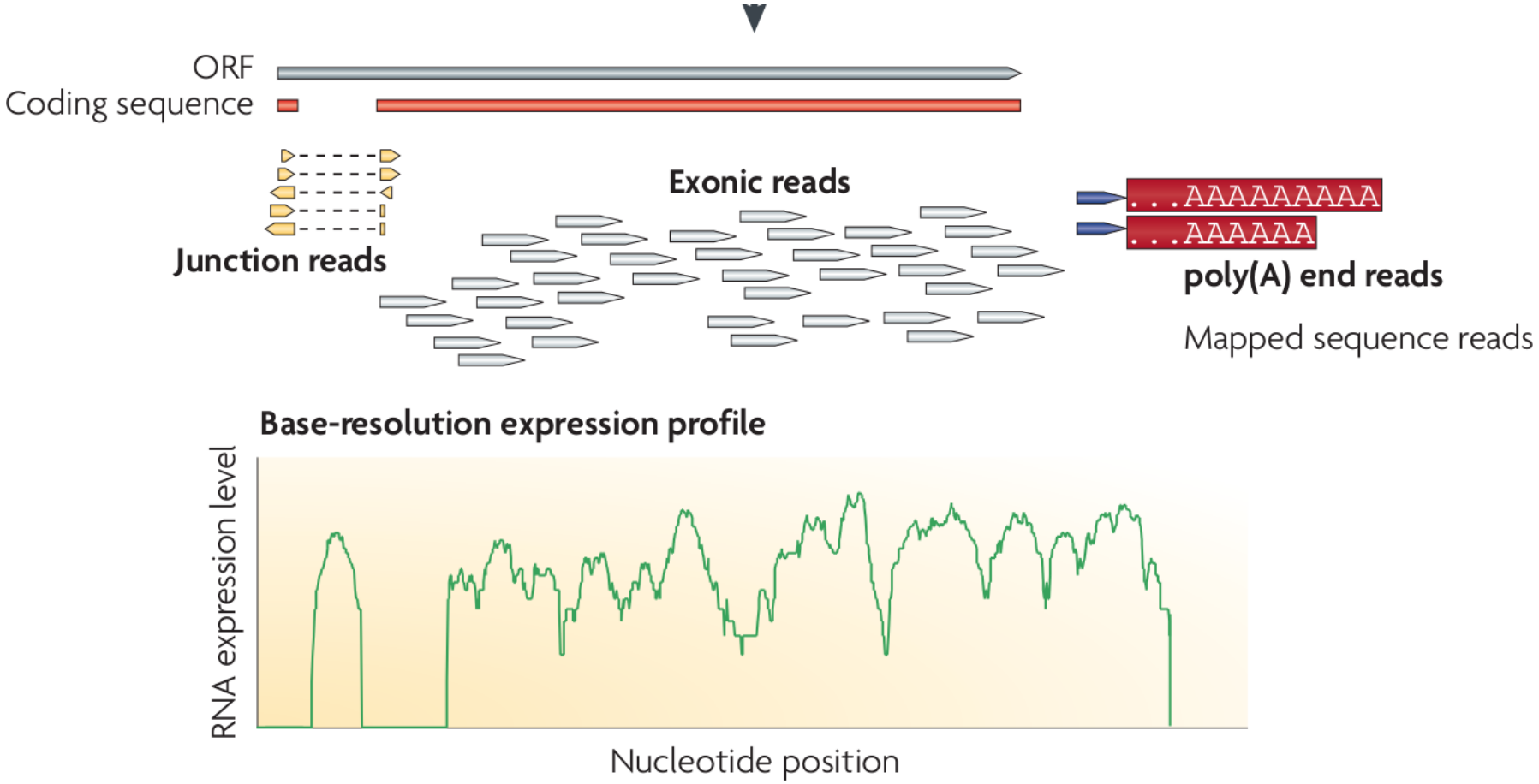
Learning about RNA-seq data through realistic simulations

Botond Sipos

Goldman group, EMBL-EBI
http://sbotond.github.io
@4a6a5a

October 24, 2013

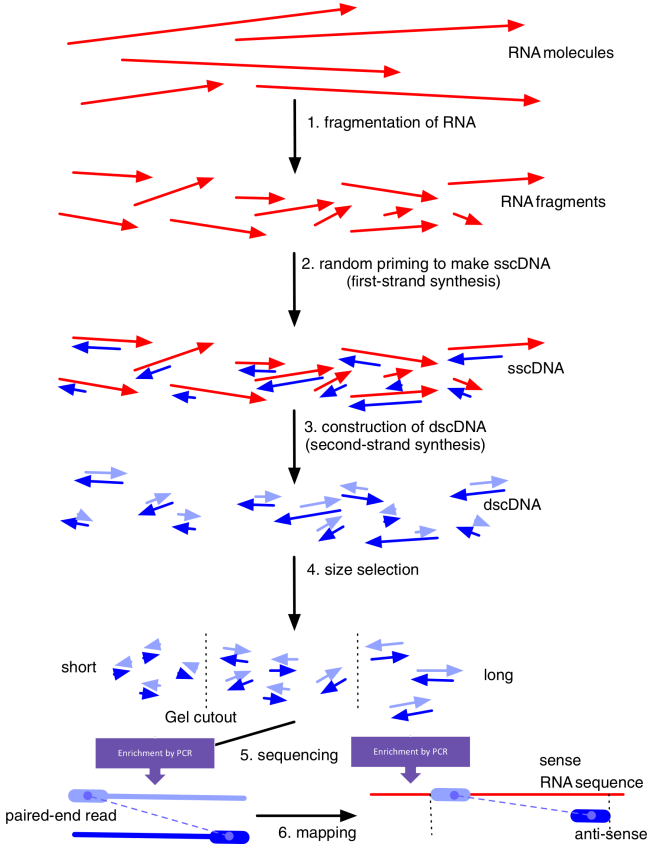
RNA-seq: the Swiss Army knife of transcriptomics



(Wang et al.; Nat Rev Gen 2009)

RNA-seq is not perfect

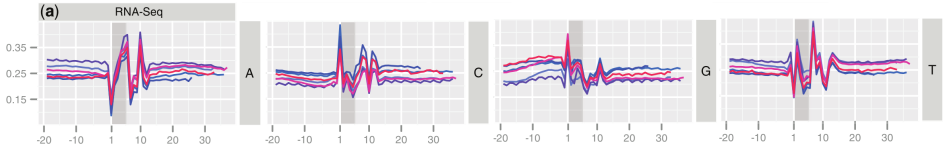
“Standard” protocol



(Roberts et al. 2011; Genome Biol)

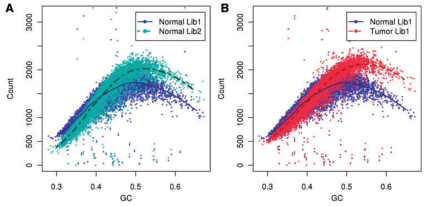
Biased coverage

- Hexamer priming → fragment start/end bias:



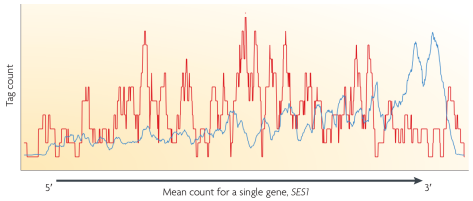
(Hansen et al. 2009; NAR)

- PCR → fragment GC bias:



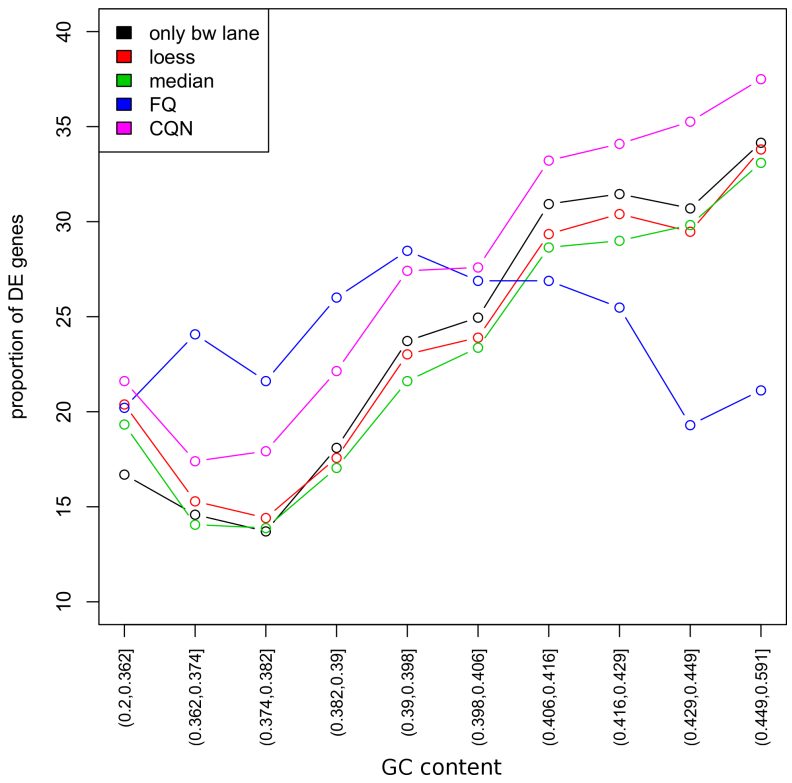
(Benjamini and Speed 2012; NAR)

- Interacting biases → uneven coverage:



Should we care about the biases?

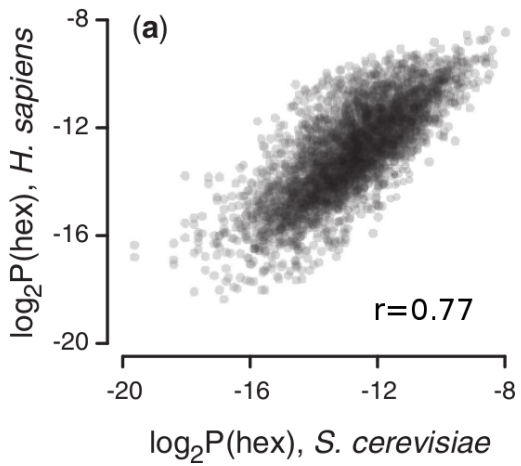
GC biases and differential expression



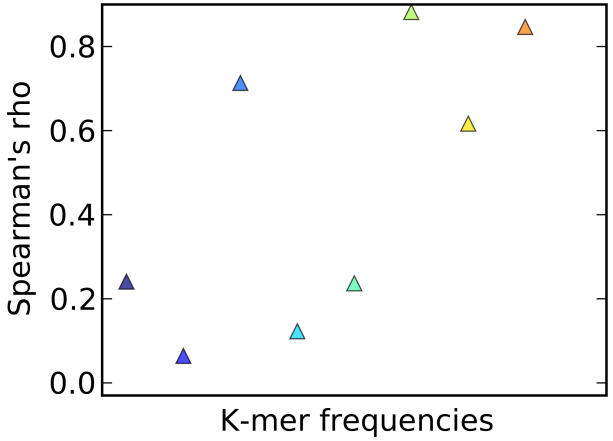
(Risso et al. 2011; BMC Bioinf)

See also Hansen and Irizarry (2012; Biostatistics).

Reproducible priming biases?



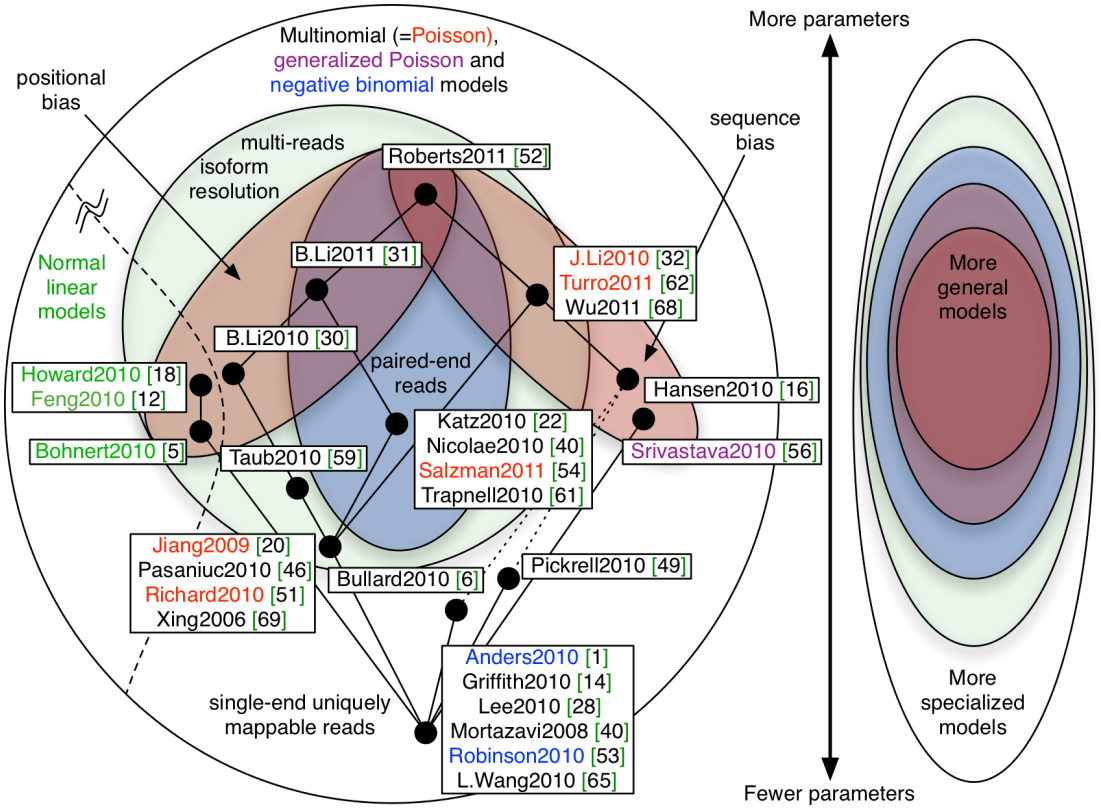
(Hansen et al. 2010; NAR)



Our survey: between and within species comparisons

A cornucopia of tools?

Inference methods are booming



(Pachter; arXiv 2011)

Notable simulators

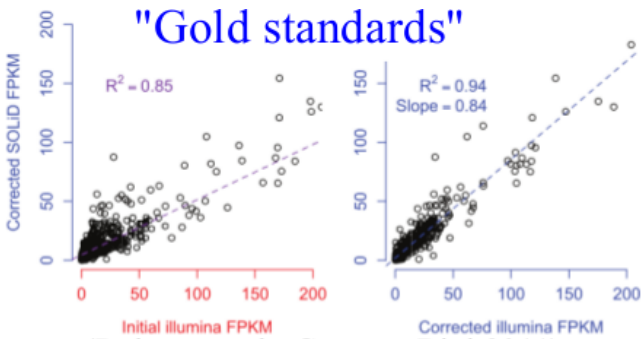
FluxSimulator - great inspiration, lacks inference tools

Why do we need (realistic) simulations anyway?

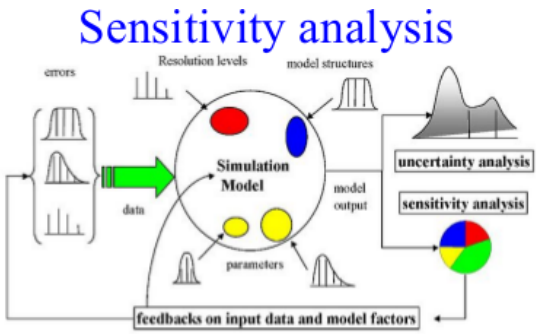
- Simulation is an essential tool in the modelling toolbox.
- Testing software correctness, debugging:

```
rmultinom(n=1,size=nr_reads,prob=expr_levels) ...
```

- Assessing robustness of quantification and correction methods:



(Roberts et al.; Genome Biol 2011)



(Wikipedia)

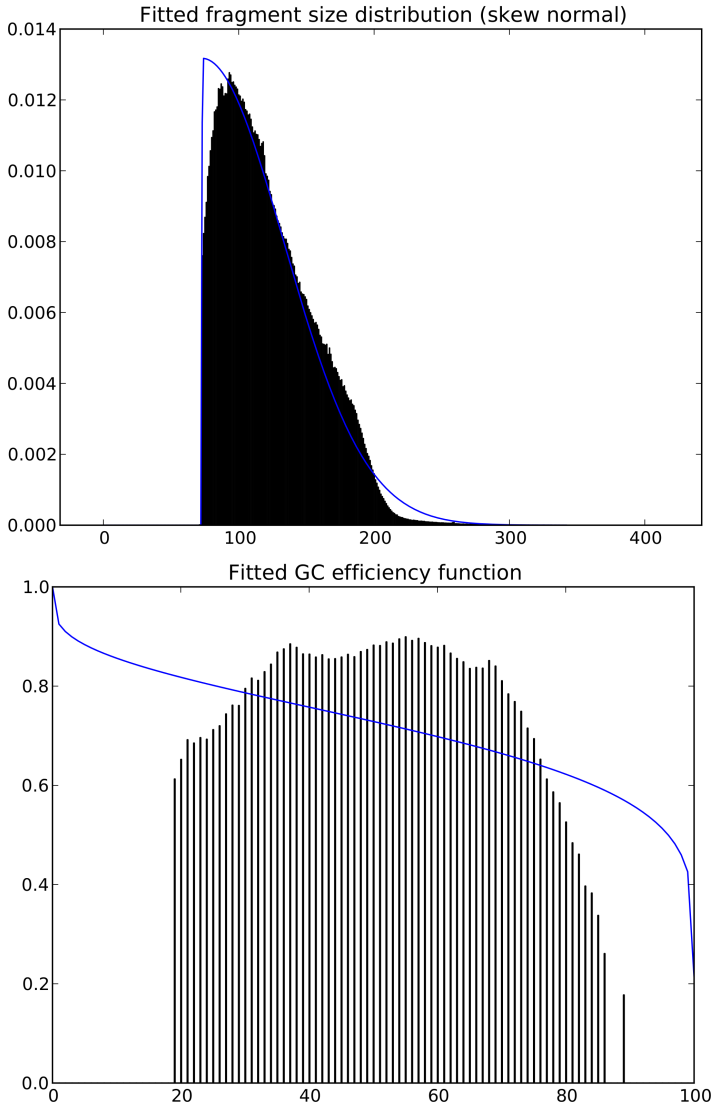
- Realistic simulations provide both control and a true baseline.
- "Virtual laboratory" for *in silico* exploration of new protocols.
- Learning about the data generating process by "synthesis".

The `rlsim` package: estimation

The `effest` tool

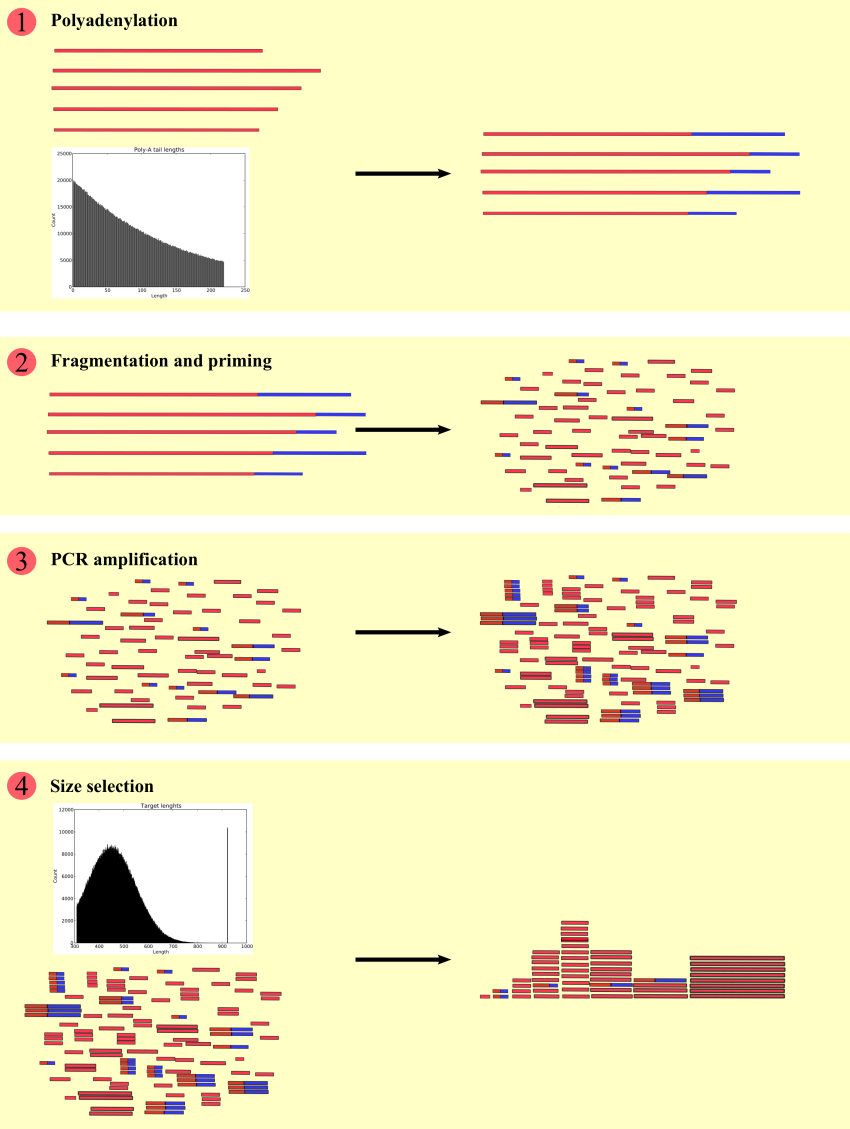
Estimates parameters from reads mapped to canonical transcriptome:

- Insert size distribution
- (Relative) GC-dependent amplification efficiencies
- Corrected/uncorrected expression levels



Git repository & homepage: http://bit.ly/rlsim-git

The `rlsim` package: simulation

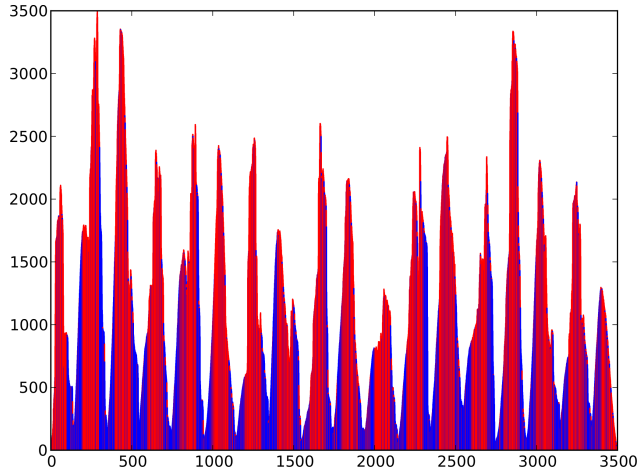


The `rlsim` tool

Simulate based on:

- Parameters estimated by `effest`
- Priming affinities based on oligo binding energies
- See documentation at http://bit.ly/rlsim-doc for other simulated factors.

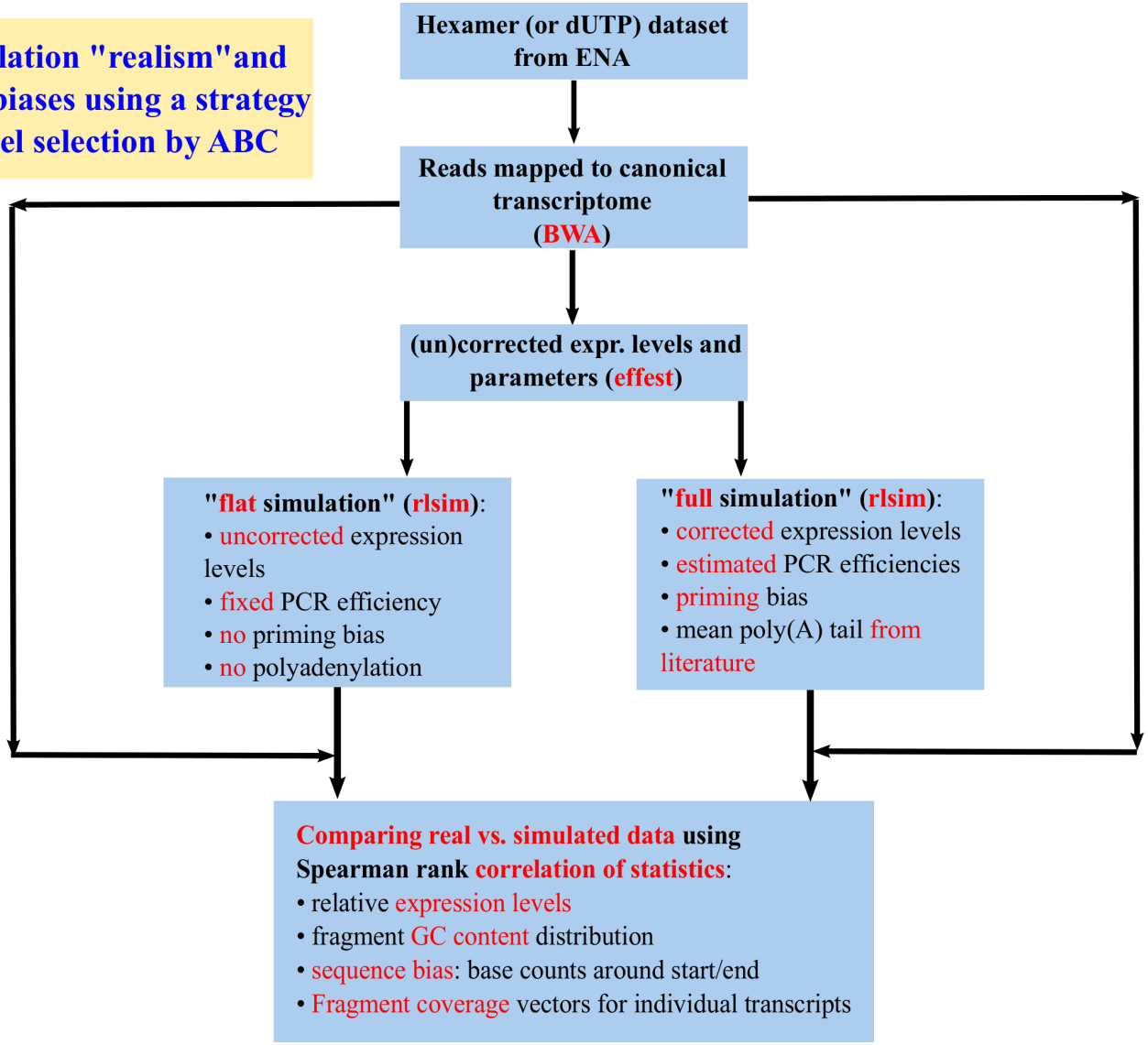
Complex simulated coverage



Git repository & homepage: http://bit.ly/rlsim-git

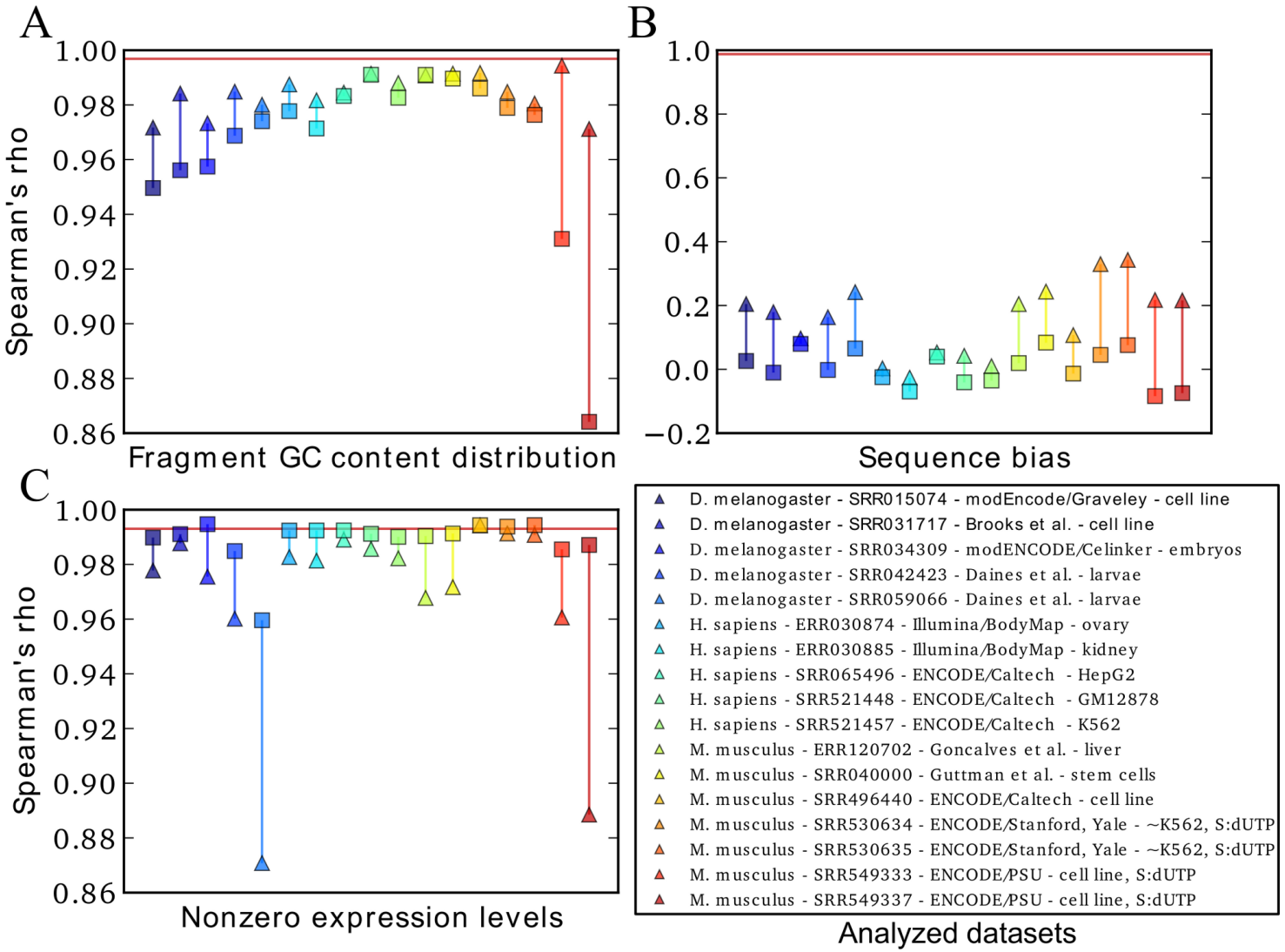
Learning about data through simulations

Quantify simulation "realism"and magnitude of biases using a strategy similar to model selection by ABC



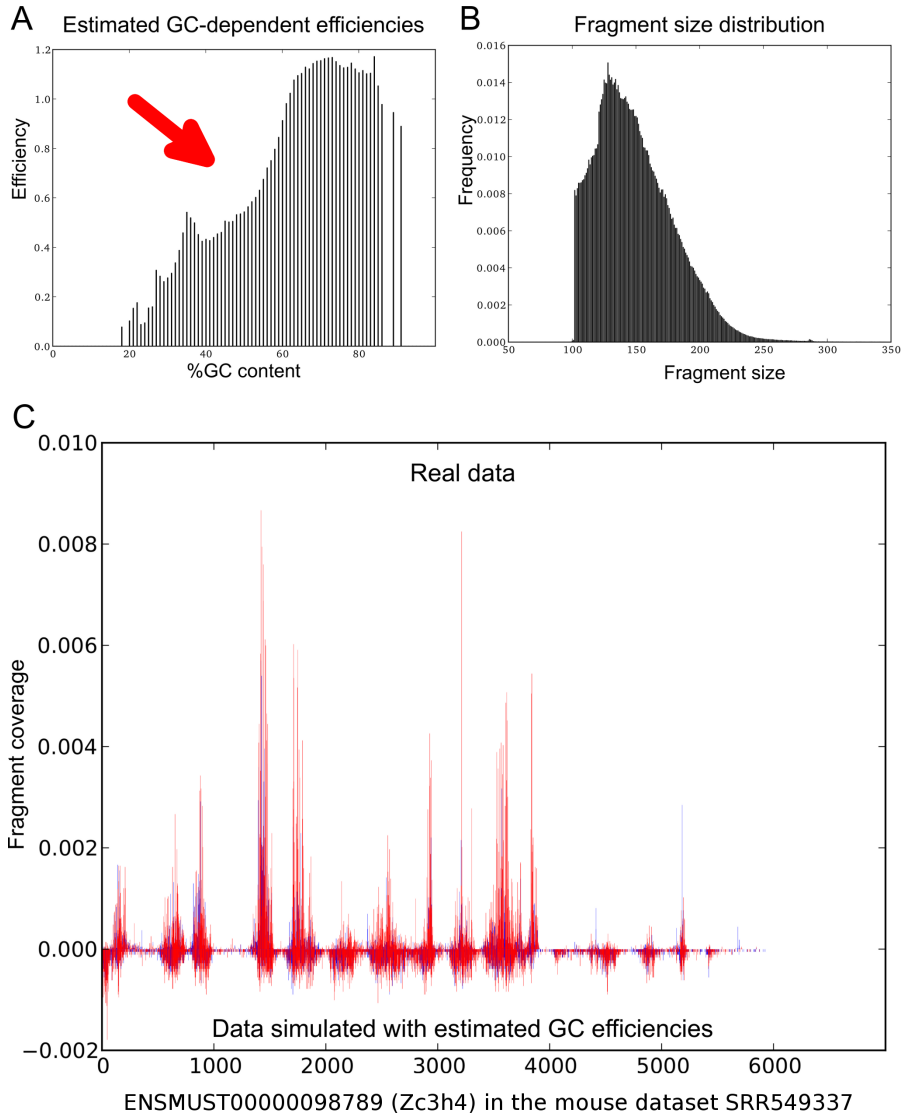
Runnable pipeline with results & documentation: http://bit.ly/rlsim-pl

RNA-seq biases are highly sample-specific



Botond Sipos, Greg Slodkowicz, Tim Massingham, Nick Goldman (2013) *Realistic simulations reveal extensive sample-specificity of RNA-seq biases* **arXiv** http://bit.ly/rlsax

Biases reproduced on the coverage level



Botond Sipos, Greg Slodkowicz, Tim Massingham, Nick Goldman (2013) *Realistic simulations reveal extensive sample-specificity of RNA-seq biases* **arXiv** http://bit.ly/rlsax

Conclusions and conundrums

Conclusions

- Considering the practical importance of Illumina RNA-seq, the development and assessment of analysis methods will remain relevant.
- RNA-seq biases are highly sample-specific, this has to be dealt with in both estimation and simulation settings.
- Parameterise from real data to get both control & realism: parameter files available at http://bit.ly/rlsim-pa

Conundrums

- Do you find simulations useful? Can we rely on cross-platform comparisons only?
- We have an ever growing vast “corpus” of RNA-seq data. Does this make simulations unnecessary?
- Do you expect future technologies to be less biased?