Computational analysis of cell-to-cell heterogeneity in single-cell RNA-seq data

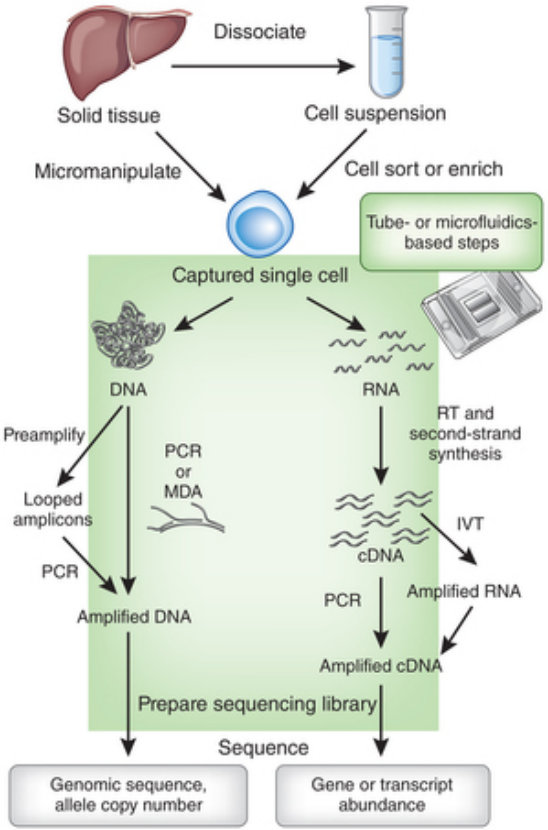
Buettner et al, 2015 Nature 33:3

Critical paper review by

Jasleen Grewal

Strides in sequencing

- Traditional approach – bulk of cells, ground up and sequenced.
- Single cell sequencing[1]
  - Genome
  - Transcriptome (talk the talk!)
- Transcriptomics approaches[2]
  - Microarrays (eQTLs)
  - mRNA in situ hybridization (# genes =< 32)
  - Single cell quantitative PCR (gene panel)
  - (sc)RNA-seq



1. Tal Nawy. Single-cell sequencing. Nature methods 11,18 (2014).
2. Kanter, I & Kalisky, T. Single cell transcriptomics: methods and applications. Front. Oncol., 10 March 2015.

We are all unique snowflakes

- Tissues
- New ~~snowflakes~~ cell types?
- Hidden unknown covariates
  - Unobserved covariates
  - Unknown subtle environmental perturbations

1. Deng Q, Ramskold D, Reinius B, Sandberg R. Single cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* (2014) **343**:193–6. doi:10.1126/science.1245316

Finding hidden covariates

- Bulk sequencing data
  - Sample handling and protocol adjustments
  - SNP profiles
  - BAF Fractions
  - PANAMA, ICE, PEER (eQTLs)
- Single cell sequencing data
  - ???

Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells

Florian Buettner[1,2,5], Kedar N Natarajan[2,3,5], F Paolo Casale[2], Valentina Proserpio[2,3], Antonio Scialdone[2,3], Fabian J Theis[1,4], Sarah A Teichmann[2,3], John C Marioni[2,3] & Oliver Stegle[2]

- Single cell gene expression measurement (scRNA-seq)
- Hidden confounders affect global gene expression
- Identify these ‘hidden’ expression signatures
- Interpret gene expressions without these confounding biological signatures
- Use Gaussian Latent Variable Models

Cell cycle is regulated by multiple(!) genes

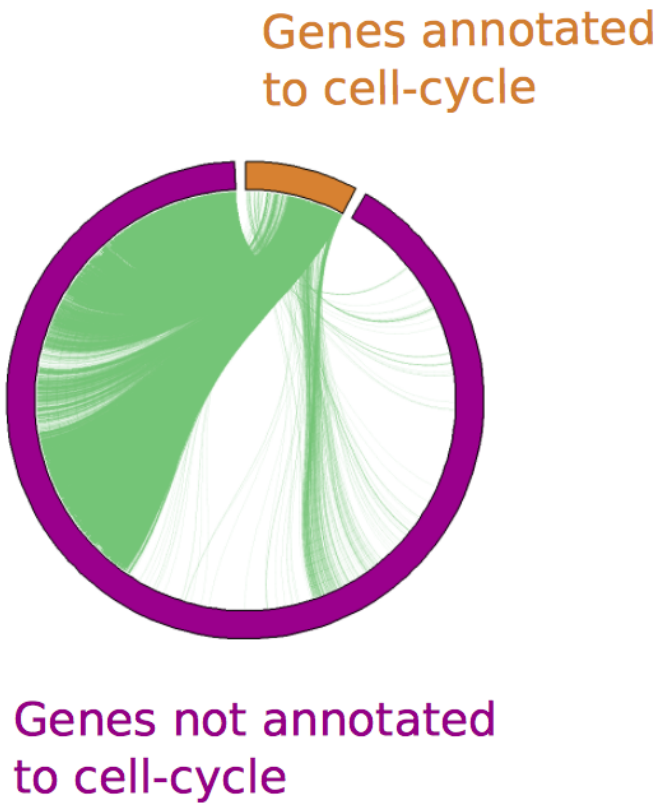
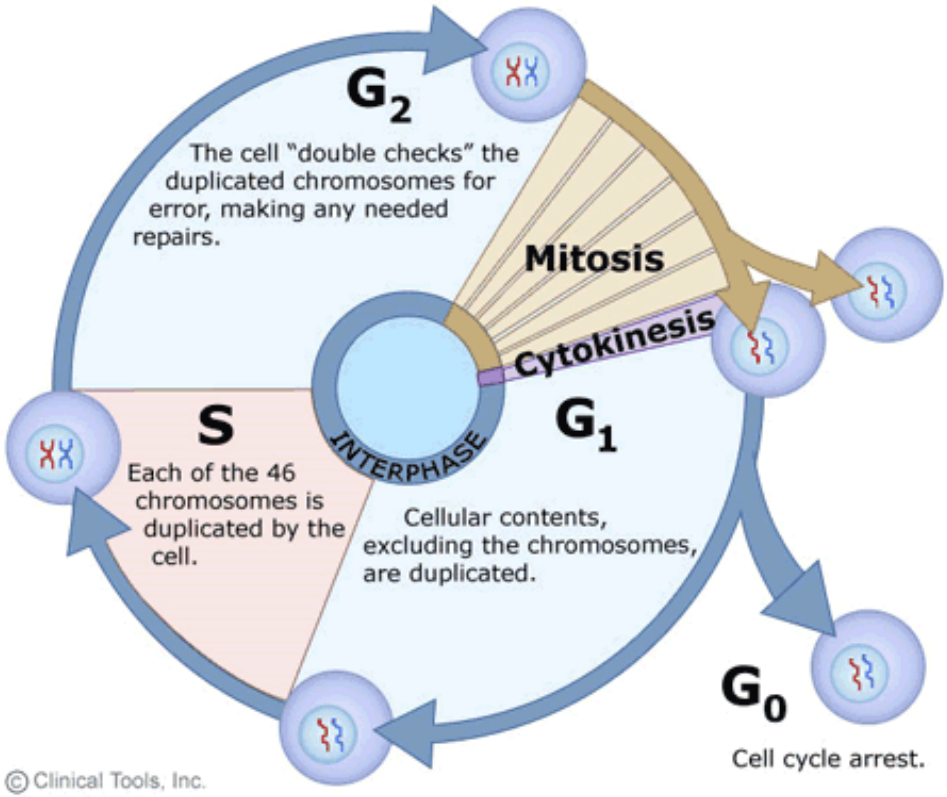


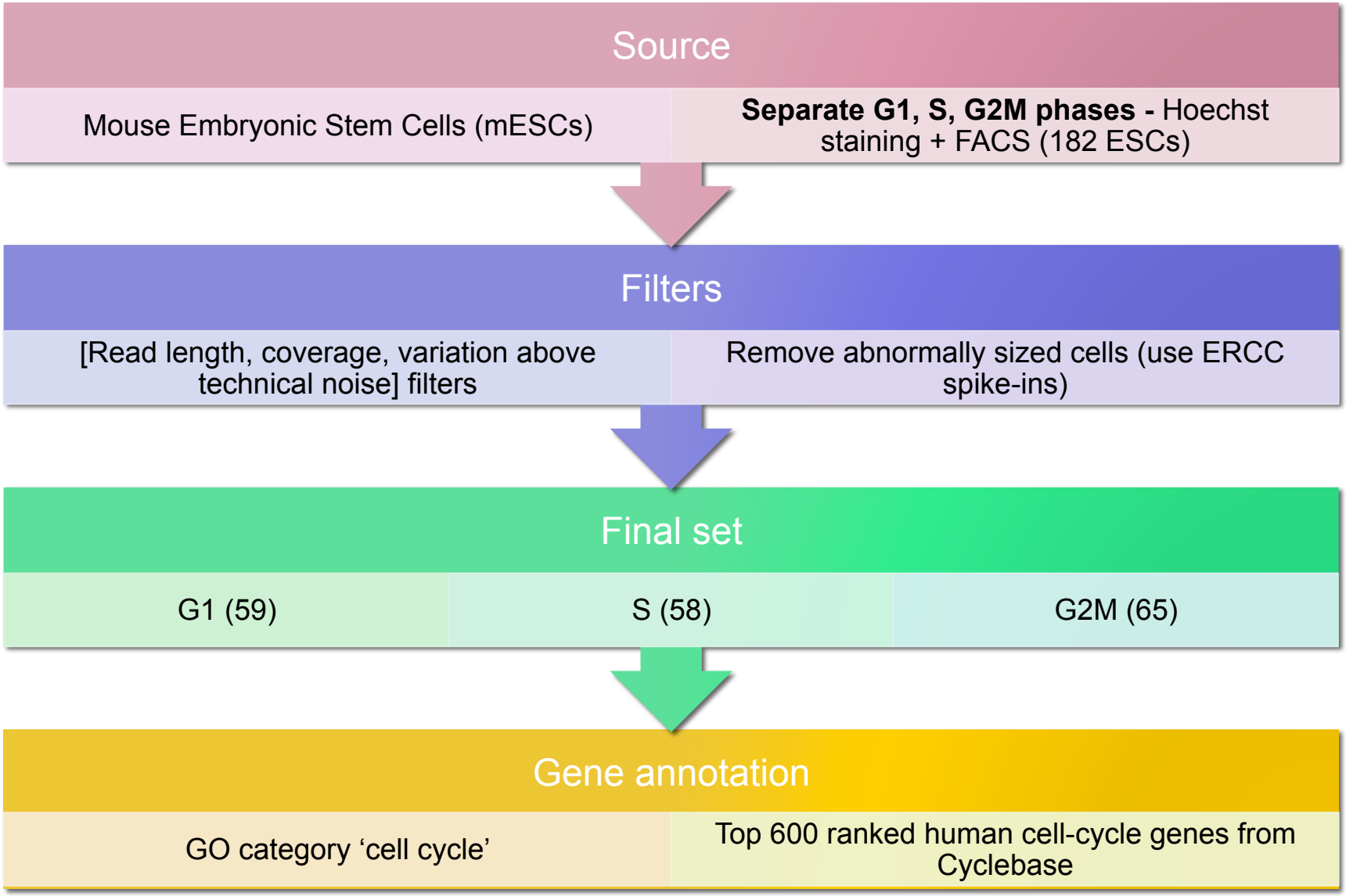
Figure 2. Correlation of cell-cycle annotated genes (549) with other genes' expression levels (6524).

How to study this in single cells?

- Have cells at different cellular time points
- Annotate different genes based on ‘latent variable’



Validation Data



Got the data, now what?

squeeze your eyes shut and pray real hard

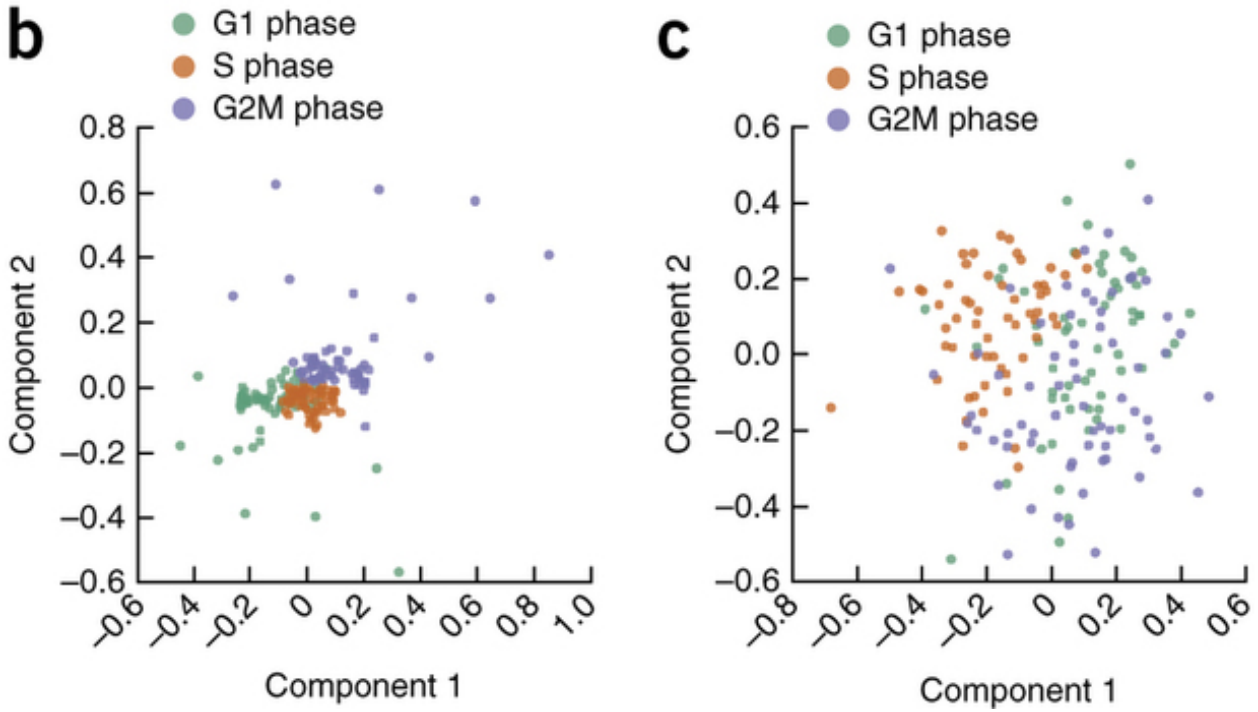
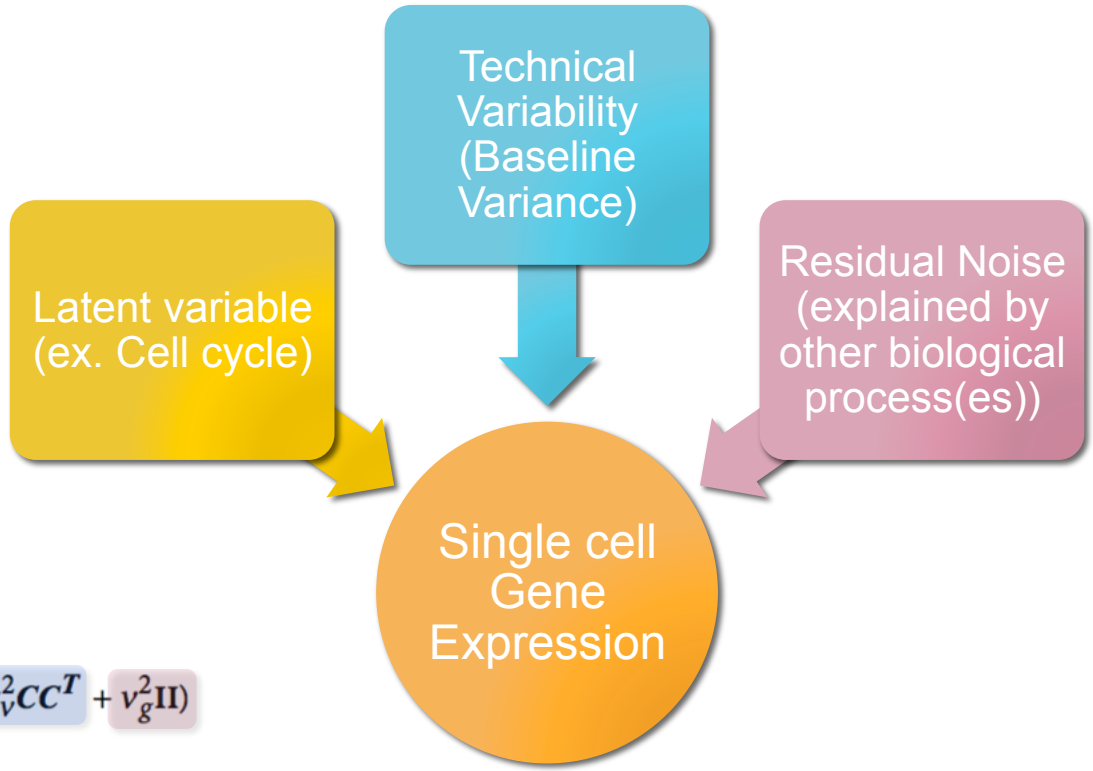


Figure 2 (paper). Nonlinear PCA based on genes not annotated as cell cycle.
(b) Without correction for cell cycle
(c) After correction for cell cycle, with scLVM

Single cell Latent Variable Model Algorithm

- Two Stage Procedure
  - Covariance Matrix (cell – to – cell) -> Learning
  - Account for *hidden factors,* and achieve great things
    - Reveal new/true associations (cells, genes)



$$y_g \sim \mathcal{N}(\mu_g \mathbf{1}, XX^T + \sigma_v^2 CC^T + v_g^2 \mathrm{II})$$

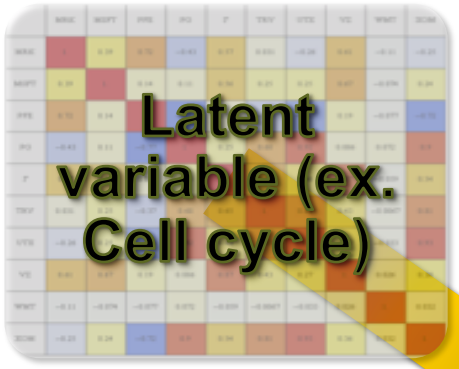
‘Hidden’/latent variable (ex. Cell cycle)

Annotated marker genes for Hidden Variable

Gene expression value in each cell for each marker gene

	MRK	MSFT	PFE	PG	T	TRV	UTX	VZ	WMT	XOM
MRK	1.	0.39	0.72	−0.43	0.57	0.031	−0.26	0.61	−0.11	−0.25
MSFT	0.39	1.	0.14	0.11	0.56	0.25	0.25	0.67	−0.074	0.24
PFE	0.72	0.14	1.	−0.77	0.08	−0.37	−0.65	0.19	−0.077	−0.72
PG	−0.43	0.11	−0.77	1.	0.25	0.68	0.92	0.086	0.072	0.9
T	0.57	0.56	0.08	0.25	1.	0.65	0.46	0.87	−0.059	0.54
TRV	0.031	0.25	−0.37	0.68	0.65	1.	0.83	0.43	−0.0067	0.81
UTX	−0.26	0.25	−0.65	0.92	0.46	0.83	1.	0.27	−0.033	0.93
VZ	0.61	0.67	0.19	0.086	0.87	0.43	0.27	1.	0.026	0.36
WMT	−0.11	−0.074	−0.077	0.072	−0.059	−0.0067	−0.033	0.026	1.	0.032
XOM	−0.25	0.24	−0.72	0.9	0.54	0.81	0.93	0.36	0.032	1.

Covariance matrix accounting for cell-to-cell differences in hidden variable



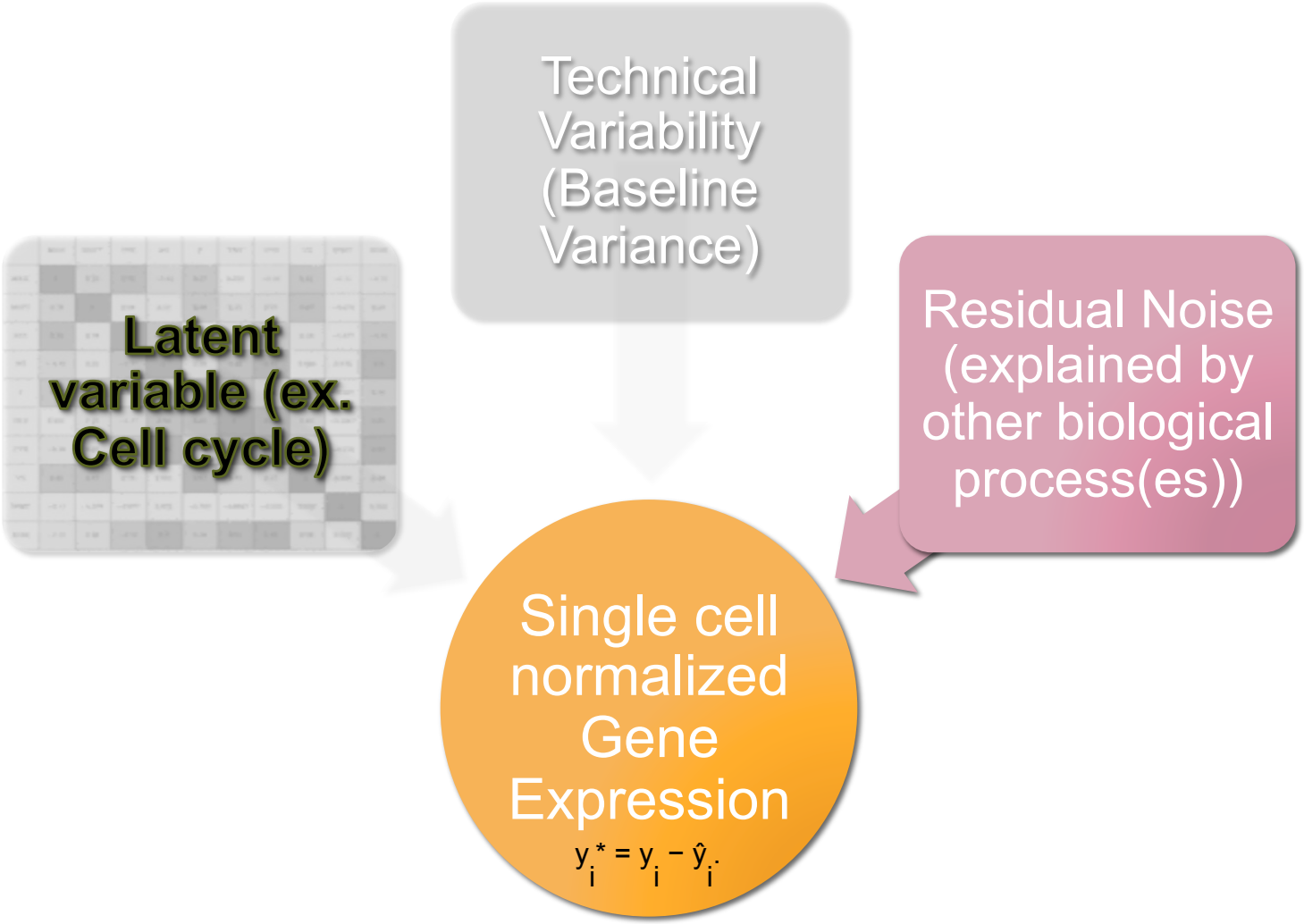
Latent variable (ex. Cell cycle)

Technical Variability (Baseline Variance)

Residual Noise (explained by other biological process(es))

Single cell Gene Expression

$$y_g \sim \mathcal{N}\left(\mu_g \mathbf{1}, \sum_{h=1}^{H} \sigma_{gh}^2 \Sigma_h + v_g^2 \mathbf{I} + \delta_g^2 \mathbf{I}\right)$$


Technical Variability (Baseline Variance)
Latent variable (ex. Cell cycle)
Residual Noise (explained by other biological process(es))
Single cell normalized Gene Expression
$y_i^* = y_i - \hat{y}_i$.

Comparison of scLVM correction with?

- **Remove all cell-cycle genes from the analysis, and then compare PCA Distribution of each cell with/ without correcting with scLVM**
- Negative control: Compare to a cell type which is noncycling (terminally differentiated neurons!), and one which is cycling (T cells)

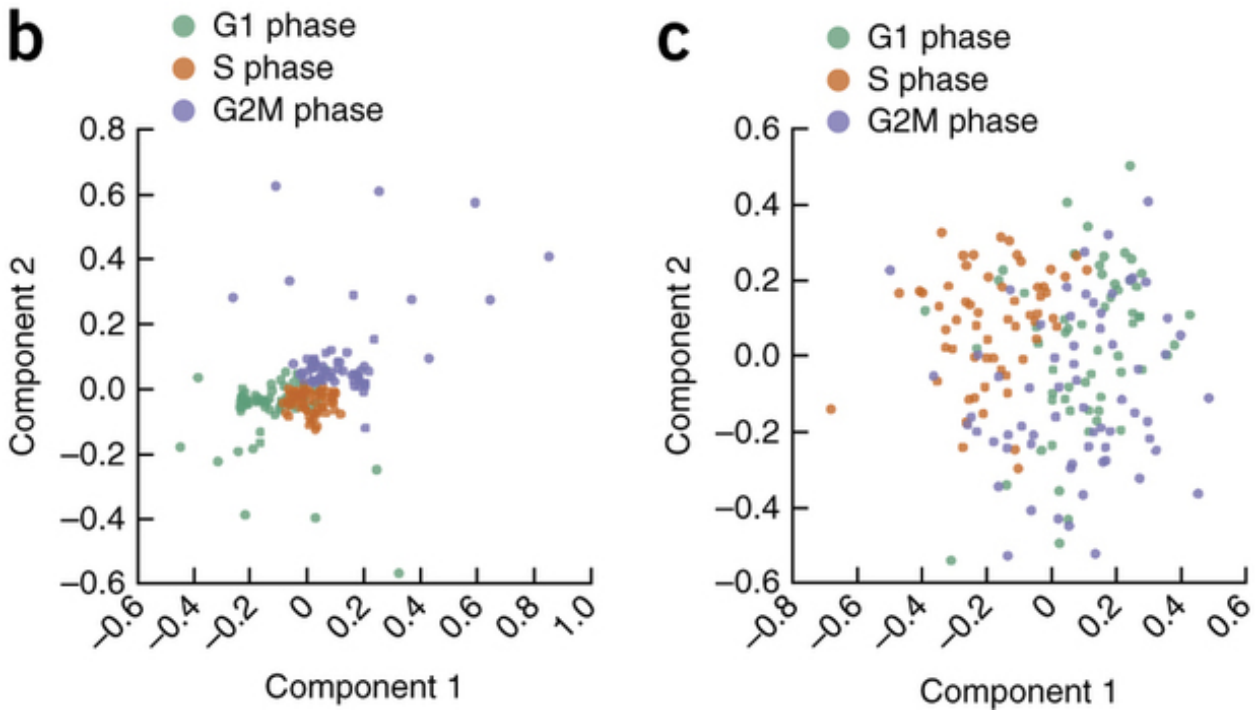


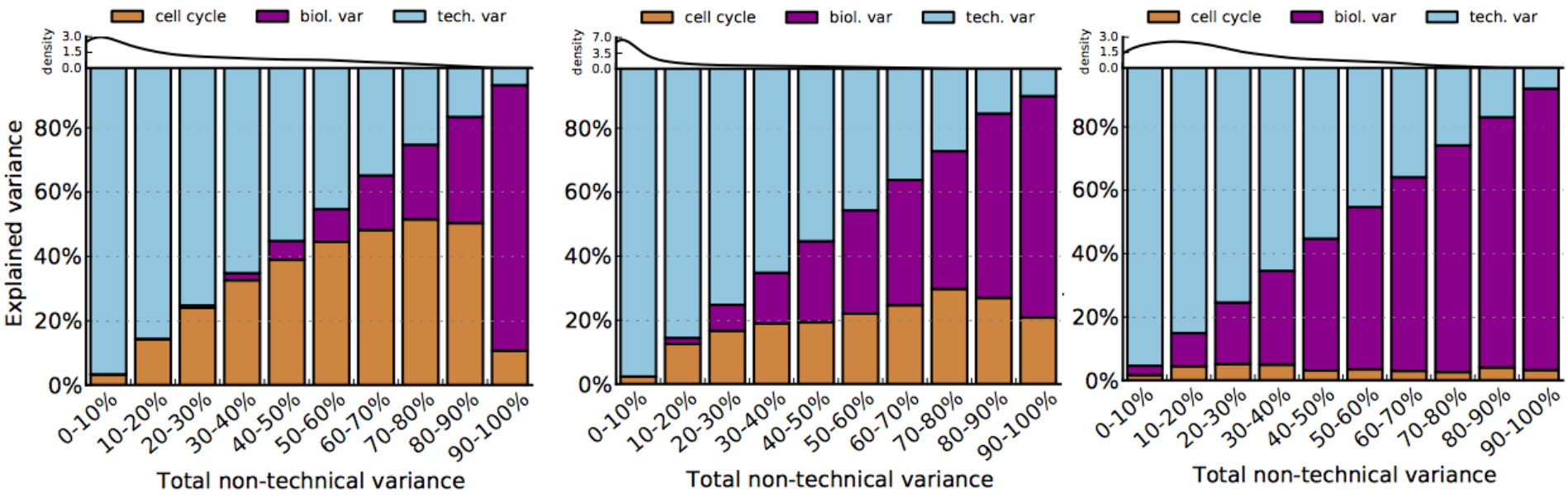
Figure 2 (paper). Nonlinear PCA based on non-cellcycle genes.
(b) Without correction for cell cycle
(c) After correction for cell cycle, with scLVM

Comparison of scLVM correction with?

- Remove all cell-cycle genes from the analysis, and then compare PCA Distribution of each cell with/without correcting with scLVM
- **Negative control: Compare to a cell type which is noncycling (terminally differentiated neurons!), and one which is cycling (T cells)**

Negative controls : Variance Decomposition by scLVM

(Supplementary Figure 7)



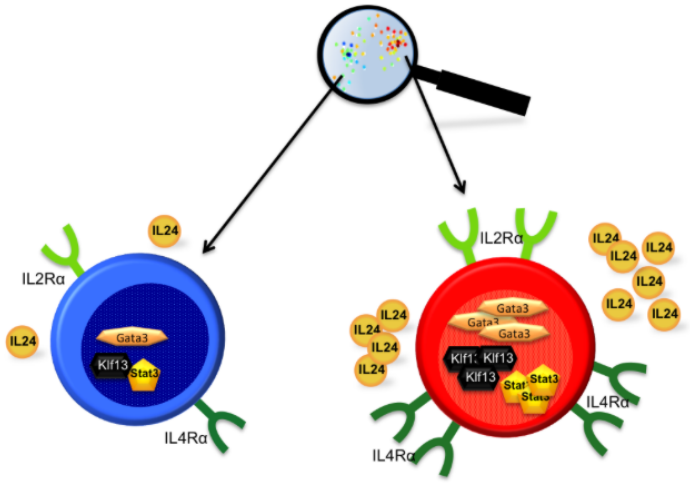
T-cells

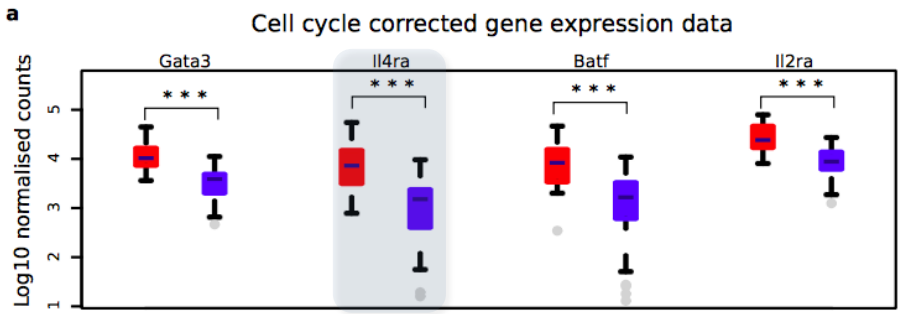
Mice ESCs

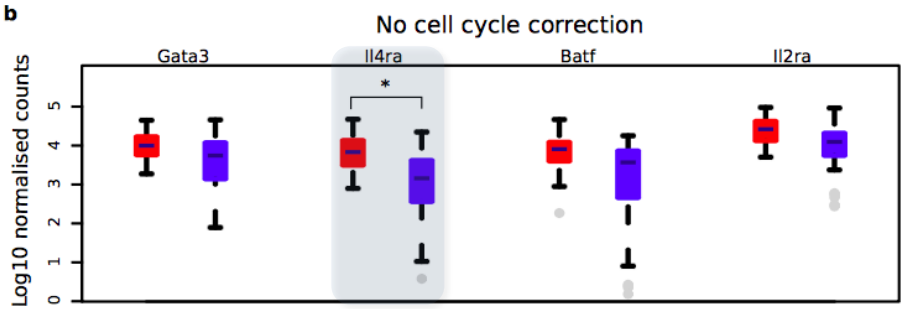
30 terminally differentiated mice neural cells

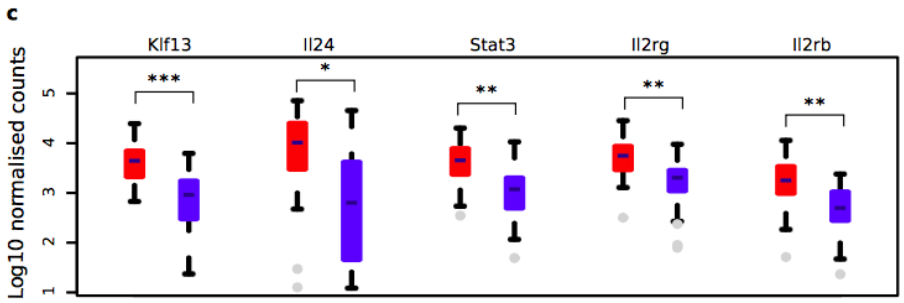
Revelations?

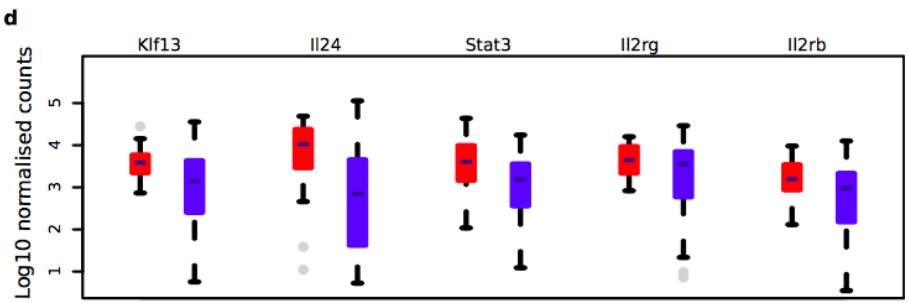
- Cell cycle has a huge impact on overall cell expression state in early development
- Once you remove the impact of such latent variables, new subpopulations are revealed





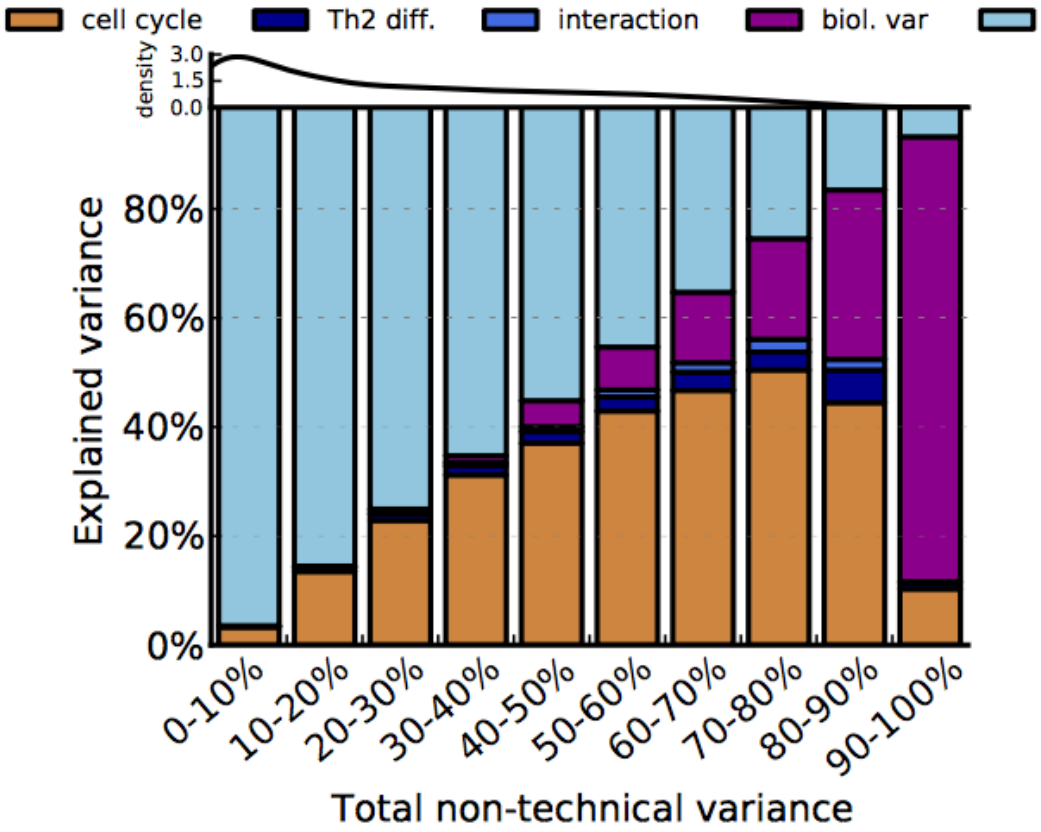






Things they already thought of

- Methodology adaption: Similar data obtained from a different protocol (Quartz-Seq), cultured under different media conditions that help reduce variability in expression of cell-cycle genes.
  - Robust to sequencing protocols, studies, experimental batches
- Does the size of the set of marker genes have a huge impact? (No... n >= 50)
- Can we model multiple latent variables
  - T_H2 differentiation
  - Tested viability of subgroup identification using hierarchical and k-means clustering as well (data not shown)



Supplementary Figure 15
For each gene, the proportion of variance explained by cell cycle, T_H2 differentiation, a multiplicative interaction between cell cycle and T_H2 differentiation as well as technical noise and residual biological variance was estimated.

Caveats & Roadblocks

- Need to understand (or atleast have some premonition) of the latent variable(s)
  - Built up on massive wetlab + bioinformatics research (and not just from scRNA-seq)
- Datasets used in this study were from *in vitro* samples
  - Several environmental influences *in vivo*
- Formal statistical methods
  - Currently assume all genes' expression set follows Gaussian distribution
  - what is technical noise? Is normalization protocol-specific?
  - Sample size (number of cells)?

Extensions

- Apply latent variable modeling on terminally differentiated cells
  - For ex., do apoptotis pathways lead to cell-cell heterogeneity in terminally differentiated cells?
  - Extension of interaction modeling may reveal significant associations between cell processes hitherto understood to be mutually exclusive
- Profile tumour heterogeneity
  - Tumour stem cells
  - Occam's razor – is each tumour cell a new subpopulation?
- Generate 'pseudo-temporal' trajectories
  - Hematopoetic Stem Cells in Zebrafish – larval stage dependent phenotypes

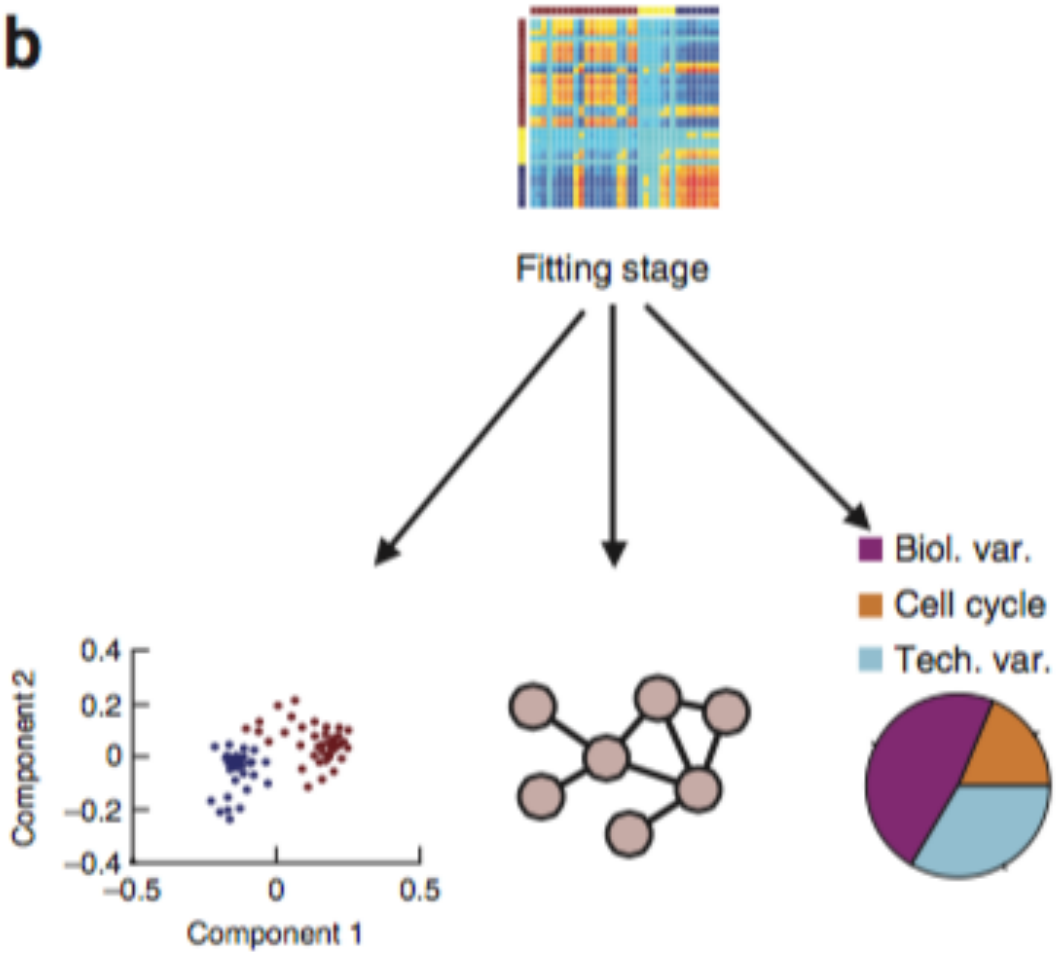
Questions?


WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS

ERCC spike-ins

- External RNA Controls Consortium
- Common set of RNA controls (92 unlabeled, polyadenylated transcripts)
- Added to RNA analysis experiment after sample isolation
- Ratio of endogenous reads to total number of reads

Apply data to model



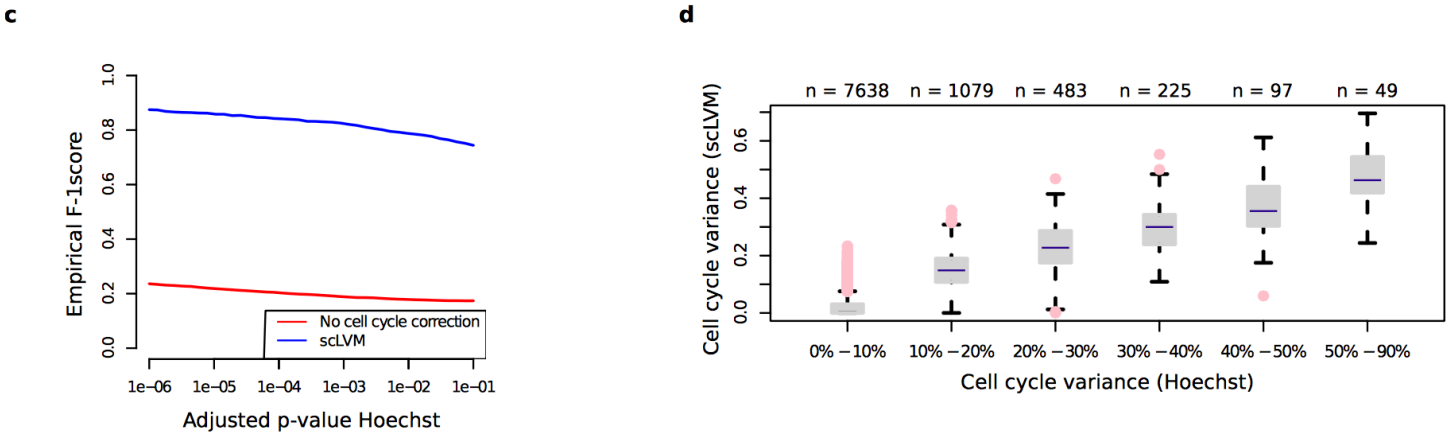
Mouse ESC data Bad Cells Filtering

Criteria for bad cells	
C1 chip capture	*Empty/Debris/double cell*
Bioinformatic quality control for excluding bad cells	
Total number of reads per cell	*Less than 0.5 million*
Percentage of reads mapping to known exons	*Less than 20%*
Number of genes detected per cell	*Less than 6,000*
Percentage of reads mapping to exons	*Greater than 60%*
Percentage of mitochondrial reads	*Greater than 15%*
Percentage of low quality reads (after GSNAP)	*Greater than 10%*

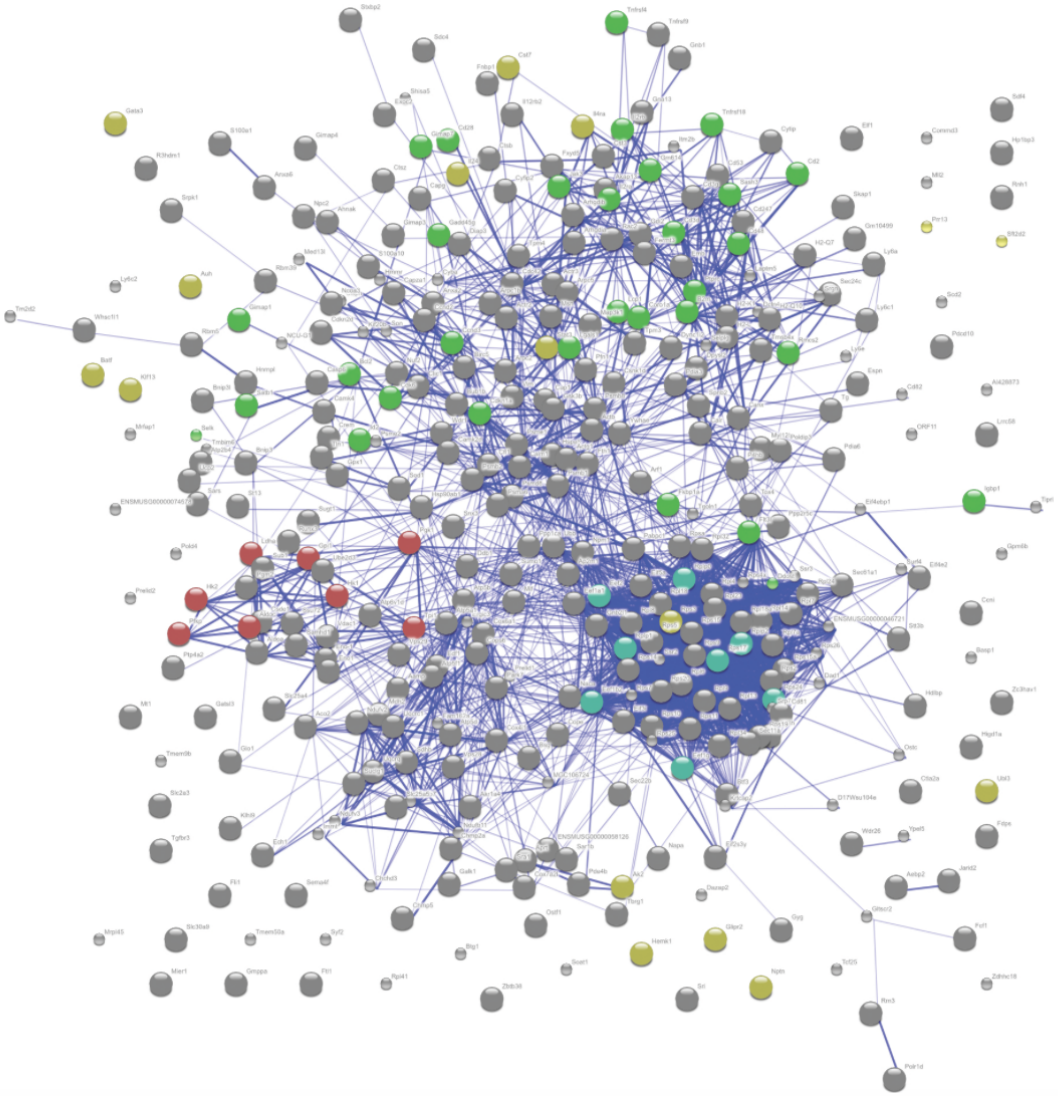
Fraction	Bad cells	**Good single cells**
G1	13 cells	**83 cells**
S	23 cells	**73 cells**
G2M	7 cells	**89 cells**

*Maximum capture per chip is 96 cells

Validation results



Supplementary Figure 6: **Additional analyses of the Mouse ESC data. a-b**, Significance of gene-gene correlations, comparing p-values derived from scLVM (**a**) or without correction (**b**) (x-axis) to p-values derived from the gold standard based on the Hoechst staining (y-axis) (Newly generated using the C1 protocol). Without cell-cycle correction, the significance of gene-gene correlations was inflated relative to correlation p-values computed with scLVM. Comparison with the gold standard data revealed that many of the gene-gene correlations detected by the model without cell-cycle correction were false positives (FP, points in the lower right quadrant defined by the dashed green lines). **c**, precision and recall for significant correlations inferred with scLVM and without cell-cycle correction was quantified over a range of Bonferroni-corrected p-values using the F1 score (harmonic mean of precision and recall). Precision and recall were calculated relative to significant "gold-standard" correlations derived based on the Hoechst staining. **d**, comparison of the estimated proportion of gene expression variance using the scLVM and gold standard estimates from the Hoechst staining in the Mouse ESC dataset. Genes were binned by the proportion of variance attributable to cell-cycle based on the Hoechst staining data.



Literature derived networks of genes that are differentially expressed between subpopulations (Supplementary Figure 14)

Cell cycle is regulated by multiple(!) genes

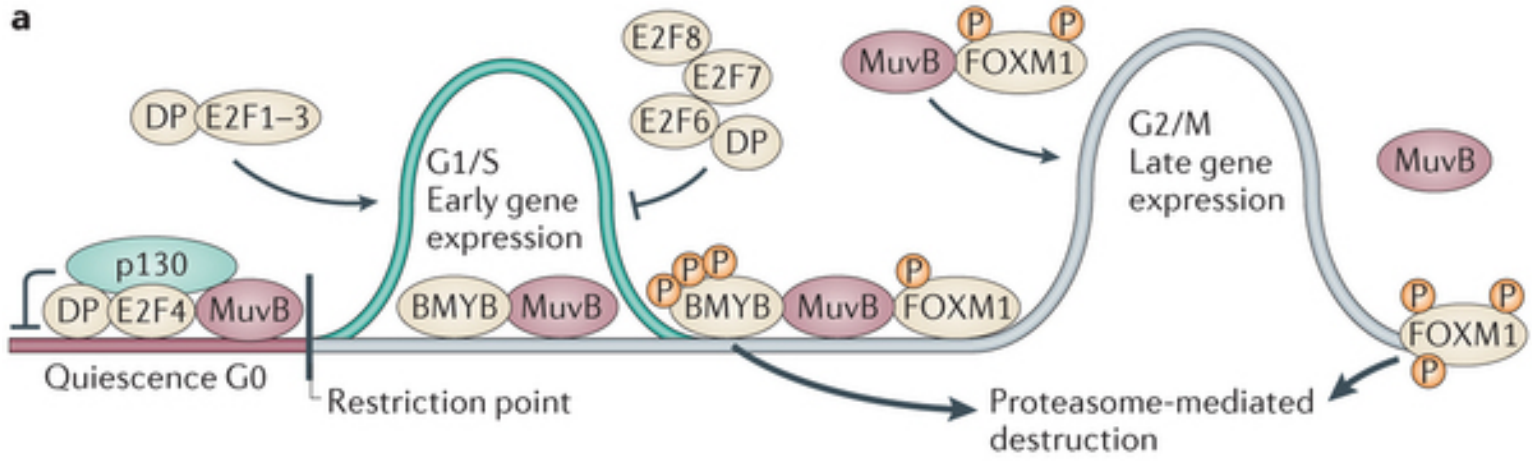


Figure 1. The differential regulation of early and late cell cycle gene expression.

Subhashini Sadasivam & James A. DeCaprio. The DREAM complex: master coordinator of cell cycle-dependent gene expression. *Nature Reviews Cancer* 13, 585-595 (2013)

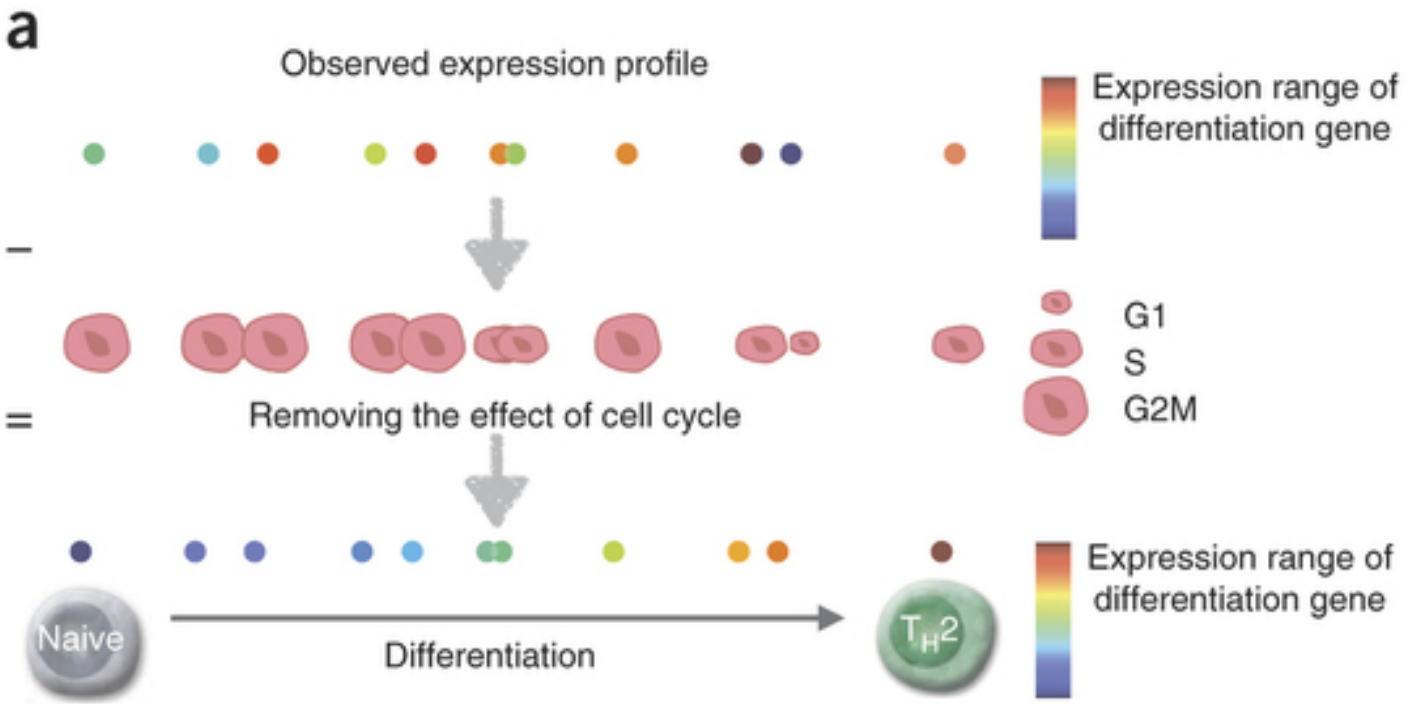
Experimental Data

Source		
scRNA-seq experiment	Differentiation of native T cells into T_H2 cells	Stimulate $CD4^+$ T helper cells, sort 4.5 days later

Mechanical Filters		
Pool into G4P (**IL-13-GFP$^+$**) and G2N (**IL-13-GFP$^-$**) groups	Pool with equal proportion from each group	Fluidigm C1 cell sorting (96 asynchronously dividing cells)

Post-sequencing Filters		
Quality Control	Technical Noise Filter	Threshold expression level above technical noise

Final set		
Single Cell RNA-seq	81 cells	7073 genes


a
Observed expression profile
Expression range of differentiation gene
–
G1
S
G2M
=
Removing the effect of cell cycle
Naive
Differentiation
T_H2
Expression range of differentiation gene

Extensions to current approach

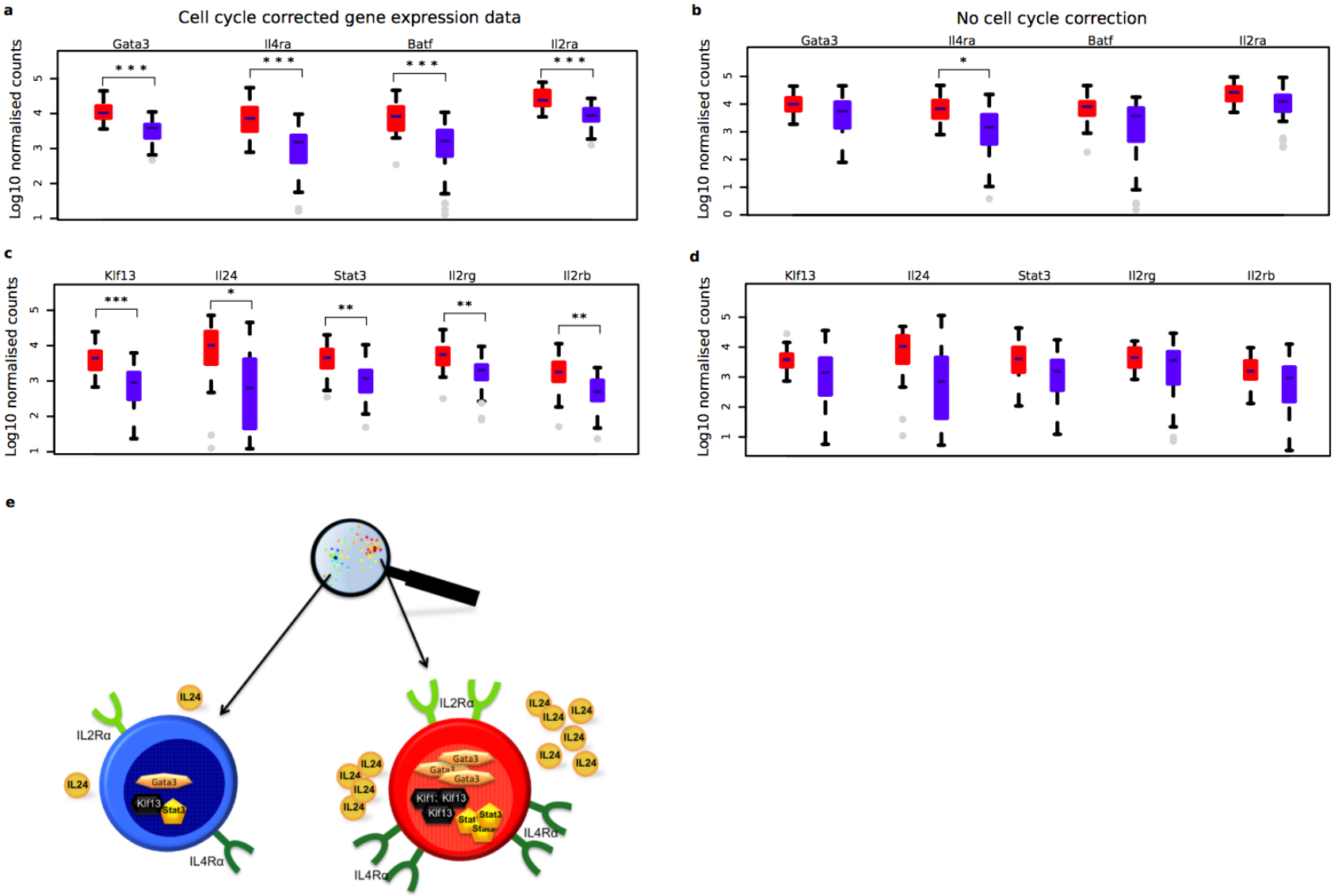
- Reduced alphabet set
  - Learn that through unsupervised clustering analyses
- Prediction of co-interacting proteins
  - Learn about new functional attributes of proteins
  - Learn about new protein classes?
- Multi-class separation using functionally characteristic alphabet sets

Why study single cell heterogeneity

- 6-7% of eukaryotic proteins are DNA binding[1]
- Replication, transcription, DNA packaging
- Chemotherapy drugs: DNA damaging agents
- Identify non-genotoxic DNA binding proteins

CAP is a gene regulatory protein from *E.coli.* In the absence of the bound protein, this DNA helix is straight.[2]

1. Kumar et al. Identification of DNA binding proteins using SVMs and evolutionary profiles. 2007, BMC bioinformatics.
2. Molecular Biology of the Cell. 4th edition. Alberts B, Johnson A, Lewis J, et al. New York: Garland Science; 2002.



Supplementary Figure 13: **Differences in gene expression between the discovered subpopulations. a** and **c**, boxplots for corrected gene expression, and **b**and **d**, boxplots for uncorrected gene expression, **e**, differentially expressed genes include receptors, cytokines and transcription factors. All comparisons were evaluated using rank sums tests (Bonferroni adjusted). Adjusted p-values <0.05 were considered significant. *p<0.05; **p<0.01; ***p<0.001. When uncorrected gene expression are compared between the sub-populations, only Il4ra remains significantly differentially expressed between the sub-populations.