Construction of a Cellular Mouse Small RNA Atlas

James Diao, Joel Rozowsky, Rob Kitchen, Mark Gerstein

Computational Biology & Bioinformatics Program | Bass Center for Molecular & Structural Biology | Yale University | New Haven, CT

Abstract

Tissue convolution is when different tissues of origin bias a combined RNA-seq signature. Deconvoluting heterogeneous samples is commonly done by the R package DeconRNASeq, which calculates a breakdown of tissue proportions from RNA-Seq inputs, given the tissue RNA signatures. This parameter—the signature matrix (S), or cellular atlas—is important for further study of mouse models, which comprise a significant fraction of sequence read data available to researchers, and uniquely capture embryonic and fetal development. My project develops this atlas from public sequence data, identifies its driving RNAs, and determines its predictive power to be between 70-87%.

Background

Small RNAs

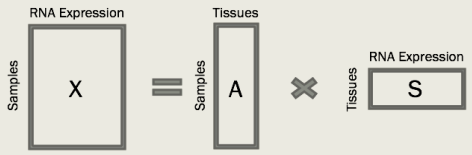
- These are: highly conserved, non-protein coding RNAs (<200 nt)
- Include: miRNAs, siRNAs, piRNAs, tRNAs
- Function: regulation of gene expression, cell-cell signaling

Extracellular/Circulating RNA

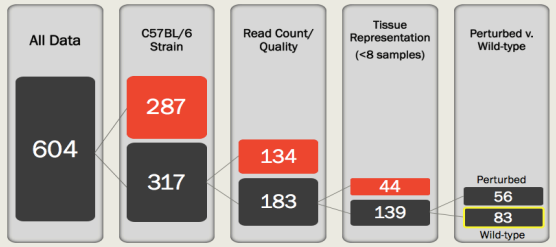
- Clinical applications: biomarkers, diagnostic tools
- Main point of interest: deconvolution problem
  Many sources: heterogeneous samples

DeconRNASeq

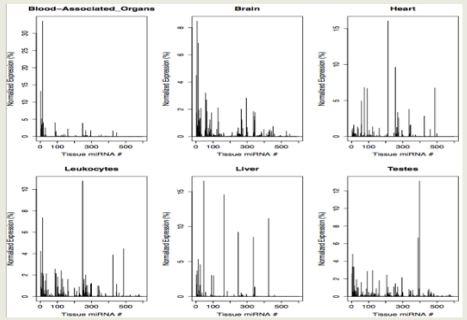
- R package: deconvolves heterogeneous RNA-Seq data
- Based on a linear model: $X = AS$, or $x_{jk} = \sum_{i=1}^{N} a_{ki} s_{ij}$
  - *X – observed expressions (samples v. RNAs)*
  - *A – proportions (samples v. tissues)*
  - S – *tissue signatures (tissues v. RNAs)*



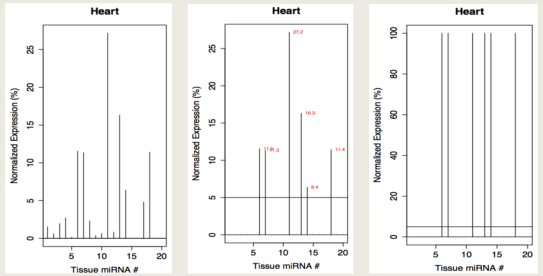
Inclusion and Exclusion of Sample Data



miRNA Signature Matrix Plot by Tissue



Top-20-Variance, Reduced, Binary



Condensed S: miRNA Table

	miRNA	Tissue	Expression
Unique:	mmu-miR-1a-3p	Heart	100%
*Expression > 99%	mmu-miR-133a-3p	Heart	100%
	mmu-miR-124-3p	Brain	100%
	mmu-miR-3471	Testes	100%
	mmu-miR-199a/b-3p	Testes	100%
Distinctive:	mmu-miR-34c-5p	Testes	96%
*Expression > 90%	mmu-miR-192-5p	Liver	94%
	mmu-miR-122-5p	Liver	93%
	mmu-miR-9-5p	Brain	92%
Majority:	mmu-miR-22-3p	Heart	70%
*Expression > 50%	mmu-miR-21a-5p	Leukocytes	54%
	mmu-let-7f-5p	Blood-Associated Organs	52%
	mmu-miR-486a/b-5p	Heart, (Leukocytes)	52% (48%)
Other:	mmu-miR-378a-3p	Liver, Leukocytes	42%, 30%
*Expression >25%	mmu-miR-30a-5p	Liver	42%
	mmu-let-7a-5p	Blood-Associated Organs	40%
	mmu-let-7c-5p	Brain	40%
	mmu-miR-16-5p	Testes, Leukocytes	40%, 37%
	mmu-let-7b-5p	Testes	37%
	mmu-miR-143-3p	Heart	37%

Tissue Identification Model

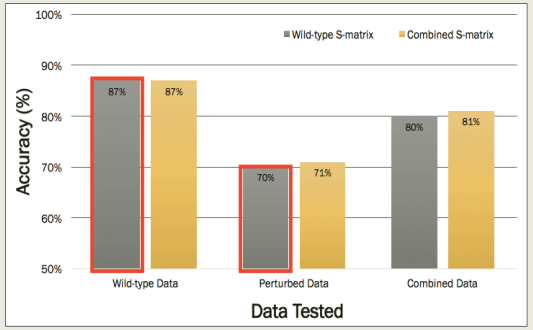
- The output of DeconRNASeq is the A matrix: shown below
- This assigns each sample a tissue breakdown based on how its miRNAs match the signature matrix S.
- If one tissue accounts for at least 50% more than any other tissue, it is considered the dominant tissue.

A Matrix: Tissue Proportions

miRNA Identification Accuracy

	Predictions	Accuracy	Total Correct
Full S	92%	95%	87%
Top 20 miRNAs S	95%	91%	87%
Binary S	81%	88%	71%

ID Accuracy of S-matrices with Different Inputs



Follow-up

- More data: independent normal samples and greater tissue coverage
- Sub-signatures in tissues

Acknowledgements

- Joel Rozowsky: Mentor
- Rob Kitchen: Advice, Pipeline, Debugging
- Mark Gerstein: Principal Investigator