Comparison of Stranded and Non-Stranded RNA-Seq in Predicting Small RNAs in a Non-Model Bacterium

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Karel Sedlar, Ralf Zimmer

Institute of BioInformatics

small RNAs

- were shown to play important regulatory roles in diverse cellular processes by participating in post-transcriptional regulation of gene expression
- two types of sRNAs: *cis*-encoded and *trans*-encoded

- *cis*-encoded (perfect base pairing): transcription terminators, potential inhibitors of translation initiation, or modulators of mRNA degradation
- *trans*-encoded (imperfect base pairing): a wider range of regulatory mechanisms - repressors of expression but also activators

sRNAs in bacteria

- former studies suggested conservation of sRNA (*E. coli* vs. *S. enterica*)
  - June 2022: 1 199 199 genome assemblies of 43 669 bacterial species
  - no. of predicted ncRNAs per genome: lower units (PGAP - Rfam cmsearch)

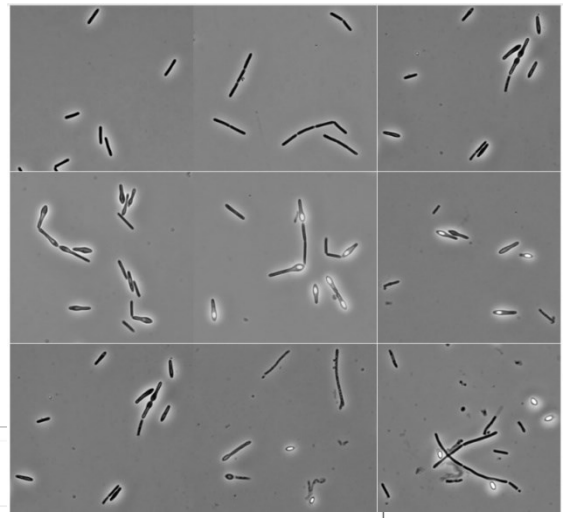
- specialized lab techniques: GRIL-Seq, RIP-Seq, RIL-Seq, …

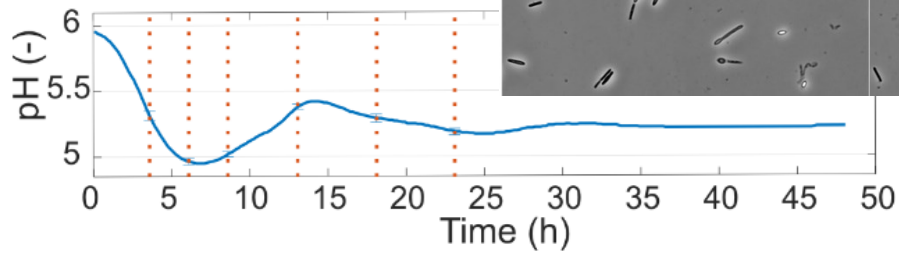
- use of standard RNA-Seq
  - stranded vs. non-stranded
  - in combination with homology based searches
  - direct prediction: APERO, Rockhopper, baerhunter,…

application in biotech

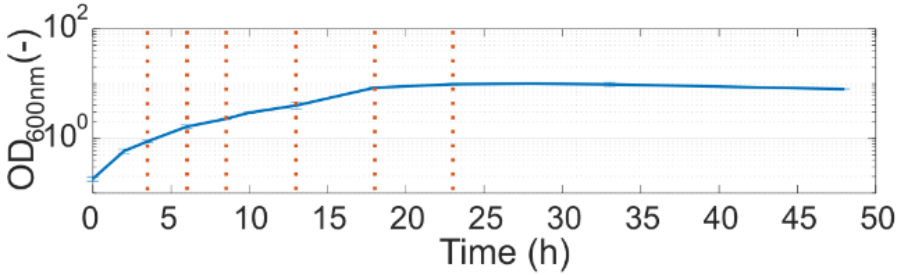
- sRNA was proved can improve bacterial phenotype, for example, tolerance to acids

- *Clostridium beijerinckii* NRRL B-598
  - gram-positive anaerobe, ABE fermentation, hydrogen producer
  - bi-phasic fermentation: acidogenic and solventogenic
  - sRNAs are unknown
  - non-stranded and stranded RNA-Seq available





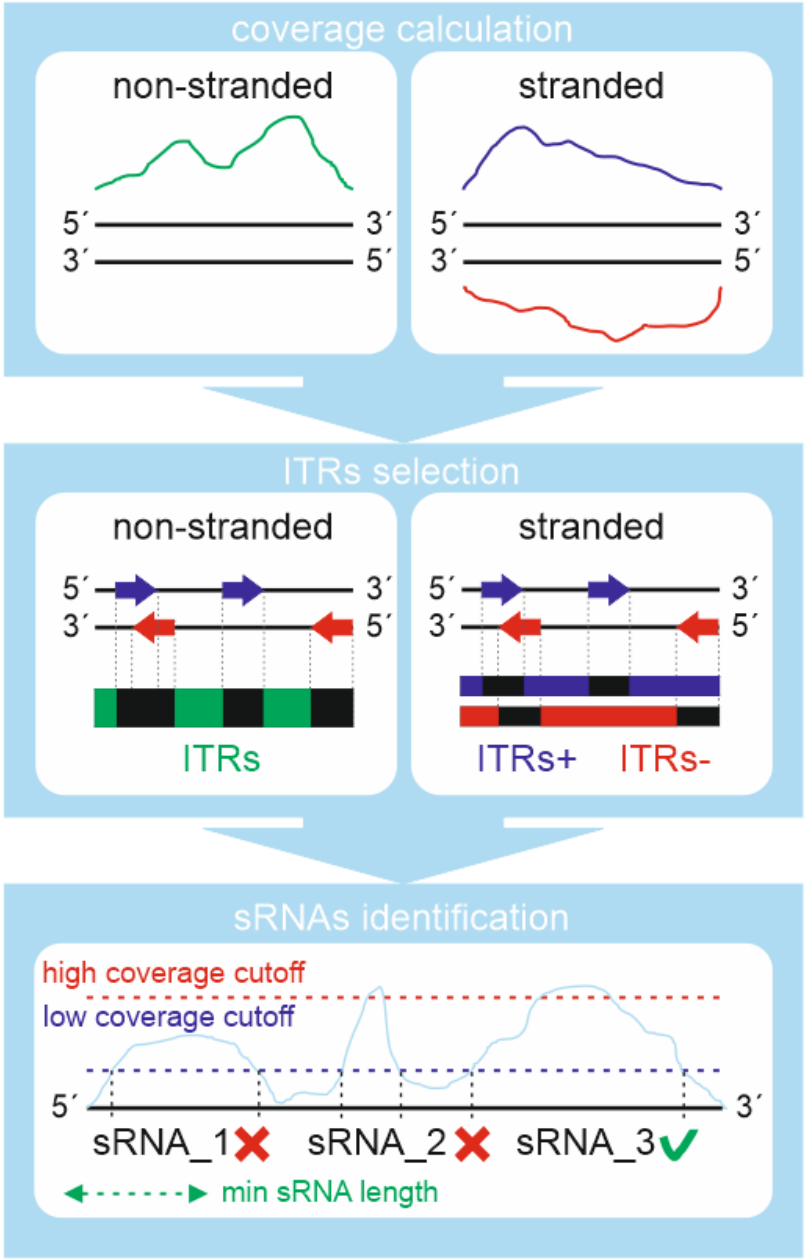
- hydrogen: 95% production from fossil fuels
- grey vs. green vs. biological H_2



materials and methods

- library A:
  - cDNA was synthesized by using a random hexamer-primer (non-stranded data)
  - Illumina HiSeq 4000, single-end, 50 bp
- library B:
  - NEBNext Ultra II stranded kit (reversely stranded)
  - Illumina NextSeq500, single-end, 75 bp

- preprocessing:
  - rRNA not removed vs. rRNA removed
  - settings 1: min PHRED 3, window 4 bp – average quality ≥ 15, minimum length 36 bp
  - settings 2: min PHRED 10, window 4 bp – average quality ≥ 25, minimum length 20 bp



preprocessing

Sample	Trimming settings	rRNA removal	No. of reads in a sample (million)	No. of mapped reads (million)
A1	1	No	21.0	11.9
A2	2	No	20.6	11.7
A1r	1	Yes	12.3	11.8
A2r	2	Yes	12.2	11.6
B1	1	No	52.5	15.3
B2	2	No	48.9	14.3
B1r	1	Yes	15.2	14.6
B2r	2	Yes	15.7	13.7

- considering the number of mapped reads and their length, library A contains only half of the sequenced bases in comparison to B

stranded predictions

- = baerhunter
  - low coverage cutoff: 10
  - high coverage cutoff: 50
  - min sRNA length: 40

Sample	No. of sRNA genes		
	trans-encoded	*cis*-encoded	total number
B1	121	115	236
B2	115	99	214
B1r	121	101	222
B2r	115	87	202

- *trans*-encoded sRNAs detection: rRNA removal has no effect
- *cis*-encoded sRNAs affected by quality trimming as well as computational ribodepletion

non-stranded predictions

- only trans-encoded sRNAs can be predicted
- library B data were handled as non-stranded

Sample	A	B	A∩B
X1	76	109	32
X2	75	108	30
X1r	76	109	32
X2r	75	108	30

- independence of ribodepletion confirmed
- predicted sRNA differed between libraries

evaluation

- baerhunter‘s stranded prediction as a reference

Sample	A		B	
	Precision	Recall	Precision	Recall
X1/X1r	44.7%	28.1%	97.2%	87.6%
X2/X2r	42.7%	27.8%	94.4%	88.7%

- after adjustment to different sequencing depth
  - low coverage cutoff: 10
  - high coverage cutoff: 25
  - min sRNA length: 40

Sample	sRNAs	Precision	Recall
A1/A1r	113	62.8%	93.4%
A2/A2r	114	63.3%	99.1%

conclusions

- direct prediction from standard RNA-Seq data seems to be advantageous
- current tools require the stranded RNA-Seq, but sRNAs can also be identified using non-stranded RNA-Seq with comparable sensitivity
- although the detection is „independent“ of computational ribodepletion, it is highly influenced by sequencing depth that needs to be calculated from mRNA (and sRNA) sequences only
- results depend on a threshold that has to be set up manually in current tools, more benchmarking is needed to ensure reliable and fully automatic prediction of small RNAs in bacterial genomes

acknowledgement

- the research was conducted within the project „The Annotation and Functional Description of Non-Model Bacterial Organisms for Bio-based Engineering and Industry (HOPE-4-BEST)“

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101023766.