Transcript analysis using RNA-seq

What others call transcriptomics?!

Sven Findeiß

Bioinformatics Group, Department of Computer Science;
and Interdisciplinary Center for Bioinformatics,
University of Leipzig

Bled, February 2010

[Sorek and Cossart, 2010]

"Transcriptomics is a powerful tool for understanding gene structures and RNA-based regulation in any organism."

[Sorek and Cossart, 2010]

"Transcriptomics is a powerful tool for understanding gene structures and RNA-based regulation in any organism."

What it might be:

[Sorek and Cossart, 2010]

"Transcriptomics is a powerful tool for understanding gene structures and RNA-based regulation in any organism."

What it might be:

Wang:2009

"The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition."

[Sorek and Cossart, 2010]

"Transcriptomics is a powerful tool for understanding gene structures and RNA-based regulation in any organism."

What it might be:

Wang:2009

"The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition."

Stadler and Prohaska in preparation

"In cellular and molecular biology, the suffix *-ome* refers to 'all constituents considered' collectively."

[Sorek and Cossart, 2010]

"Transcriptomics is a powerful tool for understanding gene structures and RNA-based regulation in any organism."

What it might be:

Wang:2009

"The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition."

Stadler and Prohaska in preparation

"In cellular and molecular biology, the suffix *-ome* refers to 'all constituents considered' collectively."

Lederberg:2001

"$\cdots$-OM signifies fullness, completeness as in divinity $\cdots$, it encompasses the entire universe in its unlimitedness."

Transcriptomics in Prokaryotes

Whole transcriptome studies have been started only recently.

- microbial gene structure was regraded as simple
  $\rightarrow$ no introns $\curvearrowright$ no splicing
  $\rightarrow$ no editing
- technical difficulties e.g. for the mRNA enrichment
  $\rightarrow$ lack poly(A) tails
  $\rightarrow > 95\%$ of cellular RNA is composed of rRNA and tRNA

[Sorek and Cossart, 2010]

Transcriptomics in Prokaryotes

Whole transcriptome studies have been started only recently.

- microbial gene structure was regraded as simple
  $\rightarrow$ no introns $\curvearrowright$ no splicing
  $\rightarrow$ no editing
- technical difficulties e.g. for the mRNA enrichment
  $\rightarrow$ lack poly(A) tails
  $\rightarrow > 95\%$ of cellular RNA is composed of rRNA and tRNA

Gained knowledge:

$\Rightarrow$ 5'UTR annotation

$\Rightarrow$ novel untranslated regulatory elements

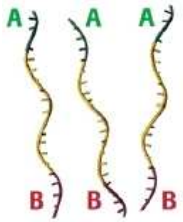
$\Rightarrow$ alternative operon structures

$\Rightarrow$ discovery of novel ncRNAs

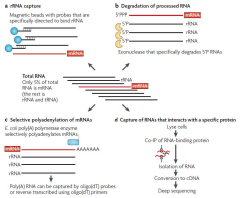
[Sorek and Cossart, 2010]

Transcriptome Sequencing (454)

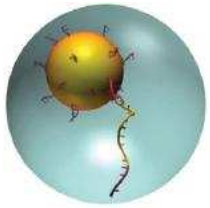
2) Fragment Preparation



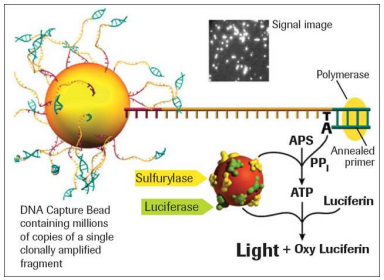
1) RNA Isolation



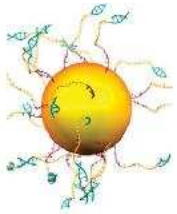
3) One Bead = One Fragment



6) One Bead = One Read



4) Amplification



5) PicoTiterPlate Fixation

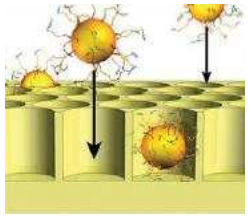
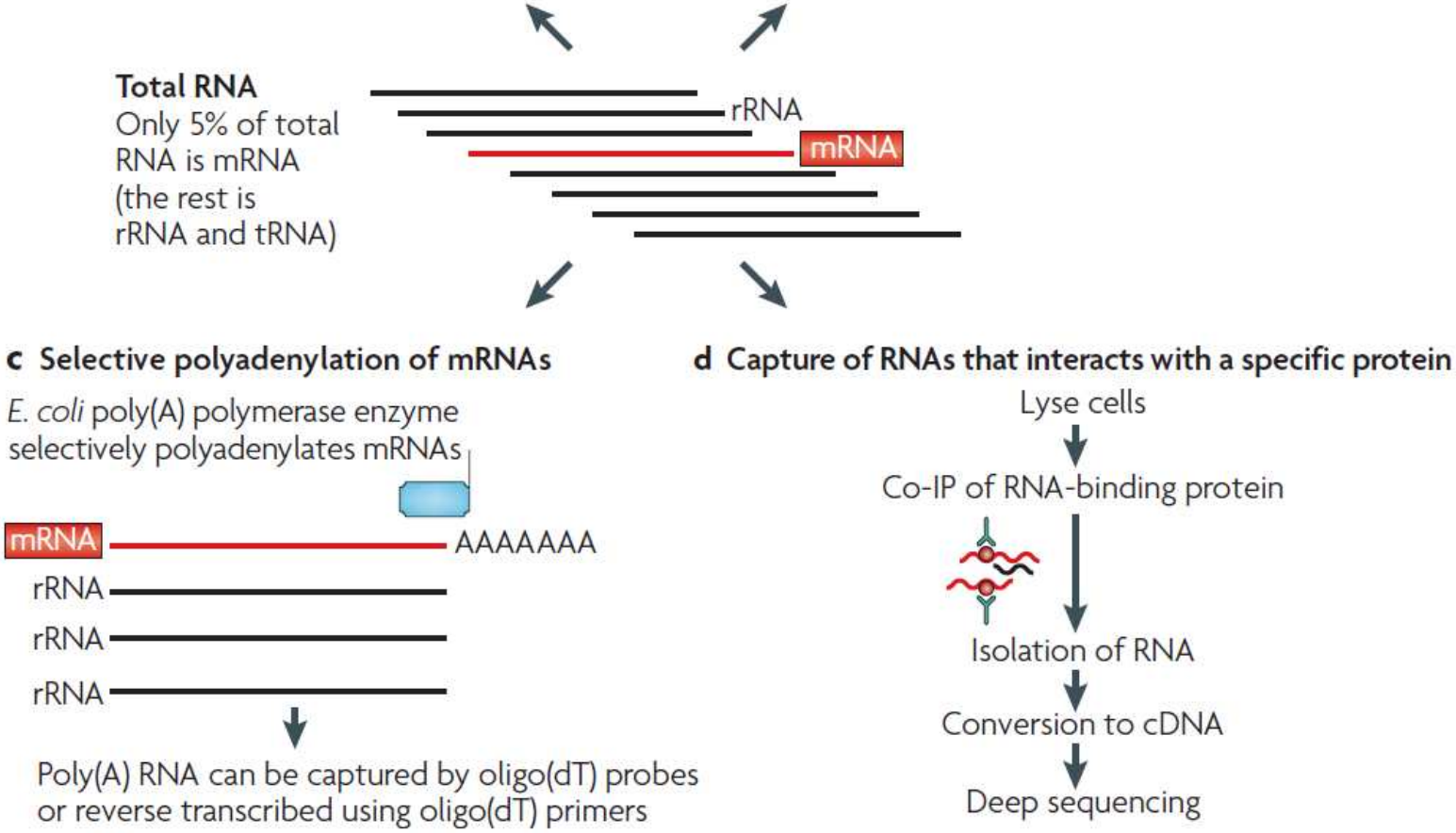


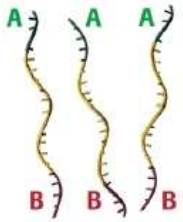
Image for step 1 is taken from [Sorek and Cossart, 2010] and images for step 2 to 6 are from [Marcel Margulies et al., 2005]

Transcriptome Sequencing (454)

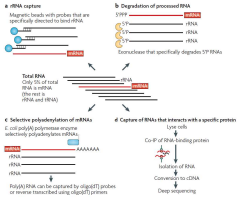


Transcriptome Sequencing (454)

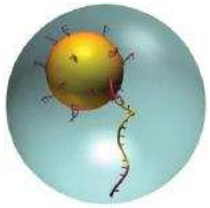
2) Fragment Preparation



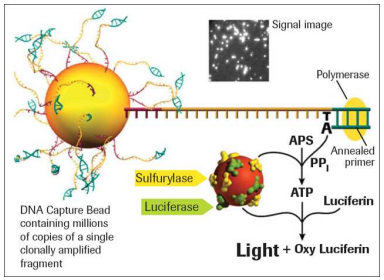
1) RNA Isolation



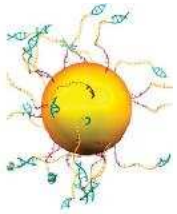
3) One Bead = One Fragment



6) One Bead = One Read



4) Amplification



5) PicoTiterPlate Fixation

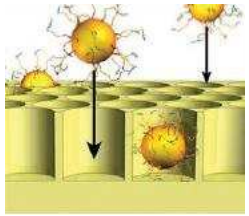
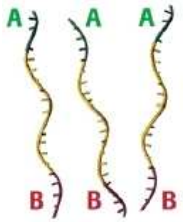


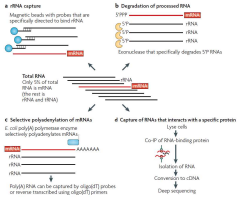
Image for step 1 is taken from [Sorek and Cossart, 2010] and images for step 2 to 6 are from [Marcel Margulies et al., 2005]

Transcriptome Sequencing (454)

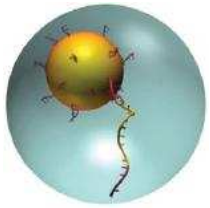
2) Fragment Preparation



1) RNA Isolation



3) One Bead = One Fragment



Thousands of sequencing reads

⇓

Post processing steps like clipping, poly-A filtering, etc.

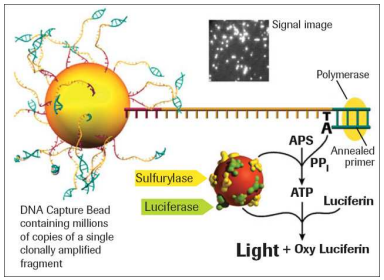
⇓

Mapping with a certain method like Blast, Segemehl, etc.

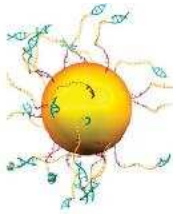
⇓

90% of mapped sequencing reads

6) One Bead = One Read



4) Amplification



5) PicoTiterPlate Fixation

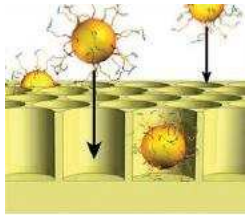
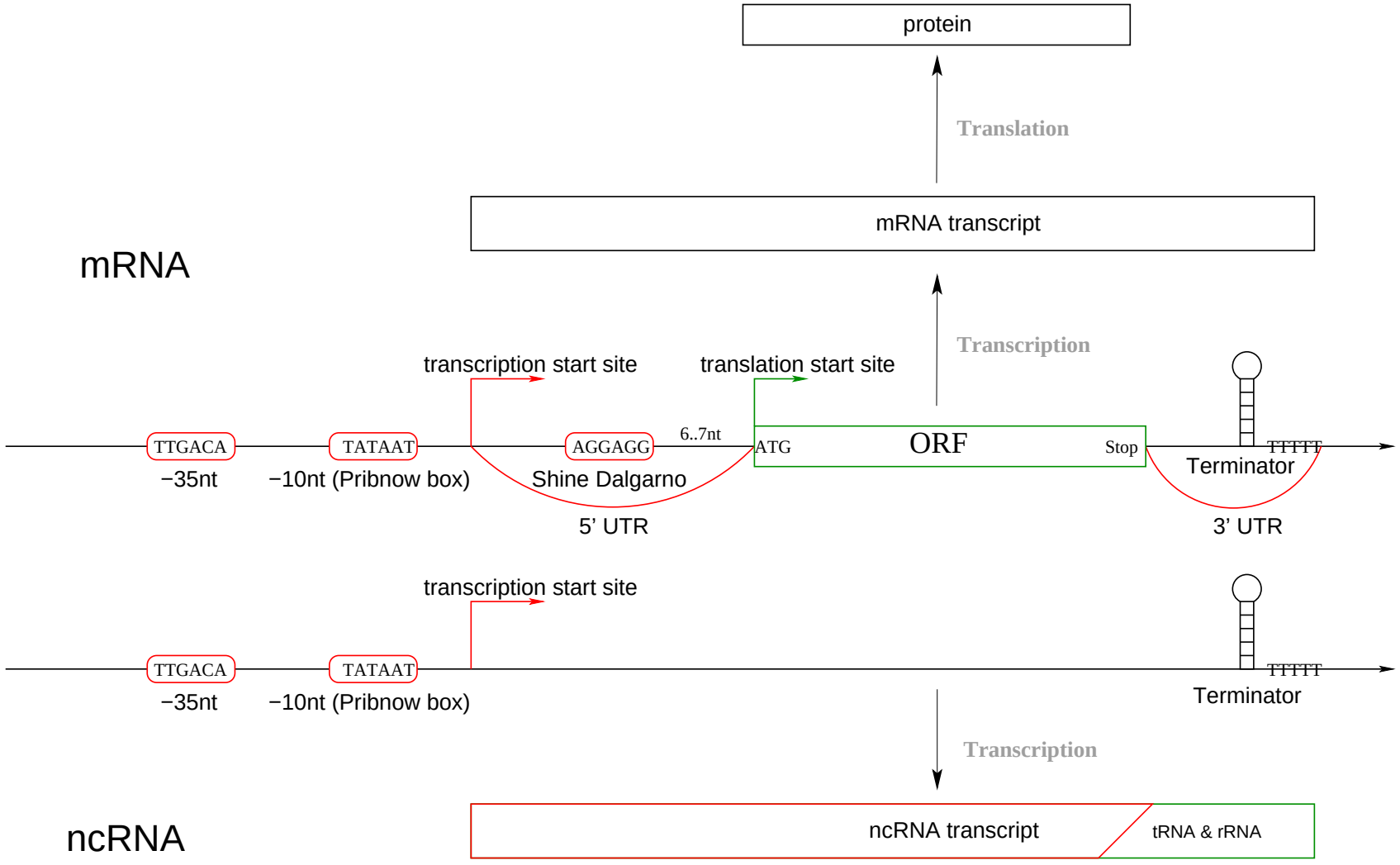


Image for step 1 is taken from [Sorek and Cossart, 2010] and images for step 2 to 6 are from [Marcel Margulies et al., 2005]

What we have and What we want



Enriched Library

- library is enriched for primary transcripts
- 5'end of the transcripts are enriched over "normal" RNA

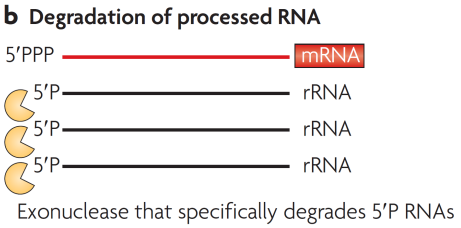
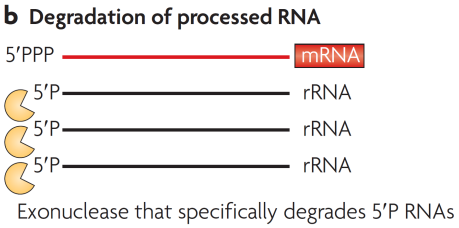


Image taken from [Sorek and Cossart, 2010]

Enriched Library

- library is enriched for primary transcripts
- 5'end of the transcripts are enriched over "normal" RNA

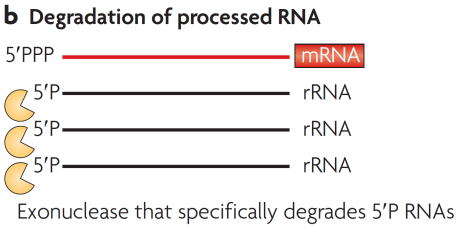


TSS annotation:

Image taken from [Sorek and Cossart, 2010]

Enriched Library

- library is enriched for primary transcripts
- 5'end of the transcripts are enriched over "normal" RNA



TSS annotation:

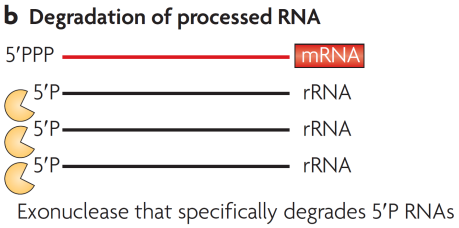
1. manual inspection

   [Sharma et al., 2010]; [Albrecht et al., 2009]; [Jäger et al., 2009] Take the IGB or UCSC and go through the whole genome, basically click by click → base by base, and annotate start sites.

Image taken from [Sorek and Cossart, 2010]

Enriched Library

- library is enriched for primary transcripts
- 5'end of the transcripts are enriched over "normal" RNA



TSS annotation:

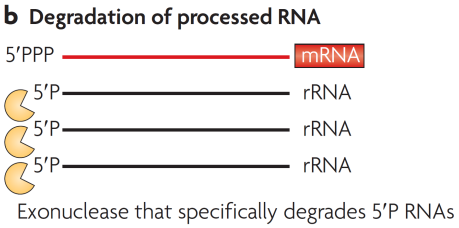
1. manual inspection

   [Sharma et al., 2010]; [Albrecht et al., 2009]; [Jäger et al., 2009] Take the IGB or UCSC and go through the whole genome, basically click by click → base by base, and annotate start sites.

Image taken from [Sorek and Cossart, 2010]

Enriched Library

- library is enriched for primary transcripts
- 5'end of the transcripts are enriched over "normal" RNA



TSS annotation:

1. manual inspection

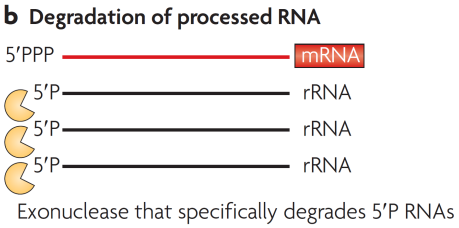
   [Sharma et al., 2010]; [Albrecht et al., 2009]; [Jäger et al., 2009] Take the IGB or UCSC and go through the whole genome, basically click by click → base by base, and annotate start sites.

   + you get in touch with the data
   − biased towards annotated genes
   − it is not reproducible
   − takes a lot of time

Image taken from [Sorek and Cossart, 2010]

Enriched Library

- library is enriched for primary transcripts
- 5'end of the transcripts are enriched over "normal" RNA



TSS annotation:

1. manual inspection

   [Sharma et al., 2010]; [Albrecht et al., 2009]; [Jäger et al., 2009] Take the IGB or UCSC and go through the whole genome, basically click by click → base by base, and annotate start sites.

   + you get in touch with the data
   − biased towards annotated genes
   − it is not reproducible
   − takes a lot of time

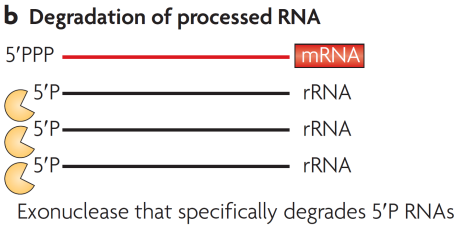
2. automated methods:

   [Wurtzel et al., 2009]; Own Methods
   Establish a method that searches for start sites. Inspect some of the annotated start sites and maybe refine the method.

Image taken from [Sorek and Cossart, 2010]

Enriched Library

- library is enriched for primary transcripts
- 5'end of the transcripts are enriched over "normal" RNA



TSS annotation:

1. manual inspection

   [Sharma et al., 2010]; [Albrecht et al., 2009]; [Jäger et al., 2009] Take the IGB or UCSC and go through the whole genome, basically click by click → base by base, and annotate start sites.

   + you get in touch with the data
   − biased towards annotated genes
   − it is not reproducible
   − takes a lot of time

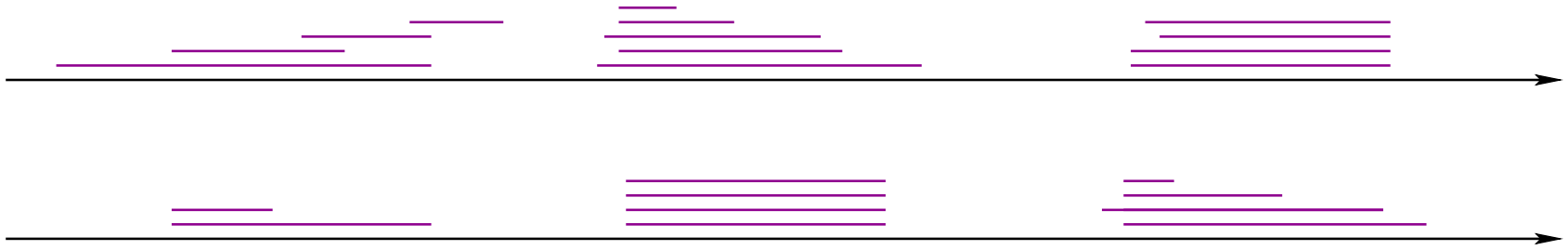
2. automated methods:

   [Wurtzel et al., 2009]; Own Methods
   Establish a method that searches for start sites. Inspect some of the annotated start sites and maybe refine the method.

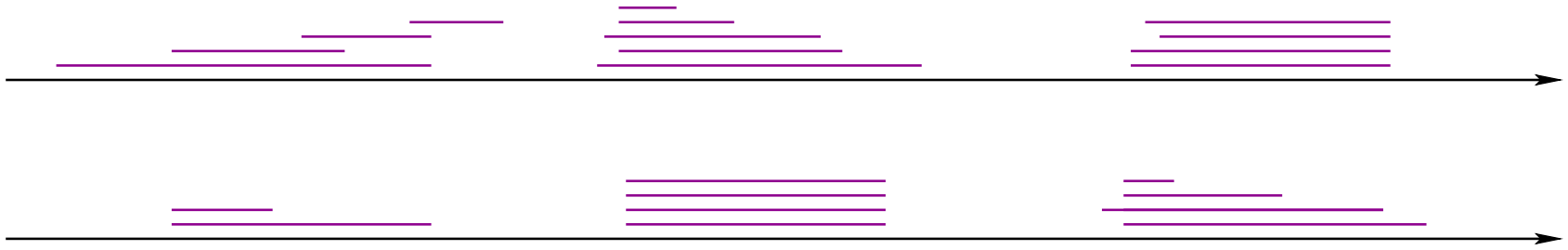
   − you do not get in touch with the complete data
   + unbiased if no annotation is used
   + it is easy to refine and reproducible
   + once the method is established it takes seconds to annotate start sites
   − you could have TSS annotated that do not fit the idea

Image taken from [Sorek and Cossart, 2010]

A simple approach

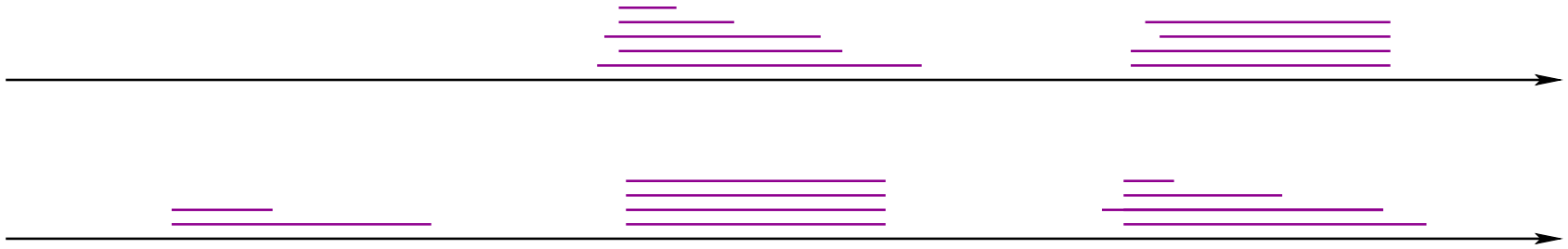


A simple approach



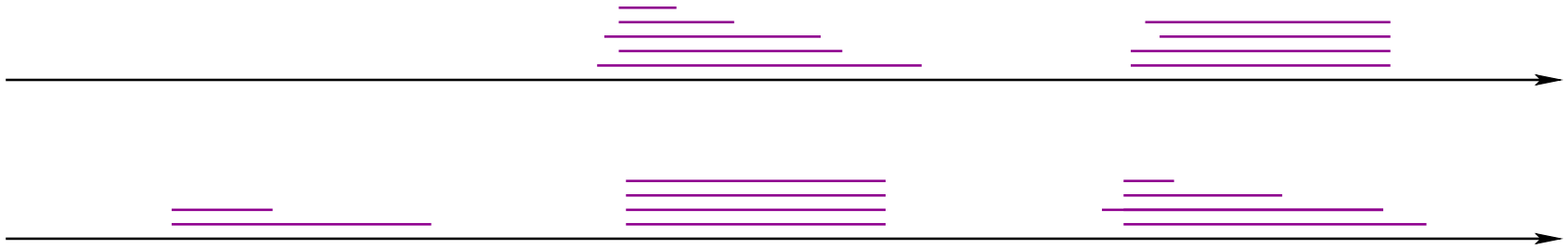
1. within a small window some reads should start

A simple approach



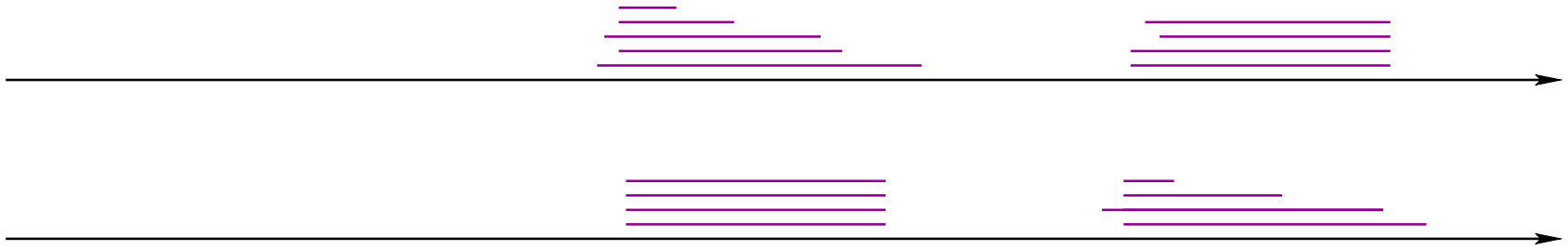
1. within a small window some reads should start

A simple approach



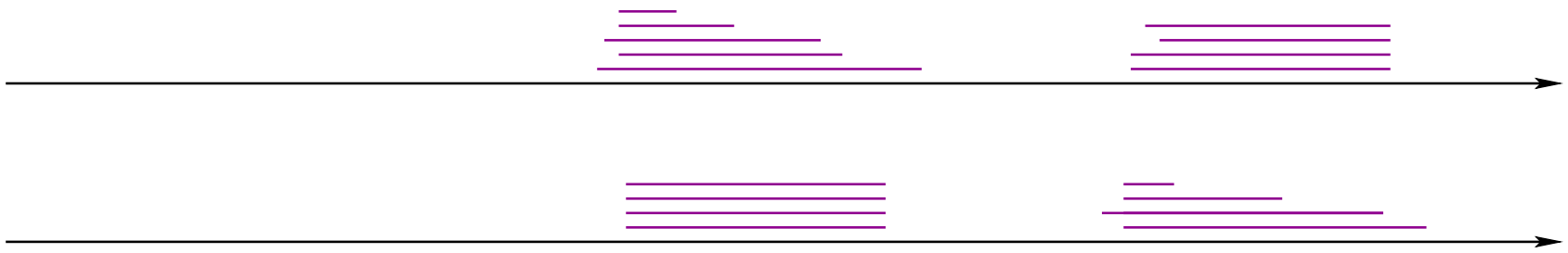
1. within a small window some reads should start
2. some := at least three reads

A simple approach



1. within a small window some reads should start
2. some := at least three reads

A simple approach

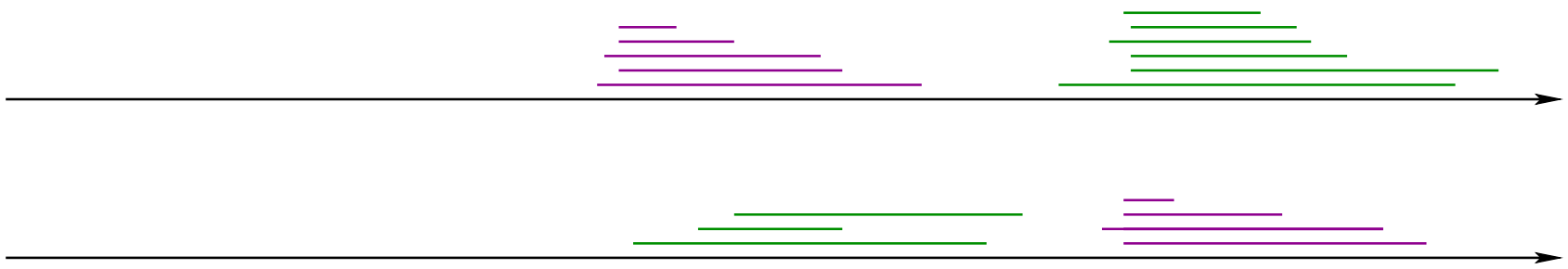


1. within a small window some reads should start
2. some := at least three reads
3. reads have to end on different positions

A simple approach

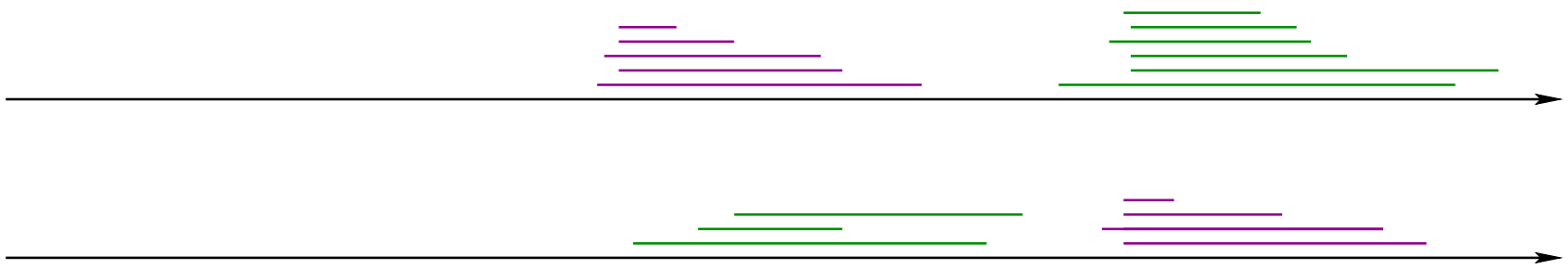
1. within a small window some reads should start
2. some := at least three reads
3. reads have to end on different positions

A simple approach



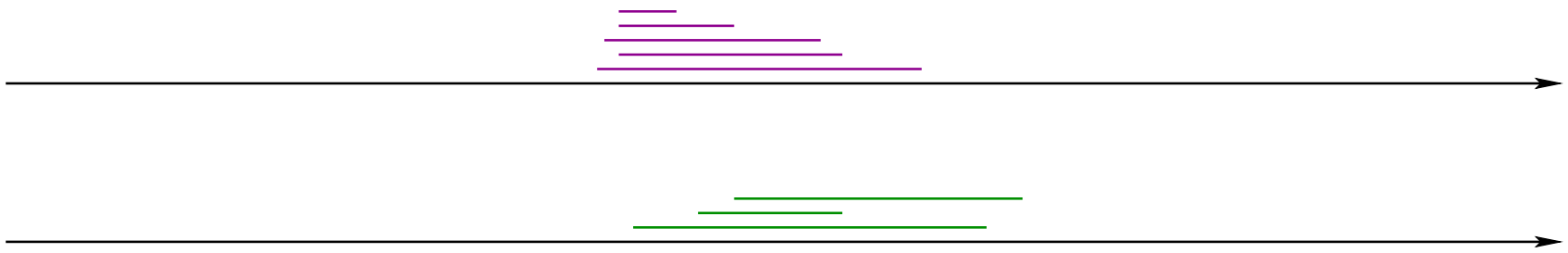
1. within a small window some reads should start
2. some := at least three reads
3. reads have to end on different positions

A simple approach



1. within a small window some reads should start
2. some := at least three reads
3. reads have to end on different positions
4. #reads in a treated library $>$ #reads in an untreated library

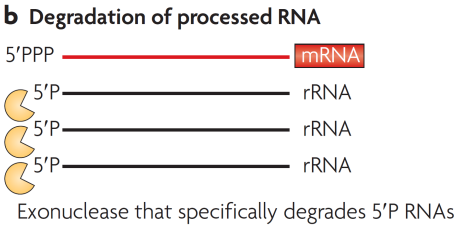
A simple approach



1. within a small window some reads should start
2. some := at least three reads
3. reads have to end on different positions
4. #reads in a treated library $>$ #reads in an untreated library

Enriched Library

- library is enriched for primary transcripts
- 5'end of the transcripts are enriched over "normal" RNA



TSS annotation:

1. manual inspection

   [Sharma et al., 2010]; [Albrecht et al., 2009]; [Jäger et al., 2009] Take the IGB or UCSC and go through the whole genome, basically click by click → base by base, and annotate start sites.

   + you get in touch with the data
   − biased towards annotated genes
   − it is not reproducible
   − takes a lot of time

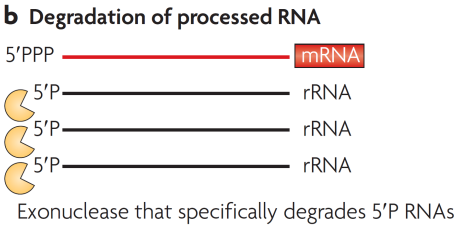
2. automated methods:

   [Wurtzel et al., 2009]; Own Methods
   Establish a method that searches for start sites. Inspect some of the annotated start sites and maybe refine the method.

Image taken from [Sorek and Cossart, 2010]

Enriched Library

- library is enriched for primary transcripts
- 5'end of the transcripts are enriched over "normal" RNA



TSS annotation:

1. manual inspection

   [Sharma et al., 2010]; [Albrecht et al., 2009]; [Jäger et al., 2009] Take the IGB or UCSC and go through the whole genome, basically click by click → base by base, and annotate start sites.

   + you get in touch with the data
   − biased towards annotated genes
   − it is not reproducible
   − takes a lot of time

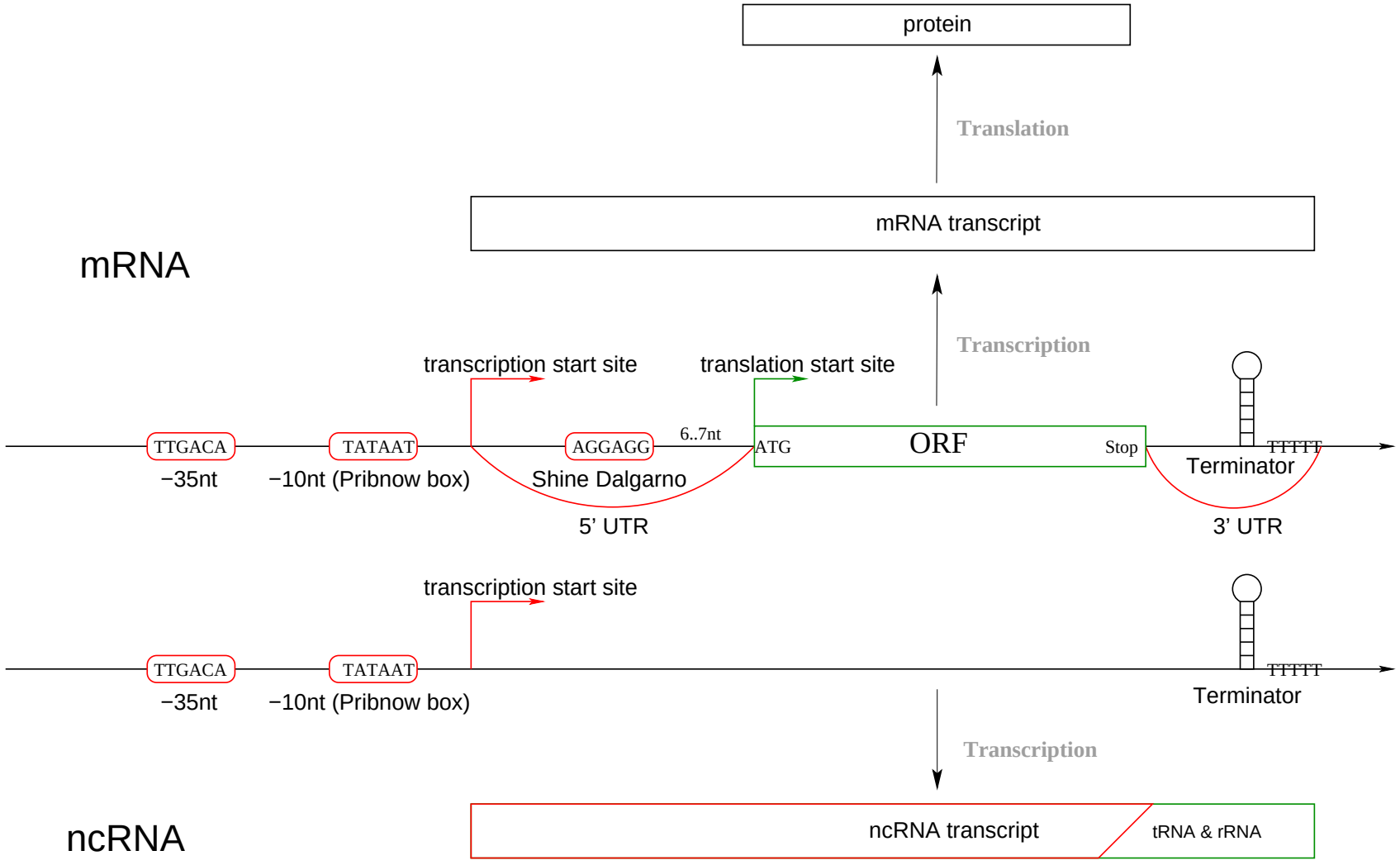
2. automated methods:

   [Wurtzel et al., 2009]; Own Methods
   Establish a method that searches for start sites. Inspect some of the annotated start sites and maybe refine the method.

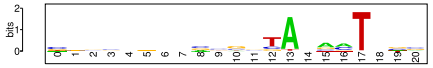
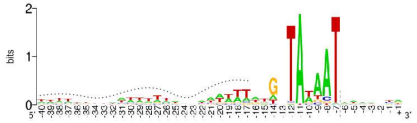
   − you do not get in touch with the complete data
   + unbiased if no annotation is used
   + it is easy to refine and reproducible
   + once the method is established it takes seconds to annotate start sites
   − you could have TSS annotated that do not fit the idea

Image taken from [Sorek and Cossart, 2010]

What we have and What we want



Regulatory motif Detection

Publication	# Library	Enriched/Normal Total	Core Promoter	5’ UTR
Own data	1	45,419/56,257 101,676		
[Albrecht et al., 2009]	2	?/? 249,432		
[Sharma et al., 2010]	5	1,435,974/1,384,949 2,820,923		

For Further Reading

Rotem Sorek and Pascale Cossart
Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity.
Nat Rev Genet, 2010

Zhong Wang and Mark Gerstein and Michael Snyder
RNA-Seq: a revolutionary tool for transcriptomics.
Nat Rev Genet, 2009

Joshua Lederberg and Alexa T. McCray
Ome Sweet 'Omics– A Genealogical Treasury of Words
Commentary by The Scientist 15[7]:8, Apr. 2, 2001

Marcel Margulies et al.
Genome sequencing in microfabricated high-density picolitre reactors.
Nature, 2005

Cynthia M Sharma et al.
The primary transcriptome of the major human pathogen Helicobacter pylori
Nature, 2010 doi:10.1038/nature08756

Marco Albrecht et al.
Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome.
Nucleic Acids Res, 2009

For Further Reading

Dominik Jäger et al.
Deep sequencing analysis of the Methanosarcina mazei Gö1 transcriptome in response to nitrogen availability.
Proc Natl Acad Sci U S A, 2009

Omri Wurtzel et al.
A single-base resolution map of an archaeal transcriptome.
Genome Res, 2009

Thanks to...

All the Bleden