Introduction to RNA-Seq

Roman Cheplyaka

Questions about RNA

1. Is this gene transcribed at all?
2. In what quantity?
3. Which isoforms?

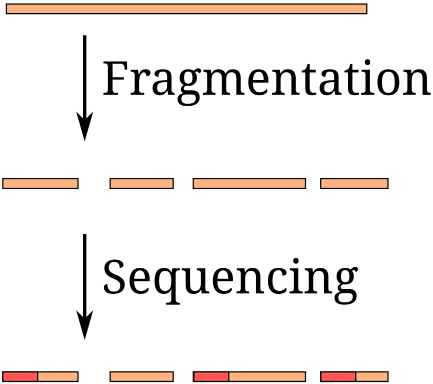
RNA analysis methods

- Microarrays
- RT-qPCR (Quantitative Reverse Transcription PCR)
- SAGE (Serial Analysis of Gene Expression)
- CAGE (Cap Analysis of Gene Expression)
- MPSS (Massively Parallel Signature Sequencing)
- **RNA-Seq (RNA Sequencing)**

RNA-Seq steps

1. RNA extraction
2. mRNA enrichment/rRNA removal
3. Reverse transcription (RNA to DNA)
4. Complementary DNA sequencing
5. **Data analysis**

Sequencing

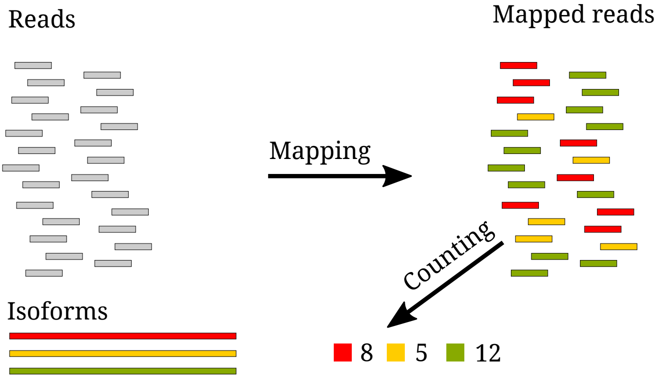


Raw data

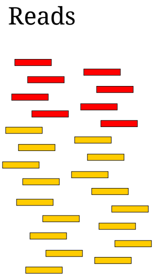
Sequencing reads in the FASTQ format:

```
@SRR1515105.2 OBIWAN:27:D1G3PACXX:6:1101:1626:1911/1
NTGAATTGGCCTGGTTTCAGATTTGGTTAGCTGCGGATTGGCTGCCTTAGCTCAT
+
#1:=BDDDDFFFBGFGFBA:CFHHHEFGFDDFFFFF0@FGCBFF79=).8BFDFE
```

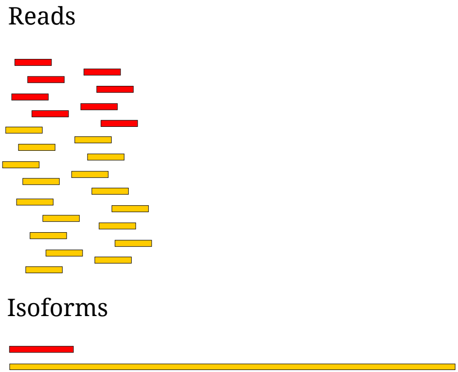
Mapping



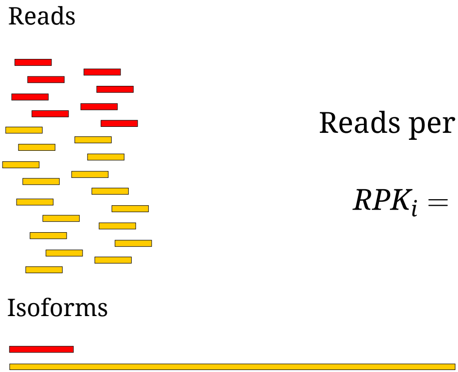
Read counts



Read counts



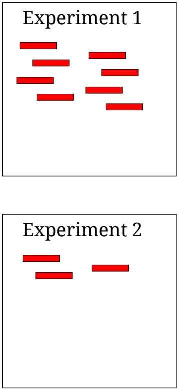
Read counts



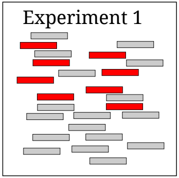
Reads per kilobase:

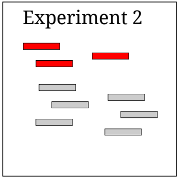
$$RPK_i = 10^3 \cdot \frac{n_i}{l_i}$$

Read counts

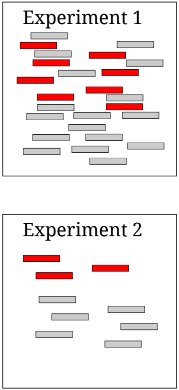


Read counts





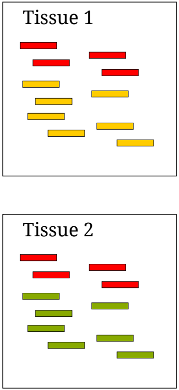
Read counts



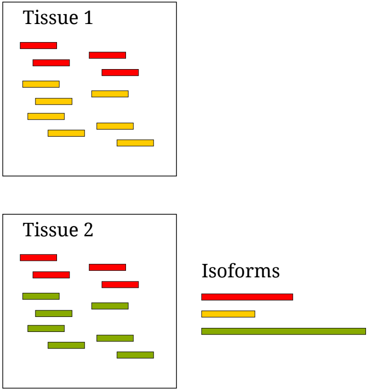
Reads per kilobase per million reads:

$$RPKM_i = 10^9 \cdot \frac{n_i}{l_i \cdot \sum_j n_j}$$

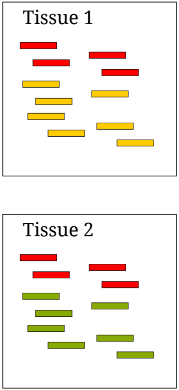
Read counts



Read counts



Read counts

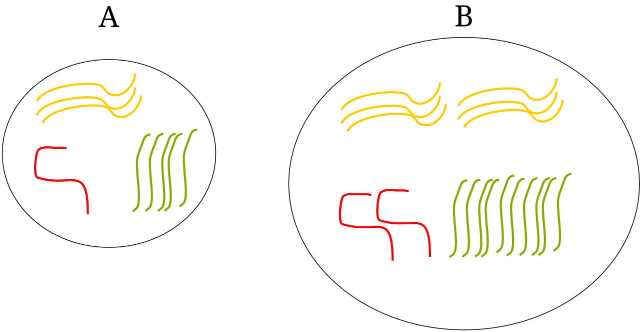


Transcripts per million:

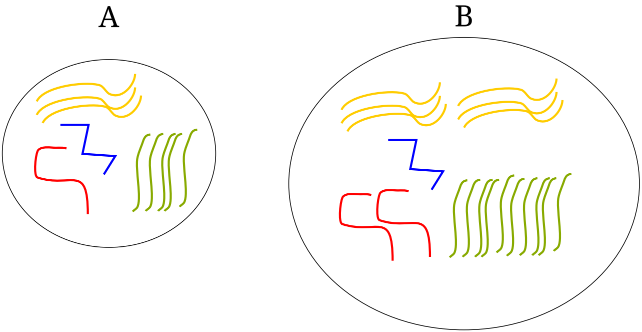
$$TPM_i = 10^6 \cdot \frac{t_i}{\sum_j t_j} = 10^6 \cdot \frac{n_i / l_i}{\sum_j n_j / l_j}$$

Isoforms

Absolute vs relative expression

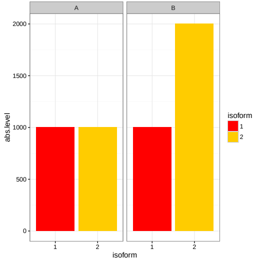


Absolute vs relative expression

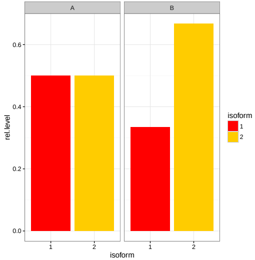


Danger of relative expression

Absolute expression



Relative expression

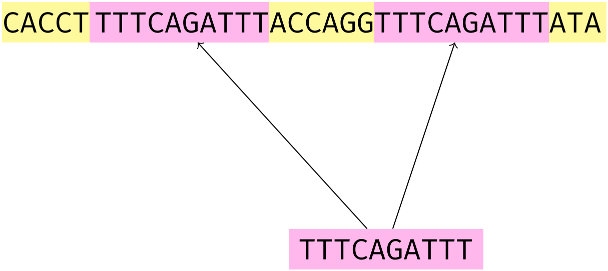


Mapping

ACCAGGTTTCAGATTTATA

TTTCAGATTT

Mapping: multiple hits



Mapping: splice junction

CACCTTTTCAACCAGGGATTTATA

TTTCAGATTT

Mapping: poly(A) tails

ACCAGGTTTCAGATTTATA

TTTCAGATTTAAAA

Mapping: errors

ACCAGGTTTCAGATTTATA

TTACAGATTT

How many reads are ambiguously mapped?

250 000 RNA-Seq reads from *Drosophila melanogaster*

Reference	# of hits	reads	
Genome	> 1 hit	13 657	5.5%
Genome	> 10 hits	1 638	0.7%
Transcriptome	> 1 hit	139 410	55.8%
Transcriptome	> 10 hits	6 197	2.5%

How to handle ambiguous reads?

Can we discard the reads with multiple hits?

How to handle ambiguous reads?

Can we discard the reads with multiple hits?

No! This may introduce a bias.

How to handle ambiguous reads?

```
Isoform 1        TTACAGATTT
Isoform 2        TTGCAGATTT

Reads            TTACA
                 TTACA
                   ACAGA
                    CAGAT
                    CAGAT
                      GATTT
```

RNA-Seq data analysis challenges

1. Mapping (or not!)
2. Counting
3. Normalization
4. Statistical inference