

Outline

Riboprofiling samples

Trimming reads

Removing contaminants

Genome mapping STAR

Comparing stats with references

Quantification

clustering by expression

- spearman

- pearson

By datatype

- spearman

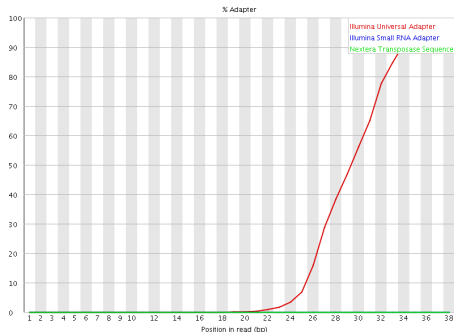
- pearson

Riboprofiling samples

	6h	18h	120h
Reads	42,130,444	45,982,977	44,077,266
Sequence length	50	50	50
%GC	59	55	56

Trimming

✖ Adapter Content



- ▶ min length: 25
- ▶ min adapter alignment length: 5
- ▶ unclipped discarded
- ▶ first base discarded

Trimming - Cutadapt

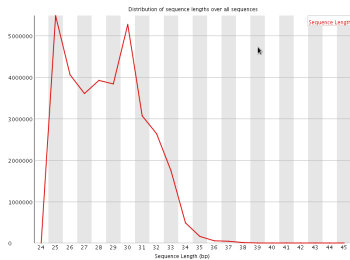
`-quality-cutoff=10` Trim low-quality bases from 3' ends of each read before adapter removal.

id	input	output	discarded TooShort	reads with Adapter
H006	42,130,444	34,439,444 (81.7%)	6,933,352 (16.5%)	41,271,051 (98.0%)
H018	45,982,977	39,441,134 (85.8%)	5,682,240 (12.4%)	44,907,869 (97.7%)
H120	44,077,266	32,501,482 (73.7%)	10,638,847 (24.1%)	42,985,050 (97.5%)

Sequence length distribution after trimming - Cutadapt

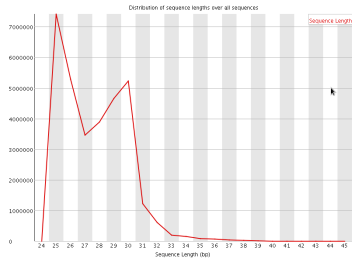
① Sequence Length Distribution

H006



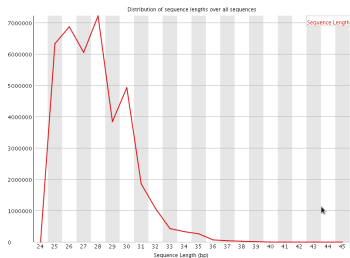
① Sequence Length Distribution

H120



① Sequence Length Distribution

H018



time	sequence length
H006	25-44
H018	25-44
H120	25-44

Removing contaminants - rRNA

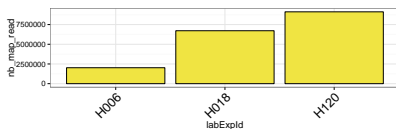
- ▶ STAR: without split mapping; max 10 multimaps; outFilterMatchNmin 16

id	reads processed	uniquely mapped	multiple loci	too many loci	discarded too short
H006	34,439,444	29,465,892 (85.56%)	45,281 (0.13%)	810 (0.00%)	14.31%
H018	39,441,134	12,559,738 (31.84%)	193,421 (0.49%)	4,945 (0.01%)	67.65%
H120	32,501,482	13,389,260 (41.20%)	119,345 (0.37%)	12,214 (0.04%)	58.39%

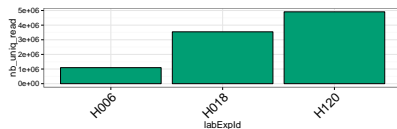
Genome mapping STAR - max 10 multimaps

- ▶ Unaligned reads from rRNA mapping
- ▶ `-outFilterMatchNmin 16`
- ▶ **max 10 multimaps**

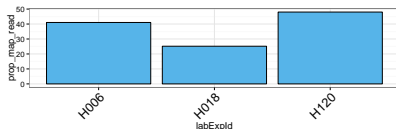
Number of mapped reads



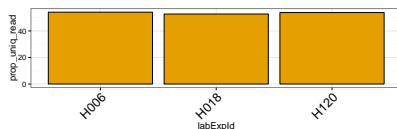
Number of uniquely mapped reads



Proportion of mapped reads



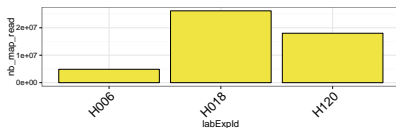
Proportion of uniquely mapped reads



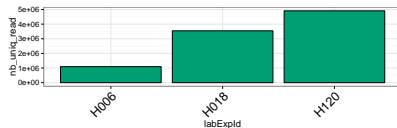
Genome mapping STAR - max 100 multimaps

- ▶ Unaligned reads from rRNA mapping
- ▶ `-outFilterMatchNmin 16`
- ▶ **max 100 multimaps**

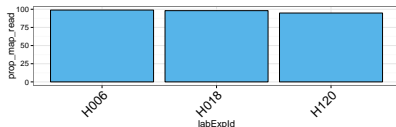
Number of mapped reads



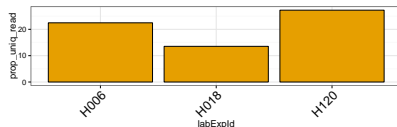
Number of uniquely mapped reads



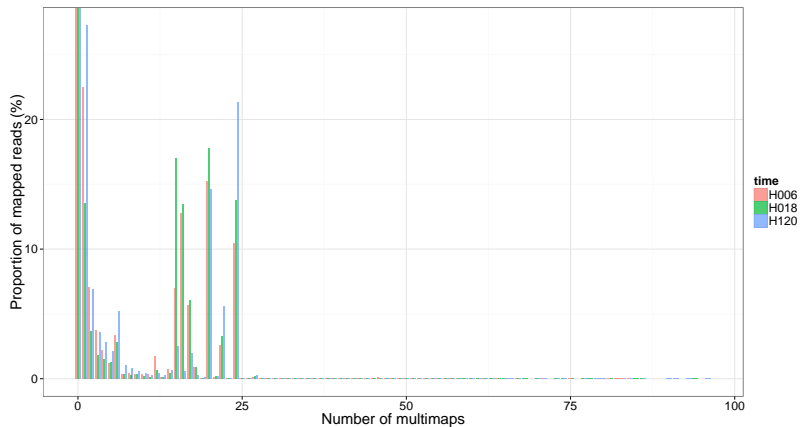
Proportion of mapped reads



Proportion of uniquely mapped reads

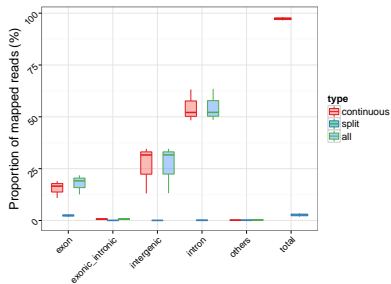


Distribution of multimaps

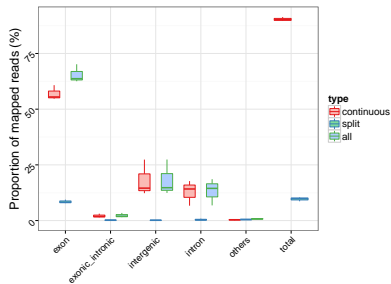


Genomic regions - max 100 multimaps

primary alignment

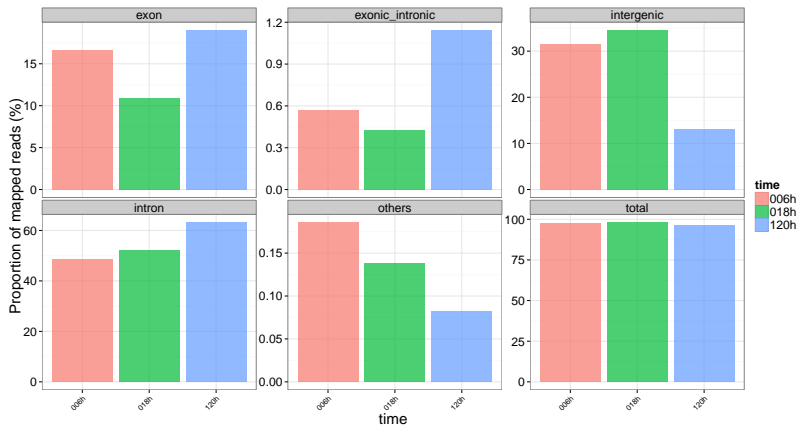


uniquely mapped reads



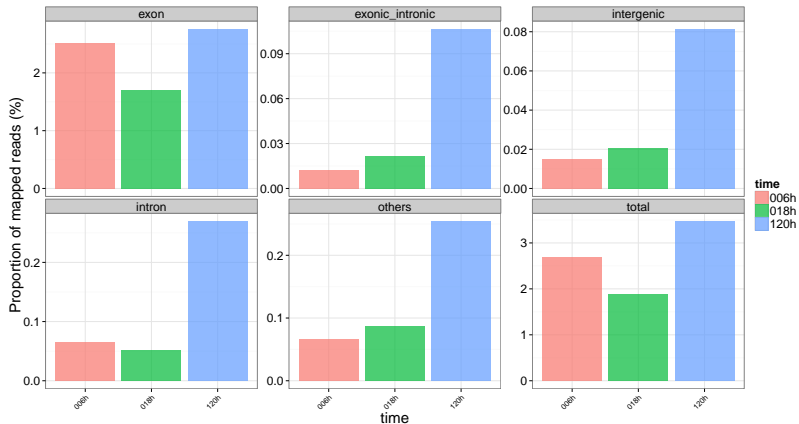
Genomic regions - continuous mapping - max 100 multimaps

primary alignments



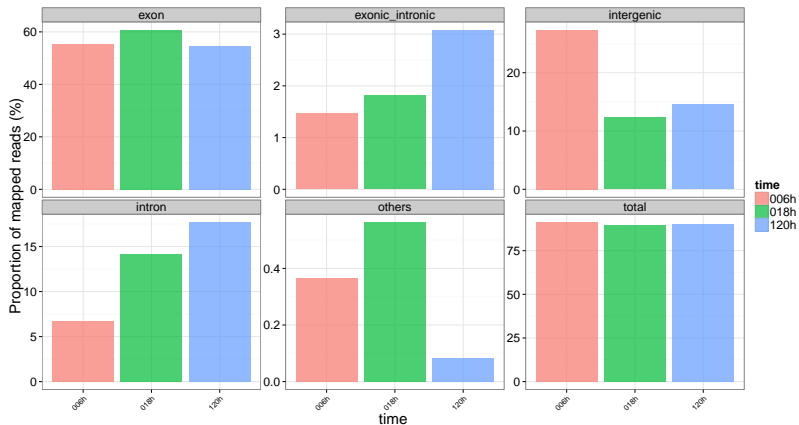
Genomic regions - split mapping - max 100 multimaps

primary alignments



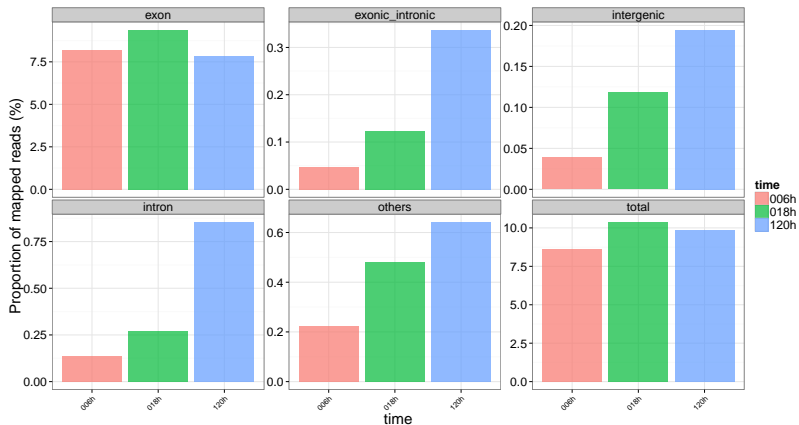
Genomic regions - continuous mapping

uniquely mapped reads



Genomic regions - split mapping - max 100 multimaps

uniquely mapped reads



Comparing stats with references

Fatima's lab

	RPF1 Mock	RPF1 KD	RPF2 Mock	RPF2 KD	RPF3 Mock	RPF3 KD
Total Reads	110,347,659	86,886,294	70,701,122	61,970,268	187,463,074	147,577,976
size-selected (22-36)	105,447,994	80,382,820	65,538,346	60,464,127	159,938,111	134,116,969
After rRNA,tRNA filtering	57,452,964	40,704,687	8,788,516	8,790,692	45,542,640	29,858,389
Aligned (-rRNA,tRNA)	22,252,759	15,717,018	4,376,899	4,232,701	23,323,320	12,759,154
In annotated CDSs	16,520,263	11,497,323	2,451,068	2,742,875	17,831,699	10,158,765

Current stats ERC

	6h	18h	120h
Reads	42,130,444	45,982,977	44,077,266
Size selected (25-44)	34,439,444	39,441,134	32,501,482
After rRNA filtering	4,928,271	26,687,975	18,992,877
Aligned (-rRNA, 10mm)	2,022,381 (41.04%)	6,718,578 (25.17%)	9,117,294 (48.00%)
Uniquely aligned (-rRNA, 10mm)	1,095,120 (54.15%)	3,543,776 (52.75%)	4910138 (53.86%)
Exonic mapping primary alignment	918,722	3,241,787	3,881,137
Exonic uniquely mapped	697,556	2,486,593	3,071,028

Evolution of Gene Regulation during Transcription and Translation

Zhe Wang^{1,†}, Xuepeng Sun^{1,2,†}, Yi Zhao^{1,3}, Xiaoxian Guo¹, Huifeng Jiang^{1,4}, Hongye Li², and Zhenglong Gu^{1,*}

Genome Biol. Evol. 7(4):1155–1167. doi:10.1093/gbe/evv059 Advance Access publication April 14, 2015

1155

Mapping statistics								
	mRNA				RFP			
	Parents rep1	Parents rep2	Hybrid rep1	Hybrid rep2	Parents rep1	Parents rep2	Hybrid rep1	Hybrid rep2
Raw reads	17,624,023	18,867,091	8,989,389	11,190,803	39,013,450	28,194,385	43,422,305	28,293,665
rRNA removed	17,498,738	18,780,244	8,922,195	11,128,519	13,695,629	8,022,019	17,881,744	10,250,293
Unique mapped	8,775,097	14,519,040	6,138,988	7,722,284	5,773,238	3,799,847	8,588,366	5,764,230
Assigned to Scer	3,711,925	6,234,419	2,875,434	3,610,498	3,492,718	2,367,021	4,318,725	2,875,927
Assigned to Sbay	5,063,172	8,284,621	3,263,554	4,111,786	2,280,520	1,432,826	4,269,641	2,888,303
Splicing Alignment(SA)	6,255	10,645	13,082	16,222	6,033	8,432	11,391	13,122
SA in Scer	4,497	7,847	11,815	14,930	4,325	5,447	7,502	8,550
SA in Sbay	1,758	2,798	1,267	1,292	1,708	2,985	3,889	4,572

- ▶ To enable comparable analysis of high-throughput sequencing data sets, we used a uniform alignment and preprocessing pipeline.
- ▶ Reads were sequentially aligned using Bowtie 2 v.2.0.5 (Langmead and Salzberg 2012).
- ▶ All reads mapping to human rRNA and tRNA sequences were filtered out.
- ▶ **The remaining reads were aligned to APPRIS principal transcripts (release 12) (Rodriguez et al. 2013) from the GENCODE mRNA annotation v.15 (Harrow et al. 2012).**
- ▶ **For all transcript level analyses, reads that map only to coding regions were used.**

- ▶ **The remaining reads were aligned using parameters “-L 18 –norc” to APPRIS principal transcripts (release 12) (Rodriguez et al., 2013) from the GENCODE mRNA annotation v.15 (Harrow et al., 2012).**
- ▶ **This step was followed by alignment to all GENCODE transcripts and finally to the human genome (hg19).**
- ▶ **This strategy was preferred to avoid any differences in mappability of the exon-exon junction spanning reads due to read length differences between ribosome profiling and RNA-seq libraries.**
- ▶ We only retained alignments with a mapping quality greater than two for subsequent analyses.
- ▶ Reads mapping to coding regions, 5'UTRs, and 3'UTRs were counted separately using bedtools (Quinlan and Hall, 2010) and custom scripts.
- ▶ For all transcript level analyses, reads that map only to coding regions were used.

Comparing stats with references

Genik et al. 2015, Genome Res.

→ How many initial reads and reads mapping to coding regions you had approximately?

This varies quite a bit based on the efficiency of rRNA depletion.

Even after using the oligo-depletion a large fraction goes to rRNAs.

Just to give you an idea, I picked one of the 100 ribosome profiling libraries we sequenced:

Reads	16,739,559
After trimming	14,866,539 (88%)
Mapped to rRNA	64.49%
Reads after rRNA removal	5,265,866
aligned uniquely to APPRIS transcriptome	3,355,633 (63.72%)
aligned > 1 times	722,762 (13.73%)
remaining	1,187,471
mapped to the genome but not the transcriptome	476,051

Current stats ERC

	6h	18h	120h
Reads	42,130,444	45,982,977	44,077,266
Size selected (25-44)	34,439,444	39,441,134	32,501,482
After rRNA filtering	4,928,271	26,687,975	18,992,877
Aligned (-rRNA, 10mm)	2,022,381 (41.04%)	6,718,578 (25.17%)	9,117,294 (48.00%)
Uniquely aligned (-rRNA, 10mm)	1,095,120 (54.15%)	3,543,776 (52.75%)	4910138 (53.86%)
Exonic mapping primary alignment	918,722	3,241,787	3,881,137
Exonic uniquely mapped	697,556	2,486,593	3,071,028

For our study, the primary objective is to provide an integrated analysis of RNA, protein and translation levels. Given the biology of ribosome protection, it is not possible to obtain long reads for ribosome profiling. **Hence, directly mapping to the entire genome penalizes ribosome profiling reads more than it does RNA-Seq reads which are much longer. Hence, we made the simplifying assumption of concentrating on a defined transcriptome.** I should note that there are very few reads that map to non-APPRIS, GENCODE transcripts.

{APPRIS}

Annotating principal splice isoforms

APPRIS Database

Access annotations for the species annotated in the database via gene name or Ensembl id.

[Access the web database](#)

APPRIS WebServer

Annotate splice isoforms for vertebrate genes that are not in the APPRIS Database.

[Run the web server](#)

APPRIS Database currently houses annotations for [vertebrate genomes](#) »



Human

Assemblies: GRCh38|hg38 (Ensembl82)
Assemblies: GRCh37|hg19 (Ensembl74)



Mouse

Assemblies: GRCm38|mm10 (Ensembl82)



Zebrafish

Assemblies: GRCz10|danRer10 (Ensembl82)
Assemblies: Zv9|danRer7 (Ensembl77)



Rat

Assemblies: Rnor_6.0|rn6 (Ensembl82)
Assemblies: Rnor_5.0|rn5 (Ensembl77)

APPRIS Database currently houses annotations for [invertebrate genomes](#)



Fruitfly

Assemblies: BDGP6|dm6 (Ensembl82)

Assemble

APPRIS

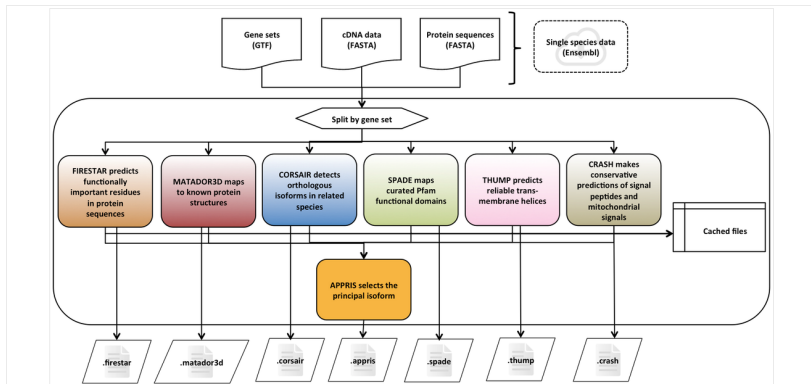
The **APPRIS WebServer** [1] executes a range of computational methods to annotate splice variants for individual genes. As part of its annotation process APPRIS selects a single CDS as the principal isoform for each gene.

Publications

[1] **APPRIS WebServer and WebServices.** [Rodriguez JM, et al.](#) Nucleic Acids Res. 2015 May 18.
PUBMED:25990727 DOI:10.1093/nar/gkv512

Alternative splicing generates different gene products. Recent studies have estimated that almost 100% of multi-exon human genes produce differently spliced mRNAs. It is important to designate one of the isoforms as the principal isoform in order to predict the potential changes in function, structure or localisation brought about by alternative splicing.

APPRIS annotates variants with biological data such as **protein structural information, functionally important residues, conservation of functional domains** and **evidence of cross-species conservation**. The APPRIS Database selects a **principal isoforms** based on this evidence.



APPRIS Database analysis

The APPRIS Database automates a range of computational methods that are used to annotate alternative splice variants and to define principal variants. The splice isoform annotations are the results of the six modules in the APPRIS Database; the final module selects the principal isoforms.

Downloads

The annotations of the following species are available.

Filter by Species: Assembly Version: Gene Dataset:

Species	Assembly version	Gene Dataset	Principal isoforms	APPRIS scores	Functional residues	Tertiary structure	Vertebrates conservation	Whole domains	Transmembrane he
Human	GRCh38	gencode23/ensembl82	TXT	TXT	GTF BED	GTF BED	GTF BED	GTF BED	GTF BED
Human	GRCh37	gencode19/ensembl74	TXT	TXT	GTF BED	GTF BED	GTF BED	GTF BED	GTF BED
Mouse	GRCh38	gencodeM7/ensembl82	TXT	TXT	GTF	GTF	GTF	GTF	GTF
Rat	Rnor_6.0	ensembl81	TXT	TXT	GTF	GTF	GTF	GTF	GTF

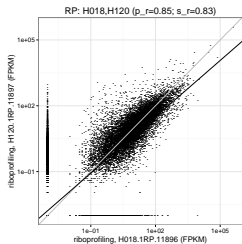
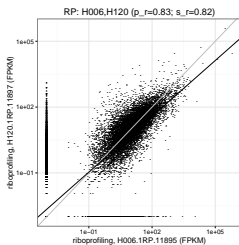
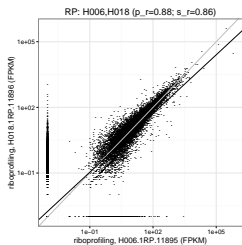
You can download our annotations files from [our server](#), the following [README file](#) explains the directory structure.

Principal Isoform flags

APPRIS selects a single CDS variant for each gene as the 'PRINCIPAL' isoform based on the range of protein features. Principal isoforms are tagged with the numbers 1 to 5, with 1 being the most reliable. The definition of the flags are as follows:

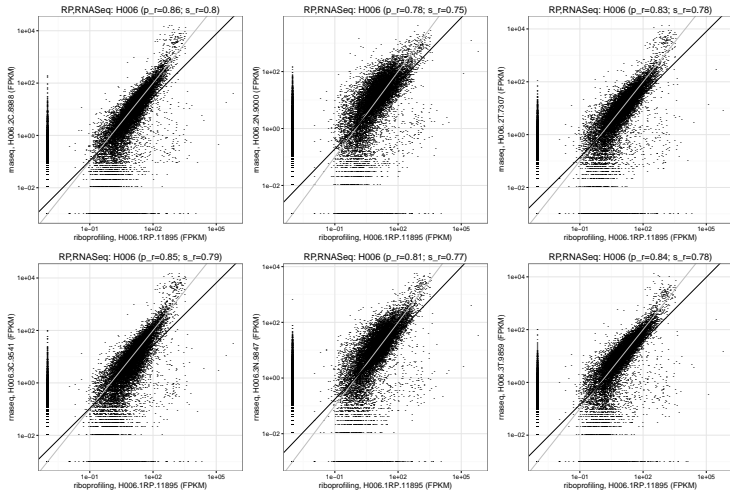
- **PRINCIPAL1**
Transcript(s) expected to code for the main functional isoform based solely on the core modules in the APPRIS database. The APPRIS core modules map protein structural and functional information and cross-species conservation to the annotated variants.
- **PRINCIPAL2**
Where the APPRIS core modules are unable to choose a clear principal variant (approximately 25% of human protein coding genes), the database chooses two or more of the CDS variants as "candidates" to be the principal variant.

Correlation between riboprofiling samples



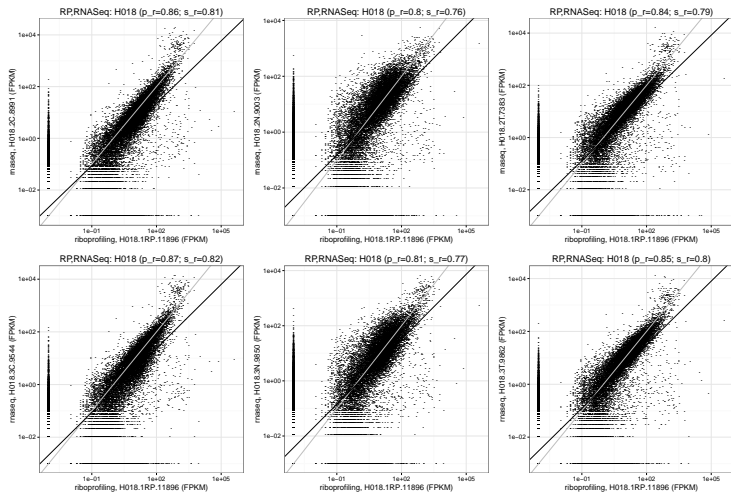
Correlation between riboprofiling and RNAseq samples

H006 - 50666 genes



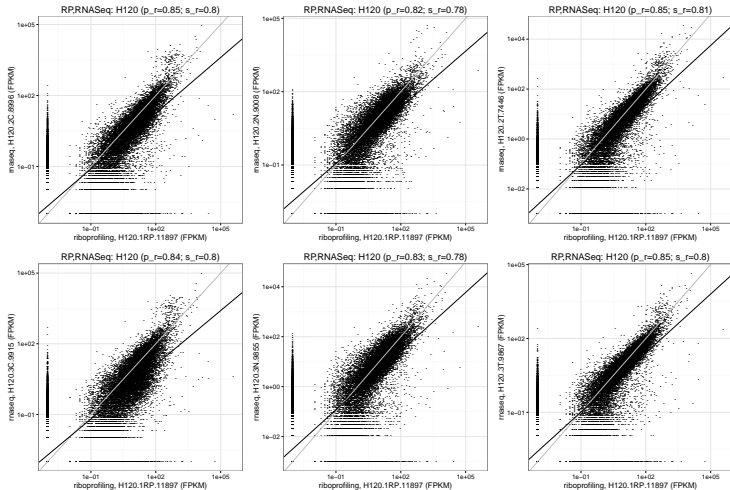
Correlation between riboprofiling and RNAseq samples

H018 - 50666 genes

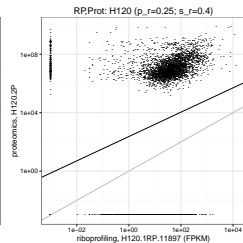
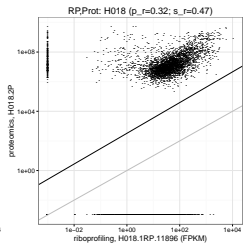
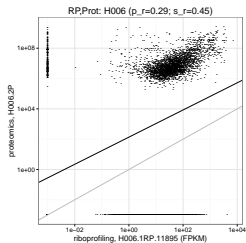


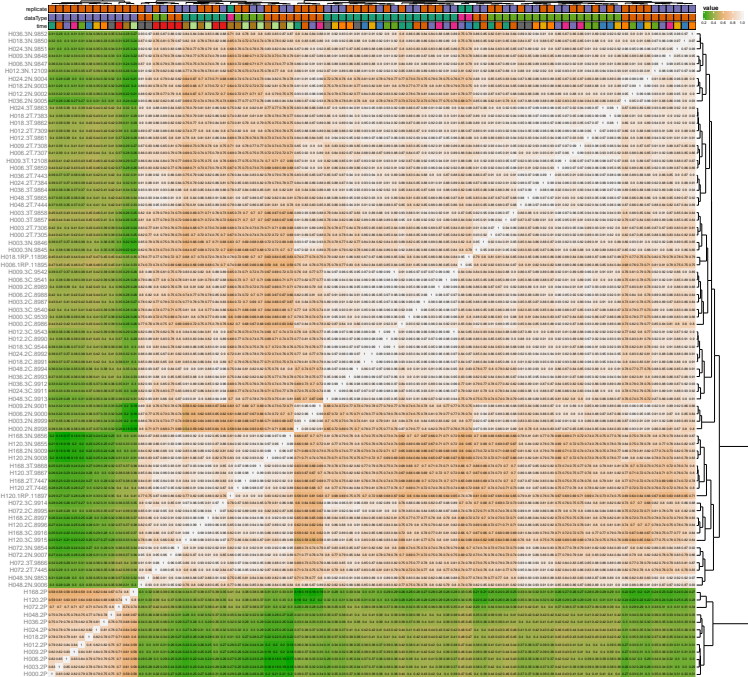
Correlation between riboprofiling and RNAseq samples

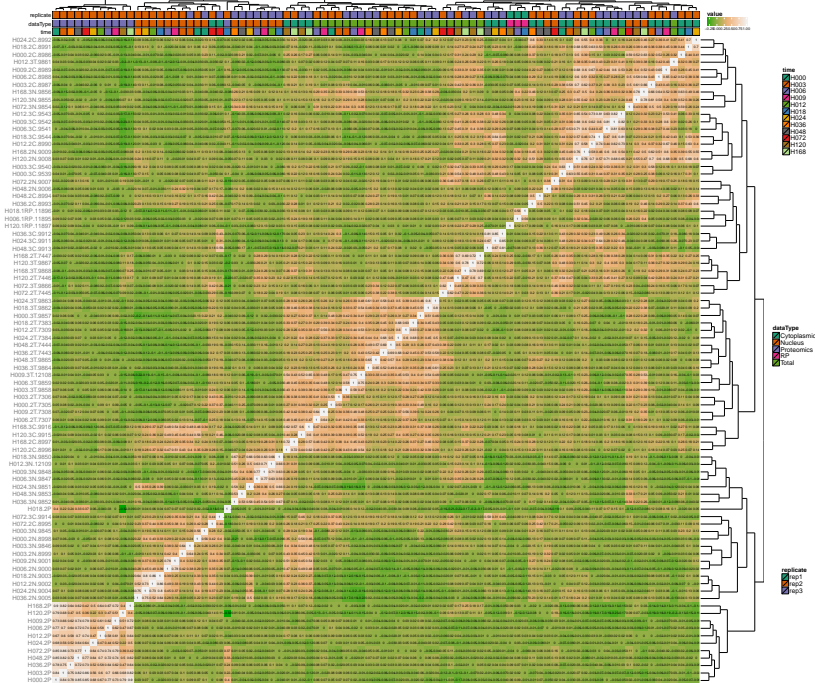
H120 - 50666 genes



Correlation between riboprofiling and proteomics samples - 6109 genes





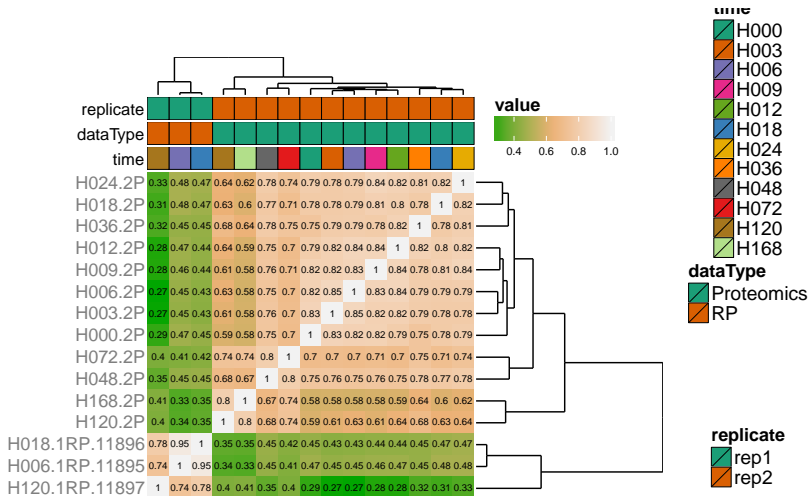


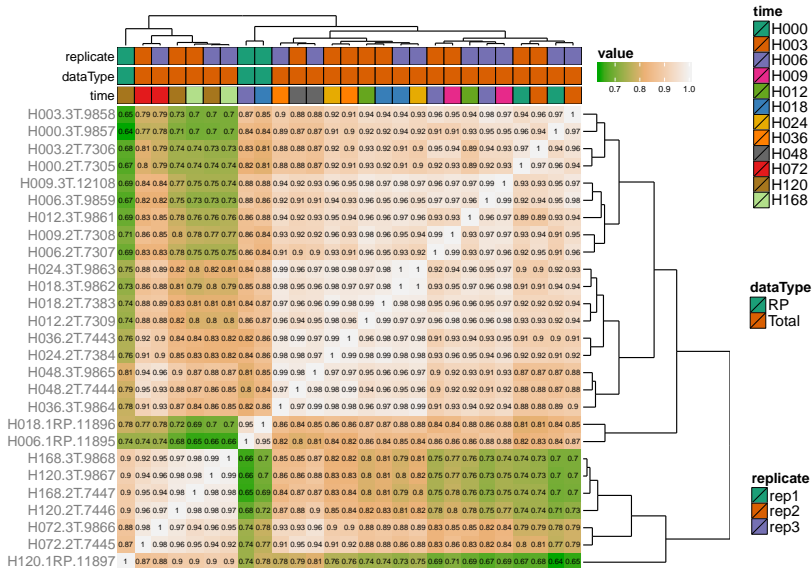


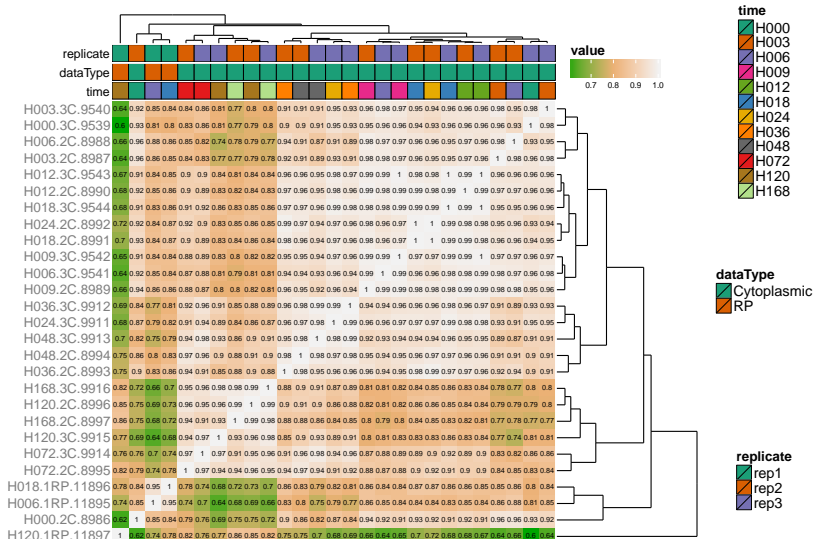
dataType
 Cytoplasmic
 Nucleus
 Mitochondria
 rIP

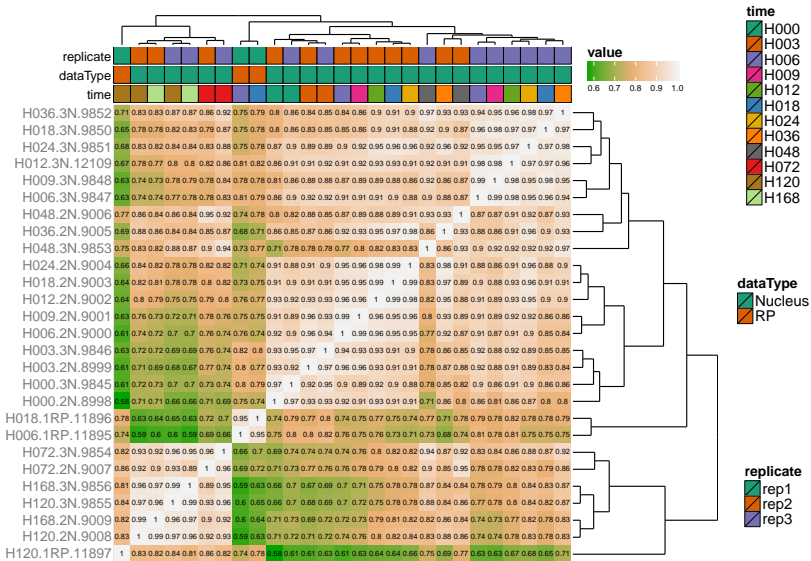
replicate
 rep1
 rep2
 rep3

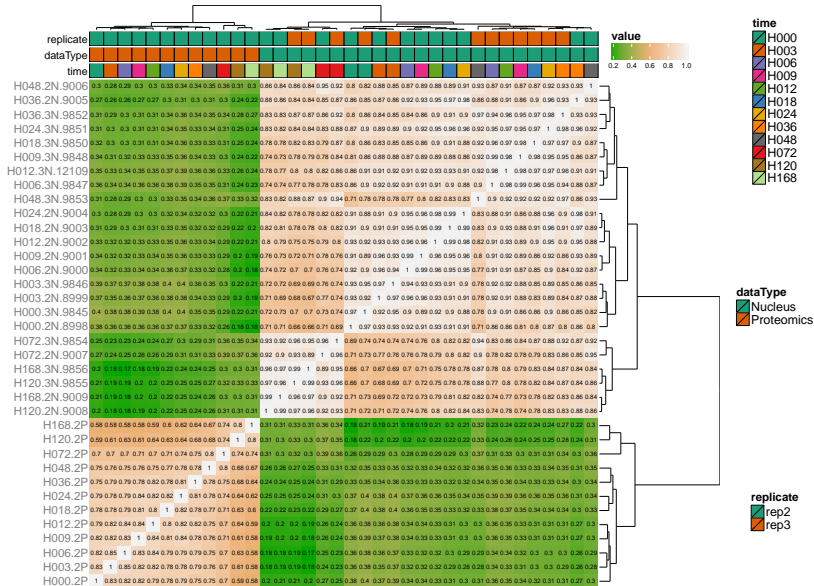
H210.1RP.11897

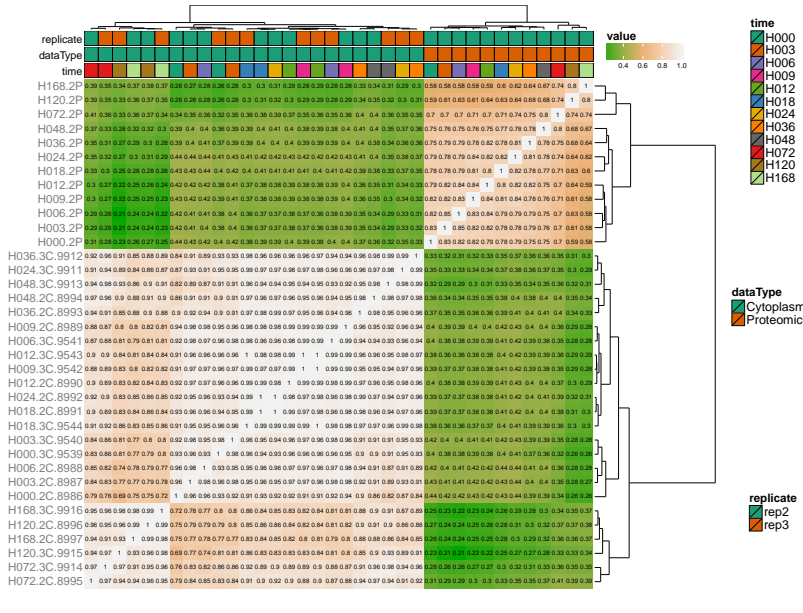


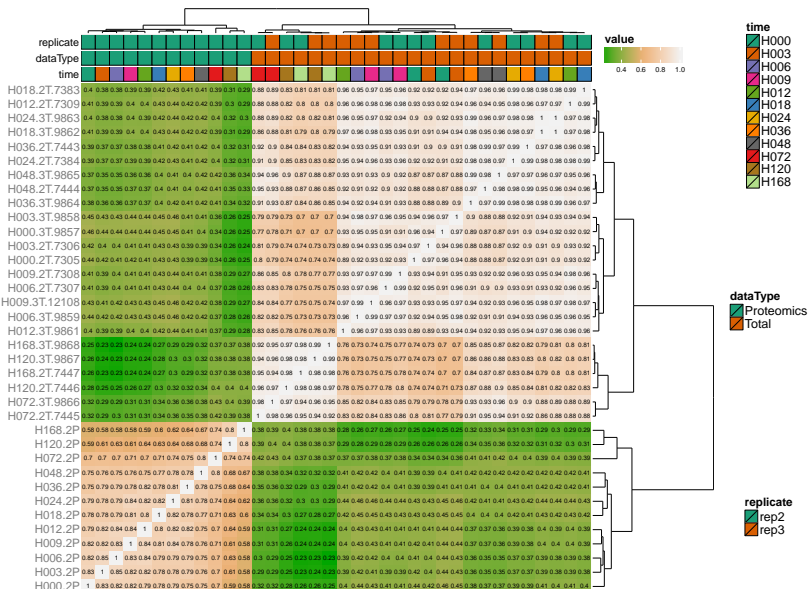


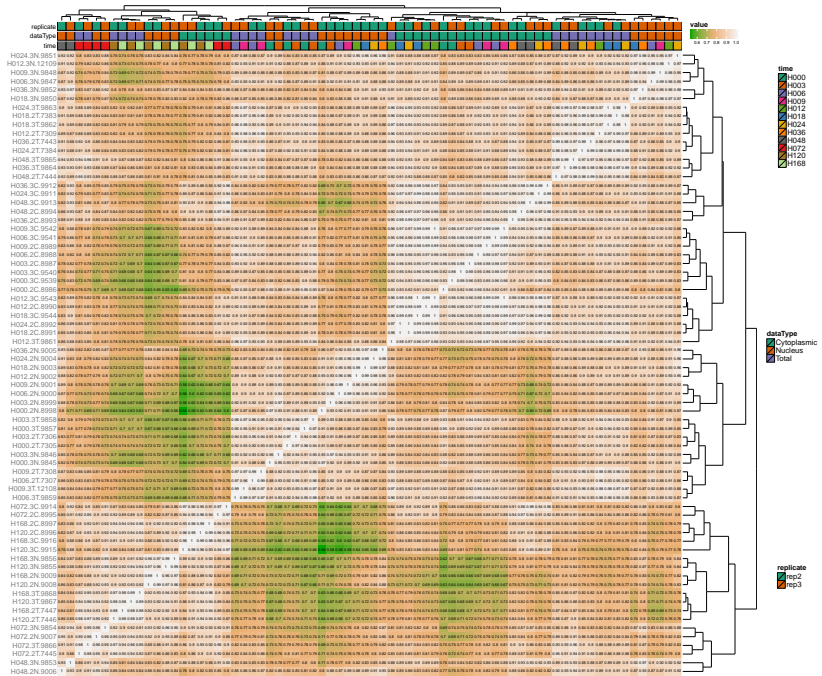






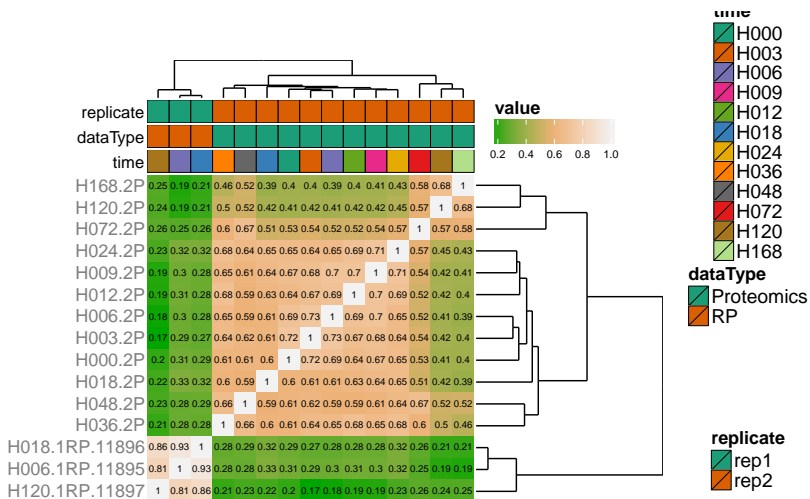


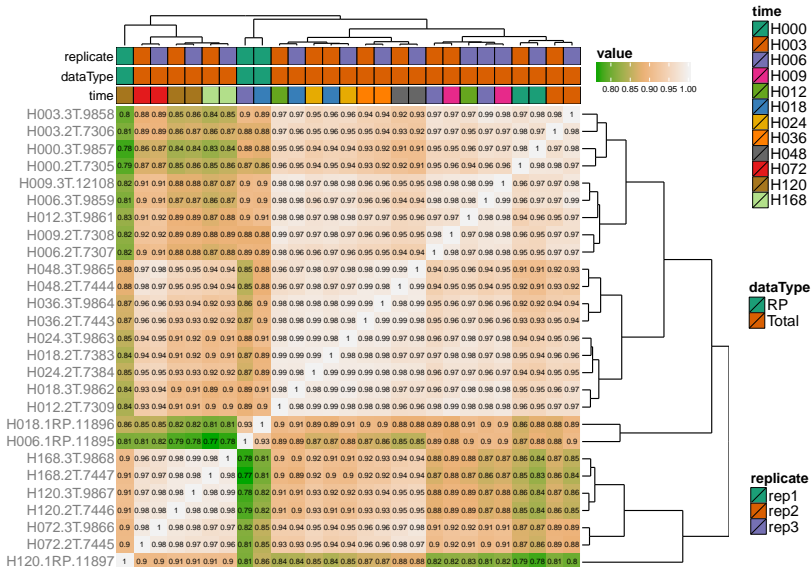


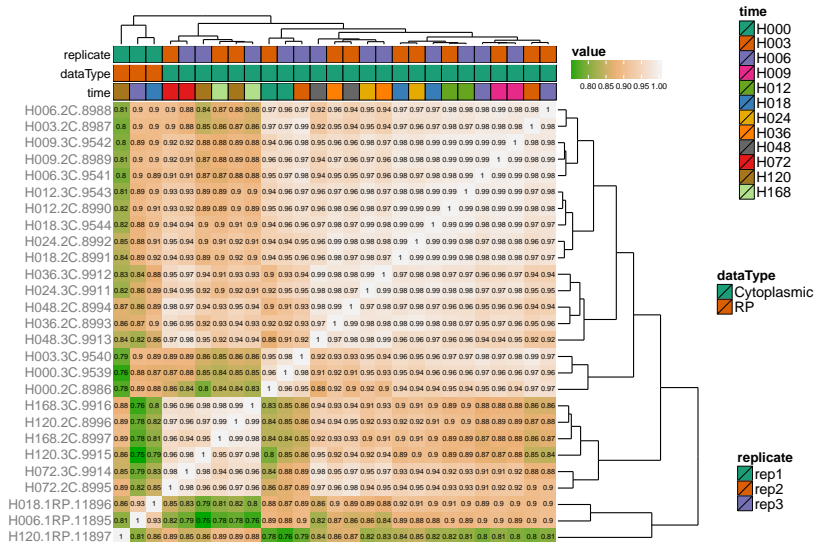


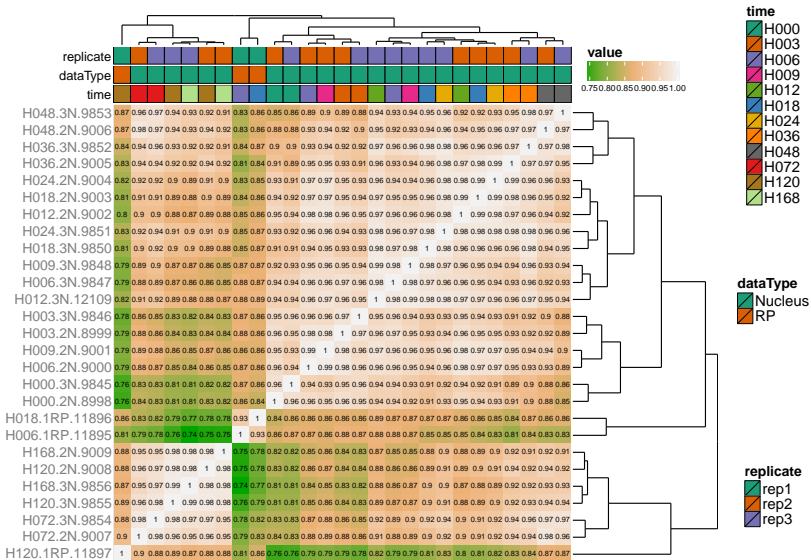


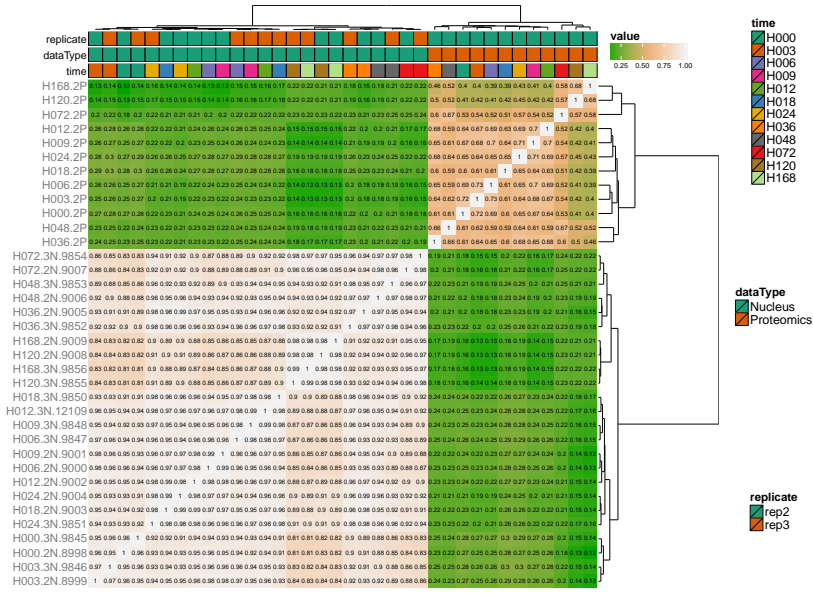
H210.TRP.1.1897

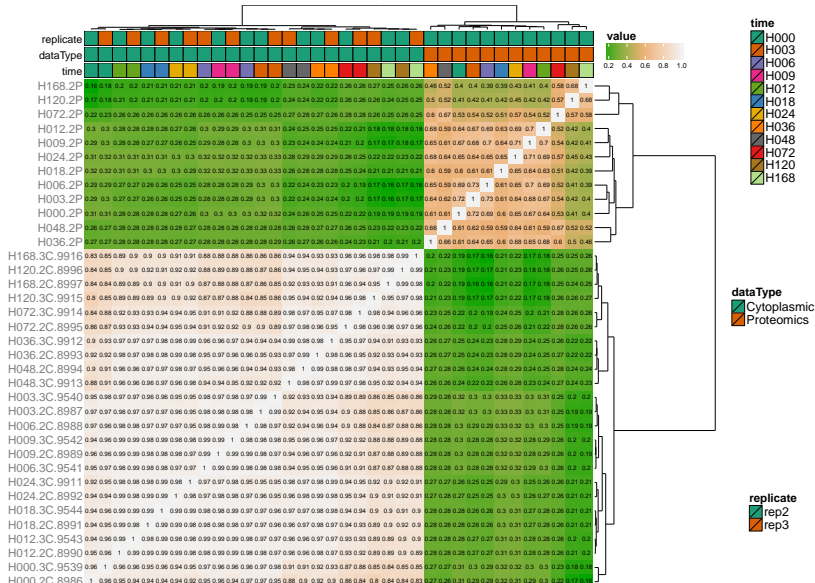


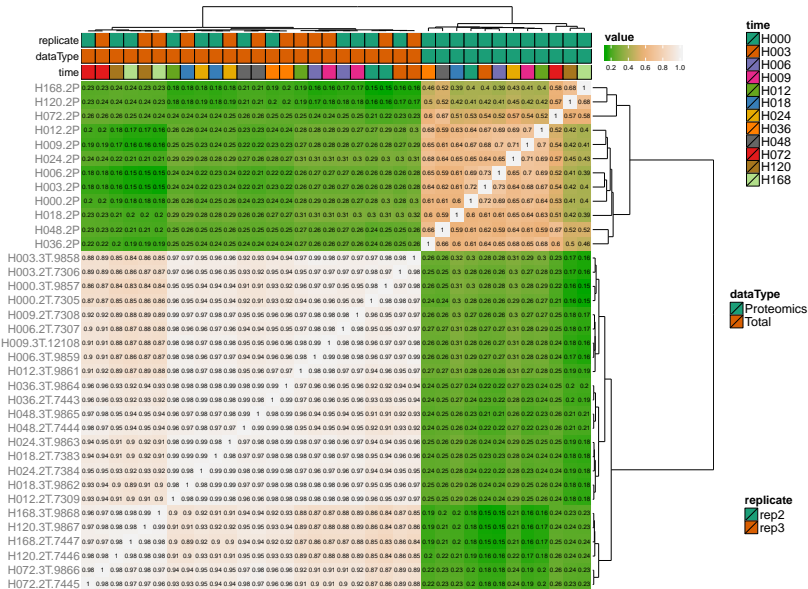


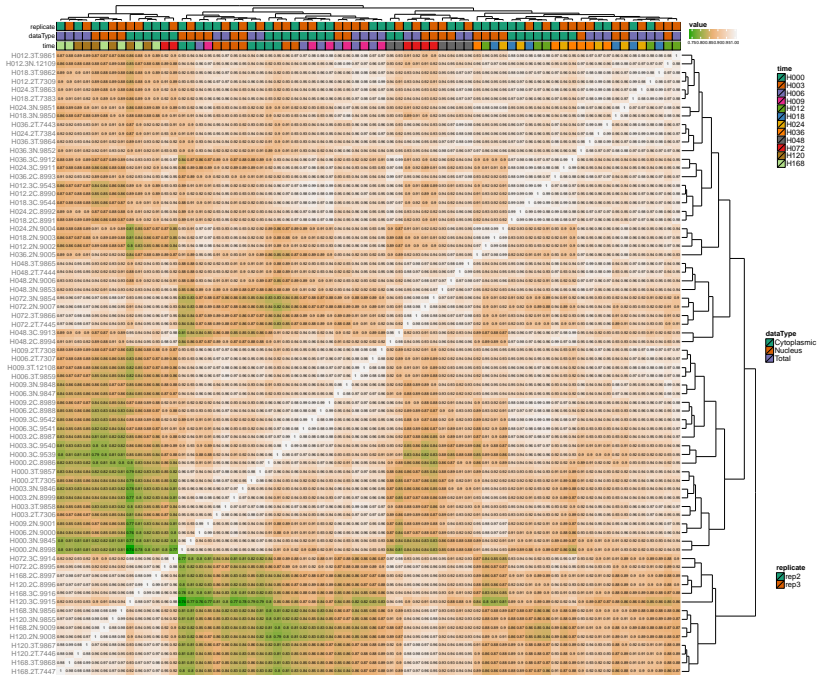








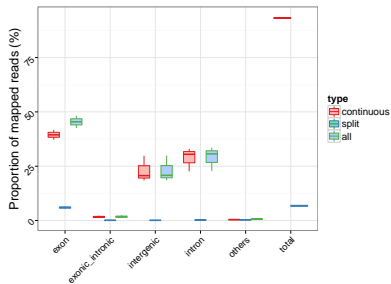




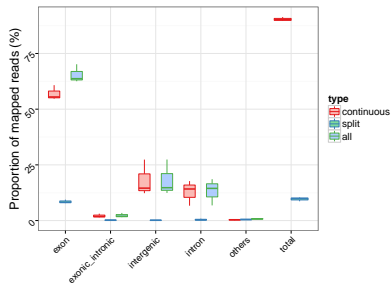
Extras

Genomic regions - max 10 multimaps

primary alignment

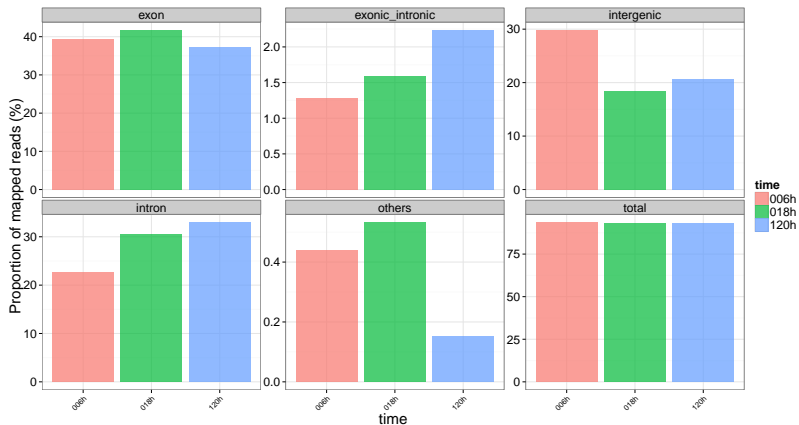


uniquely mapped reads

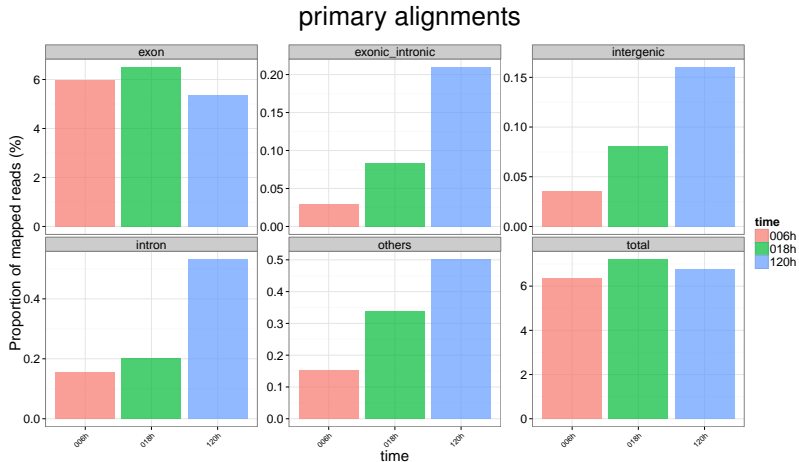


Genomic regions - continuous mapping - max 10 multimaps

primary alignments

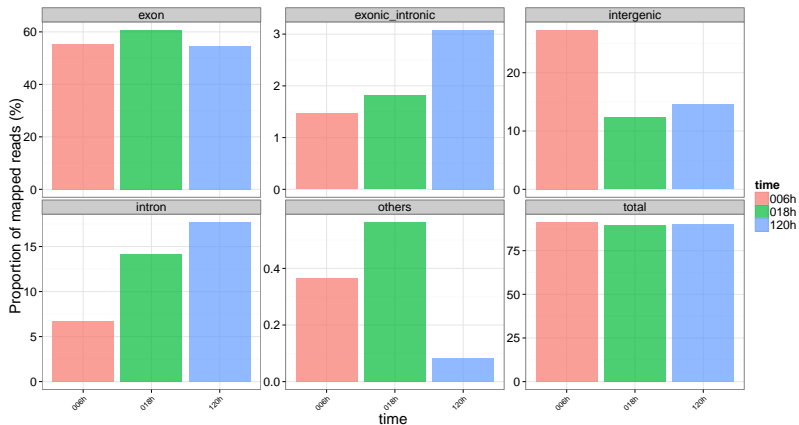


Genomic regions - split mapping - max 10 multimaps



Genomic regions - continuous mapping

uniquely mapped reads



Genomic regions - split mapping - max 10 multimaps

uniquely mapped reads

