normjam

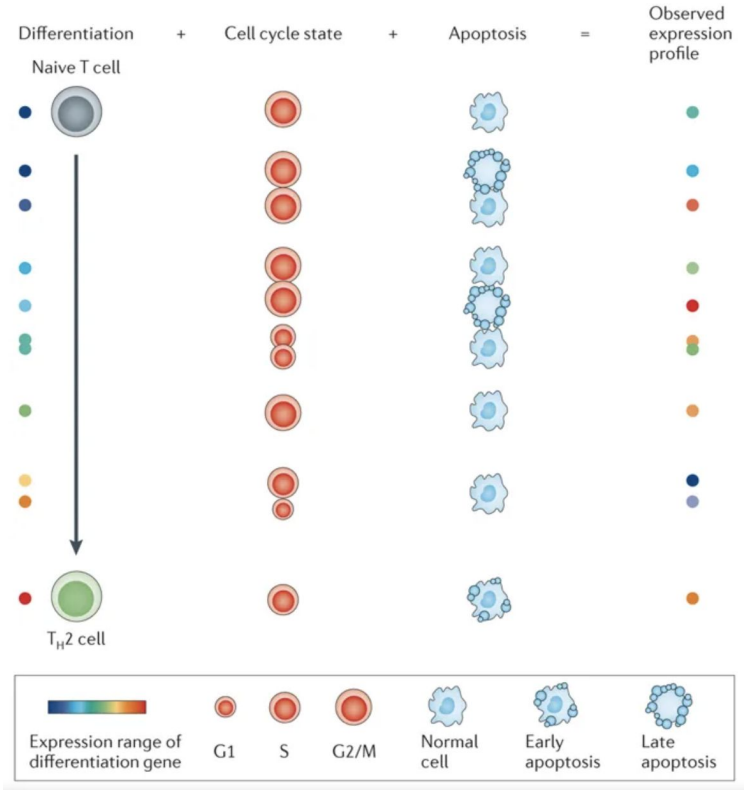
What are potential biological and negative controls we can use to help or assess normalization?

Sources of variability

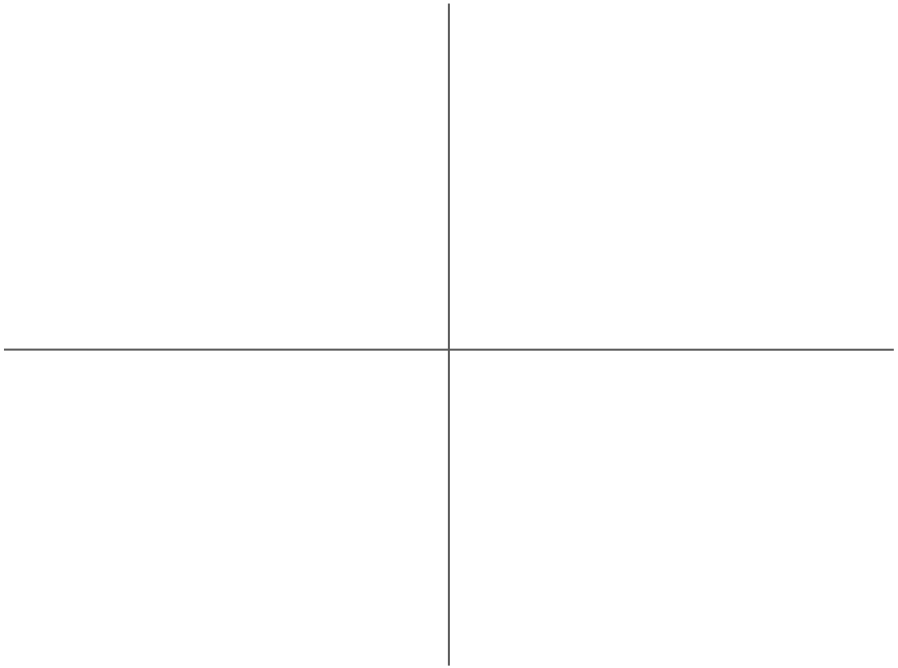
(mostly) Technical

	Cell-specific effects	Gene-specific effects	Not removed by UMIs
Sequencing depth	✓		✓
Amplification	✓	✓	
Capture and RT efficiency	✓	✓	✓
Gene length		✓	
GC content	✓	✓	✓
mRNA content	✓		✓

(mostly) Biological



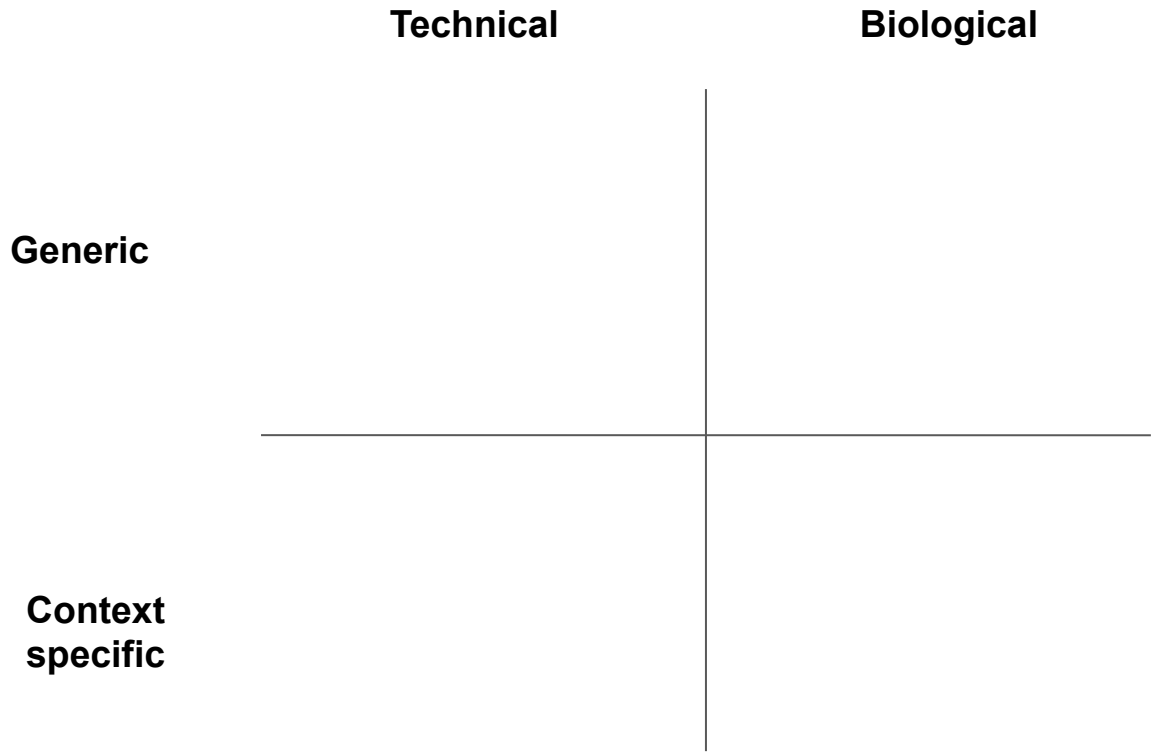
Types of controls



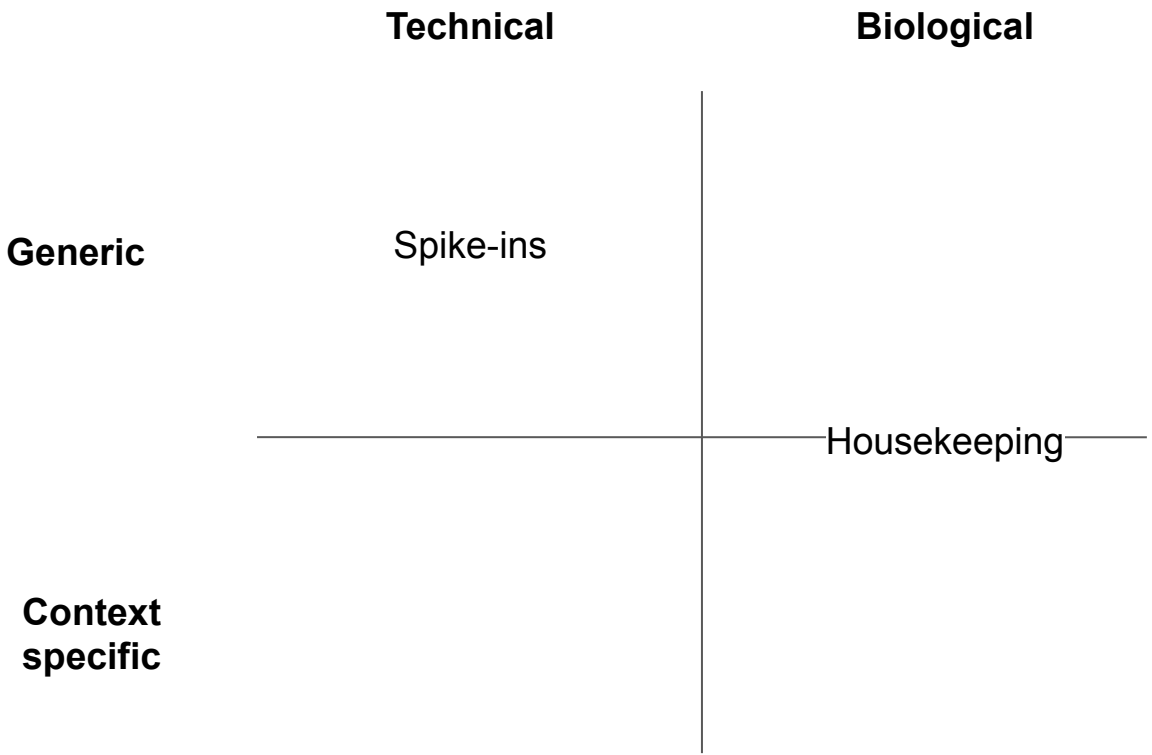
Possible quadrants

1. Biological controls
2. Technical controls
3. Within samples
4. Across samples
5. Supervised vs unsupervised

Types of controls

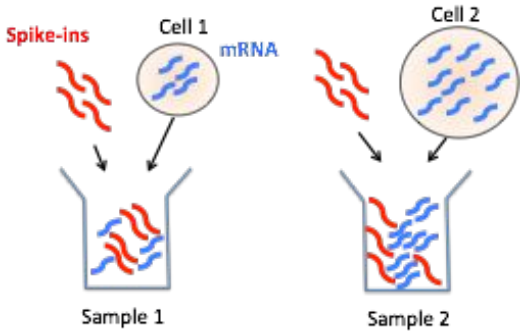


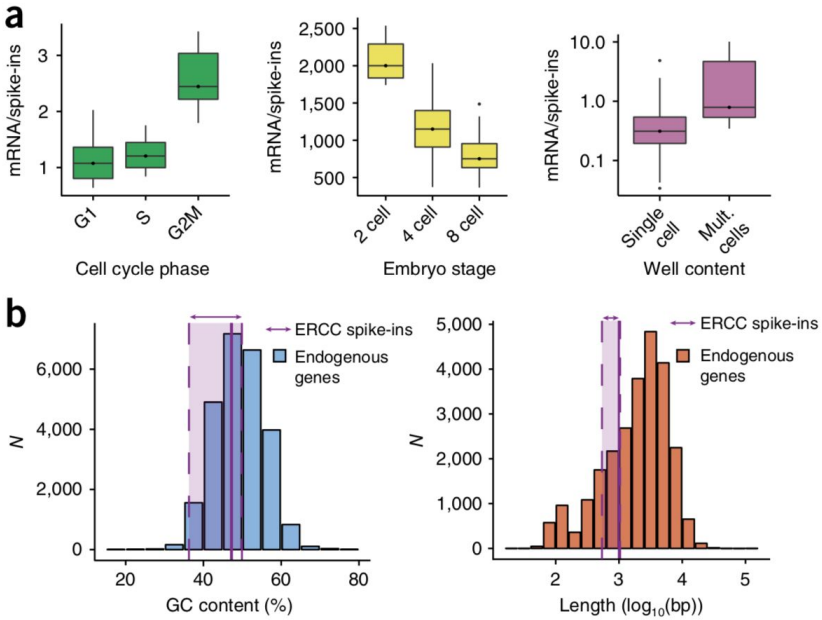
Types of controls



Spike-ins

Extrinsic molecules added in a theoretically known fixed quantity





Housekeeping genes

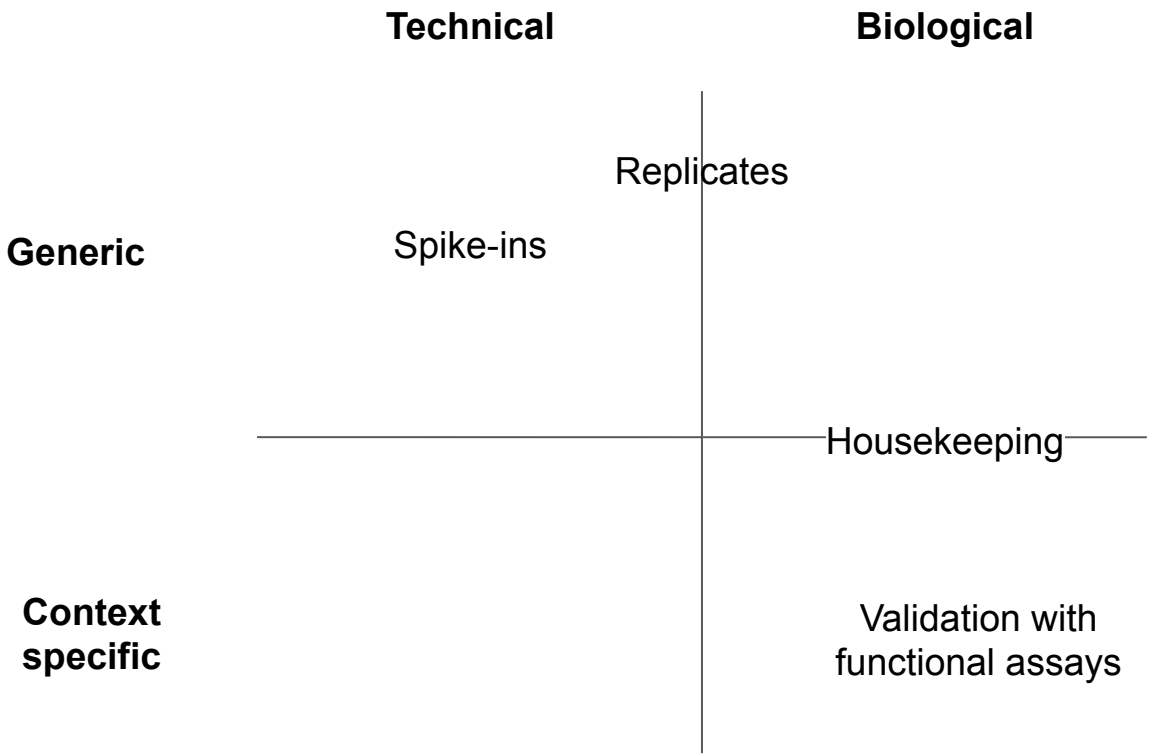
- Are there housekeeping genes at the single cell level? Does the name make sense?

- Cell-to-cell variability: individual cells are never synchronized and variability will never average-out at the single cell level

- What about finding low-variability genes?

Lin et al, Evaluating stably expressed genes in single cells, Gigascience 2019
Wang et al, Defining housekeeping genes suitable for RNA-seq analysis of the human allograft kidney biopsy tissue, BMC Medical Genomics 2019

Types of controls



Orthogonal techniques

What type of orthogonality? Different molecules, eg. transcriptome or proteome? Different way of measuring the same molecules?

Can it help normalization?

OR should we used it “just” for biological validation of the quality of the normalization?

	Validation method	Pros	Cons	Sample throughput	Number of parameters
GENOMIC	RNA flow	Fast; easily ID co-expression	Limited probes and parameters	High	Low (1–3)
	In situ hybridization (ISH)	ID novel genes in spatial context	1 probe, limited co-expression	Low	Low (1–2)
	Fluorescent in situ hybridization (FISH)	ID co-expression in spatial context	Limited probes and parameters	Low	Low (1–4)
	Target manipulation (CRISPR, siRNA, overexpression)	Causal information, functional information	Off-target effects, probe design, time-consuming	Low	Low to variable
PROTEIN	Immunohistochemistry (IHC)	ID novel target in spatial context	1 probe, limited co-expression	Low	Low (1)
	Immunofluorescence imaging	ID co-expression; can be in spatial context	Antibodies, limited parameters	Low to medium	Low (1–4)
	Flow cytometry	Fast; easily ID co-expression	Antibodies, limited parameters	High	Medium (4–10)
	Helios™	High-parameter; validates dozens of targets at the single-cell level	Antibody availability	High	High (10–50)
	Western blot	ID co-expression	Limited parameters, bulk only	Low	Low (1–3)
	Mass spectrometry	Unbiased validation of dozens of targets	Bulk only	Low	High (thousands)
OTHER	Biological assay	Definitive, functional verification of hypothesis	Difficult, non-standardized; may require development	Variable	Variable

Questions

1. Value of controls that are currently used
    a. e.g how much do we gain vs lose with spike-in reads?
2. How to include a priori knowledge references?
    a. Do we use them during normalization? Or post-hoc?
    b. Do we use a supervised method when the populations are known, e.g. access to cell type “pure” signatures
3. How to evaluate normalization with or without controls?
    a. Control experiments for method validation → case for benchmarking datasets?
    b. What about overfitting when benchmarking datasets are public?
    c. Do we need to held out datasets?

Tomorrowland

1. Superset of requirements for a benchmarking dataset
    a. Reproducible benchmarking frameworks

2. (draft of a) White paper for which controls answer which question and how to validate

3. Cheat-sheet for data processing/analysis

Questions

1. Value of controls that are currently used
    a. e.g how much do we gain vs lose with spike-in reads?
2. How to include a priori knowledge references?
    a. Do we use them during normalization? Or post-hoc?
    b. Do we use a supervised method when the populations are known, e.g. access to cell type “pure” signatures
3. How to evaluate normalization with or without controls?
    a. Control experiments for method validation → case for benchmarking datasets?
    b. What about overfitting when benchmarking datasets are public?
    c. Do we need to held out datasets?

References

https://www.nature.com/articles/s41592-019-0425-8

https://link.springer.com/protocol/10.1007%2F978-1-4939-9057-3_1

https://www-nature-com.ucsf.idm.oclc.org/articles/nmeth.4292

https://www-nature-com.ucsf.idm.oclc.org/articles/nrg3833

https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-019-0538-z

https://academic.oup.com/gigascience/article/8/9/giz106/5570567