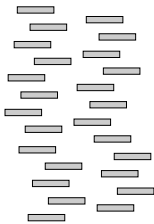


Probabilistic modeling of RNA-Seq data

Roman Cheplyaka

RNA-Seq

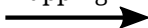
Reads



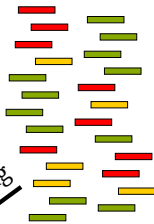
Isoforms



Mapping



Mapped reads

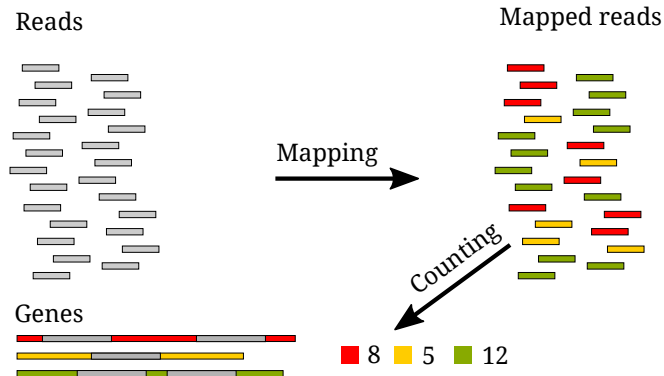


Counting



■ 8 ■ 5 ■ 12

RNA-Seq: mapping to genes



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%

Using multireads

1. Estimate the expression levels

Using multireads

1. Estimate the expression levels
2. Quantify the uncertainty

Using multireads

1. Estimate the expression levels
2. Quantify the uncertainty

$$p(\tau|R)$$

Using the distribution

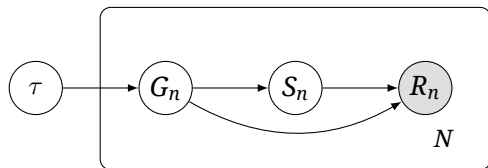
$$p(\tau_i < \tau'_i | R, R') = \iint \mathbb{I}(\tau_i < \tau'_i) p(\tau | R) p(\tau' | R') d\tau d\tau'$$

$$E(\tau_i - \tau'_i | R, R') = \iint (\tau_i - \tau'_i) p(\tau | R) p(\tau' | R') d\tau d\tau'$$

Bayes' theorem

$$p(\tau|R) = p(R|\tau) \cdot \frac{p(\tau)}{p(R)}$$

Generative model



- ▶ τ — isoform expression levels
- ▶ G_n — isoform of read n
- ▶ S_n — start position of read n in the isoform
- ▶ R_n — nucleotide sequence of read n

How to compute the integral?

Grid: $\tau^{(1)}, \tau^{(2)}, \dots, \tau^{(n)} \sim p(\tau|R)$ uniformly distributed

$$\int f(\tau)p(\tau|R)d\tau \approx \sum_{i=1}^n f(\tau^{(i)})p(\tau^{(i)}|R)\Delta\tau$$

1 TPM resolution per isoform, 30 000 isoforms $\Rightarrow$
 $\approx 10^{60\,000}$ points

Better representation

$$\tau^{(1)}, \tau^{(2)}, \dots, \tau^{(n)} \sim p(\tau|R)$$

$$\int f(\tau)p(\tau|R)d\tau \approx \frac{1}{n} \sum_{i=1}^n f(\tau^{(i)})$$

Very easy to compute with!

How to generate samples?

It is a hard problem in general.

Key idea: Markov chain Monte Carlo (MCMC)

Further reading

1. RNA-Seq gene expression estimation with read mapping uncertainty (Li et al., 2009)
2. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome (Li et al., 2011)
3. RNA-Seq data analysis through expectation-maximization (my blog post, <https://ro-che.info/articles/2017-01-29-rsem>)

Conclusions

1. Prefer isoforms over genes
2. Probabilistic modeling accounts for uncertainty
3. MCMC makes probabilistic modeling viable