Final 5 weeks

Office hours to discuss projects:

Mondays 12.00-13.00 Y11-J-16
(November 26, December 3, 10, 17)

Projects due 11 Jan 2019 18:00

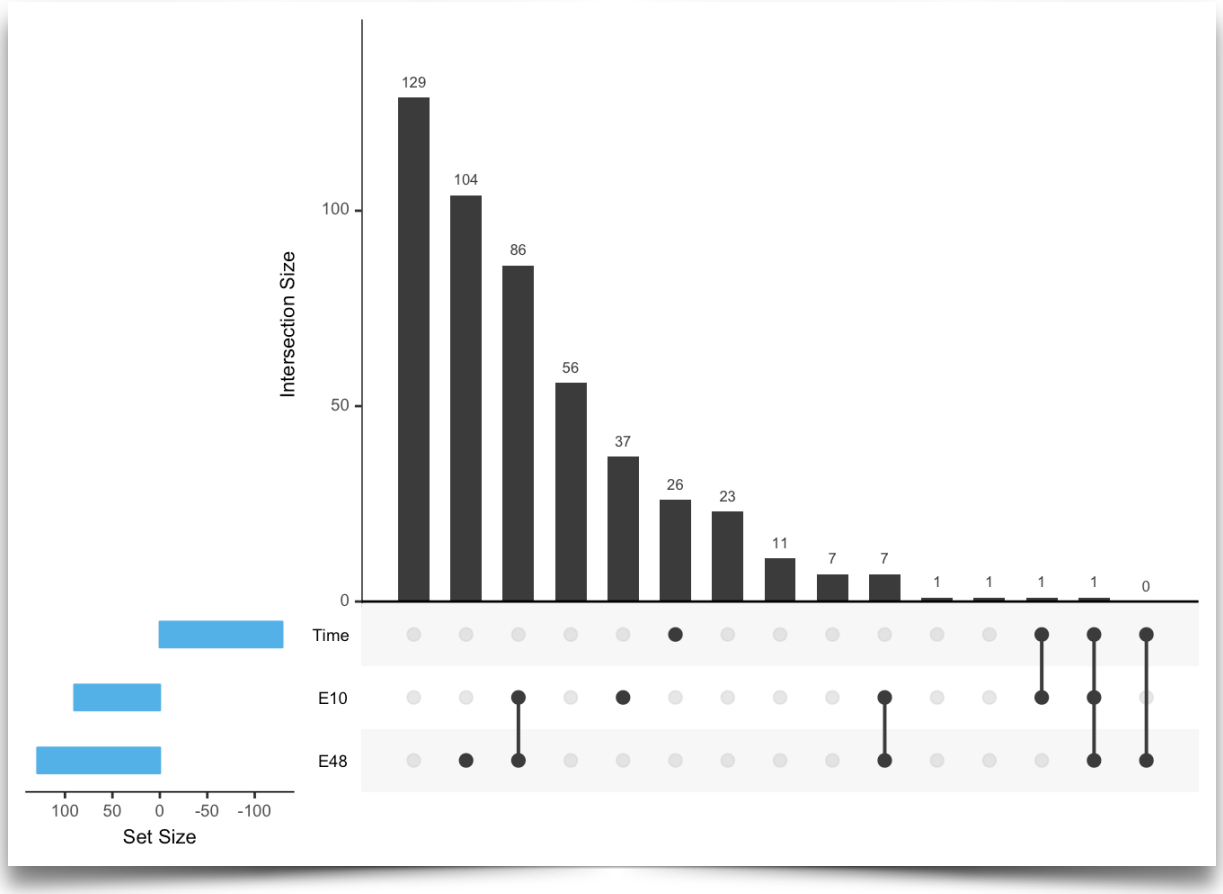
Plan due next week (give GitHub usernames; I can create private repo for the project and invite members)

Note: for reproducing analyses from papers, this is typically meant for biology papers rather than methods papers.

19.11.2018	Mark	single-cell dim. reduction + clustering; FDR	conquer	Normalization of RNA-seq data using factor analysis of control genes or samples (RM, JD, CV)	Diffusion maps for high-dimensional single-cell analysis of differentiation data (SP, GK)
26.11.2018	Lukas	hands-on session #2: cytometry	cytof null comparison	Epigenome-wide association studies without the need for cell-type composition (RL, SG)	X
03.12.2018	Hubert	classification	MLInterfaces	Predicting cell types in single cell mass cytometry data (CM, SS)	
10.12.2018	Mark	loose ends: HMM, EM, robustness	segmentation, peak finding	Differential expression analysis for sequence count data (AA, PS)	Visualizing Data using t-SNE (MJT, TB, MP)
17.12.2018	Mark	hands-on session #3: single-cell RNA-seq	full scRNA-seq pipeline	Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies (SB,ST)	x

Quick thoughts on Exercise 8

UpSet plots



Quick thoughts on Exercise 8

Design matrix

```
# do the limma modeling
f <- paste(targets$estrogen,targets$time.h,sep="")
f <- factor(f)

# create design matrix
design <- model.matrix(~0+f)
colnames(design) <- levels(f)
design
```

```
##   absent10 absent48 present10 present48
## 1        1        0         0         0
## 2        1        0         0         0
## 3        0        0         1         0
## 4        0        0         1         0
## 5        0        1         0         0
## 6        0        1         0         0
## 7        0        0         0         1
## 8        0        0         0         1
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$f
## [1] "contr.treatment"
```

```
f <- paste(targets$time.h, targets$estrogen, sep="")
design2 <- model.matrix(~0+f)
design2
```

```
##   f10absent f10present f48absent f48present
## 1         1          0         0          0
## 2         1          0         0          0
## 3         0          1         0          0
## 4         0          1         0          0
## 5         0          0         1          0
## 6         0          0         1          0
## 7         0          0         0          1
## 8         0          0         0          1
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$f
## [1] "contr.treatment"
```

Quick thoughts on Exercise 8

Design matrix

```
# do the limma modeling
f <- paste(targets$estrogen,targets$time.h,sep="")
f <- factor(f)

# create design matrix
design <- model.matrix(~0+f)
colnames(design) <- levels(f)
design
```

```
##   absent10 absent48 present10 present48
## 1        1        0         0         0
## 2        1        0         0         0
## 3        0        0         1         0
## 4        0        0         1         0
## 5        0        1         0         0
## 6        0        1         0         0
## 7        0        0         0         1
## 8        0        0         0         1
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$f
## [1] "contr.treatment"
```

```
(mm_alt <- model.matrix( ~ estrogen*time.h, data=targets ))
```

```
##   (Intercept) estrogenpresent time.h48 estrogenpresent:time.h48
## 1           1                0        0                       0
## 2           1                0        0                       0
## 3           1                1        0                       0
## 4           1                1        0                       0
## 5           1                0        1                       0
## 6           1                0        1                       0
## 7           1                1        1                       1
## 8           1                1        1                       1
```

```
> fit <- lmFit(eset, design)
> cont.matrix <- makeContrasts(E10="present10-absent10",
+                              E48="present48-absent48",
+                              Time="absent48-absent10",levels=design)
> cont.matrix
           Contrasts
Levels      E10 E48 Time
  absent10   -1   0   -1
  absent48    0  -1    1
  present10   1   0    0
  present48   0   1    0
> fit2  <- contrasts.fit(fit, cont.matrix)
```

```
(mm_alt <- model.matrix( ~ estrogen*time.h, data=targets ))
```

```
##   (Intercept) estrogenpresent time.h48 estrogenpresent:time.h48
## 1           1                0        0                       0
## 2           1                0        0                       0
## 3           1                1        0                       0
## 4           1                1        0                       0
## 5           1                0        1                       0
## 6           1                0        1                       0
## 7           1                1        1                       1
## 8           1                1        1                       1
## attr(,"assign")
## [1] 0 1 2 3
## attr(,"contrasts")
## attr(,"contrasts")$estrogen
## [1] "contr.treatment"
##
## attr(,"contrasts")$time.h
## [1] "contr.treatment"
```

```
fit_alt <- lmFit(eset, mm_alt)
fit_alt <- eBayes(fit_alt)
```

```
# do the limma modeling
f <- paste(targets$estrogen,
f <- factor(f)

# create design matrix
design <- model.matrix(~0+f)
colnames(design) <- levels(f)
design
```

```
##   absent10 absent48 present10 present48
## 1        1        0         0         0
## 2        1        0         0         0
## 3        0        0         1         0
## 4        0        0         1         0
## 5        0        1         0         0
## 6        0        1         0         0
## 7        0        0         0         1
## 8        0        0         0         1
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$f
## [1] "contr.treatment"
```

```
topTable(fit_alt,coef=2,n=3)
```

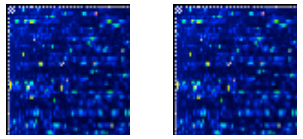
```
##                logFC   AveExpr         t      P.Value    adj.P.Val        B
## 39642_at  2.939428  7.876515 23.71715 4.741579e-09 3.128295e-05 9.966810
## 910_at    3.113733  9.660238 23.59225 4.955715e-09 3.128295e-05 9.942522
## 31798_at  2.800195 12.115778 16.38509 1.025747e-07 3.511070e-04 7.977290
```

```
topTable(fit2,coef=1,n=3)
```

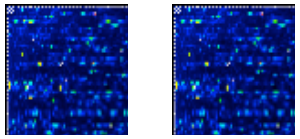
```
##                logFC   AveExpr         t      P.Value    adj.P.Val        B
## 39642_at  2.939428  7.876515 23.71715 4.741579e-09 3.128295e-05 9.966810
## 910_at    3.113733  9.660238 23.59225 4.955715e-09 3.128295e-05 9.942522
## 31798_at  2.800195 12.115778 16.38509 1.025747e-07 3.511070e-04 7.977290
```

Analysis of Variance → Linear model

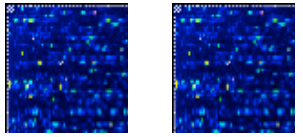
WT x 2



Cond A x 2



Cond B x 2



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

α_1 = wt log-expression

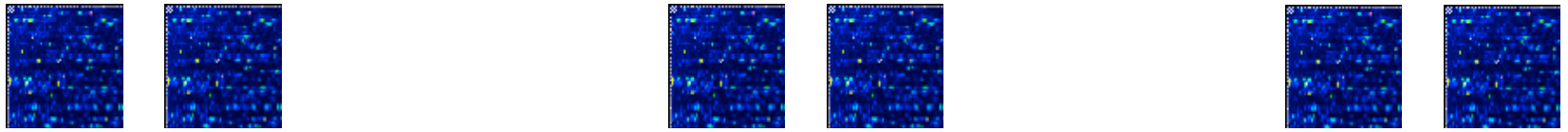
α_2 = Cond A - wt

α_3 = Cond B - wt

$E[y_1]=E[y_2]=\alpha_1$ $E[y_3]=E[y_4]=\alpha_1+\alpha_2$ $E[y_5]=E[y_6]=\alpha_1+\alpha_3$

Analysis of Variance → Linear model, alternative parameterization

WT x 2 | Cond A x 2 | Cond B x 2



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

α_1 = wt log-expression

α_2 = Cond A log-expression

α_3 = Cond B log-expression

$E[y_1]=E[y_2]=\alpha_1$ $E[y_3]=E[y_4]=\alpha_2$ $E[y_5]=E[y_6]=\alpha_3$

From Lecture 7.

Multiple testing and adjusted p-values

- Each statistical test has an associated false positive rate
- Traditional method in statistics is to control family wise error rate, e.g., by Bonferroni.
- Controlling the false discovery rate (FDR) is more **appropriate** in ~~microarray~~ transcriptomic studies
- Benjamini and Hochberg method controls expected FDR for independent or weakly dependent test statistics. Simulation studies support use for genomic data.
- All methods can be implemented in terms of adjusted p-values.

https://www.jstor.org/stable/2346101

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG
Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

Multiple testing

Statistical testing is all about controlling error rates.

FWER = family wise error rate: P(**V** > 1)

FDR = false discovery rate: E(**V** / (**V**+**S**))

TABLE 1
Number of errors committed when testing m null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

Multiple testing

Statistical testing is all about control the rate of errors.

FWER = family wise error rate: P(**V** >= 1)

FDR = false discovery rate: E(**V** / (**V**+**S**))

Hochberg (1988) has suggested a different way to utilize Simes's procedure so that it does control the FWER in the strong sense, by offering the following procedure:

let k be the largest i for which $P_{(i)} \leqslant \frac{i}{m+1-i}\alpha$;

then reject all $H_{(i)}$ $i = 1, 2, \ldots, k$.

3. FALSE DISCOVERY RATE CONTROLLING PROCEDURE

3.1. *The Procedure*

Consider testing $H_1, H_2, \ldots, H_m$ based on the corresponding p-values $P_1, P_2, \ldots, P_m$. Let $P_{(1)} \leqslant P_{(2)} \leqslant \ldots \leqslant P_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$. Define the following Bonferroni-type multiple-testing procedure:

let k be the largest i for which $P_{(i)} \leqslant \frac{i}{m}q^*$;

then reject all $H_{(i)}$ $i = 1, 2, \ldots, k$. (1)

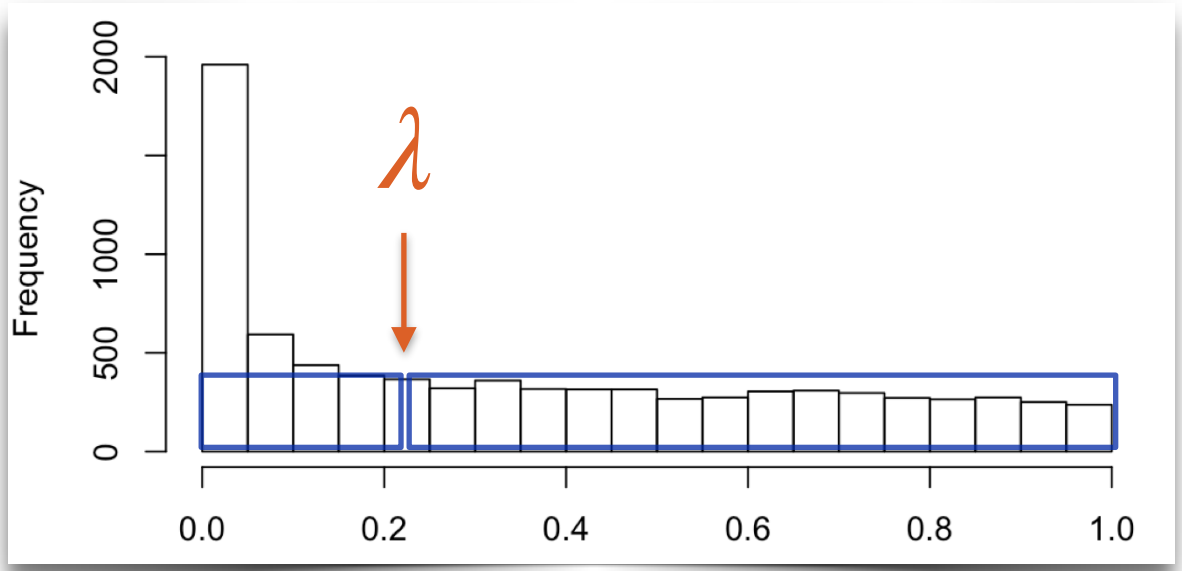
Other variations FDR

TABLE 1
Number of errors committed when testing m null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
	$m - \mathbf{R}$	R	m

- Start with distribution of P-values
- Try to approximate FDR as:

$$E\left[\frac{\mathbf{V}}{\mathbf{R}}\right] \approx \frac{E[\mathbf{V}]}{E[\mathbf{R}]}$$

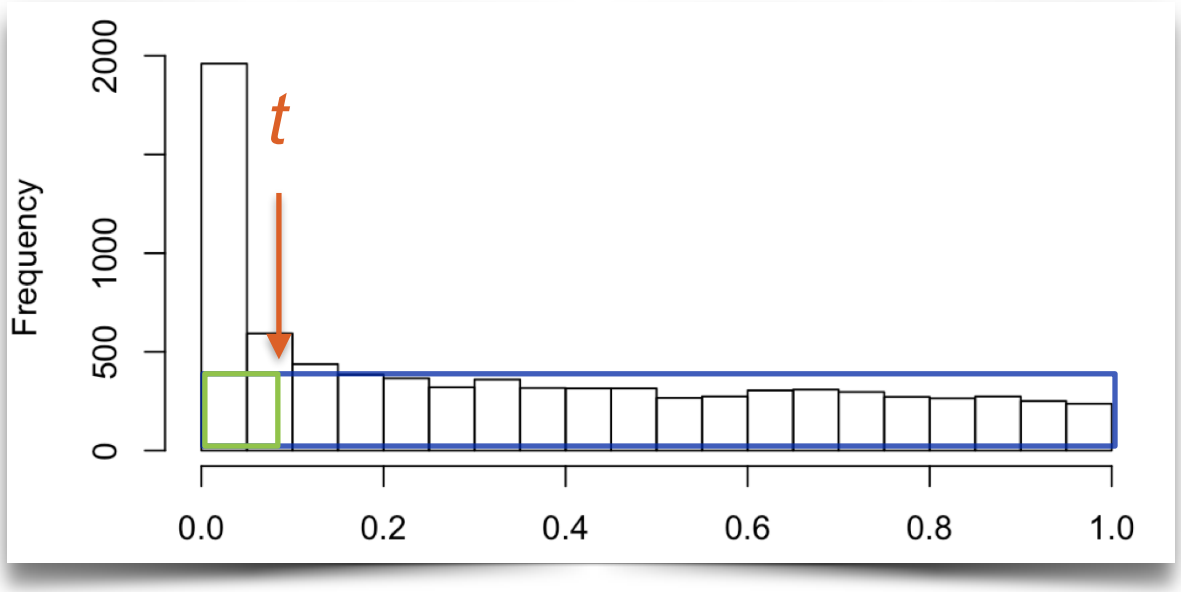


$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1,\ldots,m\}}{m(1-\lambda)}$$

$$E\left[\frac{\mathbf{V}}{\mathbf{R}}\right] \approx \frac{E[\mathbf{V}]}{E[\mathbf{R}]}$$

Find threshold that satisfies desired FDR

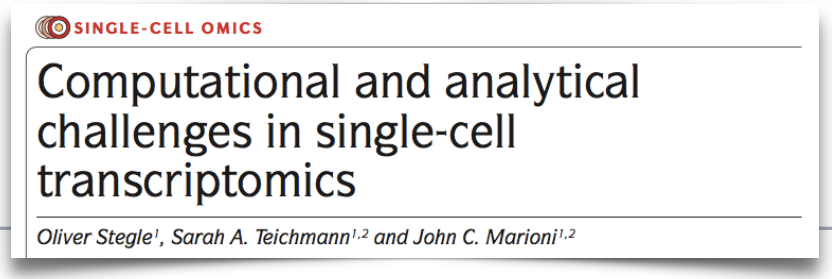
- We have estimated the proportion of nulls, use this to estimate FDR given a threshold



$$FDR = \frac{\hat{\pi}_0 m \cdot t}{S(t)} = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i < t\}}$$

Single cell analysis

- why single cell?
- single-cell RNA-seq (scRNA-seq): a few variations of protocols
- flow/mass cytometry (FACS/CyTOF)
- common themes of data analysis: dimension reduction, clustering, pseudo-time ordering, etc.



Why single cell?

"Bulk" versus single-cell

Discover and quantify abundance of (new) cell types

Study heterogeneity of gene expression

However, there are also important biological questions for which bulk measures of gene expression are insufficient[14]. For instance, during early development, there are only a small number of cells, each of which can have a distinct function and role[15–17]. Moreover, complex tissues, such as brain tissues, are composed of many distinct cell types that are typically difficult to dissect experimentally[18]. Consequently, bulk-based approaches may not provide insight into whether differences in expression between samples are driven by changes in cellular composition (that is, the abundance of different cell types) or by changes in the underlying phenotype. Finally, ensemble measures do not provide insights into the stochastic nature of gene expression[19,20].

Hypothetical situations

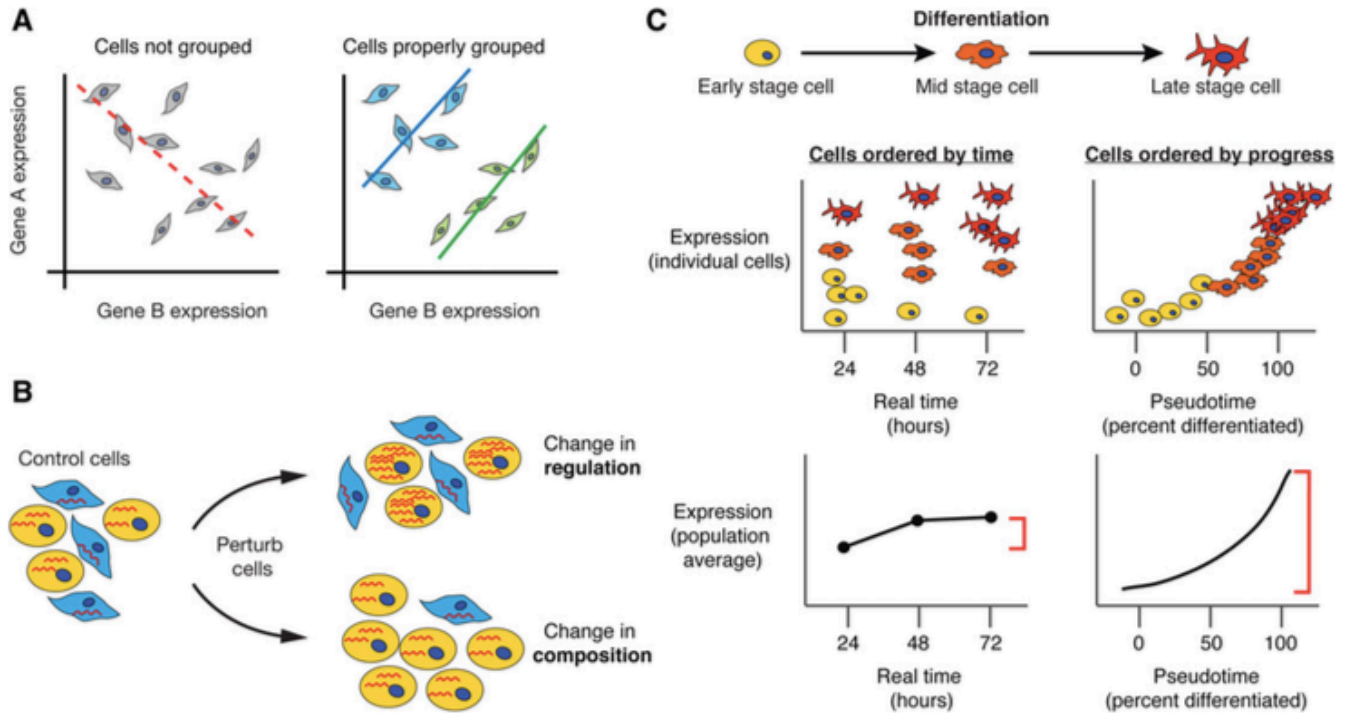


Figure 1. Single-cell measurements preserve crucial information that is lost by bulk genomics assays. (*A*) Simpson's Paradox describes the misleading effects that arise when averaging signals from multiple individuals. (*B*) Bulk measurements cannot distinguish changes due to gene regulation from those that arise due to shifts in the ratio of different cell types in a mixed sample. (*C*) Time series experiments are affected by averaging when cells proceed through a biological process in an unsynchronized manner. A single time point may contain cells from different stages in the process, obscuring the dynamics of relevant genes. Reordering the cells in "pseudotime" according to biological progress eliminates averaging and recovers the true signal in expression (Trapnell et al. 2014).

Trapnell, 2015 Genome Research

Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], Aviv Regev[2,3,5] & Nir Yosef[1,4,5]

Some terminology: Cell identity, type, state, ..

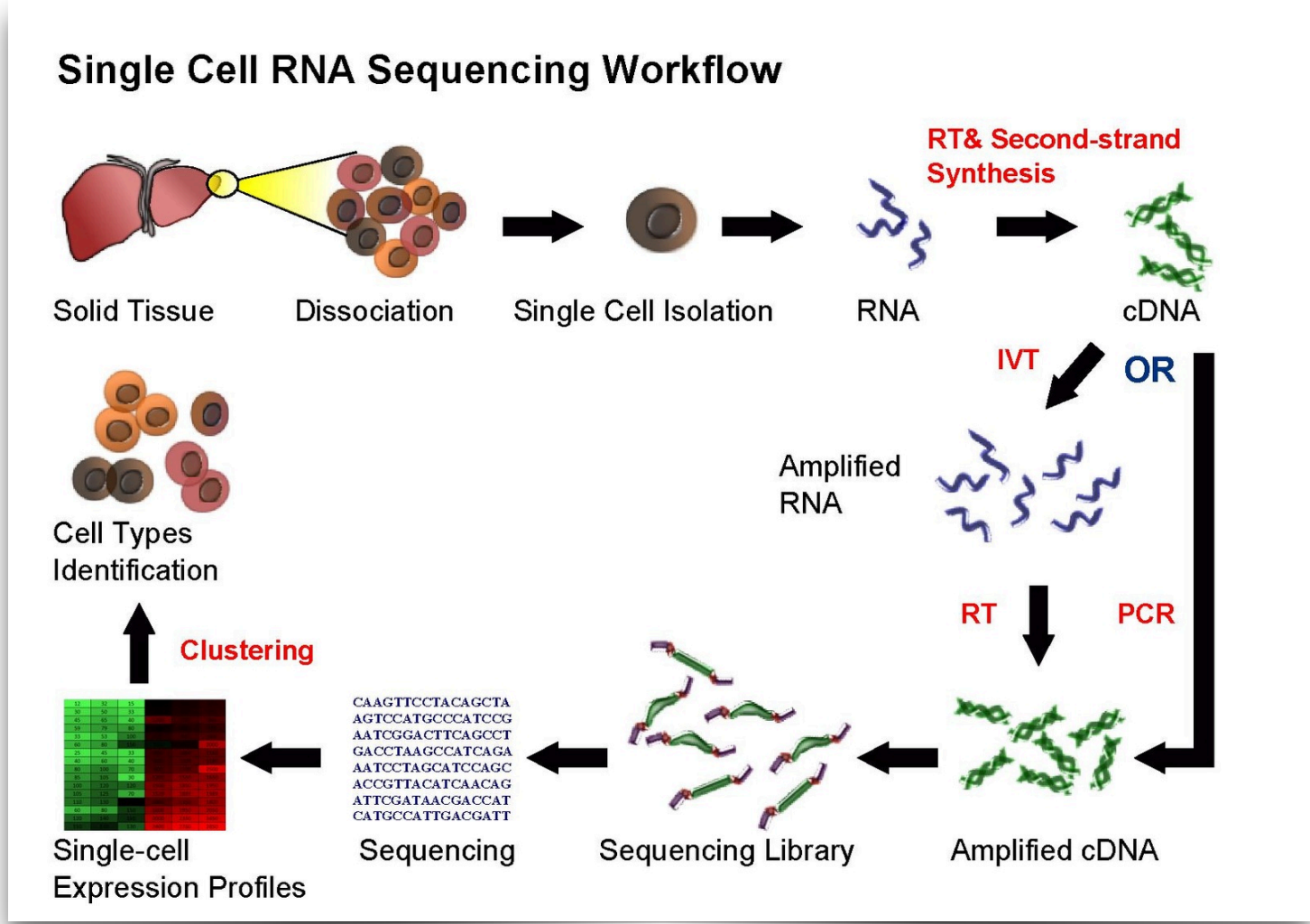
Box 1 The many facets of a cell's identity

We define a cell's *identity* as the outcome of the instantaneous intersection of all factors that affect it. We refer to the more permanent aspects in a cell's identity as its *type* (e.g., a hepatocyte typically cannot turn into a neuron) and to the more transient elements as its *state*. Cell types are often organized in a hierarchical taxonomy, as types may be further divided into finer subtypes; such taxonomies are often related to a cell fate map, reflecting key steps in differentiation. Cell *states* arise transiently during time-dependent processes, either in a *temporal progression* that is unidirectional (e.g., during differentiation, or following an environmental stimulus) or in a *state vacillation* that is not necessarily unidirectional and in which the cell may return to the origin state. Vacillating processes can be *oscillatory* (e.g., cell-cycle or circadian rhythm) or can transition between states with no predefined order (e.g., due to stochastic, or environmentally controlled, molecular events). These time-dependent processes may occur transiently within a stable cell type (as in a transient environmental response), or may lead to a new,

Type: permanent

State: transient

https://en.wikipedia.org/wiki/Single_cell_sequencing

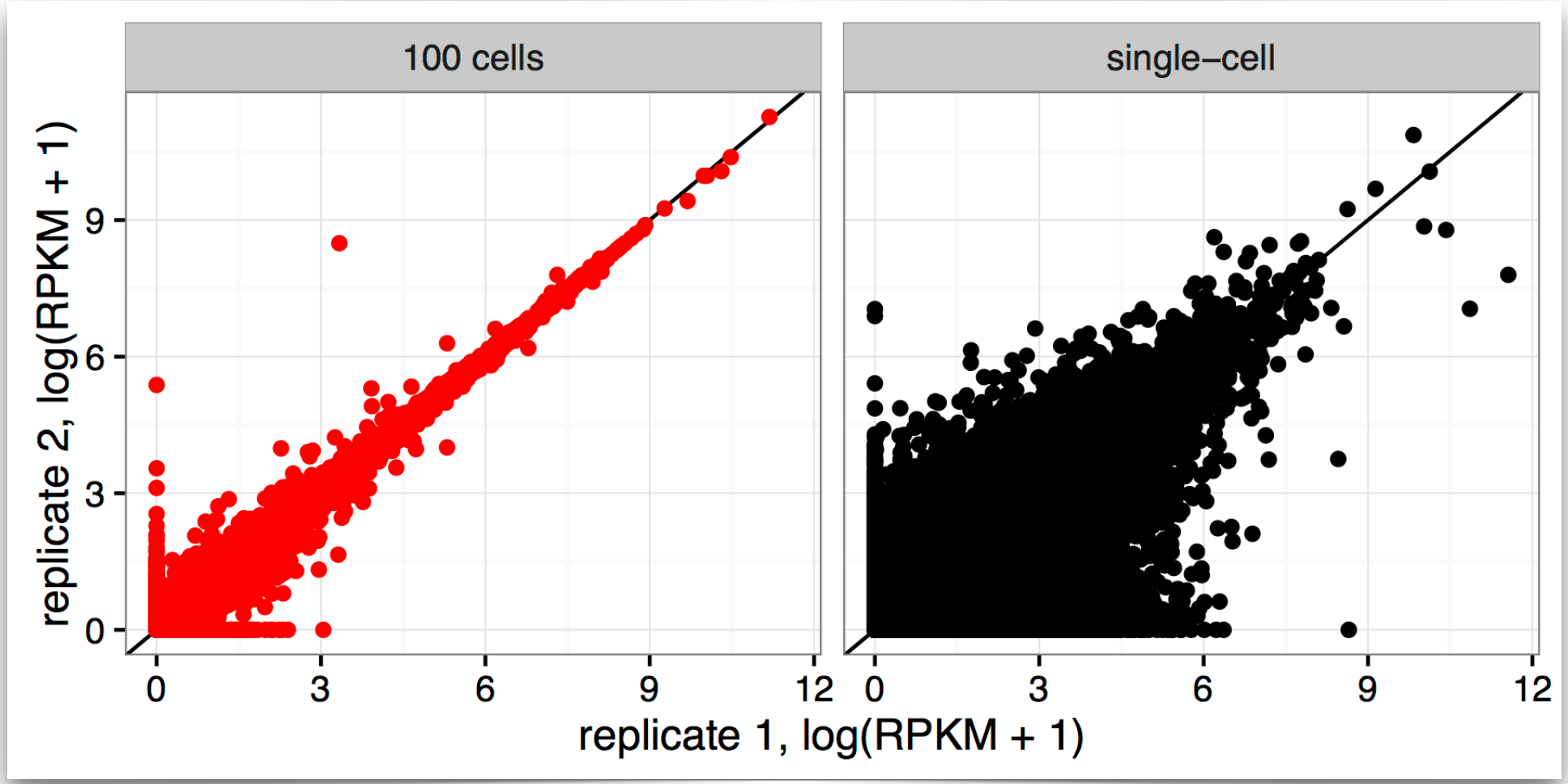


Droplet-based single cell genomics/transcriptomics

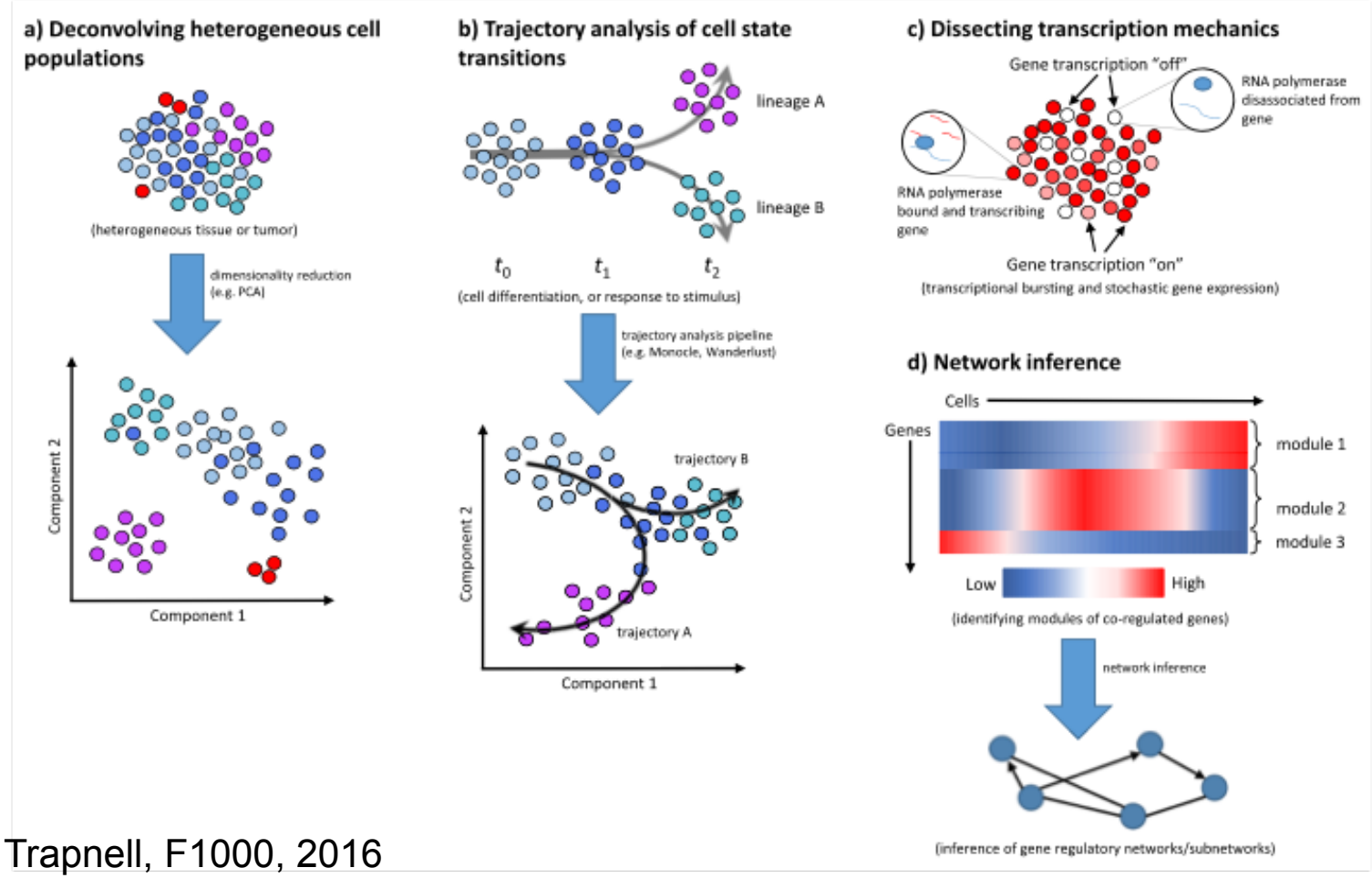
Using oil droplets loaded with reagents ..
Show 10x genomics video: https://10xgenomics.wistia.com/medias/f75ht43w1q

1:25-2:05

Basic properties: Variability levels



Tasks



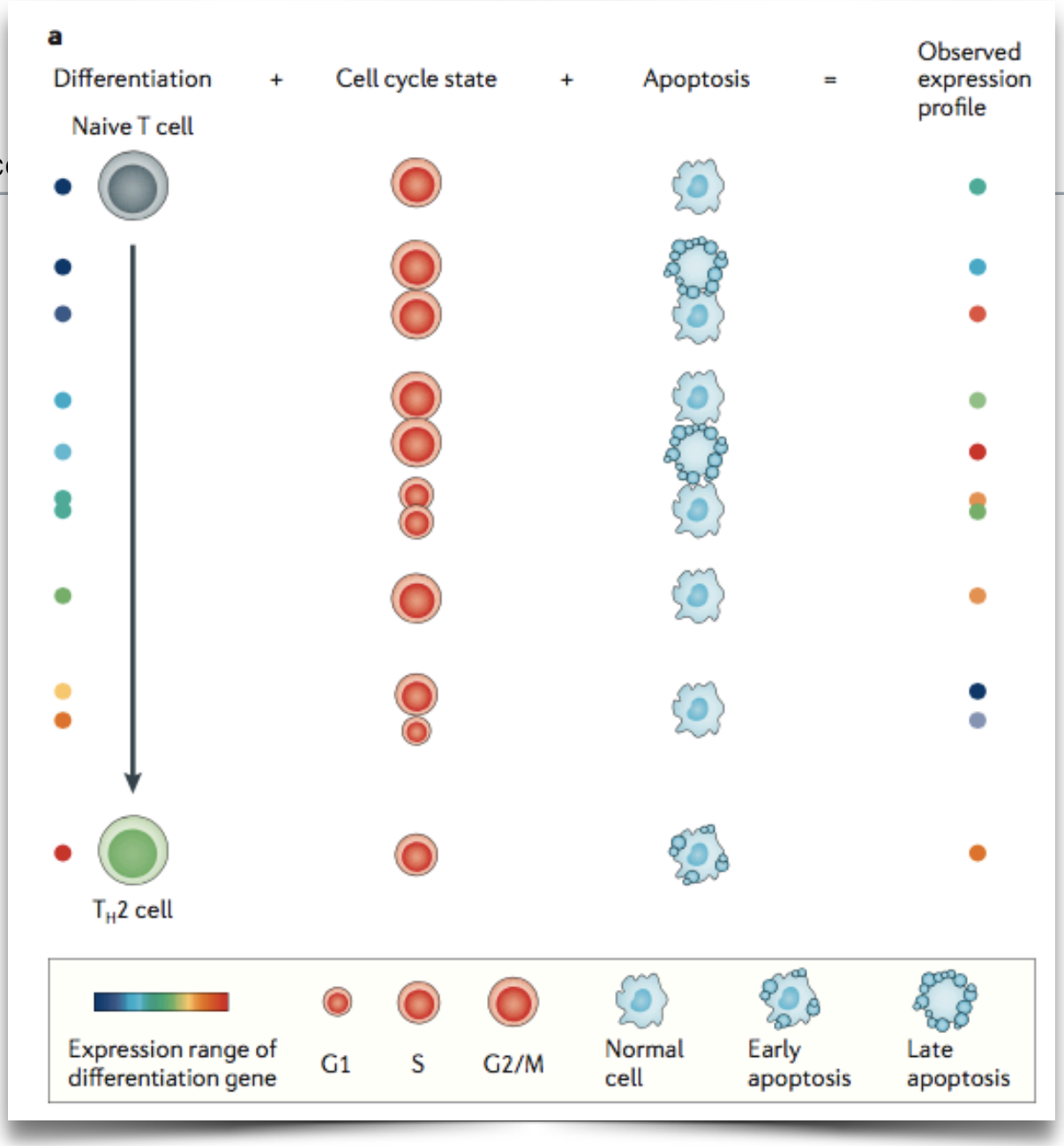
Liu and Trapnell, F1000, 2016

Measurements are a convolution of other signals

More specifically, for any gene g that is annotated to the hidden factor under consideration, its expression profile y_g across cells is modeled as

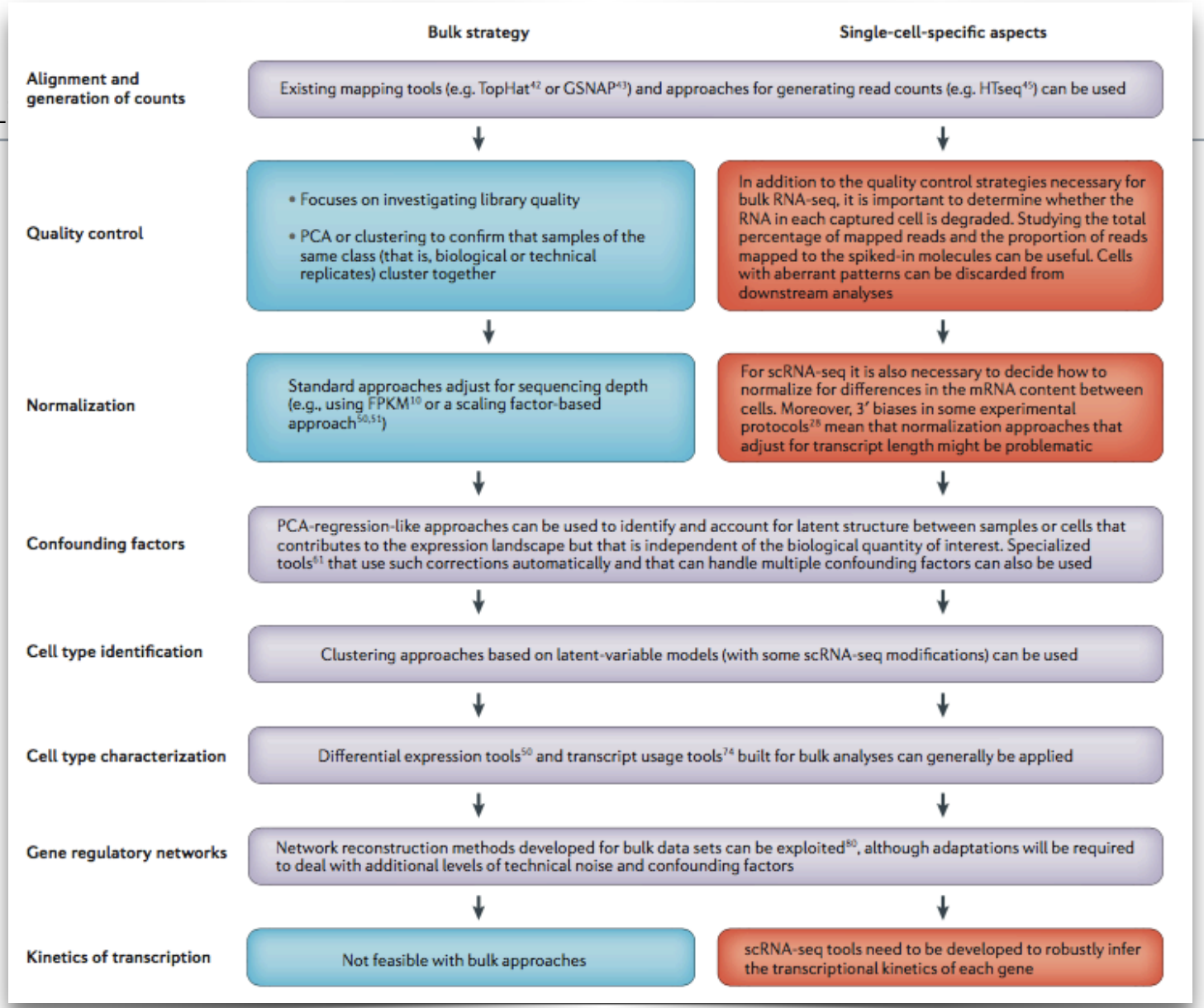
$$y_g \sim \mathcal{N}(\mu_g \mathbf{1}, XX^T + \sigma_v^2 CC^T + v_g^2 \mathrm{II}) \quad (1)$$

where X represents the hidden factor (such as cell cycle), C corresponds to additional observed covariates (if available) and v_g^2 denotes the residual variance. Because the same distributional assumptions are shared across a large set of genes in the annotated set, the state of the hidden variables X and the remaining covariance parameters can be robustly inferred by means of standard maximum likelihood approaches (**Supplementary Notes**). Once X is inferred, we calculate the covariance structure between cells, which is induced by the hidden factor as $\Sigma = XX^T$.



Stegle et al., NBT, 2015; Buettner et al., NBT, 2015

Many steps in the (scRNA-seq) pipeline are the same / similar to bulk.



Stegle et al., NBT, 2015

scRNA-seq Gotcha #1: batch effects

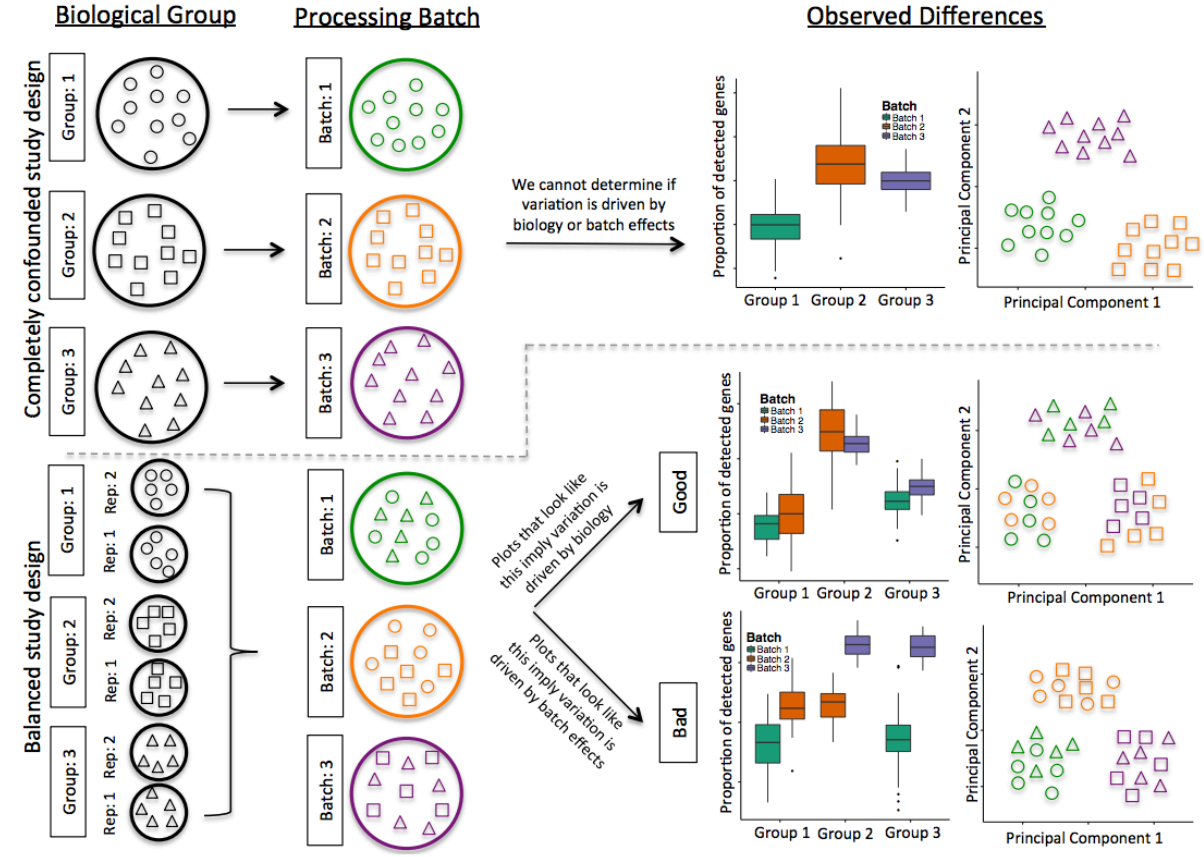
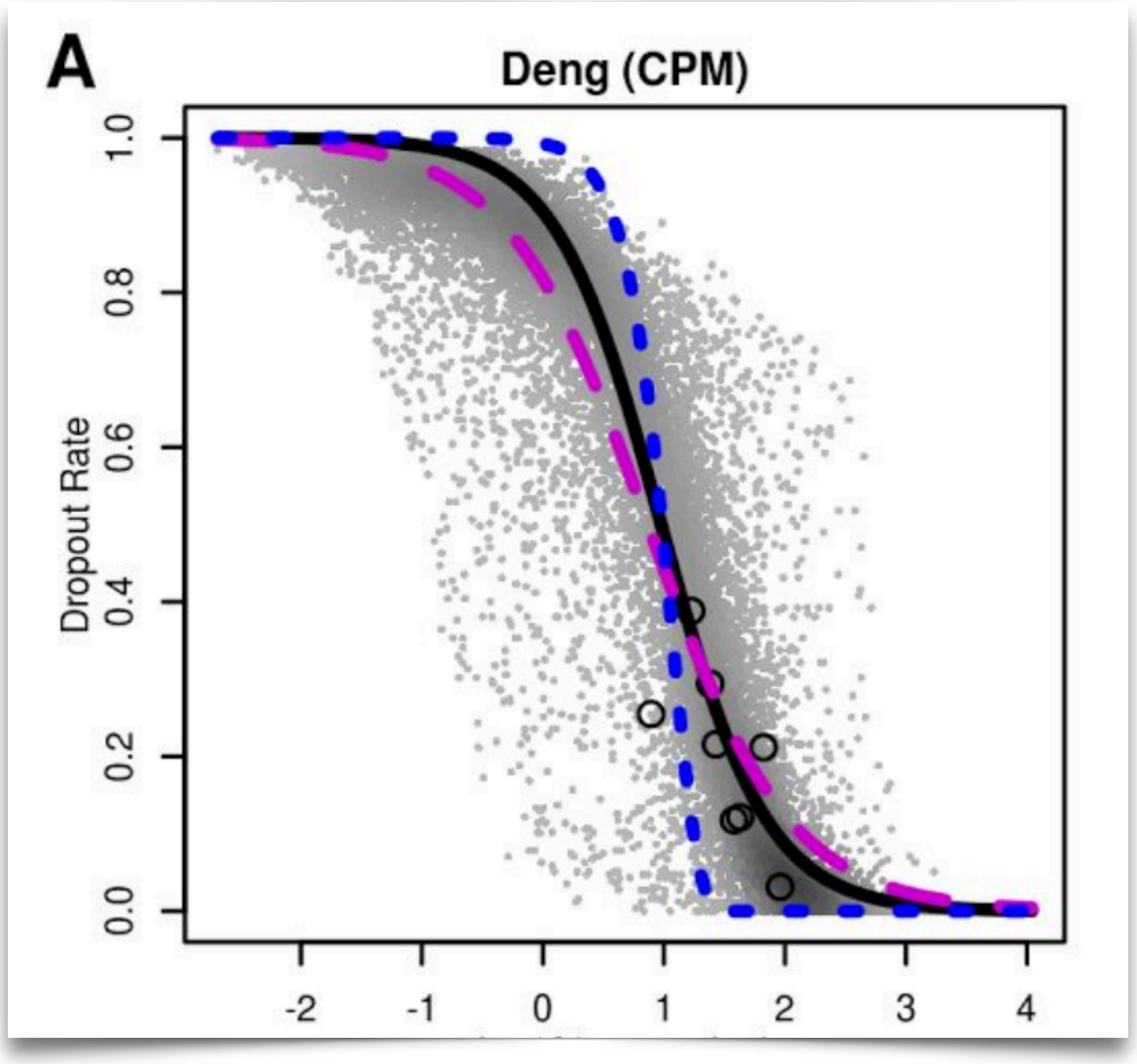


Figure 1: The problem of confounding biological variation and batch effects. The top section depicts a completely confounded study design of processing individual cells from three biological groups (represented by shapes) in three separate batches (represented by colors). In this case, we cannot determine if biology or batch effects drive the observed variation. The bottom section depicts a balanced study design consisting of multiple replicates (rep) split and processed across multiple batches. The use of multiple replicates allows observed variation be attributed to biology (cells cluster by shape) or batch effects (cells cluster by color).

n.b.: some debate recently about the prominence of dropout in the newer droplet-based platforms.

scRNA-seq #2: dropout

(A) The MichaelisMenten (solid black), logistic (dashed purple), and double exponential (dotted blue) models are fit to Deng dataset. Expression (counts per million) was averaged across all cells for each gene (points) and the proportion of expression values that were zero was calculated. ERCC spikeins are shown as open black circles.



Andrews and Hemberg, biorxiv, 2016

Differential expression: zero inflation / model dropout, mixture models, etc.

hurdle model

Single-cell RNA-seq hurdle model

We model the $\log_2(\text{TPM}+1)$ expression matrix as a two-part generalized regression model. The gene expression rate was modeled using logistic regression and, conditioning on a cell expressing the gene, the expression level was modeled as Gaussian.

Given normalized, possibly thresholded (see Additional file 1), scRNA-seq expression $Y=[y_{ig}]$, the rate of expression and the level of expression for the expressed cells are modeled conditionally independent for each gene g. Define the indicator $Z=[z_{ig}]$, indicating whether gene g is expressed in cell i (i.e., $z_{ig}=0$ if $y_{ig}=0$ and $z_{ig}=1$ if $y_{ig}>0$). We fit logistic regression models for the discrete variable Z and a Gaussian linear model for the continuous variable $(Y \mid Z=1)$ independently, as follows:

$$logit\left(Pr(Z_{ig}=1)\right)=\mathrm{X_i}\,\beta_g^{\mathrm{D}}$$

$$\Pr\left(Y_{ig}=y|Z_{ig}=1\right)=\mathrm{N}\left(\mathrm{X_i}\beta_g^{\mathrm{C}},\ \sigma_g^2\right)$$

The regression coefficients of the discrete component are regularized using a Bayesian approach as implemented in the *bayesglm* function of the *arm* R package, which uses weakly informative priors [30] to provide sensible estimates under linear separation (See Additional file 1 for details). We also perform regularization of the continuous model variance parameter, as described below, which helps to increase the robustness of gene-level differential expression analysis when a gene is only expressed in a few cells.

Finak et al., Genome Biology, 2015

mixture model

Differential expression analysis. With a Bayesian approach, the posterior probability of a gene being expressed at an average level x in a subpopulation of cells S was determined as an expected value (E) according to

$$p_S(x)=E\left[\prod_{c\in B}p(x\,|\,r_c,\Omega_c)\right]$$

where B is a bootstrap sample of S, and $p(x|r_c,\Omega_c)$ is the posterior probability for a given cell c, according to

$$p(x\,|\,r_c,\Omega_c)=p_d(x)p_{Poisson}(x)+(1-p_d(x))p_{NB}(x\,|\,r_c)$$

where p_d is the probability of observing a dropout event in cell c for a gene expressed at an average level x in S, $p_{Poisson}(x)$ and $p_{NB}(x|r_c)$ are the probabilities of observing expression magnitude of r_c in case of a dropout (Poisson) or successful amplification (NB) of a gene expressed at level x in cell c, with the parameters of the distributions determined by the Ω_c fit. For the differential expression analysis, the posterior probability that the gene shows a fold expression difference of f between subpopulations S and G was evaluated as

$$p(f)=\sum_{x\in X}p_S(x)p_G(fx)$$

where x is the valid range of expression levels. The posterior distributions were renormalized to unity, and an empirical P value was determined to test for significance of expression difference.

Kharchenko et al., Nature Methods, 2015

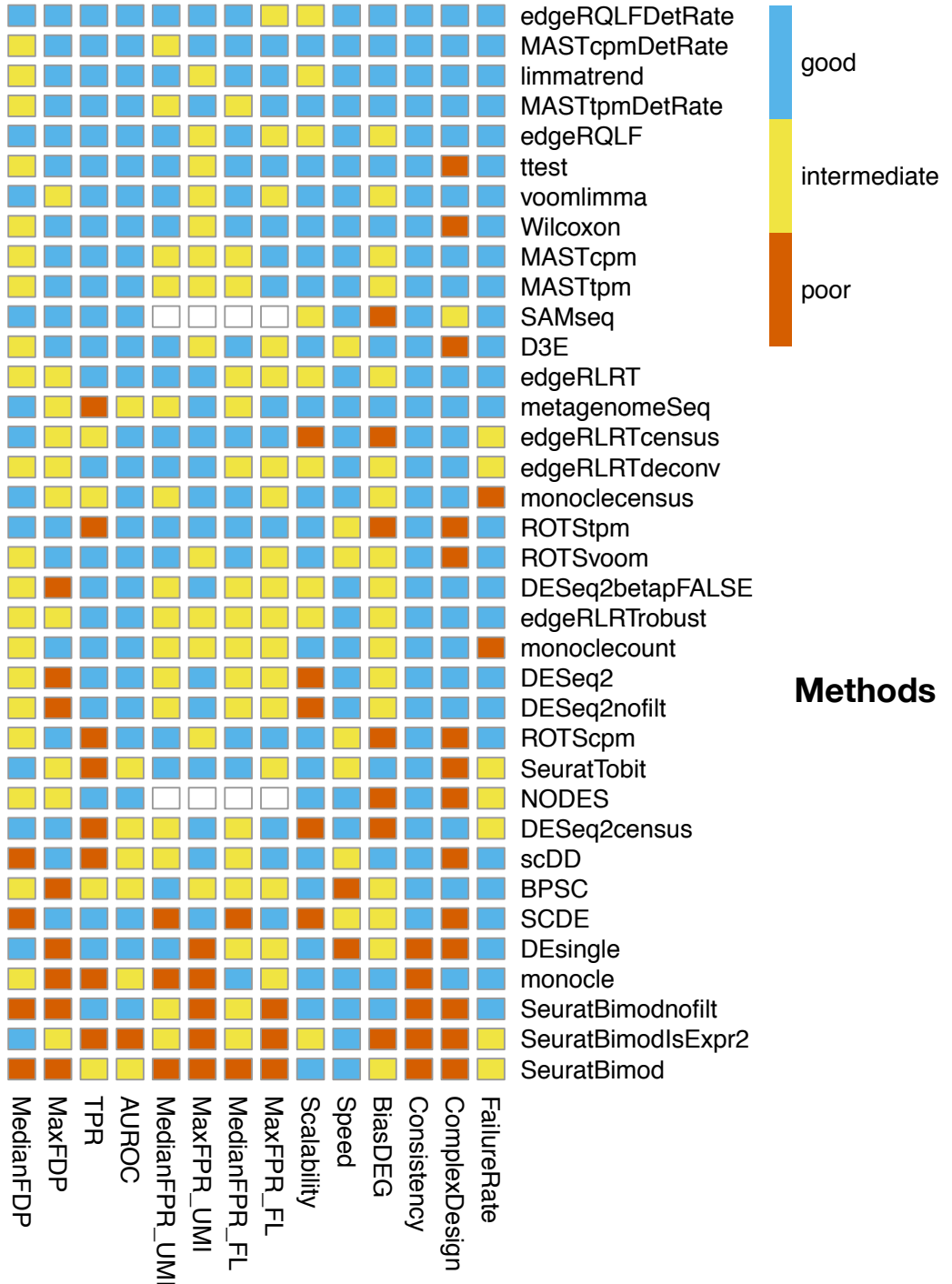
Between cell-type DE Benchmark (finding marker genes)

Bias, robustness and scalability in single-cell differential expression analysis

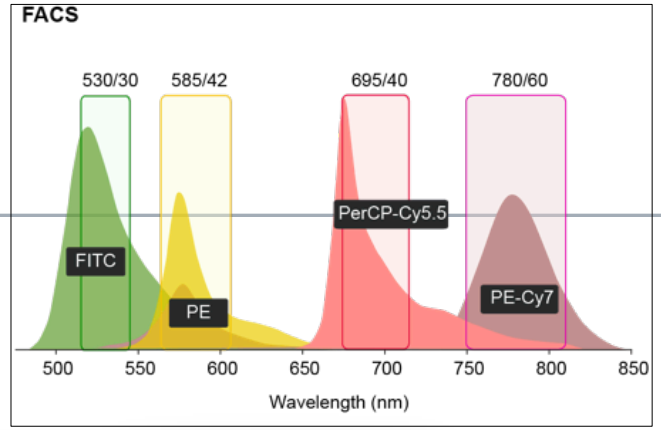
Charlotte Soneson[1,2] & Mark D Robinson[1,2]

RECEIVED 6 JUNE 2017; ACCEPTED 16 JANUARY 2018; PUBLISHED ONLINE 26 FEBRUARY 2018;

"we found that bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq."



Flow cytometry



First, cells are stained with a panel of antibodies; these antibodies have a flourescent tag attached.

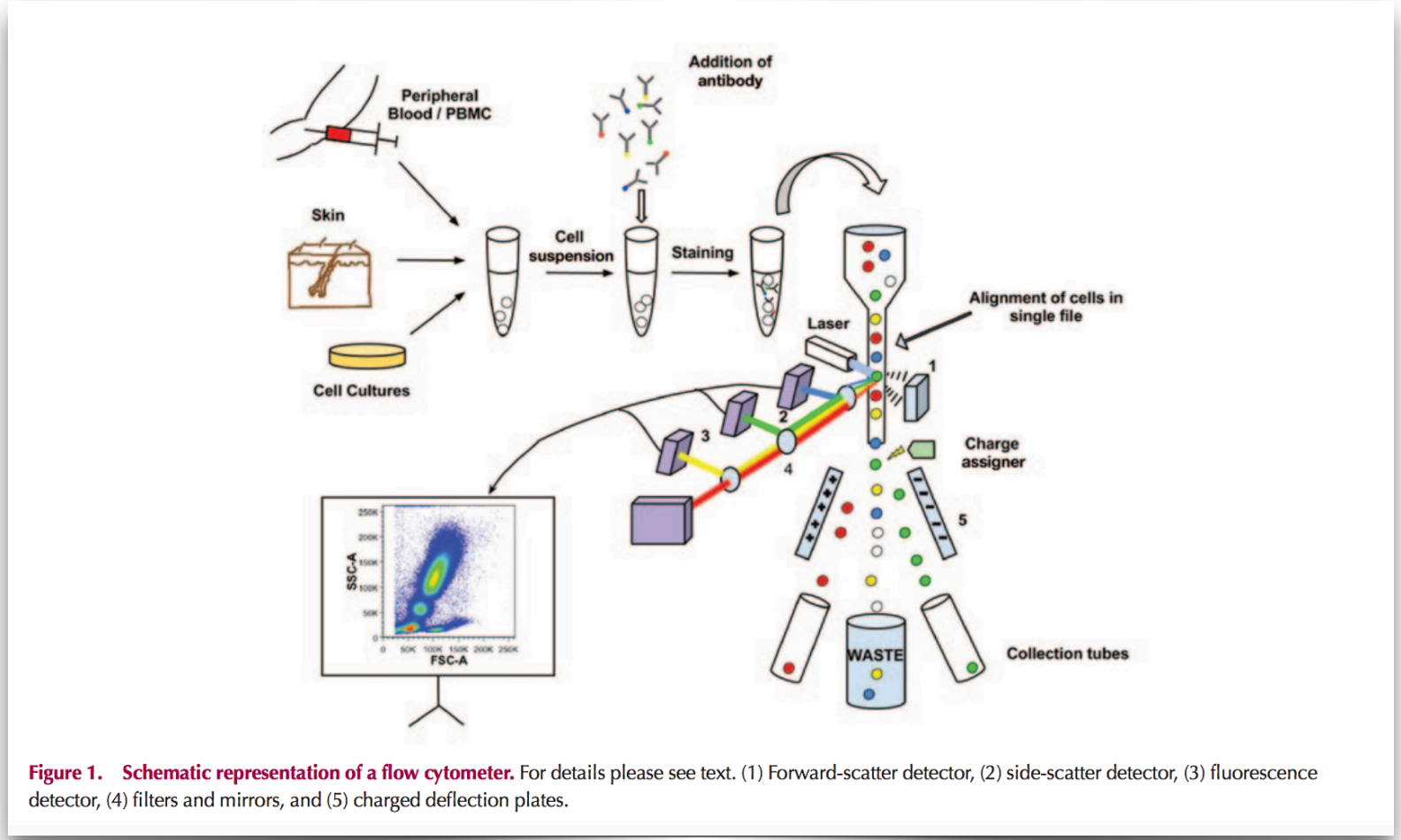
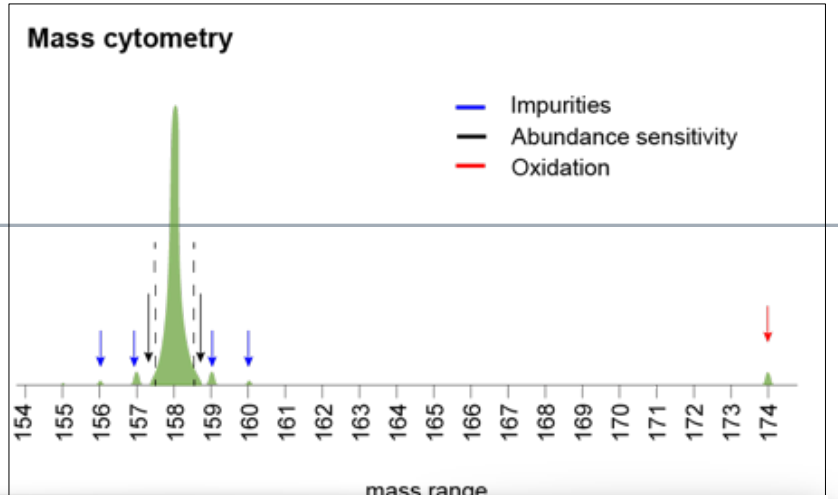
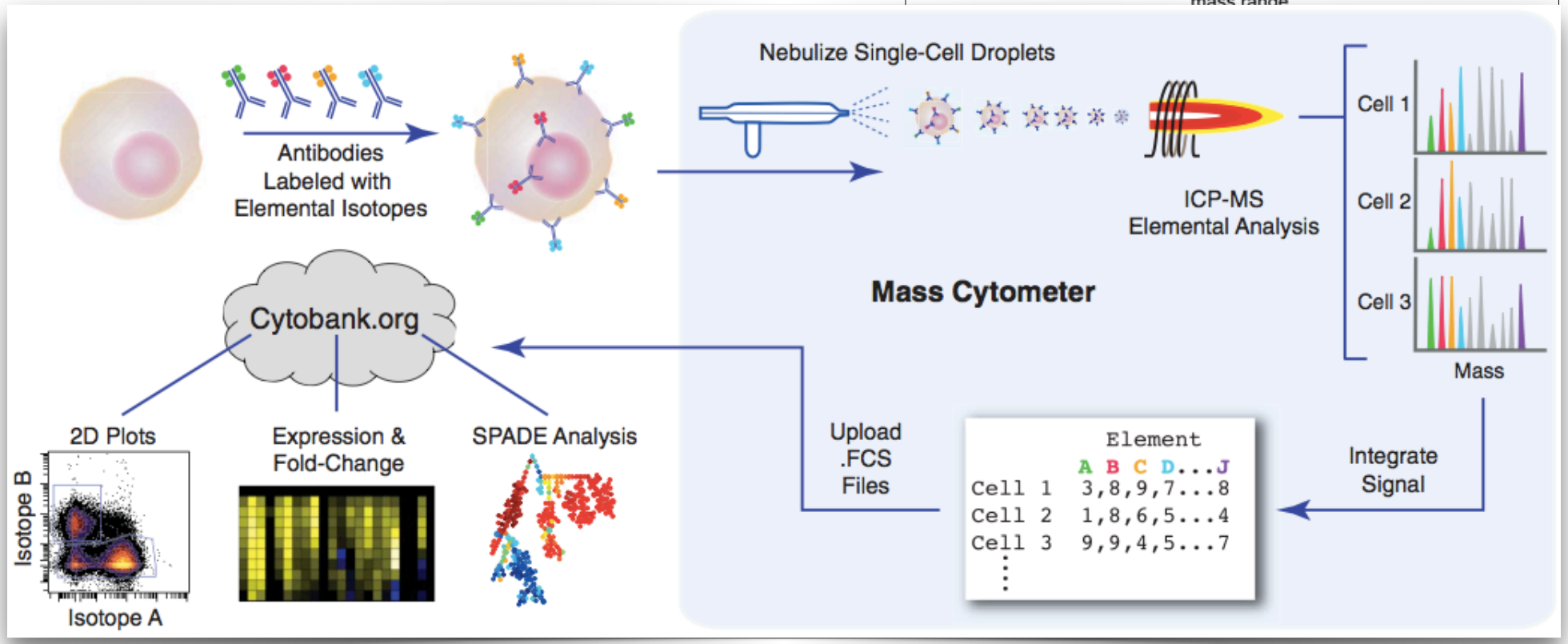


Figure 1. Schematic representation of a flow cytometer. For details please see text. (1) Forward-scatter detector, (2) side-scatter detector, (3) fluorescence detector, (4) filters and mirrors, and (5) charged deflection plates.

Jahan-Tigh et al., Journal of Investigative Dermatology, 2012

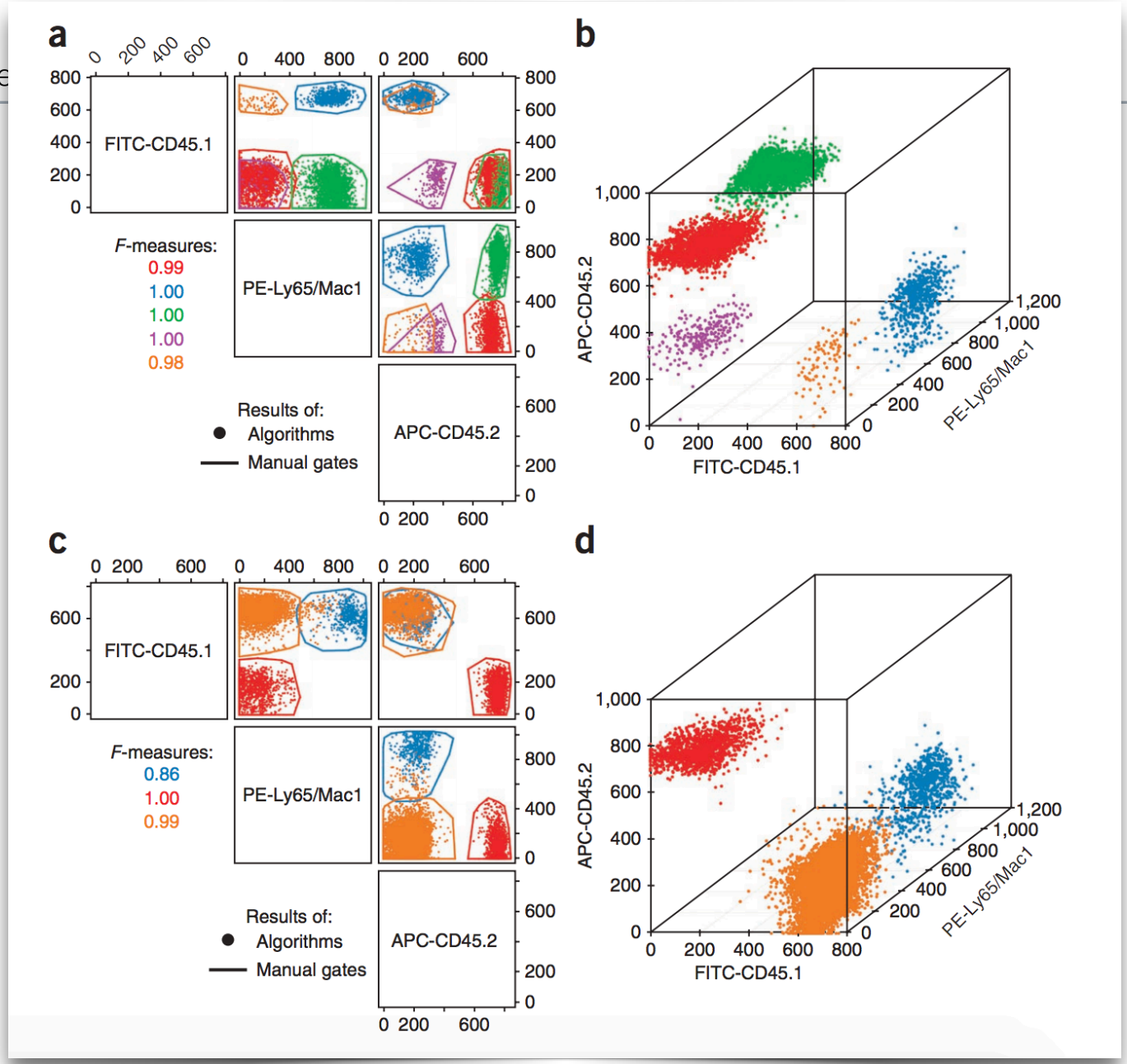
Mass cytometry (30-50 markers)





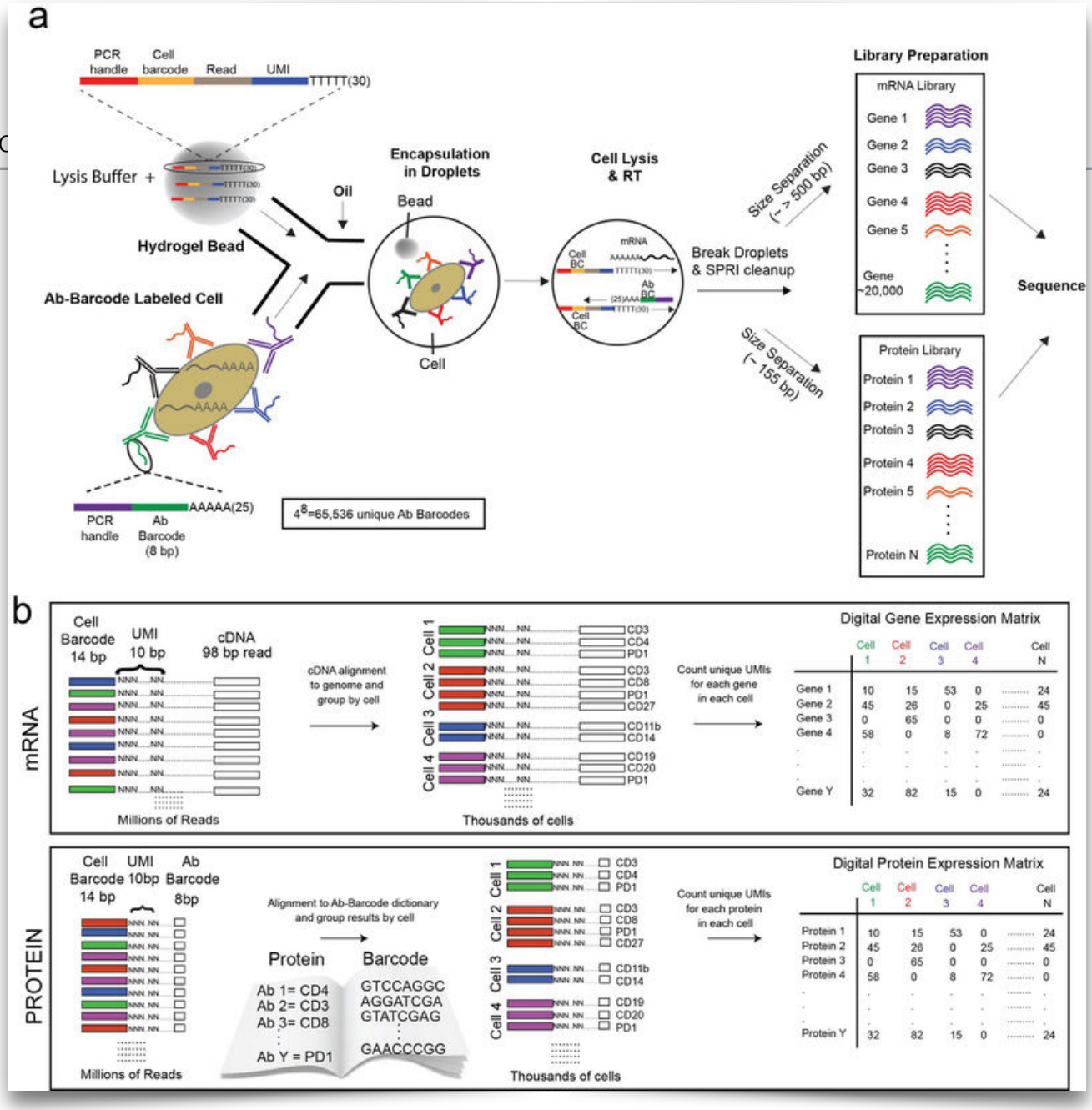
Manual gating versus clustering

Aghaeepour et al., Nature Methods, 2013

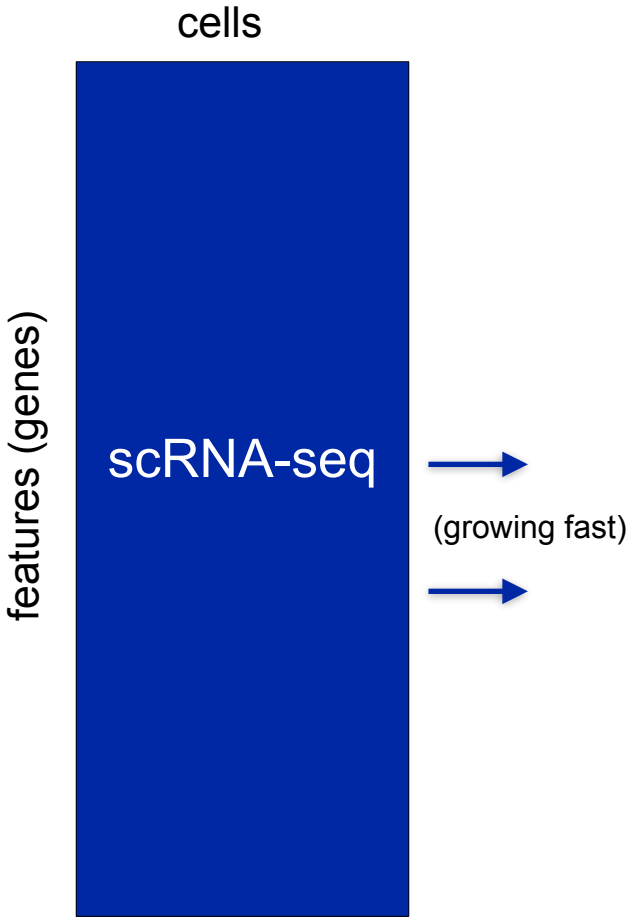


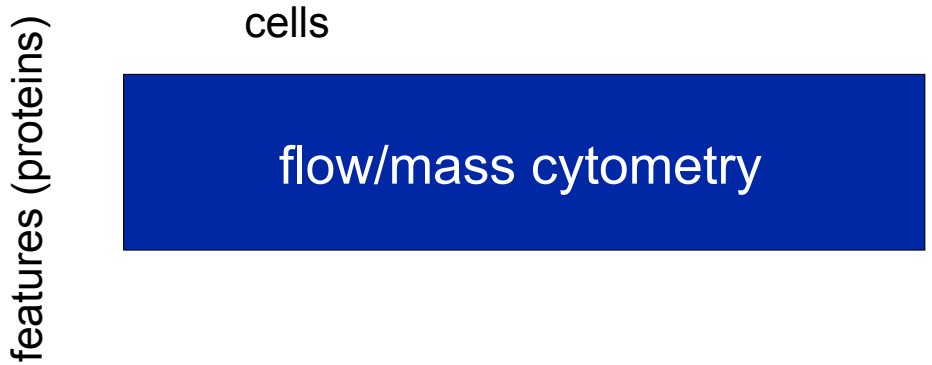
Dual assays

New assays (released in 2017) that measure both RNA and protein ..
REAP-seq
CITE-seq



Different shapes of single cell data





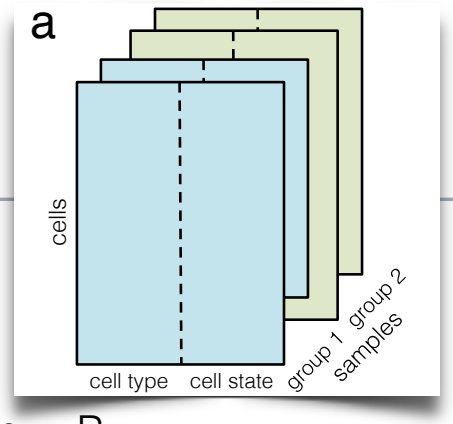
.. imaging ..

And different depths ..

Bulk RNA-seq: 10-30M reads per sample
scRNA-seq: 50k-100K (1M) reads per cell

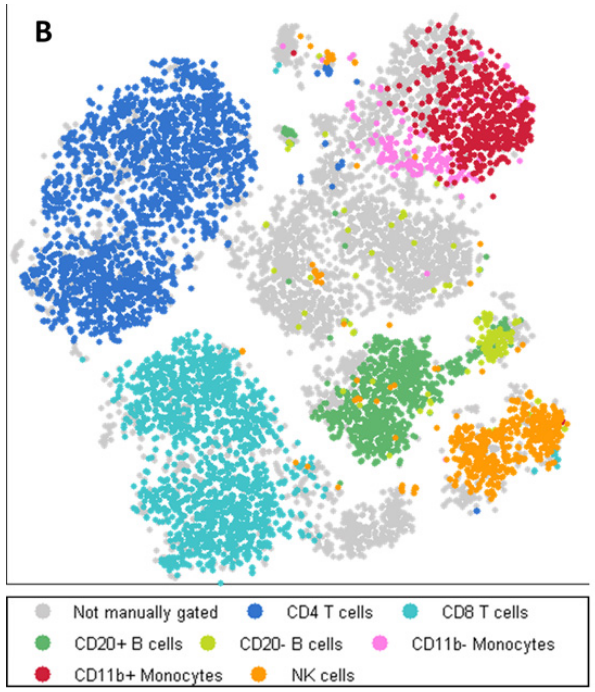
Themes common to many single-cell techniques

- Dimensionality reduction: PCA, diffusion maps, tSNE
- Clustering: hierarchical, SOMs, density-based etc.
- Inferring changes in abundance between cell types
- Trajectory analyses



Cytometry data: Dimension reduction useful in multiple directions: cells + samples

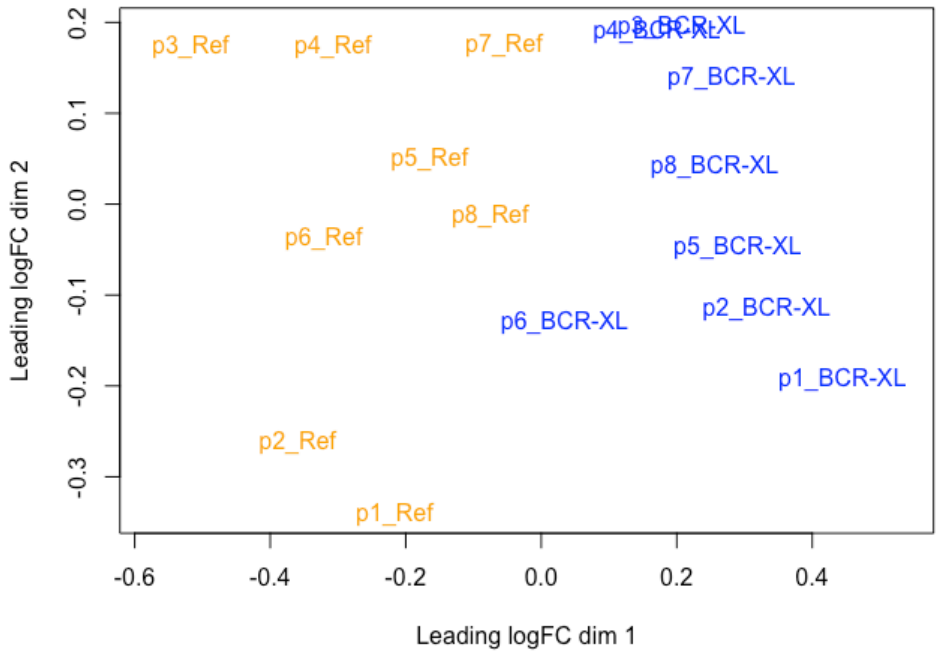
N cells x K markers —> N cells x 2 dimensions



Amir et al. 2013, Nat Biotech

Each point = **cell**

M median marker expression x P samples —> P samples x 2 dimensions



Each point = **sample**

Dimensionality reduction (generally)

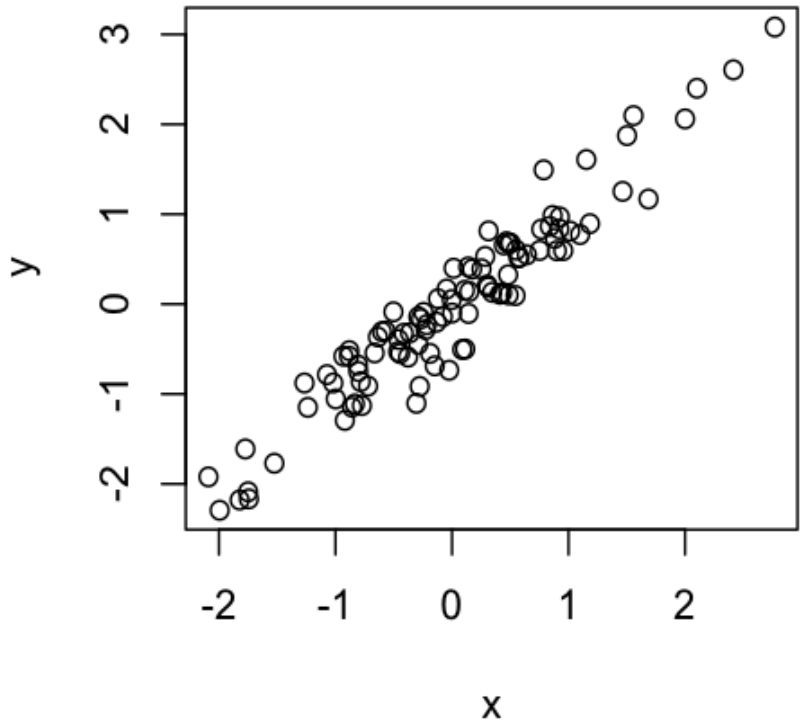
Techniques exist to **project** high-dimensional data (typical situation: 15k-20k gene expression measurements for each of N cells/samples) into a small number of dimensions (2 or 3)

Many techniques: **linear PCA**, multidimensional scaling, t-distributed stochastic neighbor embedding (**tSNE**)

Linear PCA: uses a linear combination of original variables such that the components decrease in variability (highest variance first) and are orthogonal to previous dimensions. Often, first 2 or 3 are used.

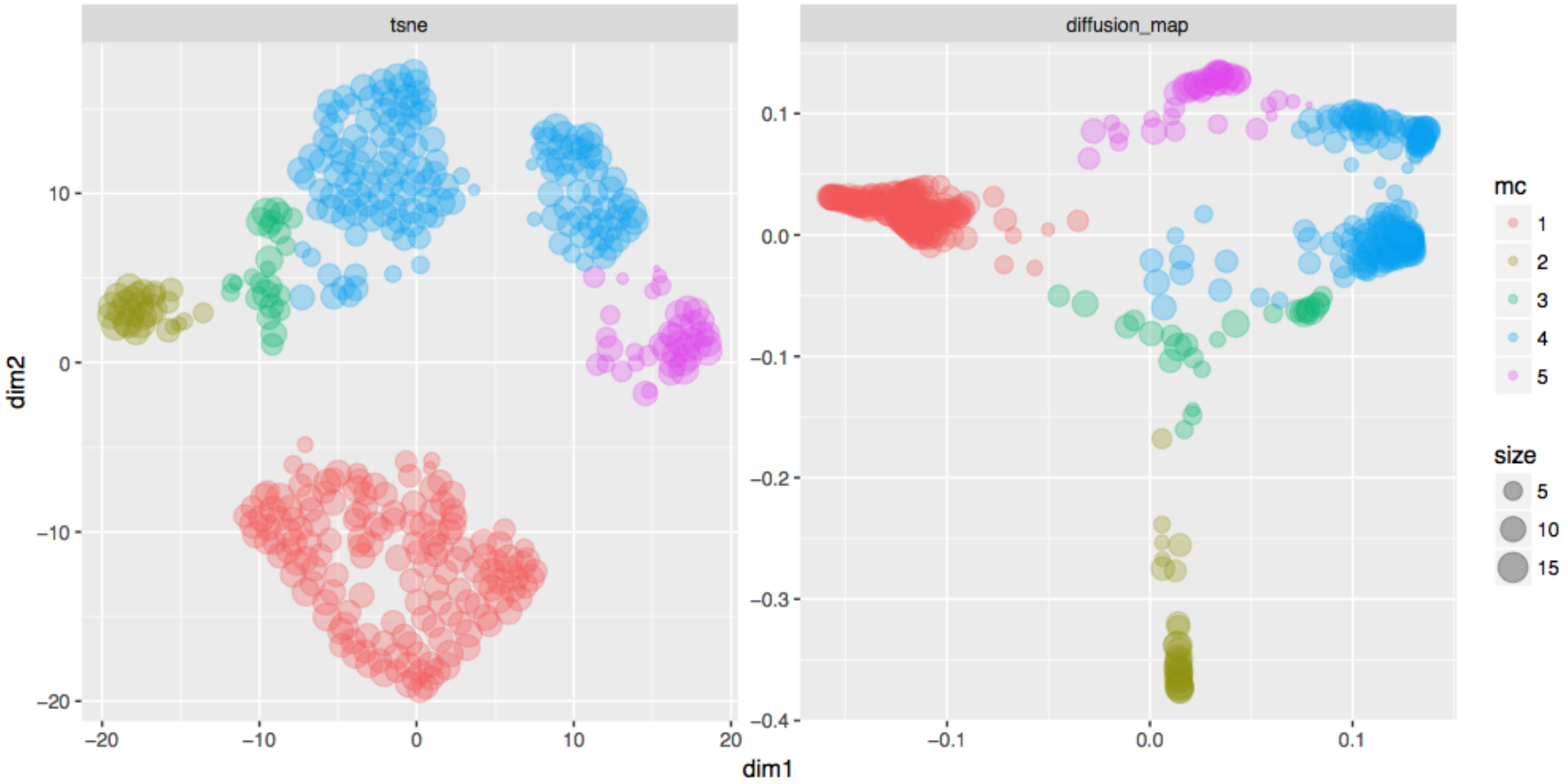
Visual explanation:

http://setosa.io/ev/principal-component-analysis/



http://distill.pub/2016/misread-tsne/

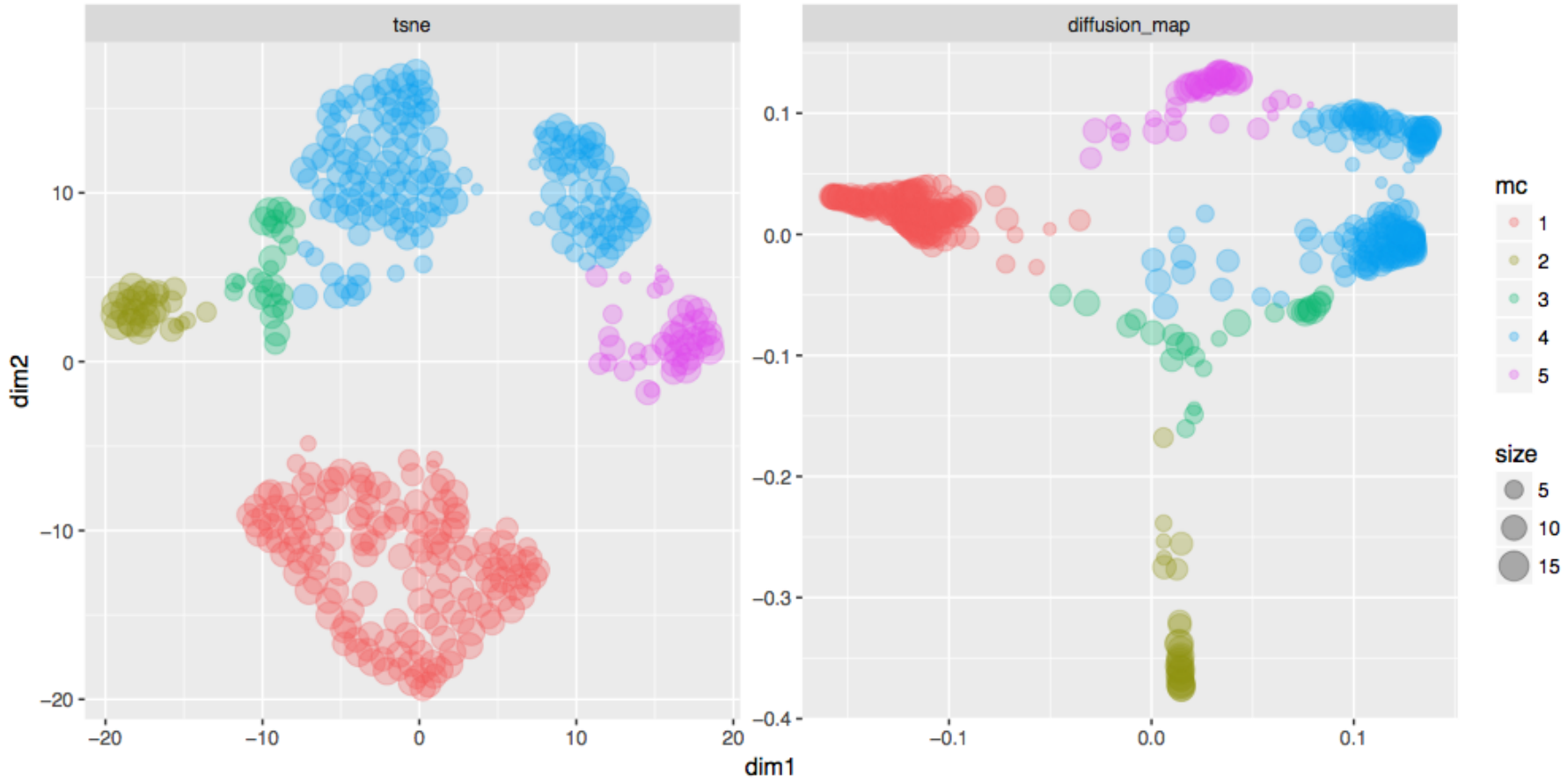
tSNE (t-dist'd stochastic neighbour embedding) + diffusion maps



"if you haven't encountered t-SNE before, here's what you need to know about the math behind it. The goal is to take a set of points in a high-dimensional space and find a faithful representation of those points in a lower-dimensional space, typically the 2D plane. The algorithm is non-linear and adapts to the underlying data, performing different transformations on different regions. Those differences can be a major source of confusion."

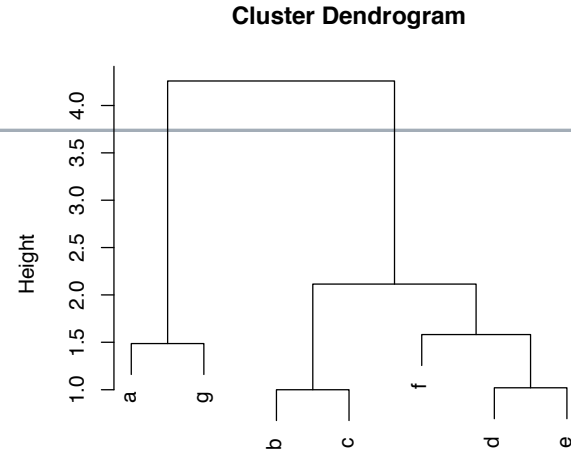
tSNE (t-dist'd stochastic neighbour embedding) + diffusion maps

"Given data in a high-dimensional space .. find parameters that describe the lower-dimensional structures of which it is comprised. Unlike other popular methods such as PCA and MDS, diffusion maps are non-linear and focus on discovering the underlying manifold (lower-dimensional constrained "surface" upon which the data is embedded). By integrating local similarities at different scales, a global description of the data-set is obtained.



http://en.wikipedia.org/wiki/Hierarchical_clustering

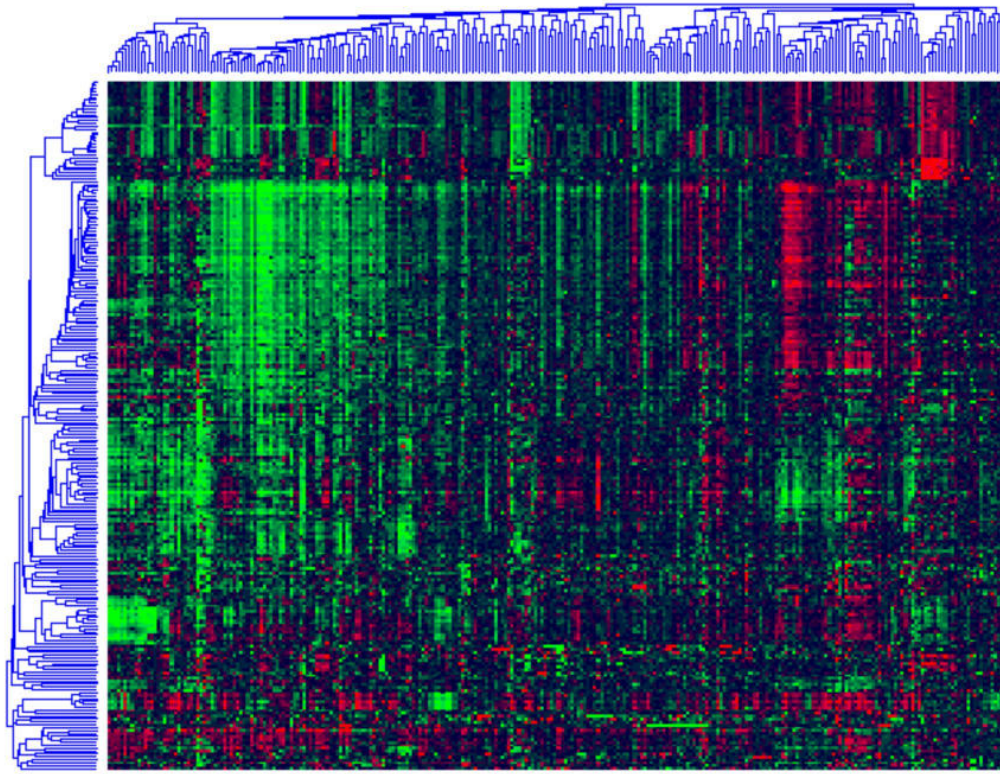
Hierarchical (Agglomerative) Clustering



Divisive (all features start as 1 cluster, then subsequently split) versus Agglomerative (every feature is it's own cluster, then subsequently merged)

Metric: to define how similar any two vectors are.

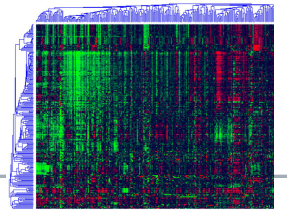
Linkage: determines how clusters are merged into a tree



http://en.wikipedia.org/wiki/Euclidean_distance

Euclidean distance:

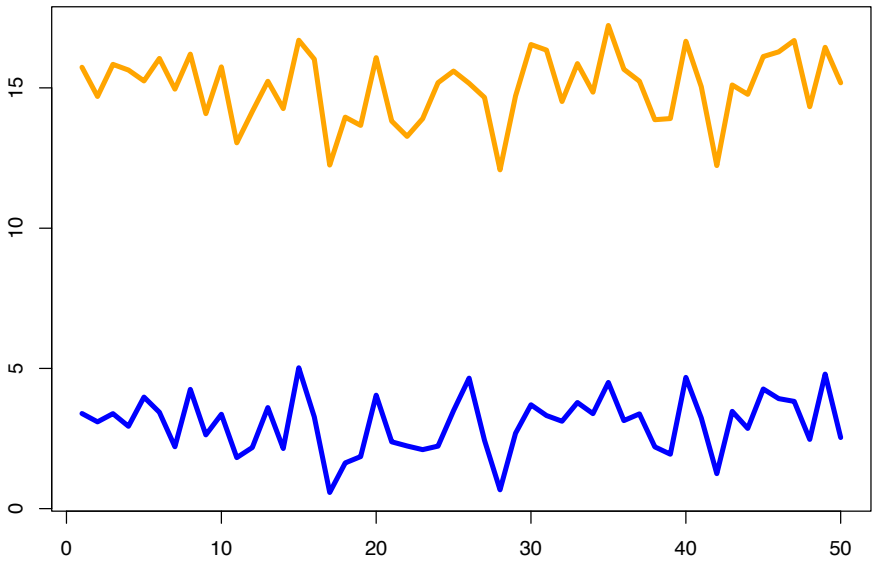
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

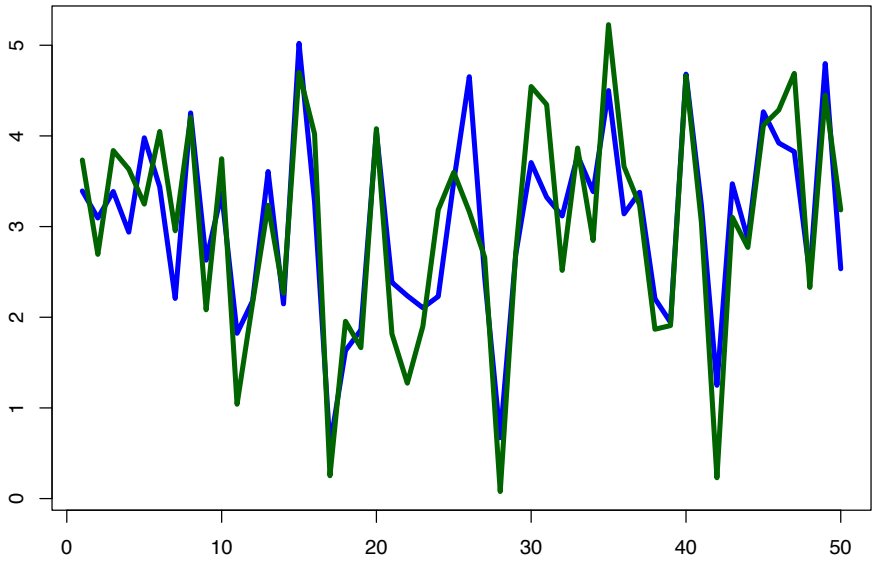


Are these “vectors” similar ?

```
> sqrt(sum((x-(y-12))^2))
[1] 3.926007
> sqrt(sum((x-y)^2))
[1] 84.84028
```

It depends how you define similar.

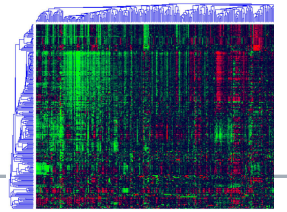




Euclidean distance: 84.84 3.92

http://en.wikipedia.org/wiki/Correlation_and_dependence

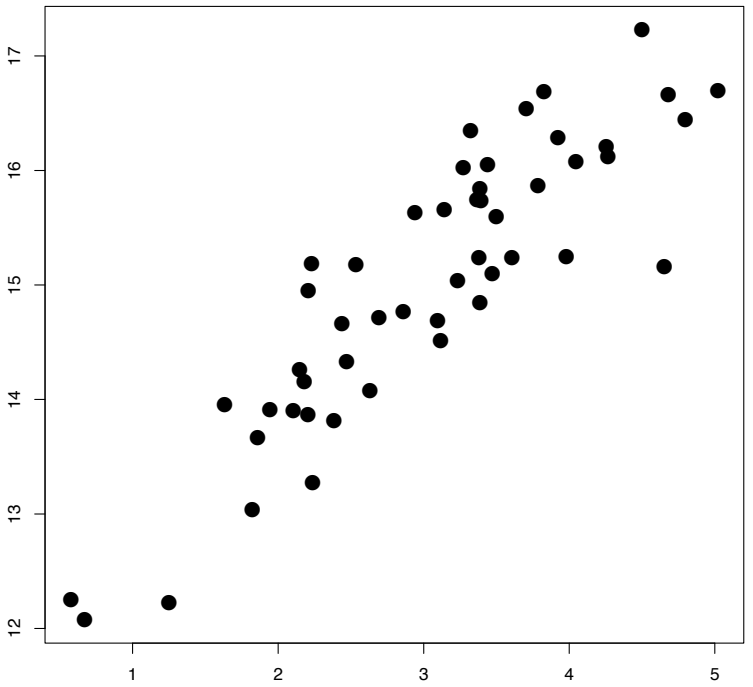
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

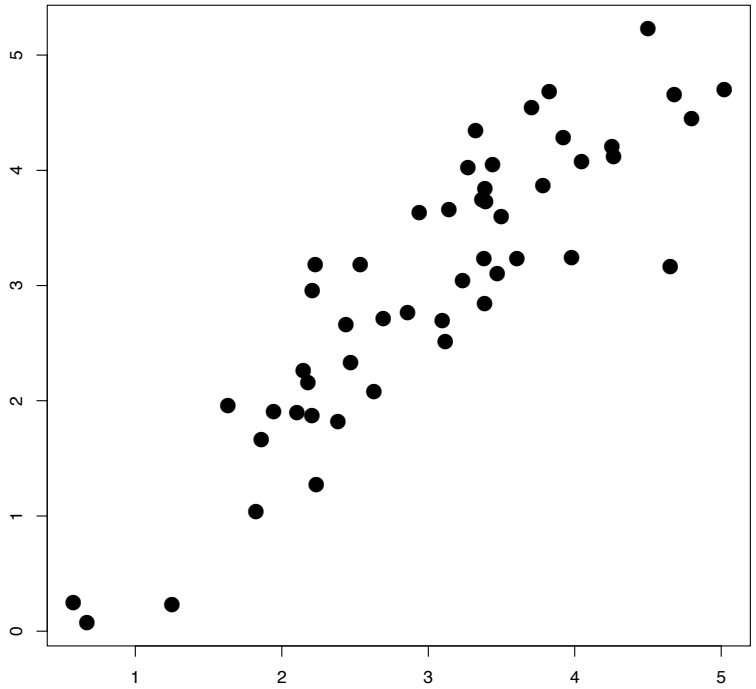


Are these “vectors” similar ?

```
> cor(x,y)
[1] 0.8901139
> cor(x,y-12)
[1] 0.8901139
```

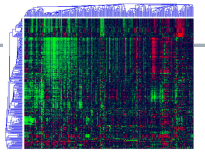
It depends how you define similar.





Correlation: 0.89 0.89

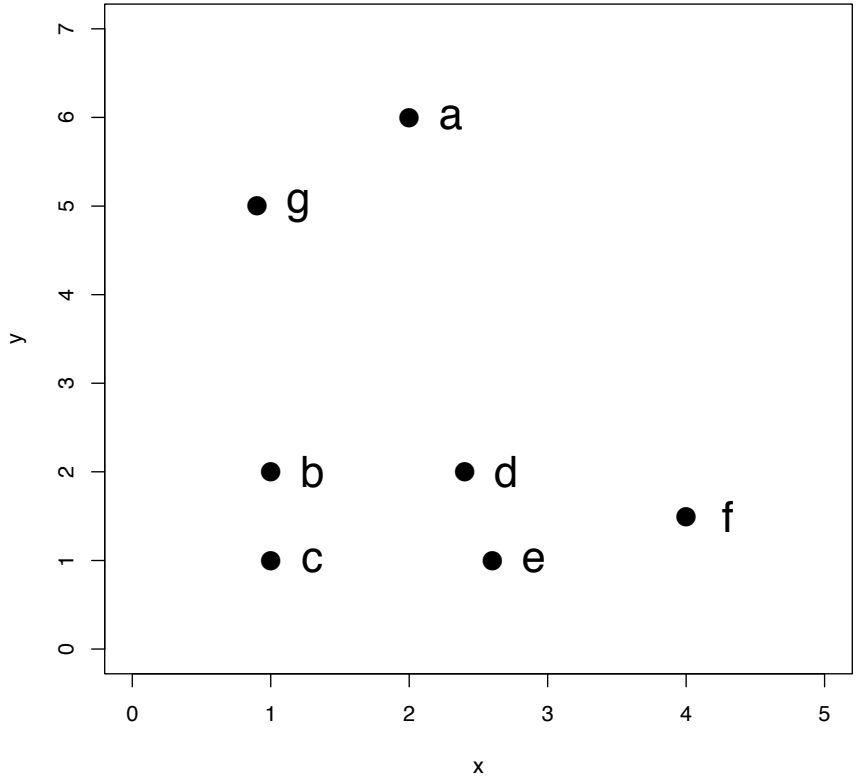
http://en.wikipedia.org/wiki/Hierarchical_clustering



Hierarchical (Agglomerative) Clustering

Start with distances.

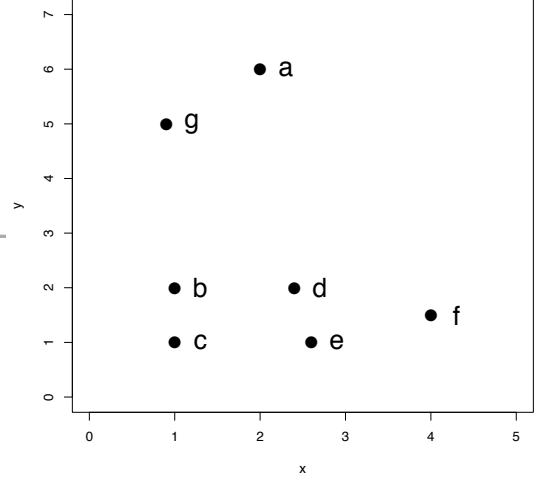
Linkage: determines how clusters are merged into a tree.

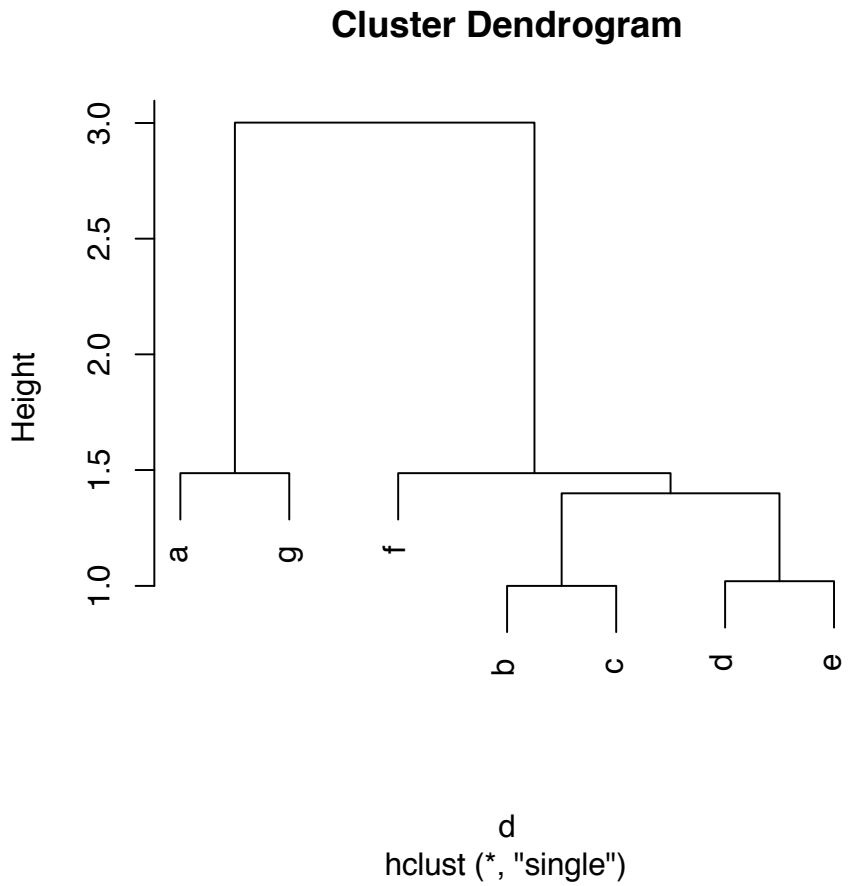


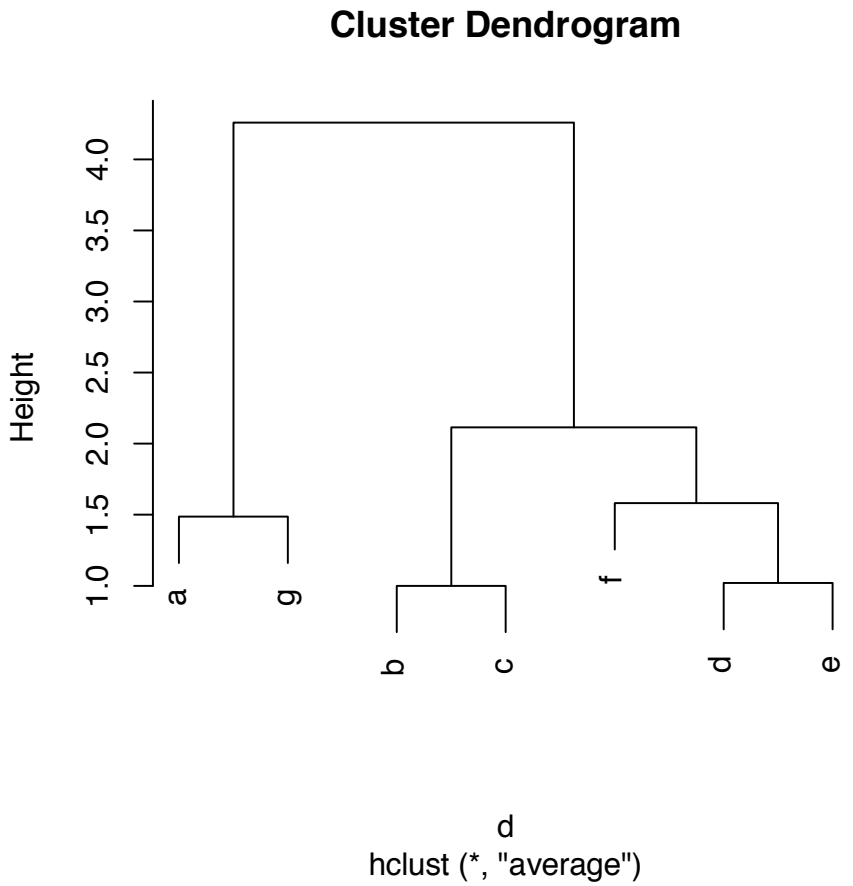
From eyeballing, here is a likely set of merges:

b,c
d,e
a,g,
(d,e),f
(b,c),((d,e),f)
ALL

Different linkages

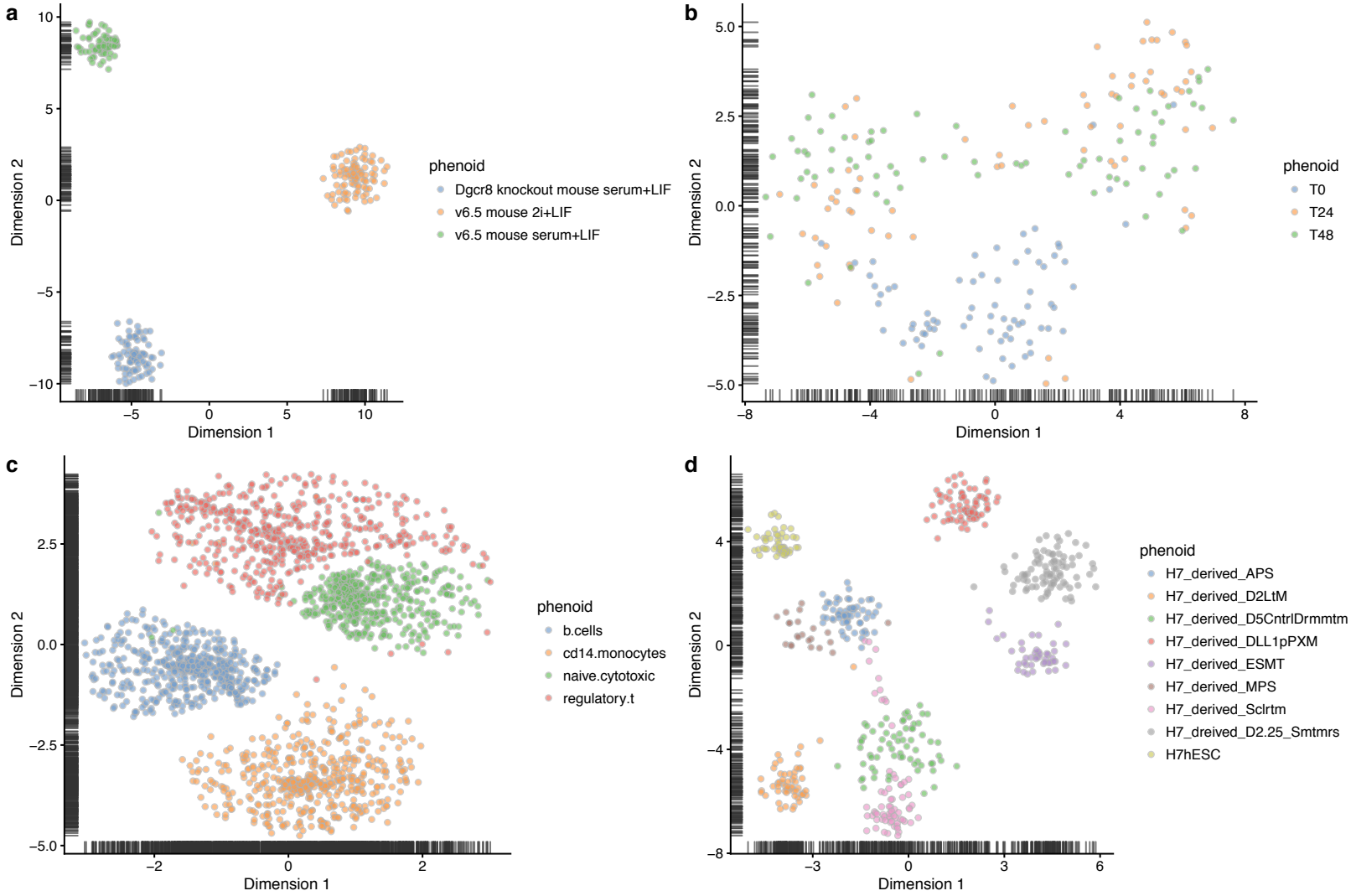






How to cluster scRNA-seq data and/or how to find cell type markers ?

- Our strategy: datasets from conquer (http://imlspenticton.uzh.ch:3838/conquer/) with **predefined groups**: range of difficulty



How to cluster scRNA-seq data?

F1000Research 2018, 7:1141 Last updated: 21 SEP 2018

Check for updates

RESEARCH ARTICLE

REVISED **A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 2; referees: 2 approved]**

Angelo Duò [1,2], Mark D. Robinson [1,2], Charlotte Soneson [1,2]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, 8057, Switzerland
[2]SIB Swiss Institute of Bioinformatics, Zurich, 8057, Switzerland

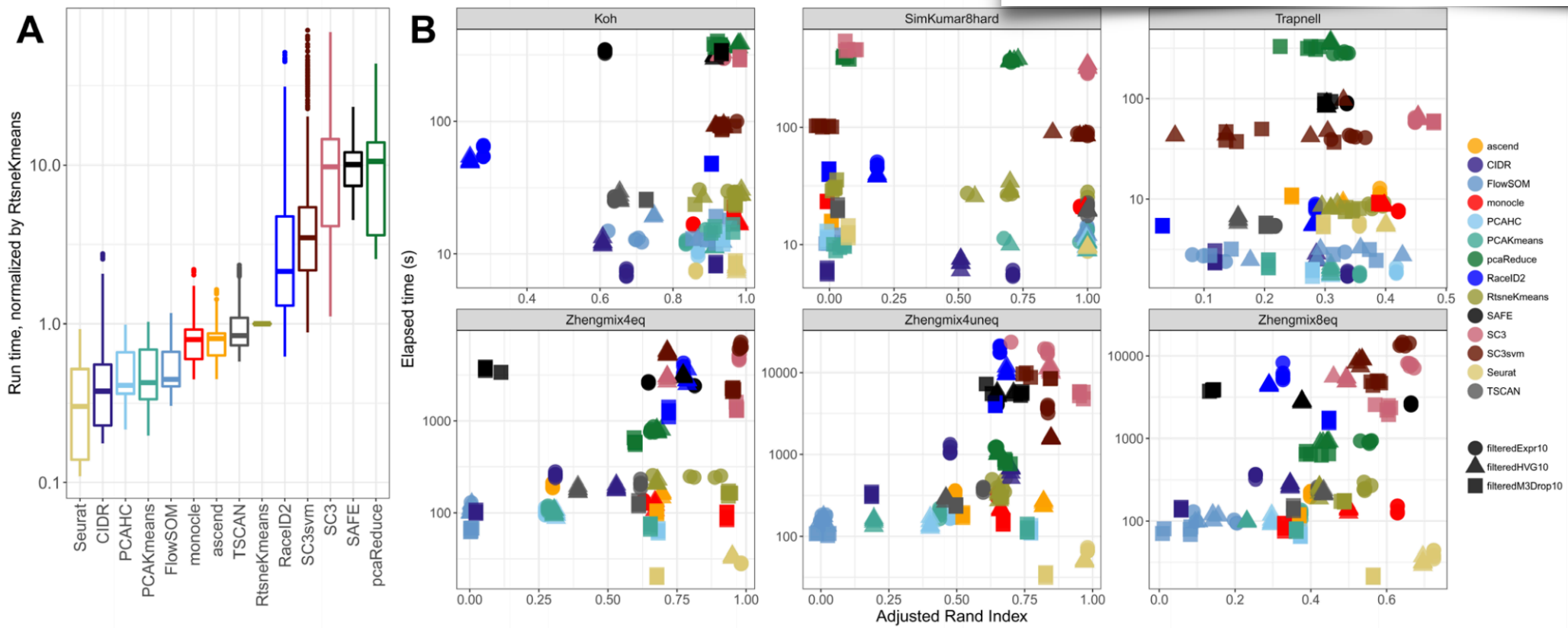


Figure 2. (**A**) Normalized run times, using `RtsneKmeans` as the reference method, across all data set instances and number of clusters. (**B**) Run time versus performance (ARI) for a subset of data sets and filterings, for the true number of clusters.

Lukas

Clustering high-dimensional flow and mass cytometry

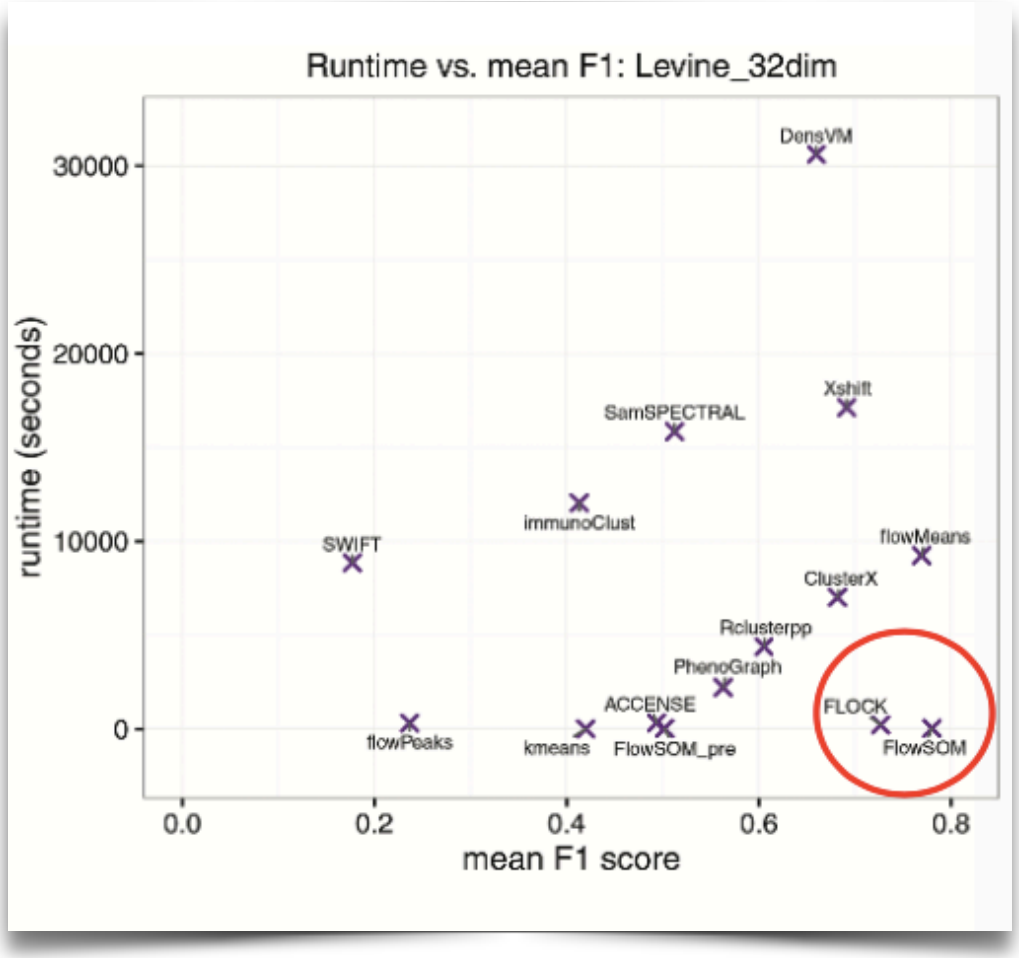
Motivation: Many new computational methods, explosion in the number of dimensions (both FACS and CyTOF) — what works “best”?

EDITOR’S CHOICE

Cytometry
PART A
Journal of the International Society for Advancement of Cytometry

Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data

Lukas M. Weber,[1,2] Mark D. Robinson[1,2]*

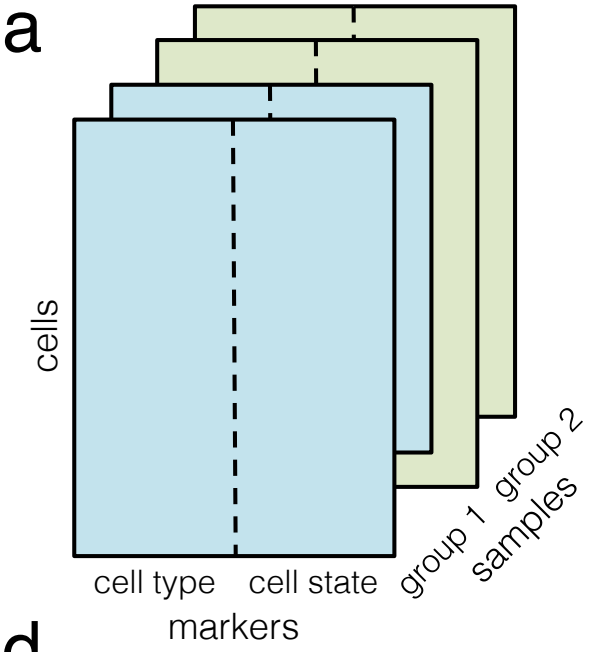


Lukas

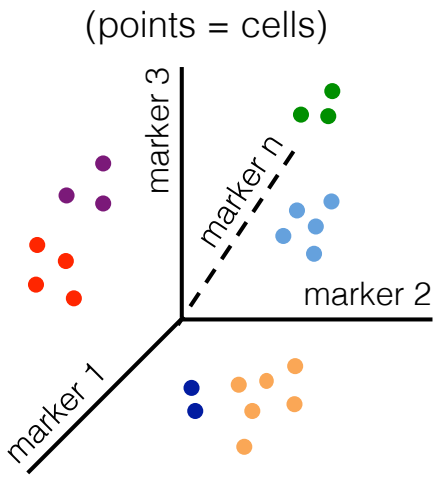
diffcyt for high-dim. cytometry (preprint on bioRxiv)

notes: over-clustering, split markers

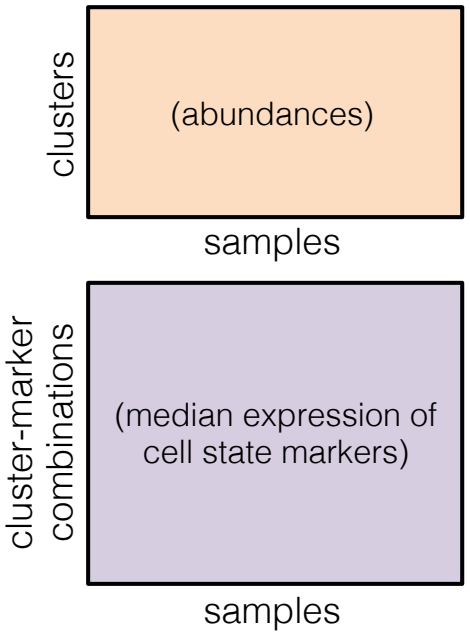
a



b



c

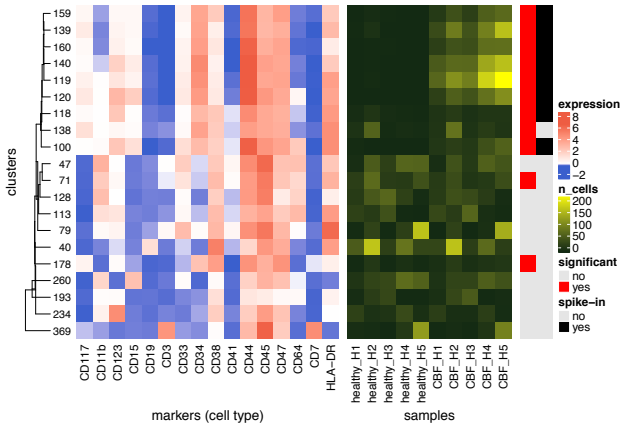


d

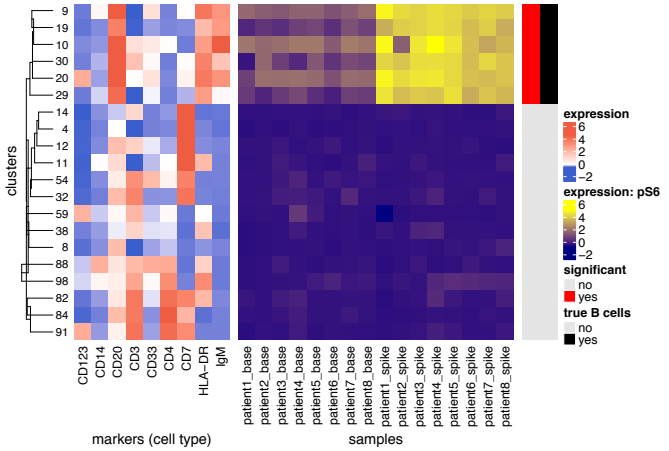
Test type	Methods
Differential abundance (DA) of cell populations	• diffcyt-DA-edgeR • diffcyt-DA-voom • diffcyt-DA-GLMM
Differential states (DS) within cell populations	• diffcyt-DS-limma • diffcyt-DS-LMM

e

Differential abundance (DA) of cell populations



Differential states (DS) within cell populations



Trajectory analysis

A comparison of single-cell trajectory inference methods: towards more accurate and robust tools

05 March, 2018

Wouter Saelens* [1 2], Robrecht Cannoodt* [1 2 3], Helena Todorov[1 2 4], Yvan Saeys[1 2]

* Equal contribution

[1] Data mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium.

[2] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium.

[3] Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium.

[4] Centre International de Recherche en Infectiologie, Inserm, U1111, Université Claude Bernard Lyon 1, CNRS, UMR5308, École Normale Supérieure de Lyon, Univ Lyon, F-69007, Lyon, France

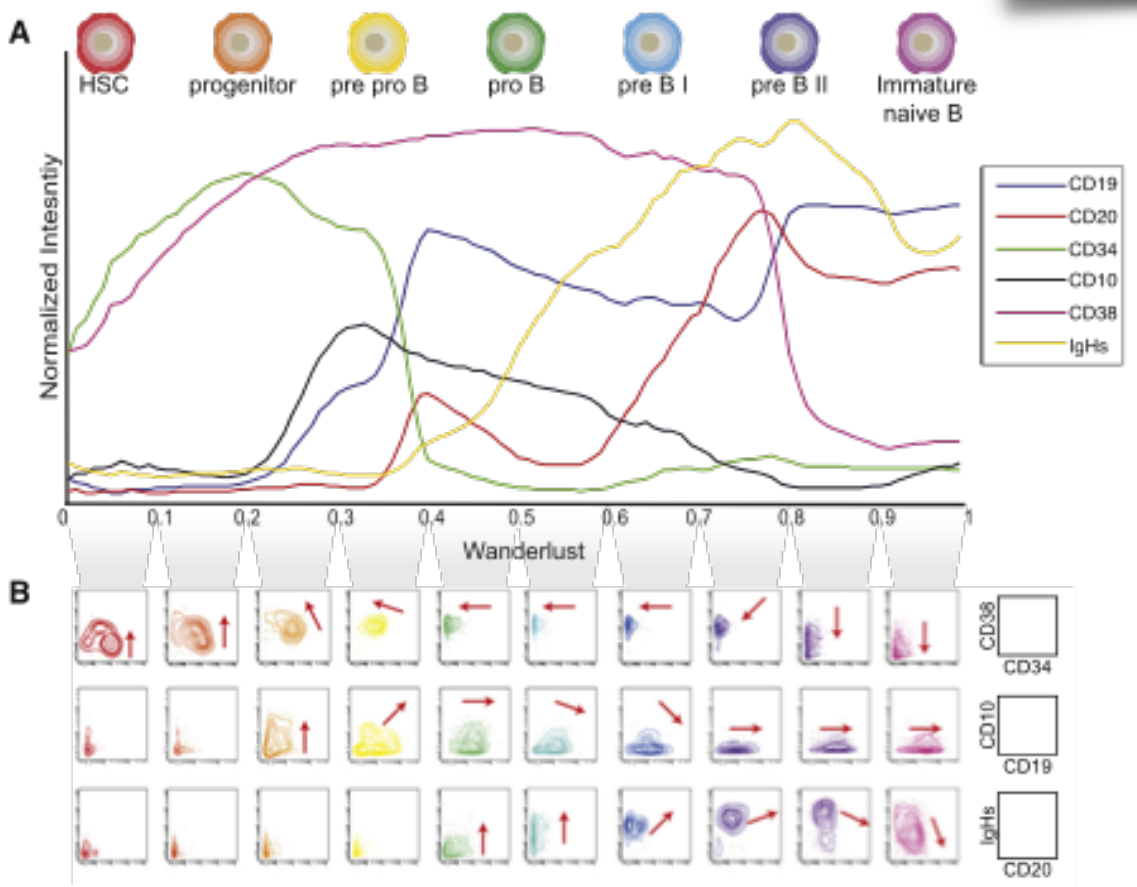


Figure 2. Wanderlust Confirms Known Hallmarks of Human B Cell Development and Is Consistent across Healthy Individuals

(A) The Wanderlust trajectory is fixed to an arbitrary scale where the most immature cells are at 0 and the most mature cells at 1. The traces (based on median marker levels within a sliding window) demonstrate the relative expression patterns of CD34, CD38, CD10, CD19, IgH (s)urface, and CD20 across development. The approximate position of progenitors and B cell fractions is indicated.

(B and C) Biaxial plots (B) demonstrate the two-dimensional progression of cellular marker expression (red arrow) across the Wanderlust trajectory taken in segments of 0.1. (C) Distribution of marker expression across the trajectory for CD24, TdT, and CD10. The green line indicates the relative standard deviation across the trajectory.

(D) Marker traces across the trajectory for four different samples (denoted a to d) aligned using cross-correlation. Pearson's $\rho > 0.9$ between the trajectories of different samples. The red box demarcates the expression of CD24, which bisects the TdT expression prior to CD10 expression across all four healthy individuals.

See also Figure S2 for traces on full marker panel

Amir et al., NBT, 2013