# RNA-seq: analysis and de novo assembly

Yakovlev Pavel

St-Petersburg Academic University
Supervisor: Segrey Nurk

June 8, 2013

# What?

We want to know everything about RNA high throughput sequencing.

- How RNA is sequenced?
- What are the features and differences from DNA sequencing data?
- How to assembly this data?
- How to validate assembled data?

# Why?

Assemble and analyze RNA is very cool, because of:

- we get only what we interested in genes (mRNA) and ncRNA,
- we can get useful information about ogranism without full-genome sequencing,
- everybody can do it, but we cannot yet!
- everybody do it awfully.

# And what do we have now?

Our results:

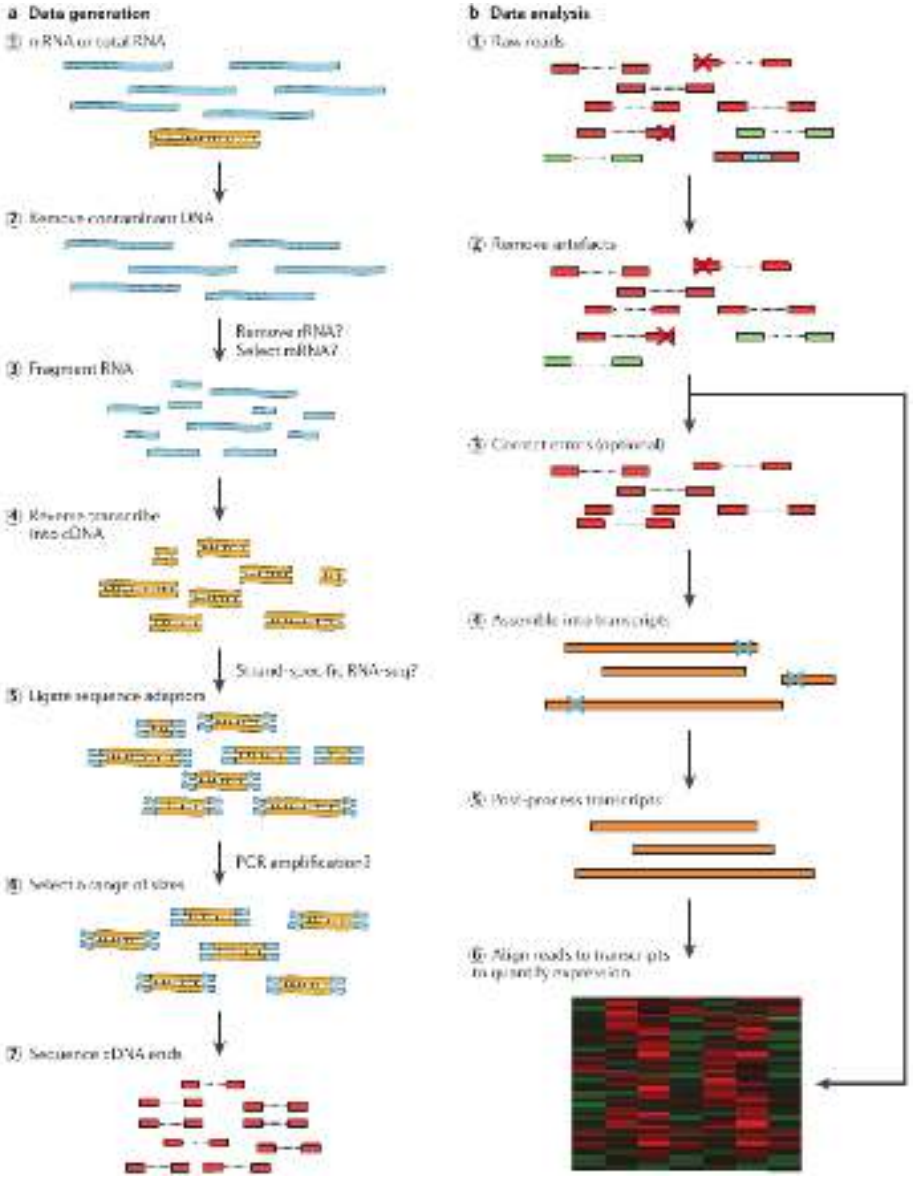
- How RNA is sequenced? & What are the features and differences from DNA sequencing data?
  - Now we are experts in RNA-Seq data
- How to assembly this data?
  - Now we know abit about this too.
- How to validate assembled data?
  - We did some validation, but it was a little magical...

# Theory

In RNA-seq:

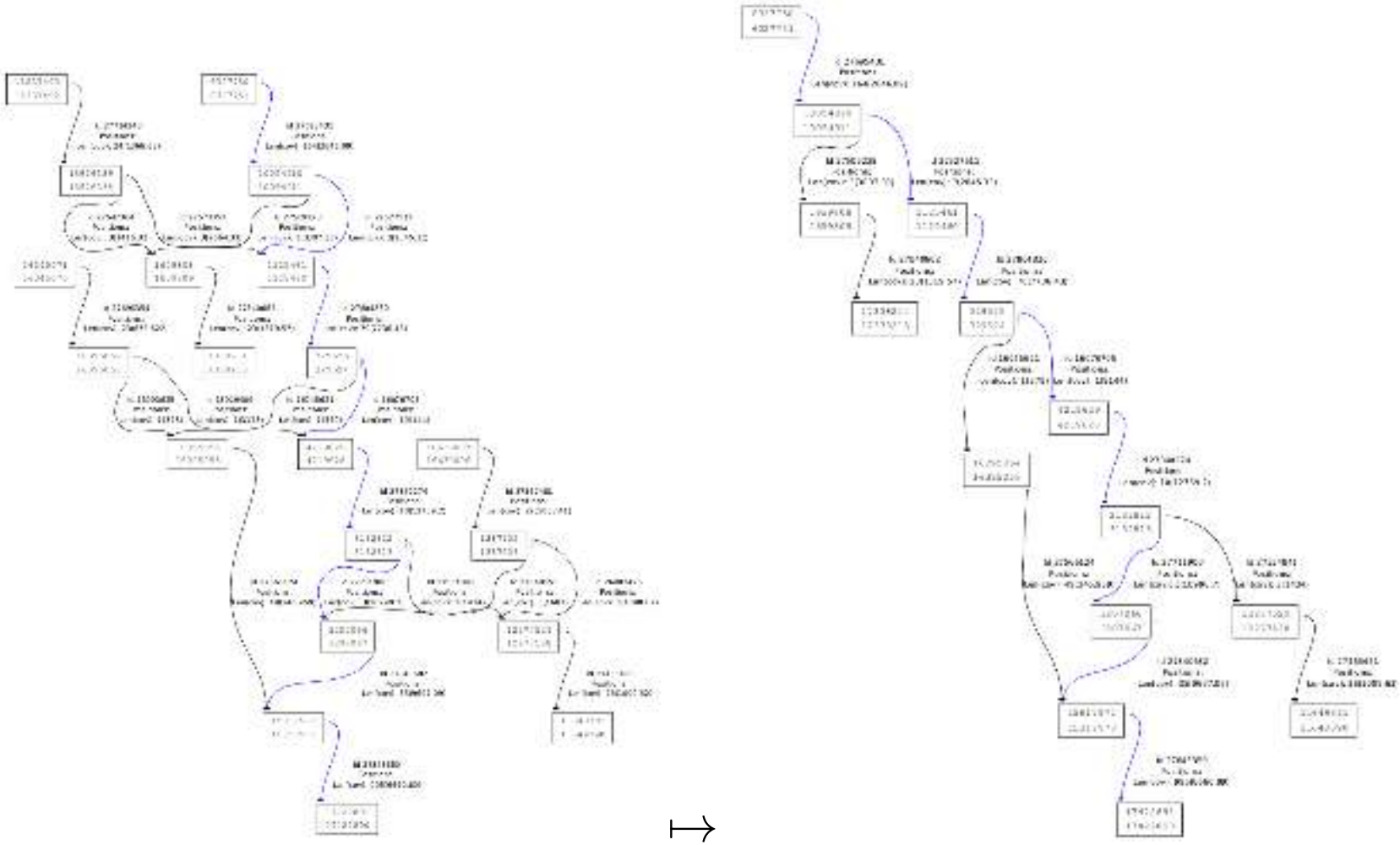
- cDNA is sequenced, so we can get paired reads
- eukaryotic mRNA has cap and poly-A tail, which help to identify start and end
- prokaryotic mRNA do not have alternative splicing, so it is easy to assemble them (**FALSE**)
- all the headache starts with alternative splicing...
  - different isoforms have different expression
  - different isoforms have same exons
  - and more, than you can imagine

# Practice
## RNA-Seq



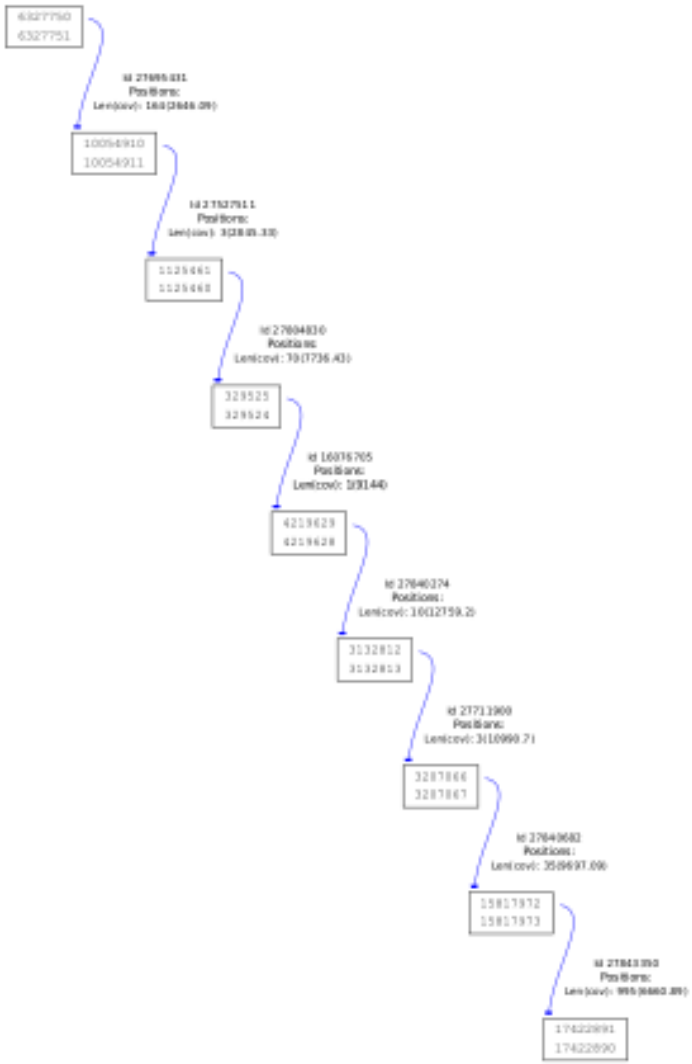

- We may be sure, that we have only mRNA reads
- We can get two-stranded pair reads from cDNA
- If we have eukaryotic mRNA, we know start (sometimes) and end (always) reads
- We have very different coverage, based on expression level

Part of result graph:



# Prokaryote assemble
## Solution

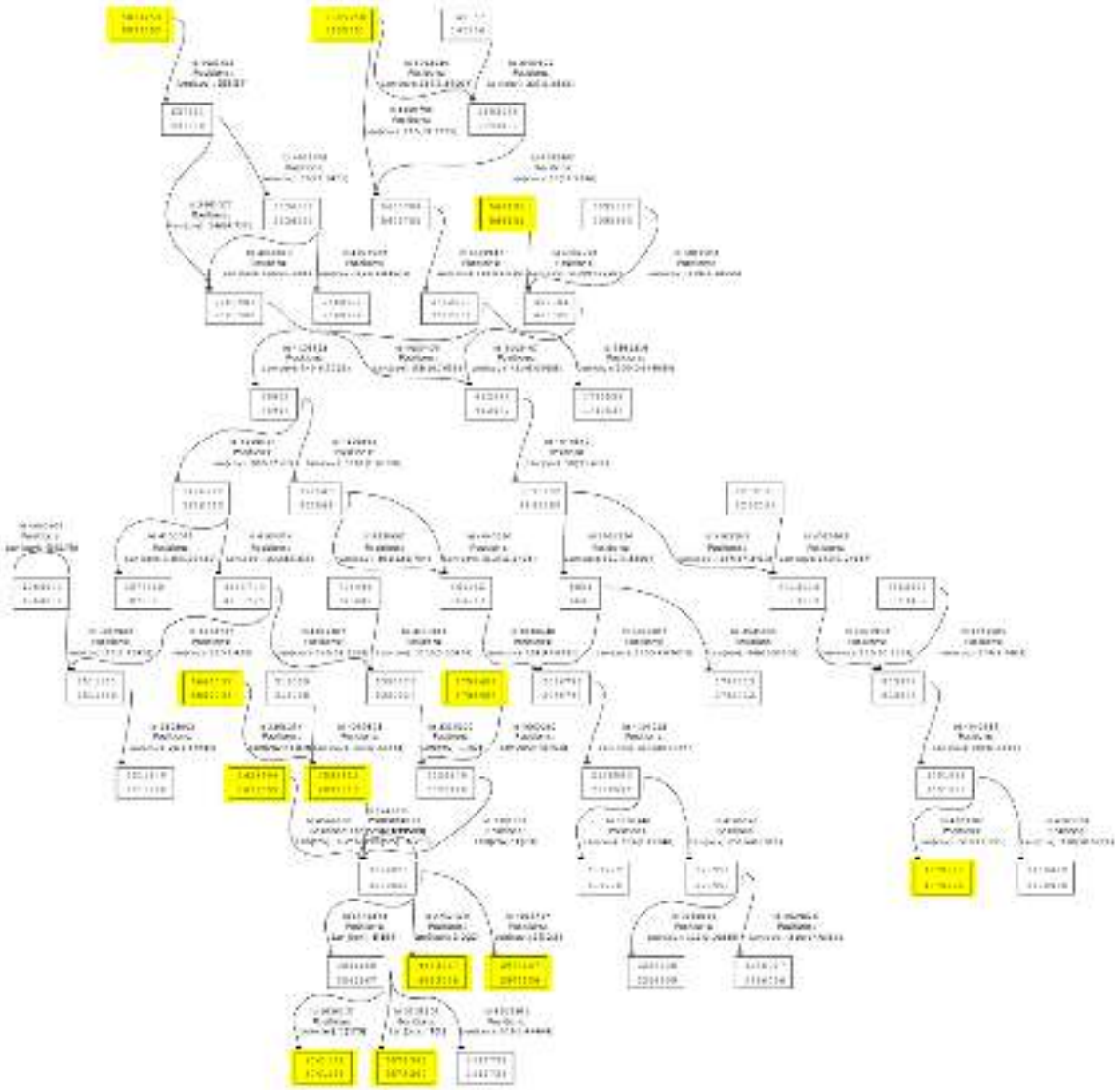
At last:



How?

- Dominanting set technique:
  - Get list of all source vertices, ordered by coverage and length of output edge
  - Find the dominant set of each vertex
  - Remove the dominant set from graph for every source vertices except the last one
- Repeat the graph simplification:
  - Tip clipper
  - Bulges remover
  - Low covered connections remover

Mouse graph part:

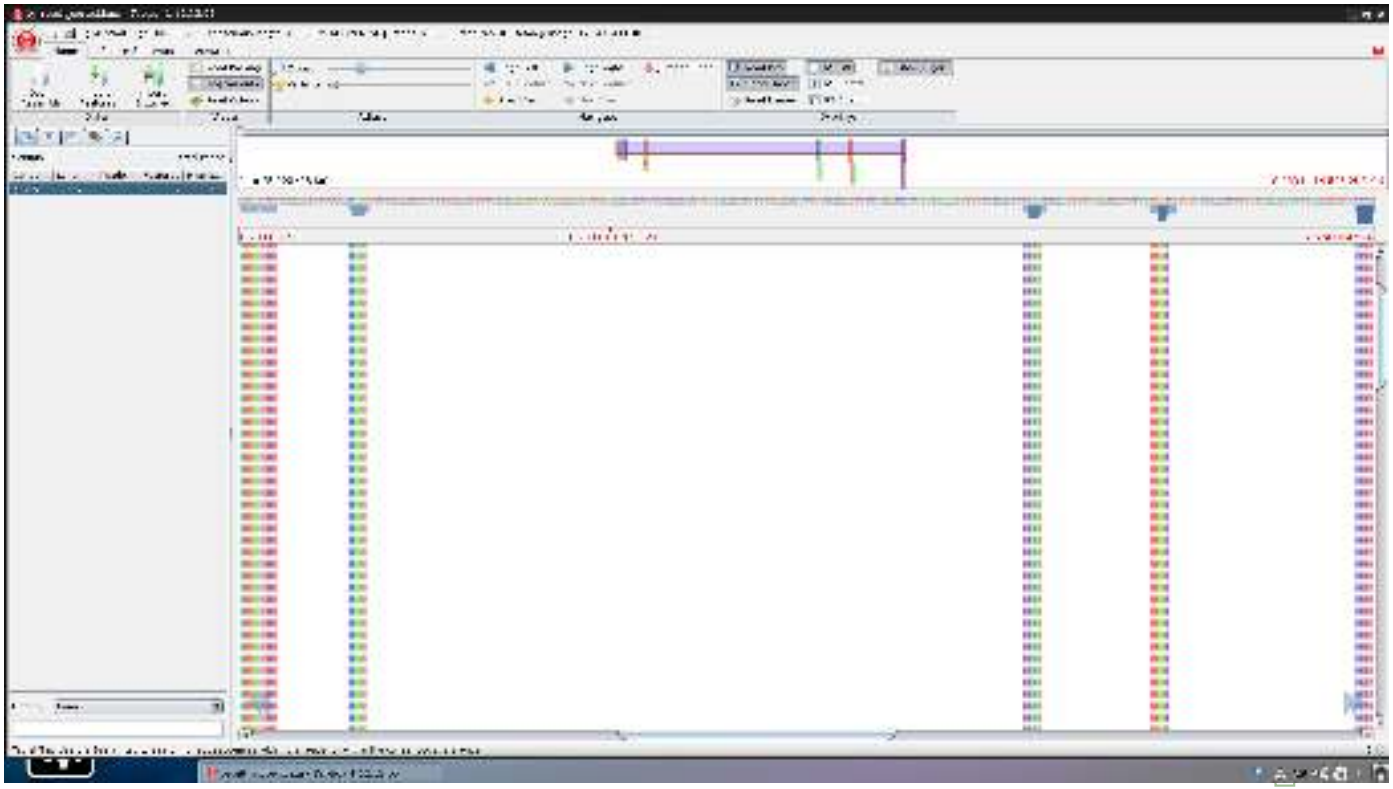


# Validation

Methods:

- Alignment
- tBLAST
- Cufflinks (reference-based assembler)

Magic:

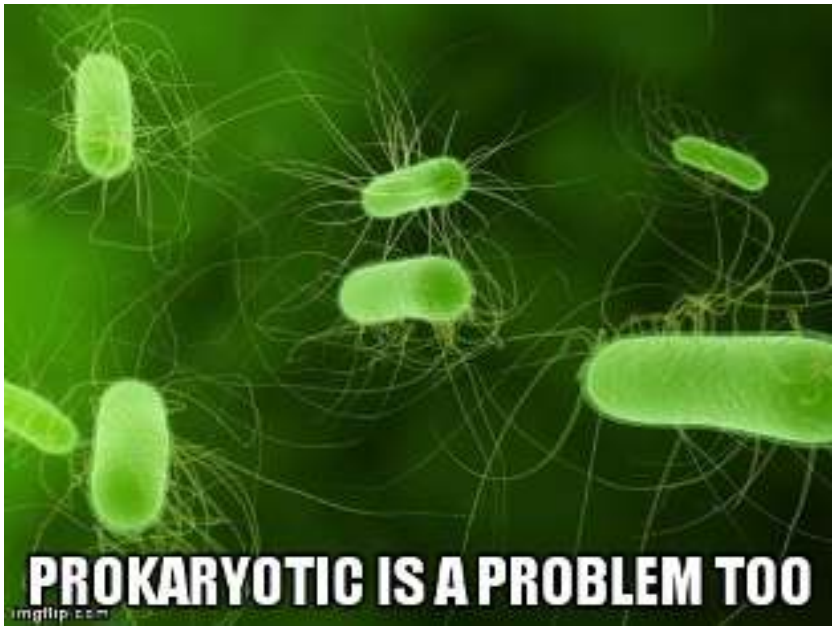


# Conclusion

- Prokaryotic transcriptome is assembled hard
  - Lots of very short edges
  - BUT dominating set technique works well
  - Selection of an edge by coverage usually works good too
- SPAdes still works bad with eukaryotic transcriptome
  - Lots of long edges with very different coverage
  - Deletion of poor covered edges make graph more readable, but do not fix the whole problem
  - Using of pair reads sends SPAdes to "Out of memory"

# The end

Questions?


Really, not simple.