De novo Assembly and
Differential GENE Expression of
Enterococcus faecium E745

Paper I – RNA-seq and Tn-seq reveal fitness determinants of vancomycin-resistant *Enterococcus faecium* during growth in human serum (Zhang *et al.* 2017)

Introduction

- The Gram-positive bacterium Enterococcus faecium mutual live in human gastrointestinal tract

Problem

- Cause bloodstream infections in hospitalized patients. (How????)
- The mechanisms by which E. faecium

Aim

- Identify genes that contribute to growth of E. faecium in human serum
  - Transcriptome profiling (RNA-seq)
  - transposon mutant library sequencing approach (Tn-seq)
- Differential expression analysis between rich medium and heat-inactivated serum

WORKFLOW

1. Quality check -FastQC	
2. Reads preprocessing: trimming + quality check (after trimmed) –Trimmomatic	
3. Genome Asembly	•PacBio +Nanopore – Canu •Illumina + Nanopore Assembly- SPAdes •Illumina+Nanopore BWA aligned PacBio + Nanopore
4. Assembly evaluation	•QUAST •MUMmerplot
5. Structural and functional annotation- Prokka	•Artemis
6. Homology Search	•Blastn •Artemis ACT
7. Mapping (RNA-Seq reads alignment against the assembled genome)	•BWA-MEM
8. Expression analyses	•Read counting-HTSeq •Differential expression analysis- DESeq2

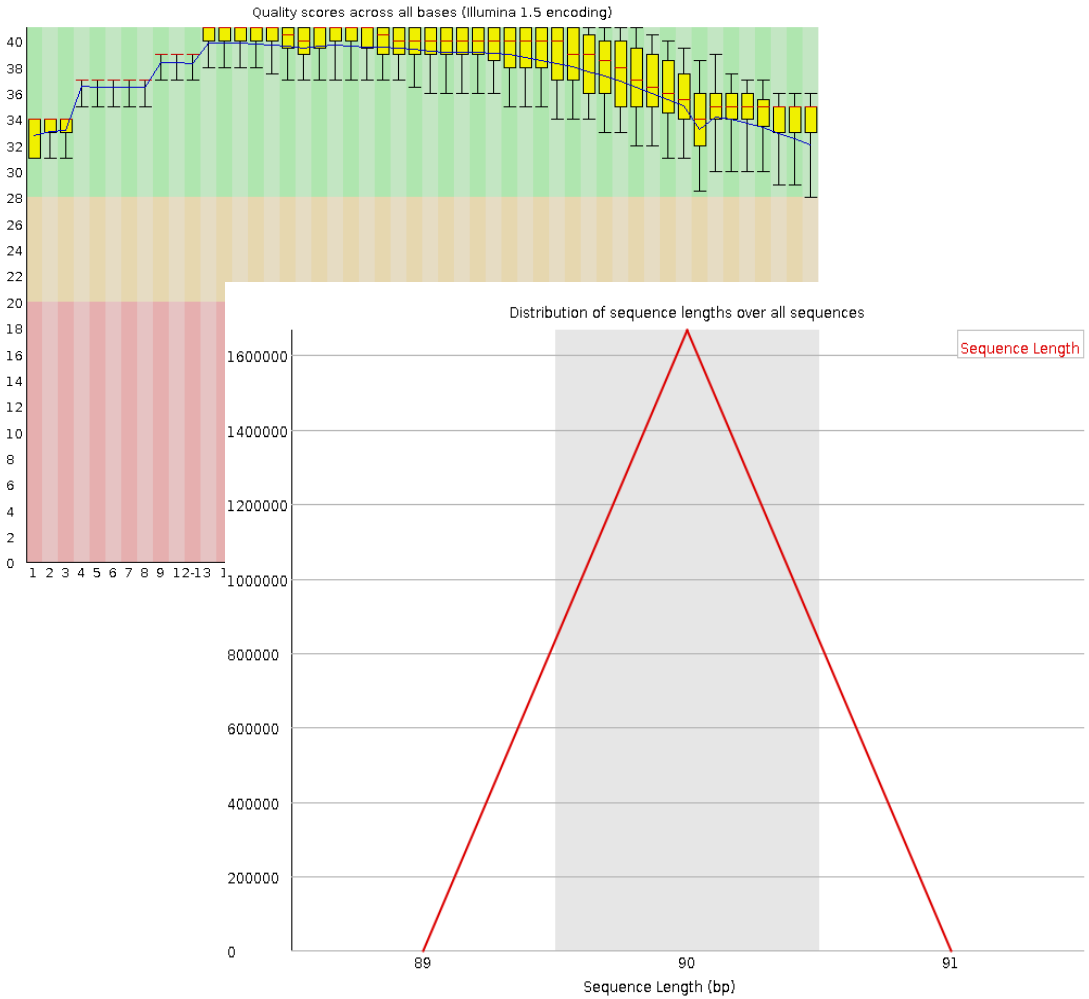
Data

- DNAseq from
  - PacBio RSII
  - illumina HiSeq 2500
  - NanoPore MinION 2D
- RNAseq (and Tn) from rich medium and human serum
  - illumina HiSeq 2500

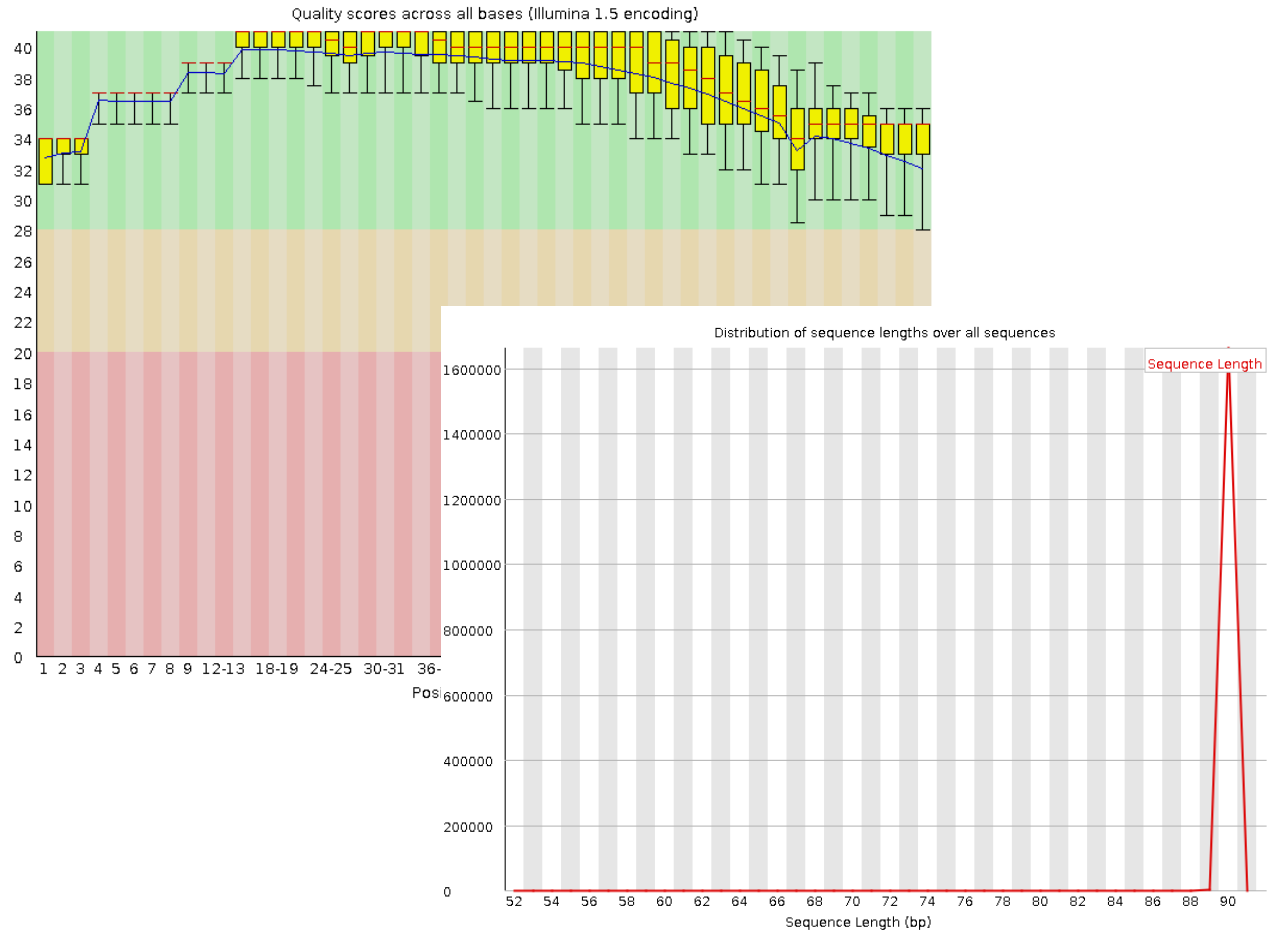
Results

- FastQC

Before



After trim



QUAST

Article:
1 circular chromosome: 2.764Mb
6 plasmids: 9.3 - 223.7kbp
Total length: 3.168Mb
3095 predicted coding sequences

Canu

Assembly	faecium.contigs
# contigs (>= 0 bp)	13
# contigs (>= 1000 bp)	13
# contigs (>= 5000 bp)	12
# contigs (>= 10000 bp)	11
# contigs (>= 25000 bp)	6
# contigs (>= 50000 bp)	3
Total length (>= 0 bp)	3179422
Total length (>= 1000 bp)	3179422
Total length (>= 5000 bp)	3175915
Total length (>= 10000 bp)	3168876
Total length (>= 25000 bp)	3083119
Total length (>= 50000 bp)	2962638
# contigs	13
Largest contig	2775856
Total length	3179422
Reference length	2919198
GC (%)	37.78
Reference GC (%)	37.88
N50	2775856
NG50	2775856
N75	2775856
NG75	2775856
L50	1
LG50	1
L75	1
LG75	1
# misassemblies	190
# misassembled contigs	1
Misassembled contigs length	2775856
# local misassemblies	37
# scaffold gap ext. mis.	0
# scaffold gap loc. mis.	0
# unaligned mis. contigs	6
# unaligned contigs	5 + 8 part
Unaligned length	580864
Genome fraction (%)	84.982
Duplication ratio	1.047
# N's per 100 kbp	0.00
# mismatches per 100 kbp	392.86
# indels per 100 kbp	24.63
Largest alignment	137651

Pilon

Assembly	E_faecium_improved
# contigs (>= 0 bp)	13
# contigs (>= 1000 bp)	13
# contigs (>= 5000 bp)	12
# contigs (>= 10000 bp)	11
# contigs (>= 25000 bp)	6
# contigs (>= 50000 bp)	3
Total length (>= 0 bp)	3179422
Total length (>= 1000 bp)	3179422
Total length (>= 5000 bp)	3175915
Total length (>= 10000 bp)	3168876
Total length (>= 25000 bp)	3083119
Total length (>= 50000 bp)	2962638
# contigs	13
Largest contig	2775856
Total length	3179422
Reference length	2919198
GC (%)	37.78
Reference GC (%)	37.88
N50	2775856
NG50	2775856
N75	2775856
NG75	2775856
L50	1
LG50	1
L75	1
LG75	1
# misassemblies	190
# misassembled contigs	1
Misassembled contigs length	2775856
# local misassemblies	37
# scaffold gap ext. mis.	0
# scaffold gap loc. mis.	0
# unaligned mis. contigs	6
# unaligned contigs	5 + 8 part
Unaligned length	580862
Genome fraction (%)	84.999
Duplication ratio	1.047
# N's per 100 kbp	0.00
# mismatches per 100 kbp	392.78
# indels per 100 kbp	24.62
Largest alignment	137651

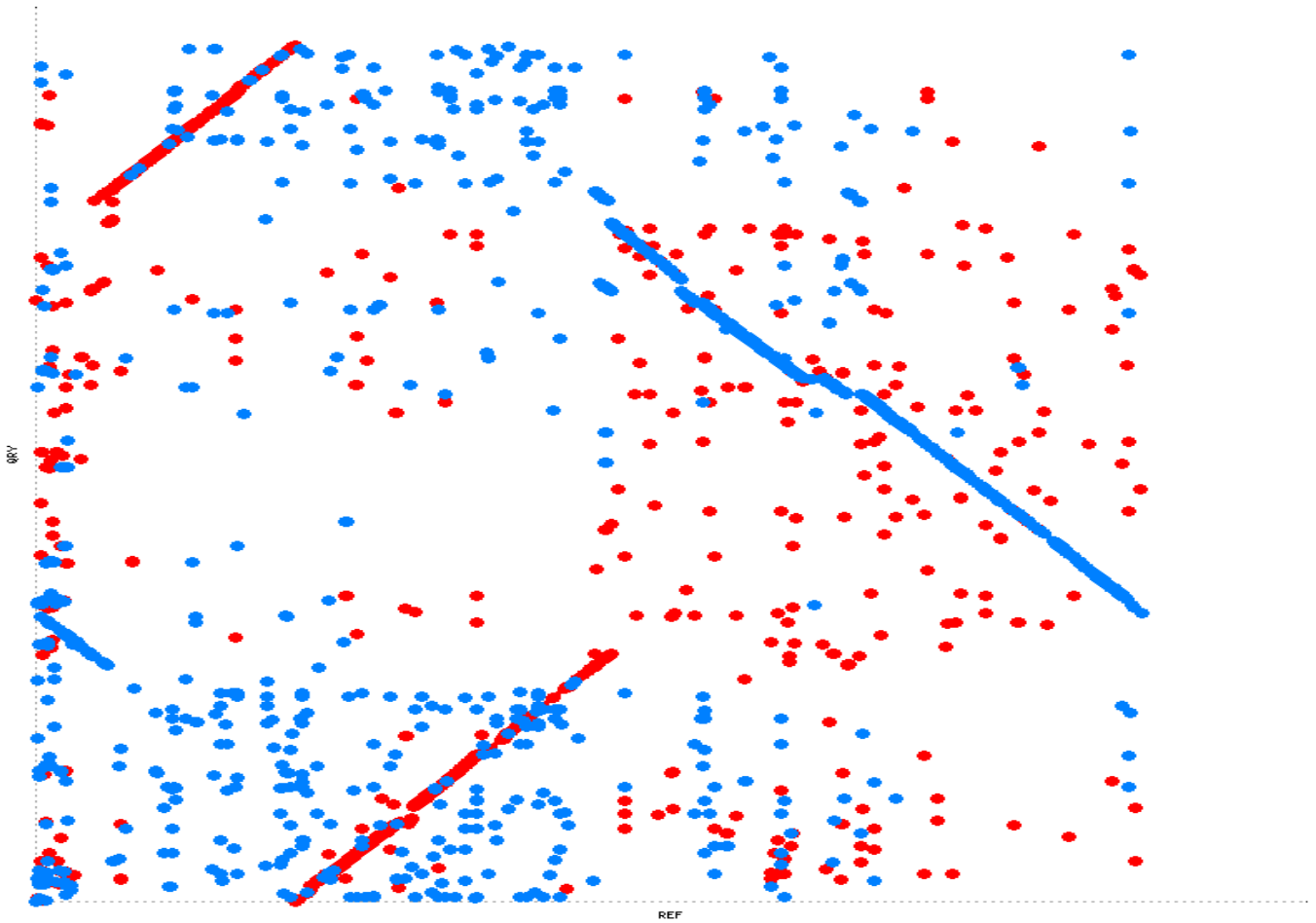
SPAdes

Assembly	scaffolds
# contigs (>= 0 bp)	71
# contigs (>= 1000 bp)	18
# contigs (>= 5000 bp)	16
# contigs (>= 10000 bp)	14
# contigs (>= 25000 bp)	13
# contigs (>= 50000 bp)	10
Total length (>= 0 bp)	3121385
Total length (>= 1000 bp)	3111126
Total length (>= 5000 bp)	3104566
Total length (>= 10000 bp)	3089939
Total length (>= 25000 bp)	3077556
Total length (>= 50000 bp)	2985873
# contigs	19
Largest contig	711421
Total length	3111645
Reference length	2919198
GC (%)	37.65
Reference GC (%)	37.88
N50	485854
NG50	485854
N75	204006
NG75	245525
L50	3
LG50	3
L75	6
LG75	5
# misassemblies	178
# misassembled contigs	8
Misassembled contigs length	2718970
# local misassemblies	38
# scaffold gap ext. mis.	1
# scaffold gap loc. mis.	1
# unaligned mis. contigs	5
# unaligned contigs	5 + 14 part
Unaligned length	559967
Genome fraction (%)	84.244
Duplication ratio	1.038
# N's per 100 kbp	27.67
# mismatches per 100 kbp	390.85
# indels per 100 kbp	14.72
Largest alignment	137660

Using plasmidfinder and Blastn I detected 5 plasmids

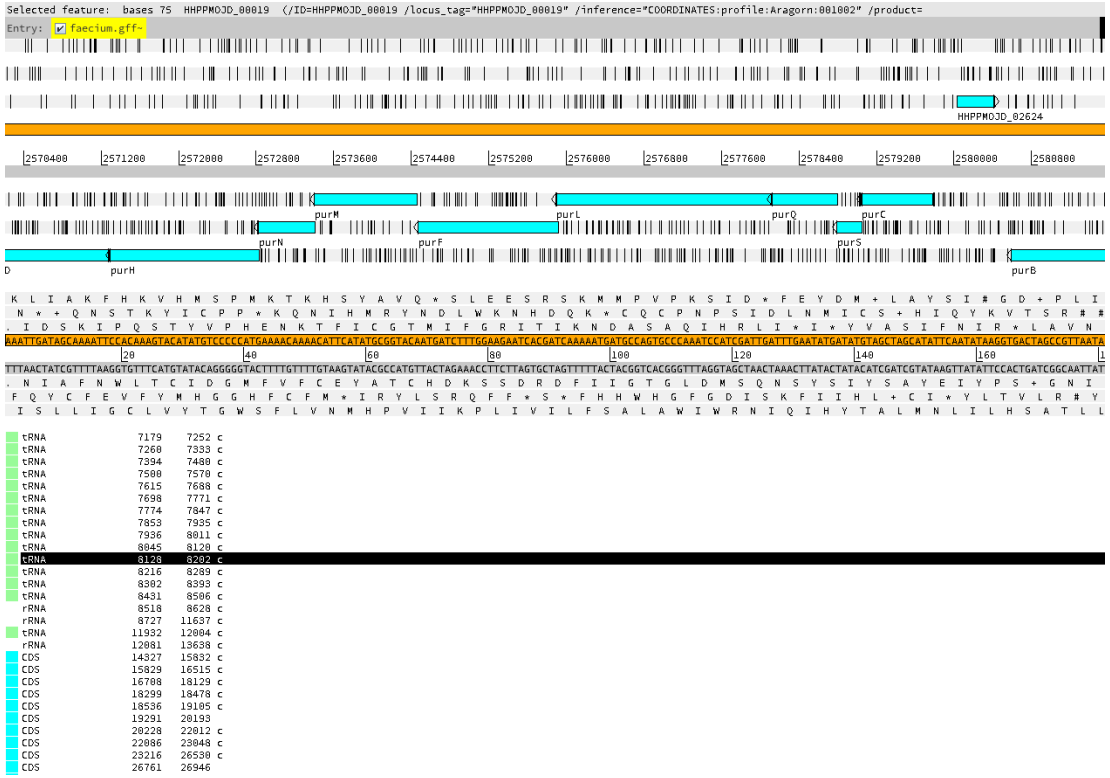
Plasmid name	Size (bp)	Accession number
Chromosome	2775856	
rep17	16563	AF507977
rep1	40010	AF007787
repUS7	14316	AB206333
repUS15	47809	CP004064
rep29	15457	Aus0085

MUmmerplot



Annotation

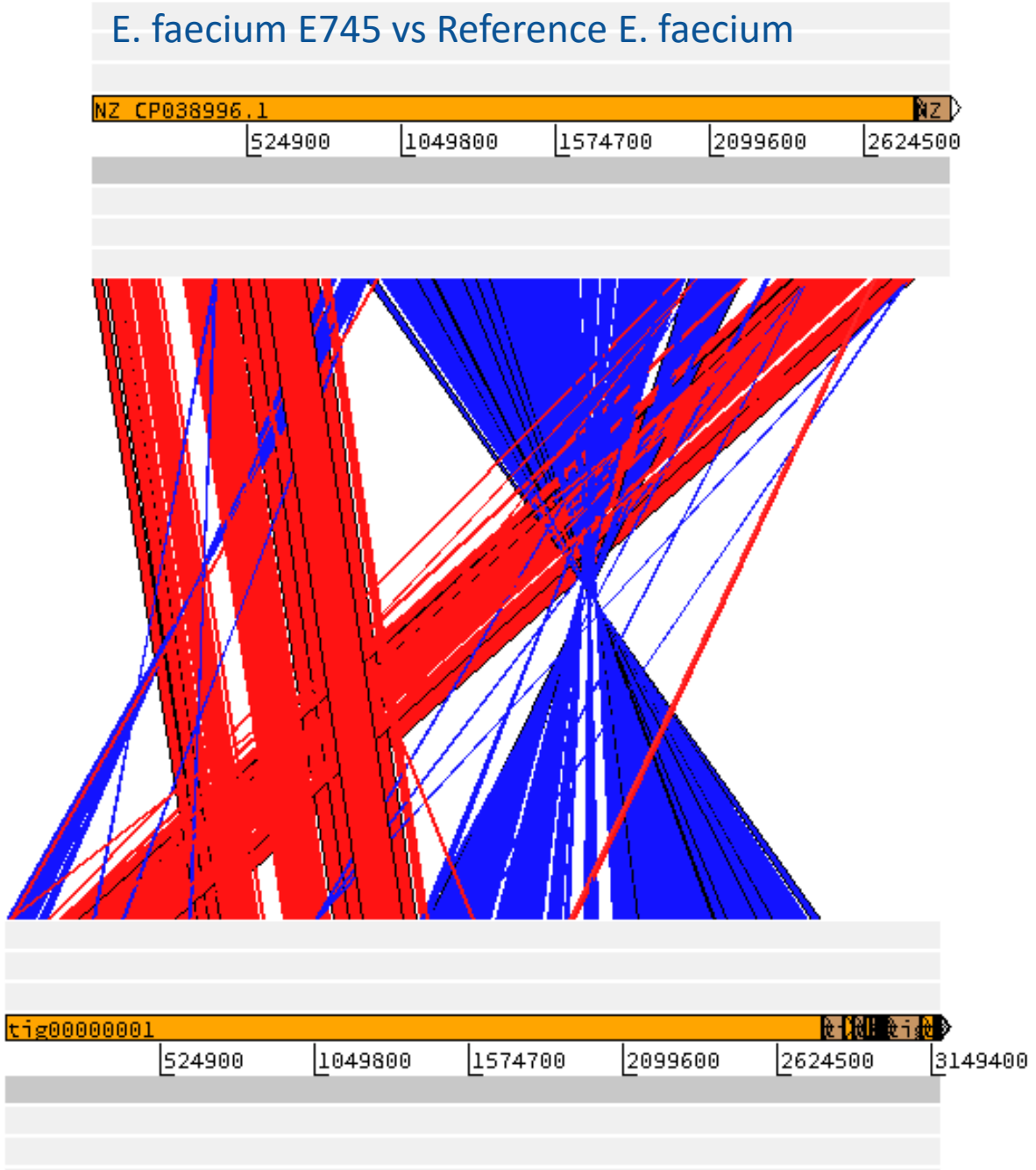
- <Feature tig00000001
- 885 244 CDS
- db_xref COG:COG1455
- gene licC_1
- inference ab initio prediction:Prodigal:002006
- inference similar to AA sequence:UniProtKB:P46317
- locus_tag HHPPMOJD_00001
- product Lichenan permease IIC component
- 1410 proteins as 'hypothetical protein



- organism: Enterococcus species strain
- contigs: 13
- bases: 3179422
- CDS: 3213
- rRNA: 20
- tRNA: 85
- tmRNA: 1

Homology Search

- Synteny comparison with a closely related genome



Transcriptomics

Mapping, counting reads, compare E. faecium in Rich Medium vs Human Serum

BWA (index, mem)
Samtools (sort, index)
Mapping RNA reads
into assembled genome

HTSeq
Count number of transcripts
mapped to each gene

DeSeq2 RStudio
Differential expression Plots

DeSeq2

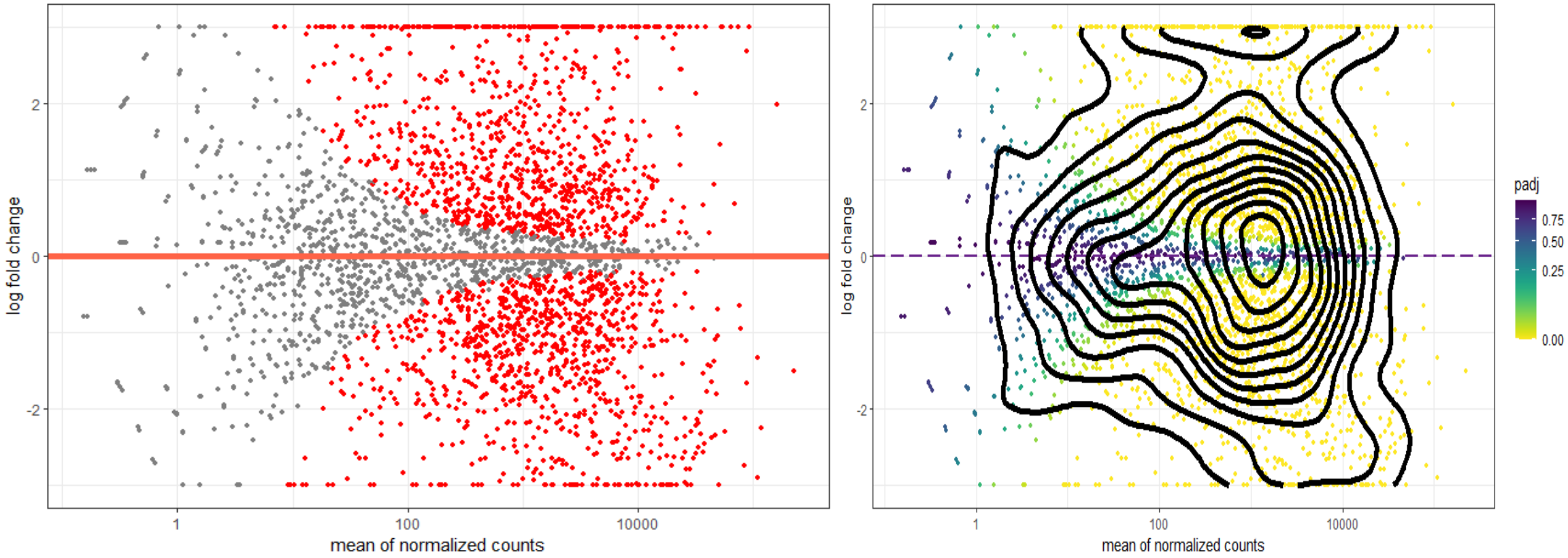
```
> summary(res001)

out of 3053 with nonzero total read count
adjusted p-value < 0.001
LFC > 0 (up)       : 1010, 33%
LFC < 0 (down)     : 973, 32%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

```
> sum(res001$padj < 0.001, na.rm=TRUE)
[1] 1983
> sum(res001$padj < 0.001 & abs(res001$log2FoldChange)>2, na.rm=TRUE)
[1] 531
> sum(res001$padj < 0.001 & abs(res001$log2FoldChange)>1, na.rm=TRUE)
[1] 1205
```

531 genes was significantly expressed between cultures at (q < 0.001 and a fold change in expression of >2)
In the published article **860 genes** exhibited significantly (q < 0.001 and a fold change in expression of >2 between cultures grown in BHI versus heat-inactivated serum) different.

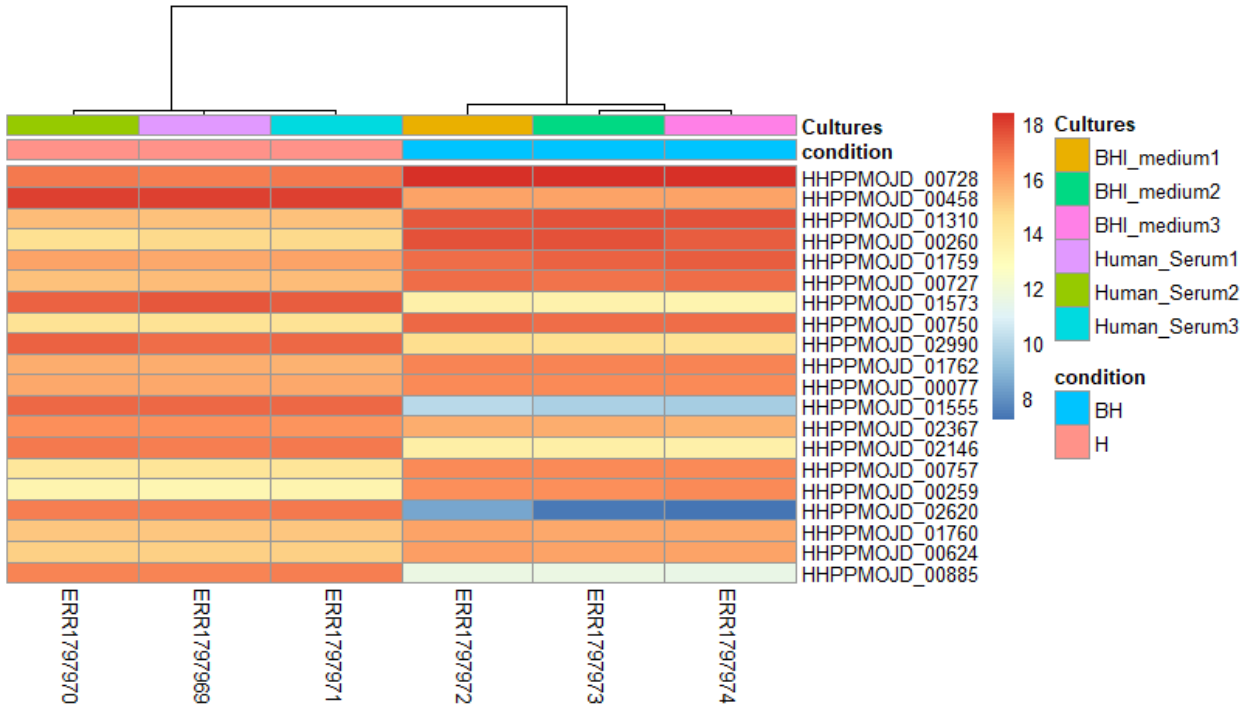
MA plot



An MA plot with a high number of data points falling above the one threshold on the y-axis would indicate a more significant number of genes being upregulated, while more below −1 would indicate high levels of downregulation in genes.

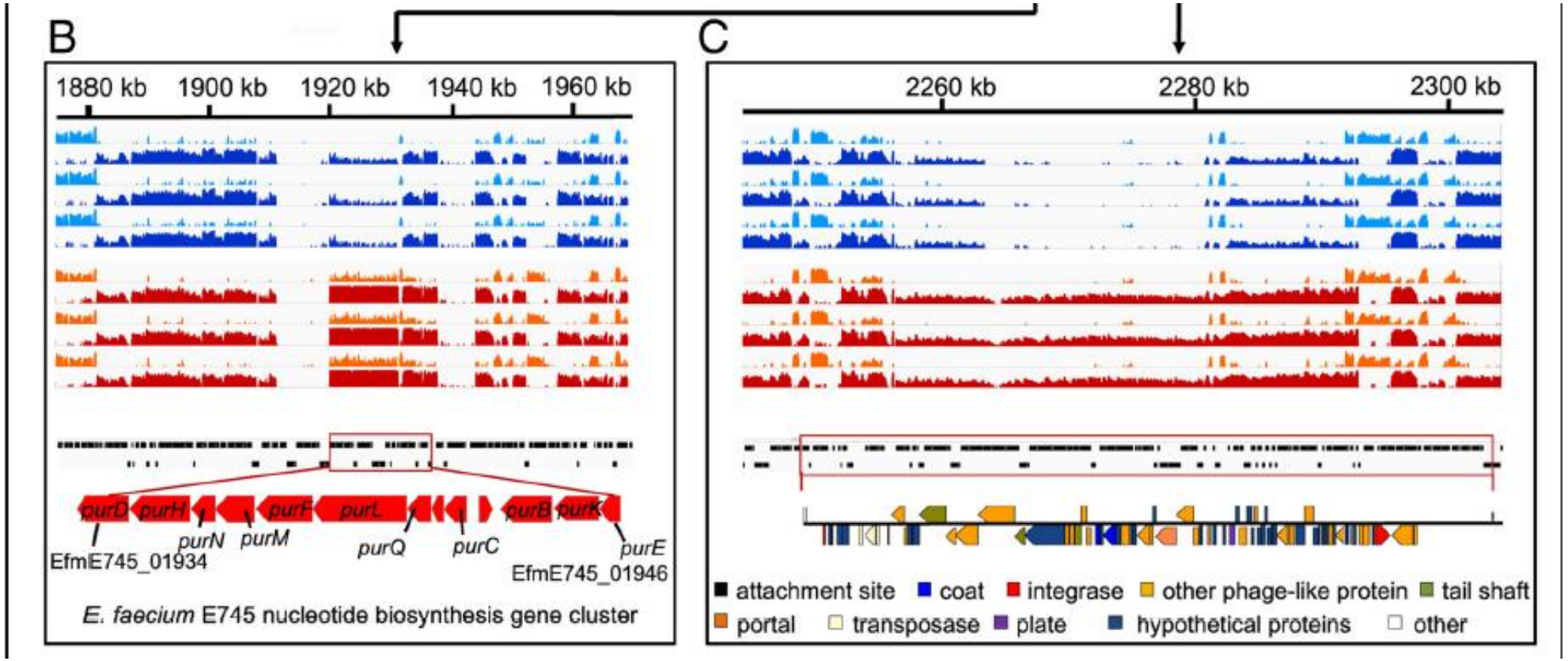
Heatmap

Top 20 genes with the largest difference in expression between the two environments



locus_tag	gene	product	GO terms
HHPPMOJD_00728	tuf_1	Elongation factor Tu	Biological function (polypeptide elongation process of protein synthesis)
HHPPMOJD_00458		hypothetical protein	
HHPPMOJD_01310	fba	Fructose-bisphosphate aldolase	Biological process (Glycolysis)
HHPPMOJD_00260	rplJ	50S ribosomal protein L10	Molecular function (Repressor, Ribosomal protein)
HHPPMOJD_01759	gap_1	Glyceraldehyde-3-phosphate dehydrogenase	Biological process (Glycolysis)
HHPPMOJD_00727	fus	Elongation factor G	Molecular function (translation elongation factor activity)
HHPPMOJD_01573	folT_1	Folate transporter FolT	Biological process (Transport of protein and molecule)
HHPPMOJD_00750	secY	Protein translocase subunit SecY	Biological process (Protein transport)
HHPPMOJD_02990		hypothetical protein	
HHPPMOJD_01762	eno	Enolase	Biological process (Glycolysis), Transcription, Transcription regulation)
HHPPMOJD_00077	ptsI	Phosphoenolpyruvate-protein phosphotransferase	Biological process (Phosphotransferase system)
HHPPMOJD_01555	dppE	Dipeptide-binding protein DppE	Dipeptide transport system
HHPPMOJD_02367	pbp	Beta-lactam-inducible penicillin-binding protein	Biological process
HHPPMOJD_02146	argS	Arginine--tRNA ligase	Biological process (Protein biosynthesis)
HHPPMOJD_00757	rplQ	50S ribosomal protein L17	Molecular function
HHPPMOJD_00259	rplL	50S ribosomal protein L7/L12	Molecular function
HHPPMOJD_02620	purL	Phosphoribosylformylglycinamidine synthase subunit PurL	Biological process (Purine biosynthesis)
HHPPMOJD_01760	pgk	Phosphoglycerate kinase	Biological process (Glycolysis)
HHPPMOJD_00624	rpoB	DNA-directed RNA polymerase subunit beta	Molecular function (Nucleotidyltransferase, Transfera
HHPPMOJD_00885	adhE	Aldehyde-alcohol dehydrogenase	Molecular function (Multifunctiona enzyme)

In Article



Biosynthetic gene clusters (BGCs) are responsible for the production of various secondary metabolites that contribute to interference competition between different microbes, and usually provide resistance against the self-produced antibiotic to protect the host cell

Conclusion

- Many genes involving in glycolysis (glycolytic genes) **gapA, pgk, fba and eno** were highly expressed in Rich Medium BHI
- In bacteria, glycolysis is an pathways by which bacteria can catabolically attack glucose.
- It involves breaking down glucose to lactate or pyruvate, resulting in energy stored in the form of adenosine triphosphate (ATP)
- It produce energy during stress environment such as nutrient limitation.

Conclusion

- The genes that showed effect on growth of E. faecium in serum included genes that are involved in carbohydrate phosphotransferase system (PTS) (ptsL), protein biosynthesis (argS) and genes involved in the biosynthesis of purine and pyrimidine nucleotides (purL)

- While in Rich Medium BHI carbohydrate degradation (gap_1, pgk,fba),gene that involves in two metabolic pathways (eno) (conversion of D-2-phosphoglycerate (2PGA) and phosphoenolpyruvate (PEP) in glycolysis and gluconeogenesis) showed effect on growth of E.faecium
- purL gene which participates in Phosphotransferase system and plays a role as a major carbohydrate transport system in bacteria was highly expressed in human serum than rich medium BHI