

Tuesday – part 4

Expression estimation

Michał Szcześniak, PhD

Faculty of Biology, Adam Mickiewicz University, Poznań
ideas4biology Ltd.

Pipelines

1. FASTQ → QC and filtering → mapping → *ab initio* assembly → (**expression estimation**)
2. FASTQ → QC and filtering → (mapping) → **expression estimation** → differential expression analysis
3. FASTQ → QC and filtering → *de novo* assembly → (**expression estimation**)

Techniques for transcript identification and expression estimation

- PCR (RT-PCR, qPCR)
- Northern blots
- EST (Expressed Sequence Tags)
- Microarrays
- SAGE (Serial Analysis of Gene Expression)
- TSS (Transcription Start Sites)
- **RNA-seq**

Selected tools for expression estimation from RNA-Seq data

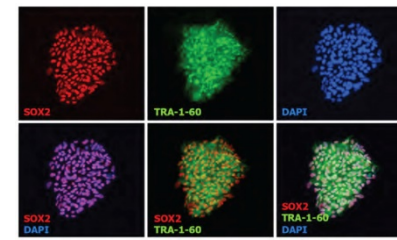
- **RSEM**

<http://deweylab.github.io/RSEM/>

- **SALMON**

<https://combine-lab.github.io/salmon/>

Input data (RSEM)



✓ **READS** (*.fastq)

- Source: ENA EBI database
- ERR990413_1.fastq, ERR990413_2.fastq
- Prepared files: **reads.1.fastq, reads.2.fastq**



✓ **GENOME** → chromosome 22

- From Ensembl
- Homo_sapiens.GRCh38.dna.chromosome.22.fa

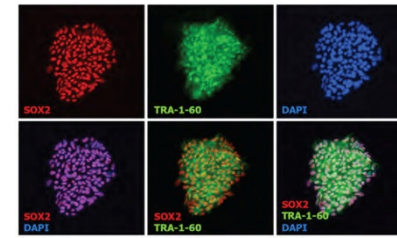


✓ **ANNOTATIONS** (*.gtf)

- From Ensembl
- Homo_sapiens.GRCh38.88.gtf



The pipeline



- **Genome (FASTA)**
- **RNA-seq reads (FASTQ)**
- **Annotations (GTF format)**

- Building an index

- Expression estimation (**RSEM**)

RSEM

- RSEM maps reads against transcriptome or genome using **Bowtie** and estimates gene and transcript expressions based on these mappings
- It does not detect novel splicing isoforms
- Advanced algorithm (including EM), high correlation of estimated expression with experimentally validated results
- Relatively slow

RSEM: preparing the data

1. Create a **rsem** directory
2. Create a **index_rsem** directory
3. Build an index for chr 22 using **rsem-prepare-reference**

```
mkdir rsem
mkdir index_rsem

rsem-prepare-reference --bowtie --gtf Homo_sapiens.GRCh38.88.gtf
Homo_sapiens.GRCh38.dna.chromosome.22.fa index_rsem/human_22_rsem
```

--bowtie → we are using Bowtie for reads mapping

-- gtf a file with genome annotations

A FASTA file with chr 22

An output index

RSEM: expression estimation

Performing expression estimation with RSEM:

rsem-calculate-expression

```
rsem-calculate-expression -p 1 --no-bam-output --paired-end  
reads.1.fastq reads.2.fastq --forward-prob 0 index_rsem/  
human_22_rsem rsem/expression
```

-p n – numer of threads

-- no-bam-output – do not save BAM files

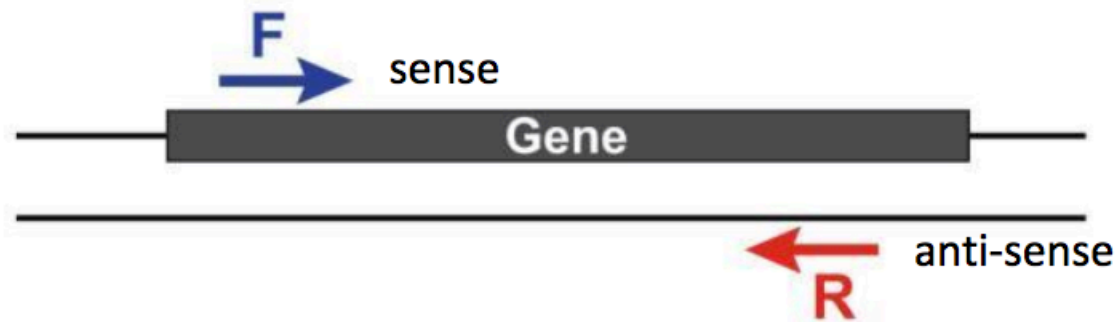
Input, paired-end reads in FASTQ format

--forward-prob 0 – type of strand specificity

Transcriptome index

Prefix for output files

Strand-specific RNA-Seq



	--SS_lib_type	Read 1	Read 2
SE cases	F		—
	R		—
PE cases	FR		
	RF		

RSEM – output files

- ✓ **expression.genes.results**
- ✓ **expression.isoforms.results**

transcript_id	gene_id	length	effective_length		expected_count		TPM	FPK
ENST00000008876	ENSG00000008735	5596	5310.13	0.00	0.00	0.00	0.00	
ENST00000329492	ENSG00000008735	3381	3095.13	2.00	19.53	14.45	100.00	
ENST00000317361	ENSG00000015475	2495	2209.13	4.48	61.33	45.39	5.12	
ENST00000342111	ENSG00000015475	854	568.13	0.00	0.00	0.00	0.00	
ENST00000399765	ENSG00000015475	1879	1593.13	0.00	0.00	0.00	0.00	
ENST00000399767	ENSG00000015475	2129	1843.13	0.00	0.00	0.00	0.00	
ENST00000399774	ENSG00000015475	2167	1881.13	22.47	360.86	267.06	30.10	
ENST00000473439	ENSG00000015475	638	352.24	3.37	289.16	213.99	24.12	
ENST00000494097	ENSG00000015475	2429	2143.13	0.00	0.00	0.00	0.00	

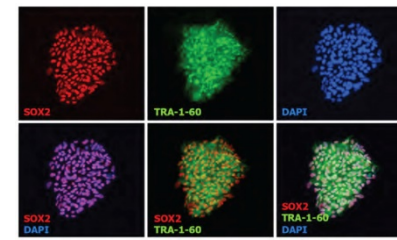
RSEM – output files

FPKM, TPM: normalized expression values

Expected_count: number of reads assigned to a given transcript / gene

IsoPct: a fraction of gene expression assigned to a given splicing isoform

Our analysis



- **Genome (FASTA)**
- **RNA-seq reads (FASTQ)**
- **Annotations**



- Transcriptome index



- Expression estimation (**RSEM**)



Expression units

RPKM -reads per kilobase per million mapped reads

number of mapped reads/locus length (kb) x all mapped reads (mln)

FPKM - fragments per kilobase per million mapped reads

TPM - Divide the read counts by the length of each gene in kilobases (RPK). Count up all the RPK values in a sample and divide this number by 1,000,000.

expected_count – numer of reads assigned to a given transcript or gene (no normalization used)

Selected tools

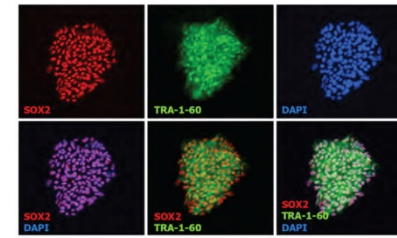
- RSEM

<http://deweylab.github.io/RSEM/>

- **SALMON**

<https://combine-lab.github.io/salmon/>

Our pipeline



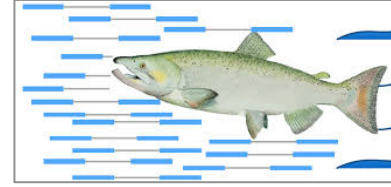
- Transcriptome (FASTA)
- RNA-seq reads (FASTQ)
- Annotations (optional)



- Transcriptome index

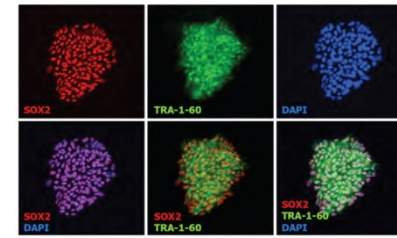
- Expression estimation (**SALMON**)

Salmon



- Incredibly fast
- Expression estimation at gene and transcript levels
- Two modes of action: indexing and quasi-mapping (no mapping required)

Input data(Salmon)



✓ **READS** (*.fastq)

- ERR990413_1.fastq, ERR990413_2.fastq
- Our files: **reads.1.fastq, reads.2.fastq**



✓ **TRANSCRIPTOME**

- Homo_sapiens.GRCh38.dna.cdna.all.fa

✓ **ANNOTATIONS** (*.gtf)

- Homo_sapiens.GRCh38.88.gtf



DATA

Ensembl → Download data via FTP → cDNA (FASTA)

```
wget ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/cdna/  
Homo_sapiens.GRCh38.cdna.all.fa.gz
```

```
gunzip Homo_sapiens.GRCh38.cdna.all.fa.gz
```

```
mv Homo_sapiens.GRCh38.cdna.all.fa human_transcriptome.fa
```

Building a transcriptome index

1. Create a **salmon** directory
2. Create a **index_salmon** directory
3. Build an index with **salmon index**

```
mkdir salmon  
mkdir index_salmon
```

```
salmon index -p 4 -t human_transcriptome.fa  
-i index_salmon/human_transcriptome_salmon
```

-p n → number of threads

A FASTA file with human transcriptome

Transcriptome index: prefix and localization

Expression estimation

```
salmon quant -p 4 --useVBOpt  
-i index_salmon/human_transcriptome_salmon  
-l ISR -1 reads.1.fastq -2 reads.2.fastq  
--geneMap Homo_sapiens.GRCh38.88.gtf -o salmon/expression
```

-p n – number of threads

--useVBOpt – used to optimize the estimations (optional)

A path to transcriptome index

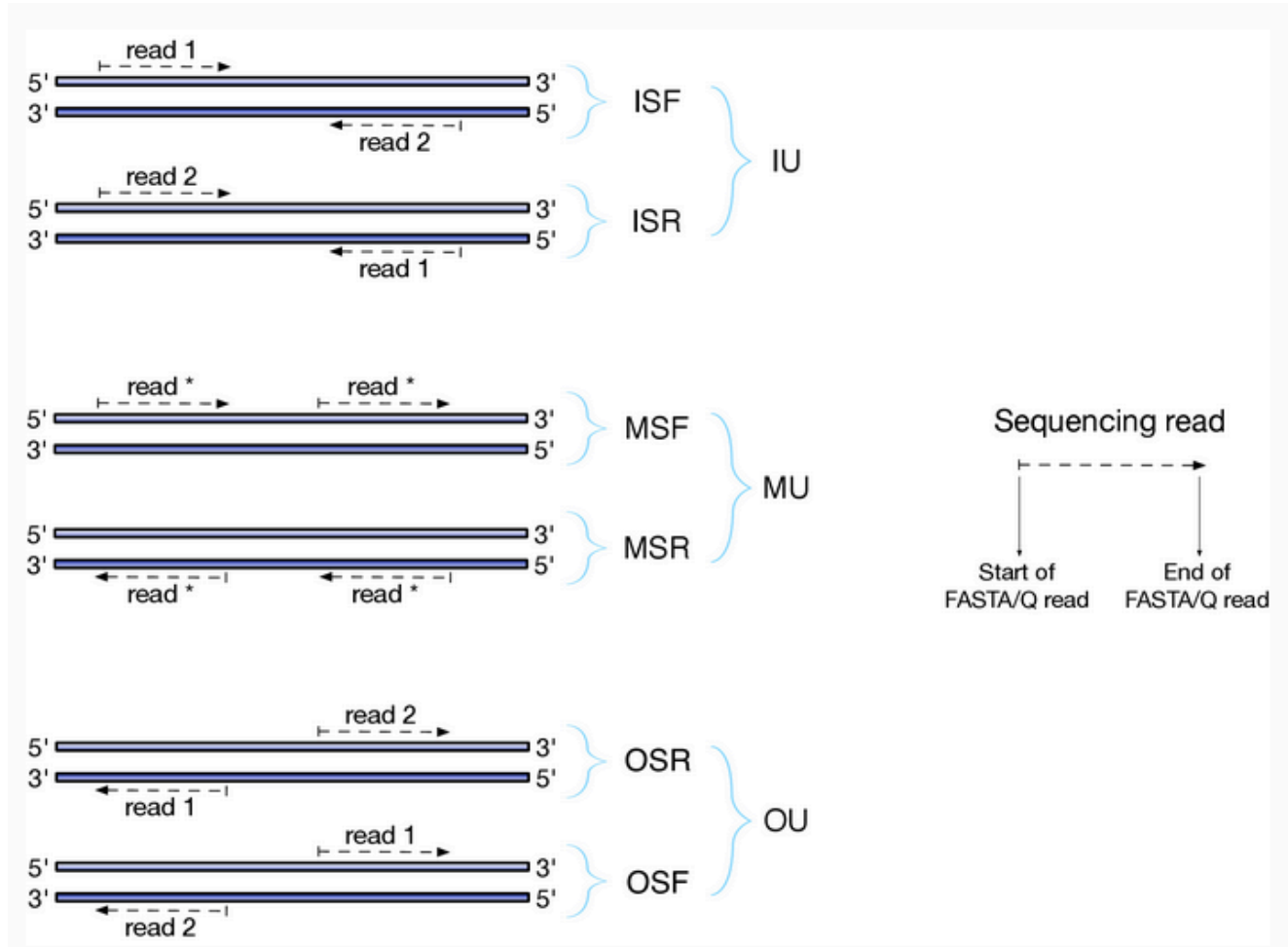
-1, -2 – FASTQ files with paired-end reads

--geneMap + annotations → used to obtain gene level estimations as well (optional)

-l ISR – strand-specificity option

Output files: prefix and localization

Salmon: types of libraries



Output files

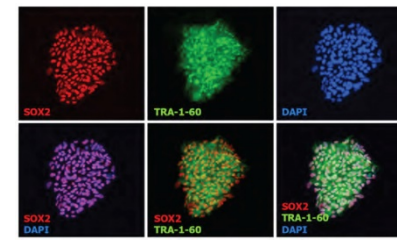
- ✓ **quant.genes.sf** → **genes**
- ✓ **quant.sf** → **transcripts**

Name	Length	EffectiveLength	TPM	NumReads
ENST00000632684.1	12	13	0	0
ENST00000434970.2	9	10	0	0
ENST00000448914.1	13	14	0	0
ENST00000415118.1	8	9	0	0
ENST00000631435.1	12	13	0	0
ENST00000390567.1	20	1	0	0
ENST00000439842.1	11	12	0	0
ENST00000454908.1	17	18	0	0
ENST00000390583.1	31	11.4537	0	0
ENST00000390572.1	28	8.45414	0	0
ENST00000390571.1	31	11.4537	0	0
ENST00000454691.1	18	19	0	0
ENST00000390588.1	20	1	0	0
ENST00000390581.1	23	3.4545	0	0
ENST00000390574.1	21	1.6	0	0
ENST00000450276.1	17	18	0	0
ENST00000431870.1	16	17	0	0
ENST00000414852.1	16	17	0	0
ENST00000390590.1	31	11.4537	0	0
ENST00000390584.1	31	11.4537	0	0
ENST00000452198.1	18	19	0	0
ENST00000634154.1	16	17	0	0
ENST00000631895.1	23	3.4545	0	0
ENST00000633030.1	19	20	0	0
ENST00000632524.1	11	12	0	0
ENST00000633009.1	20	1	0	0
ENST00000634070.1	18	19	0	0
ENST00000390591.1	31	11.4537	0	0

Output files

- **TPM** — This is salmon's estimate of the relative abundance of this transcript in units of Transcripts Per Million (TPM). TPM is the recommended relative abundance measure to use for downstream analysis.
- **NumReads** — This is salmon's estimate of the number of reads mapping to each transcript that was quantified.

Our pipeline



- Transcriptome (FASTA)
- RNA-seq reads (FASTQ)
- Annotations - optional



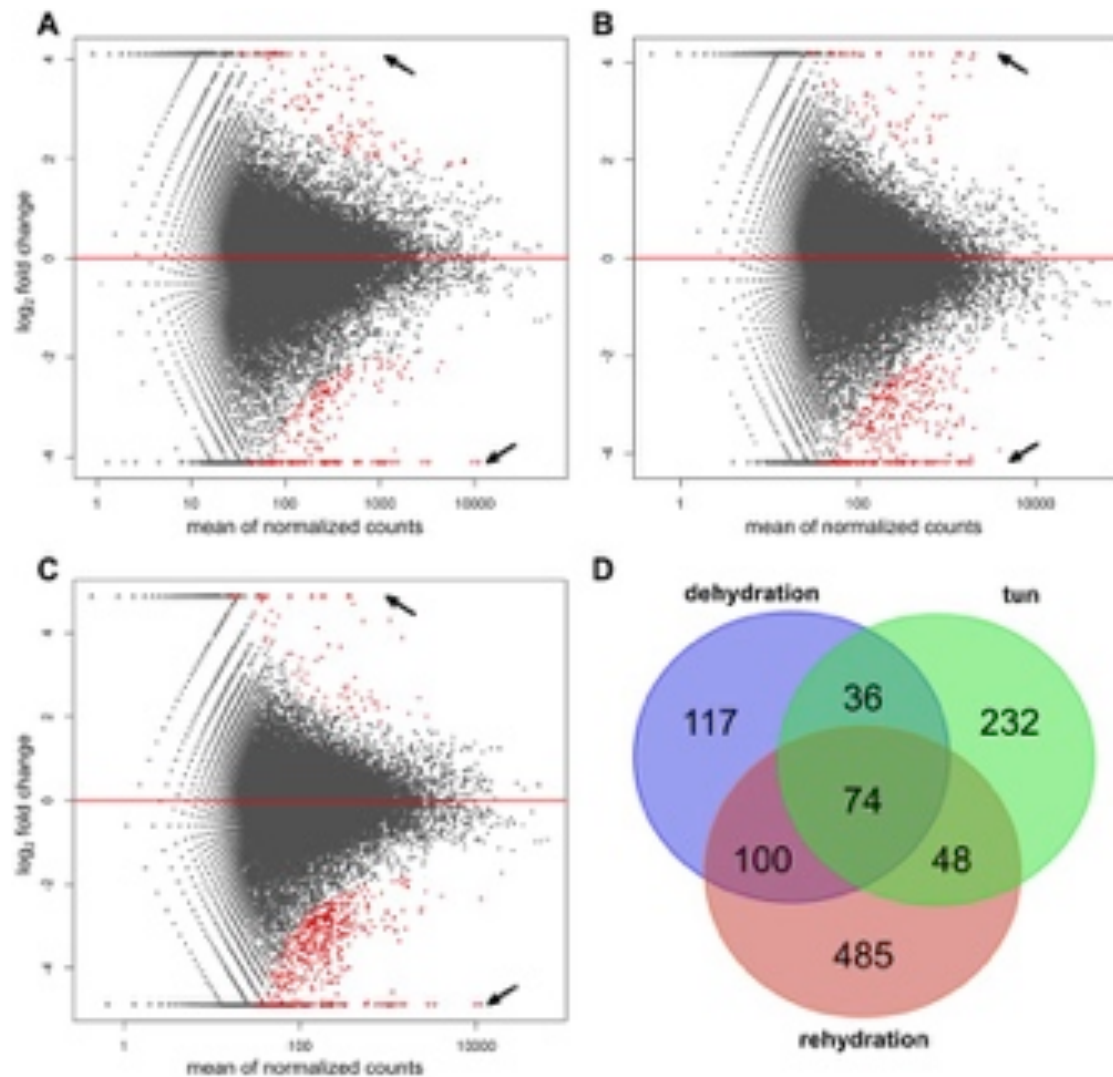
- Transcriptome index



- Expression estimation (**SALMON**)



What next?



Thank you for your attention