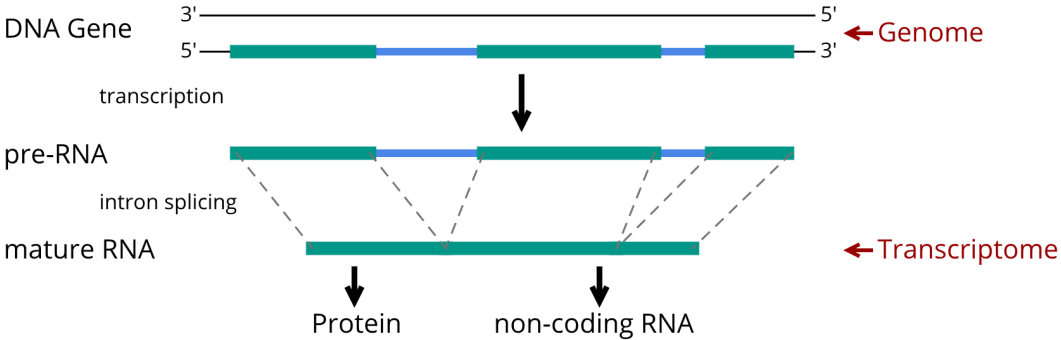
Simulating the Unknown

Thomas Gatter

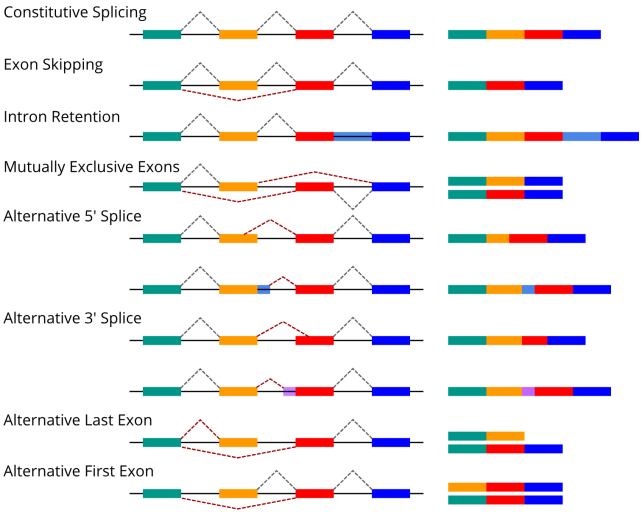
14.02.2024

Bioinformatics, Leipzig University

The Central Dogma



Alternative Splicing



we simulate what we know

$\Leftrightarrow$

what we know is based on tools conforming to simulation

But what do we measure? What is our Reality?

Technical Difficulties... Please Stand by...

Typical **technical biases** of Illumina based RNA-seq protocols:

Step	Influence
PolyA-Selection vs RiboZero	3' bias, dropout regions
Fragmentation	Fragment size distribution, 3'/5' sampling bias
Size Selection	Fragment size distribution, 3'/5' sampling bias
1st/2nd Strand Priming and Synthesis	Hexamer priming bias
Adapter Ligation	Adapter Sequences in Reads
PCR	GC bias, PCR duplicates, copy errors
Flowcell retention	downsampling
Bridge Amplification	pre-/post-phasing, copy errors
species specific influences	on most of the above
sample handling	fragment size distribution, copy errors
Sequence-By-Synthesis	read errors, quality scores

Technical Difficulties... Please Stand by...

Typical **post-processing** steps with **bias**:

Step	Influence
Quality Trimming/Adapter Trimming	over-trimming
Assembly	missed genomic duplication, ??
Mapping/(semi) Alignment	missed genomic duplications, false splice-site predictions, ???
Quantification	??

Quantification of single samples

How can we quantify the likelihood that a read originates from an isoform?

Based on:

- *de novo* assembly
- reference based annotation
- reference annotation

Complications:

- correction of technical biases
- alignment bias
- incomplete/wrong isoform base
- overlapping features

Distributions...

How do we treat isoform/exon abundances between replicates?

$\Rightarrow$ negative binomial distribution is the “gold standard”

But some still use Poisson distribution

But only for “normal” bulk RNA-seq...

Distributions...

An excerpt of current single cell models:

Tool	Year of Publication	Modeled Distribution
ESCO [1]	2020	Gamma-Poisson
hierarchicell [2]	2021	negative binomial
muscat [3]	2020	negative binomial
POWSC [4]	2020	zero-inflated, log-normal Poisson mixture
scDD [5]	2016	Bayesian negative binomial mixture
scDesign2 [6]	2021	negative binomial
SCRIP [7]	2022	Gamma-Poisson
SPARSim [8]	2020	Gamma-multivariate hypergeometric
splatter [9]	2017	Gamma-Poisson
SPsimSeq [10]	2020	log-linear + Gaussian copula
SymSim [11]	2019	Markow-Chain-Monte-Carlo
ZINB-WaVE [12]	2018	zero-inflated negative binomial

Distributions...

Tools often only validate their approach in a circular fashion...

Common choices are:

- simulate data with own expected distribution
- simulate data with own expected distribution and custom error model
- use pre-existing simulators and their distribution models and error models

Is there transcriptional noise?

Typically sized RNA-seq experiments miss out significant portions of low abundant spliceforms ([13], [14]).

This could mostly be noise ([15], [16]):

- un-mature RNA
- spliceosome failure

Is there transcriptional noise?

Can’t we just look at abundance distributions of the measurements?

No! There is systematic and unsystematic noise:

- spliceosome splice order is not random
- observed maladaptation of the spliceosome

Can’t we disregard low abundant spliceforms?

Maybe?

- rare isoforms have been associated as a key factor in diseases including cancer
- other studies suggest little correlation of sequencing depth to drawn biological conclusions [17]

In the beginning, there was an annotated reference...
And somehow measured feature abundances...

Option A:

- simulate reference as is
- provide tools with partial reference
- systematic reduction for isoform classes possible
  - ▷ assumption that there is no noise
  - ▷ assumption that reference is representative

Three Philosophies...

Option B:

- use exon chain of “dominant” isoform/consensus exons
- generate genes to presumed feature distribution
- provide tools with partial reference

▷ assumption that there is no noise

▷ partially artificial gene structures

▷ assumption that feature distribution is representative

Three Philosophies...

Option C:

- use exon chain of “dominant” isoform/consensus exons
- generate genes to presumed feature distribution
- add additional noisy transcripts
- provide tools with partial reference of true genes
  - ▷ partially artificial gene structures
  - ▷ assumption that feature distribution is representative

B or C may also use artificial distributions.

Fold change and abundance conundrum

For **either option** we need to create:

- (realistic) isoforms abundances (with replicates)
- (realistic) fold changes
- reads (with varying technical biases)
- (realistic) size differences between repeats/samples

We may define **abundances** and **fold changes** as follows:

- fully mimic a real dataset
- feature estimation to simulate and generate real-like distribution
- fully artificial mixtures

Modeling technical biases further influences simulated counts.

Annotation conundrum

Most tools will blindly follow a given reference!

We need benchmarks providing:

- full reference
- full reference + noisy transcripts
- partial references
- partial references + (related) noisy transcripts

The reference is relevant for mapping, assembly and quantification.

Alignment bias

Simulating counts is not enough.

Simulate **read sequences**:

- with technical bias
- without technical bias
- without technical bias and perfect alignment

Assembler try to maximize conformity to existing annotation and thereon based simulation.

All competitive general purpose assembler **deliberately avoid calling**:

- alternative start/end sites
- intron retention
- overlapping “shadow” genes on the opposite strand
- isoforms within introns
- low abundant isoforms in high abundance genes

Common observation: rare splice forms appear in few samples at low abundance $\rightarrow$ multi sample assembler like Taco [18], PsiCLASS [19] or `Ryūtō`[20] remove isoforms based on this property

Simulated choices have real life consequences.

But does it actually matter for your application?

(Maybe use lab testing wherever you can.)

References

[1] J. Tian, J. Wang, and K. Roeder, "Esco: Single cell expression simulation incorporating gene co-expression," *Bioinformatics*, vol. 37, no. 16, pp. 2374–2381, 2021.

[2] K. D. Zimmerman and C. D. Langefeld, "Hierarchicell: An r-package for estimating power for tests of differential expression with single-cell data," *BMC genomics*, vol. 22, no. 1, pp. 1–8, 2021.

[3] H. L. Crowell, C. Soneson, P.-L. Germain, *et al.*, "Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data," *Nature communications*, vol. 11, no. 1, p. 6077, 2020.

[4] K. Su, Z. Wu, and H. Wu, "Simulation, power evaluation and sample size recommendation for single-cell rna-seq," *Bioinformatics*, vol. 36, no. 19, pp. 4860–4868, 2020.

[5] K. D. Korthauer, L.-F. Chu, M. A. Newton, *et al.*, "A statistical approach for identifying differential distributions in single-cell rna-seq experiments," *Genome biology*, vol. 17, pp. 1–15, 2016.

[6] T. Sun, D. Song, W. V. Li, and J. J. Li, "Scdesign2: A transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured," *Genome biology*, vol. 22, no. 1, p. 163, 2021.

[7] F. Qin, X. Luo, F. Xiao, and G. Cai, "Scrip: An accurate simulator for single-cell rna sequencing data," *Bioinformatics*, vol. 38, no. 5, pp. 1304–1311, 2022.

[8] G. Baruzzo, I. Patuzzi, and B. Di Camillo, "Sparsim single cell: A count data simulator for scrna-seq data," *Bioinformatics*, vol. 36, no. 5, pp. 1468–1475, 2020.

Citations ii

[9] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: Simulation of single-cell rna sequencing data," *Genome biology*, vol. 18, no. 1, p. 174, 2017.

[10] A. T. Assefa, J. Vandesompele, and O. Thas, "Spsimseq: Semi-parametric simulation of bulk and single-cell rna-sequencing data," *Bioinformatics*, vol. 36, no. 10, pp. 3276–3278, 2020.

[11] X. Zhang, C. Xu, and N. Yosef, "Simulating multiple faceted variability in single cell rna sequencing," *Nature communications*, vol. 10, no. 1, p. 2611, 2019.

[12] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert, "A general and flexible method for signal extraction from single-cell rna-seq data," *Nature communications*, vol. 9, no. 1, p. 284, 2018.

[13] R. Sen, G. Doose, and P. F. Stadler, "Rare splice variants in long non-coding RNAs," *Non-Coding RNA*, vol. 3, no. 3, p. 23, 2017.

[14] A. Nellore, A. E. Jaffe, J.-P. Fortin, *et al.*, "Human splicing diversity and the extent of unannotated splice junctions across human rna-seq samples on the sequence read archive," *Genome biology*, vol. 17, no. 1, pp. 1–14, 2016.

[15] H. Van Bakel, C. Nislow, B. J. Blencowe, and T. R. Hughes, "Most "dark matter" transcripts are associated with known genes," *PLoS Biol*, vol. 8, no. 5, e1000371, 2010.

[16] B. Saudemont, A. Popa, J. L. Parmley, *et al.*, "The fitness cost of mis-splicing is the main determinant of alternative splicing patterns," *Genome biology*, vol. 18, no. 1, pp. 1–15, 2017.

[17] A. Conesa, P. Madrigal, S. Tarazona, *et al.*, "A survey of best practices for rna-seq data analysis," *Genome biology*, vol. 17, no. 1, pp. 1–19, 2016.

[18] Y. S. Niknafs, B. Pandian, H. K. Iyer, A. M. Chinnaiyan, and M. K. Iyer, "Taco produces robust multisample transcriptome assemblies from rna-seq," *Nature methods*, vol. 14, no. 1, pp. 68–70, 2017.

[19] L. Song, S. Sabunciyan, G. Yang, and L. Florea, "A multi-sample approach increases the accuracy of transcript assembly," *Nature communications*, vol. 10, no. 1, pp. 1–7, 2019.

[20] T. Gatter and P. F. Stadler, "Ryūtō: Improved multi-sample transcript assembly for differential transcript expression analysis and more," *Bioinformatics*, 2021.