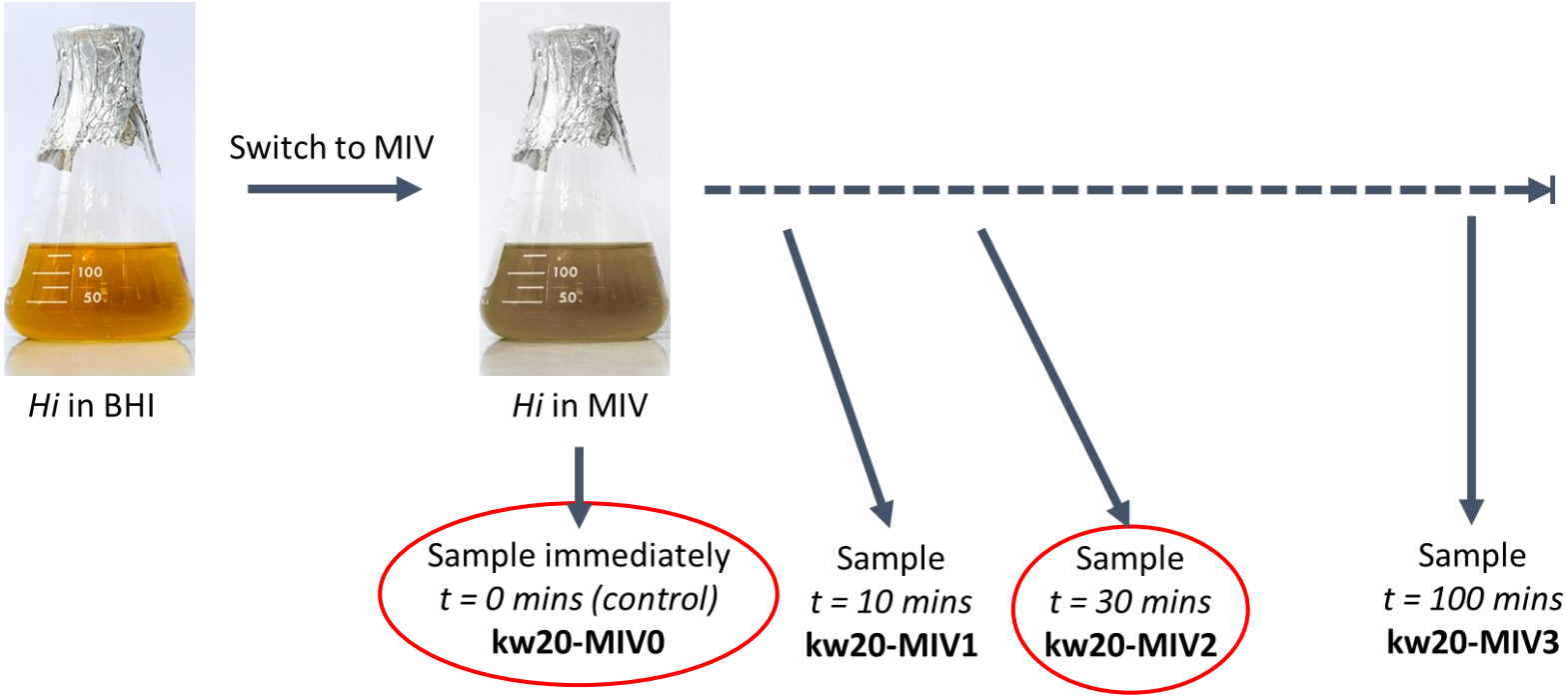
	$Sample_1$	$Sample_2$	…	$Sample_n$
$Gene_1$	4200	3217	…	4231
$Gene_2$	24	38	…	37
…	…	…	…	…
$Gene_n$	178	142	…	205

Our dataset has 33 samples and counts for 1745 gene features

- as a minimum you will analyse 6 of these samples
- analysing more will inform your understanding of competence in Hi

Core dataset

Wildtype Hi (*kw20* strain) grown in rich medium (BHI) then transferred to starvation medium (MIV), all in triplicate



Main focus on the *t*=0 (control) versus *t*=30 comparison

Additional datasets

Condition	Strain	Media	OD_{600}	time (mins)	replicates
kw20-BHI1	kw20	BHI	0.02	NA	-F,-G,-H
kw20-BHI2	kw20	BHI	0.60	NA	-F,-G,-H
kw20-BHI3	kw20	BHI	1.00	NA	-F,-G,-H
kw20-MIV0	kw20	MIV	0.25	0	-A,-B,-C
kw20-MIV1	kw20	MIV	0.25	10	-A,-B,-C
kw20-MIV2	kw20	MIV	0.25	30	-A,-B,-C
kw20-MIV3	kw20	MIV	0.25	100	-A,-B,-C
sxyx-MIV0	Sxy-	MIV	0.25	0	-A,-B,-D
sxyx-MIV1	Sxy-	MIV	0.25	10	-A,-B,-D
sxyx-MIV2	Sxy-	MIV	0.25	30	-A,-B,-D
sxyx-MIV3	Sxy-	MIV	0.25	100	-A,-B,-D

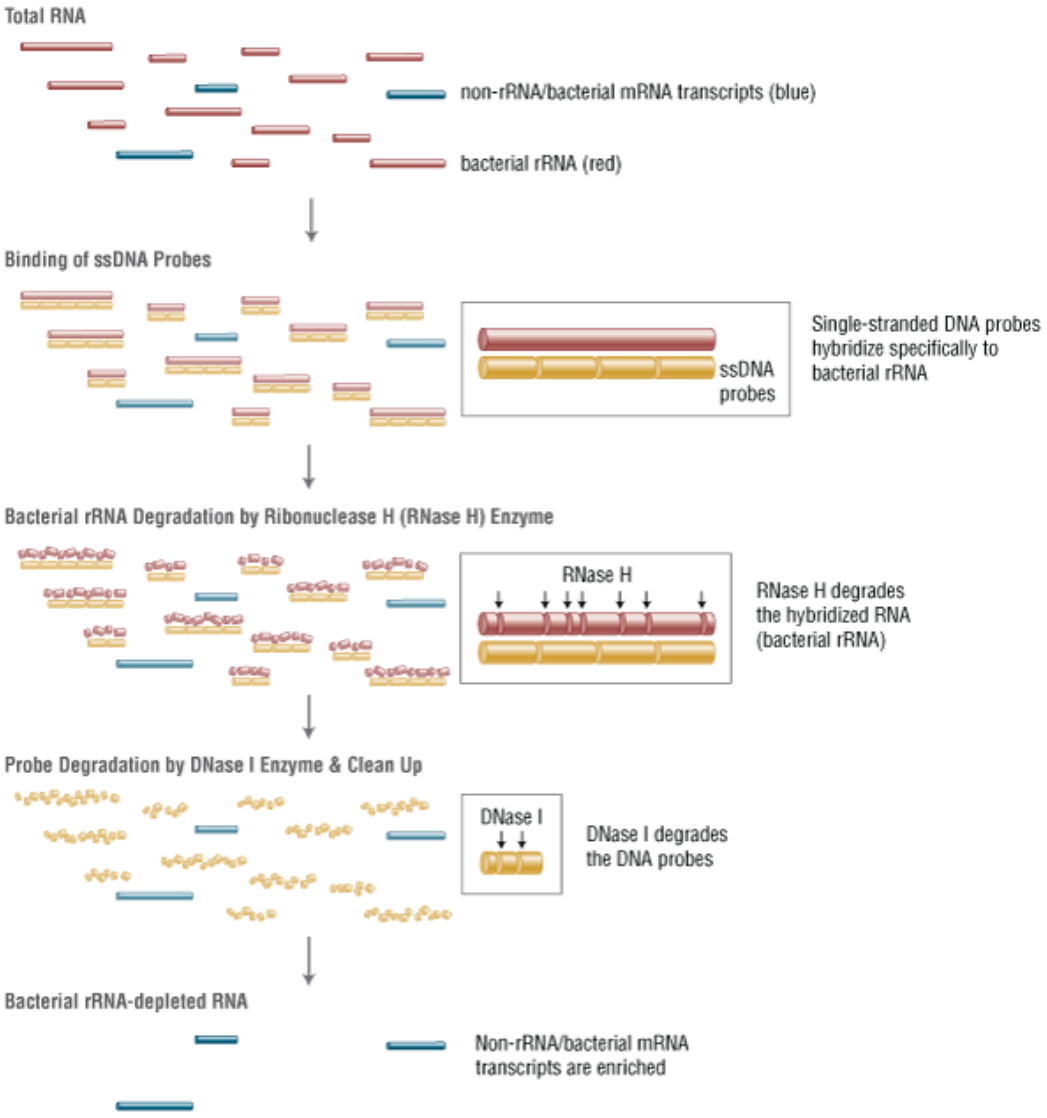
CORE

Competition?

Other timepoints?

Impact of competence regulatory mutant?

rRNA depletion libraries



Your dataset contains these transcript biotypes:

- protein-coding
- tRNA
- rRNA

rRNA is a technical leftover

- >90% of cellular RNA is rRNA

rRNA ‘expression’ in this dataset is **not quantitative**

rRNA ≠ rpL or rpS genes

- these are ribosomal protein genes

This workshop

Set up your RStudio Project *(remember that this is also assessed, not just the report)*

Load your libraries *(with any installations)*

Load and inspect your datasets

Normalise your data to get expression in transcripts per million (TPM)

Run Principal Component Analysis (PCA)

Run differential expression analysis (DEA)

Today’s datasets

Hi_PRJNA293882_counts.tsv

The RNAseq data. Tab separated data file. Sample names in the first row and feature names in the first column.

Hi_feature_names.tsv

Tab separated data file which links the Hi features IDs with known gene symbols (if possible).

Hi_feature_locations.bed

BED format file indicating locations of each feature in the Hi genome.

Data Workshop 3

BABS4 (66I) – Gene expression and biochemical interactions strand