# Overall thoughts

- Journal clubs: format
- Projects: due 13 Jan 2017
- CBB+Biostat students: rotations, M.Sc. projects

University of Zurich[UZH]

Institute of Molecular Life Sciences

# Single cell analysis

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

Mark D. Robinson, Institute of Molecular Life Sciences

# 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

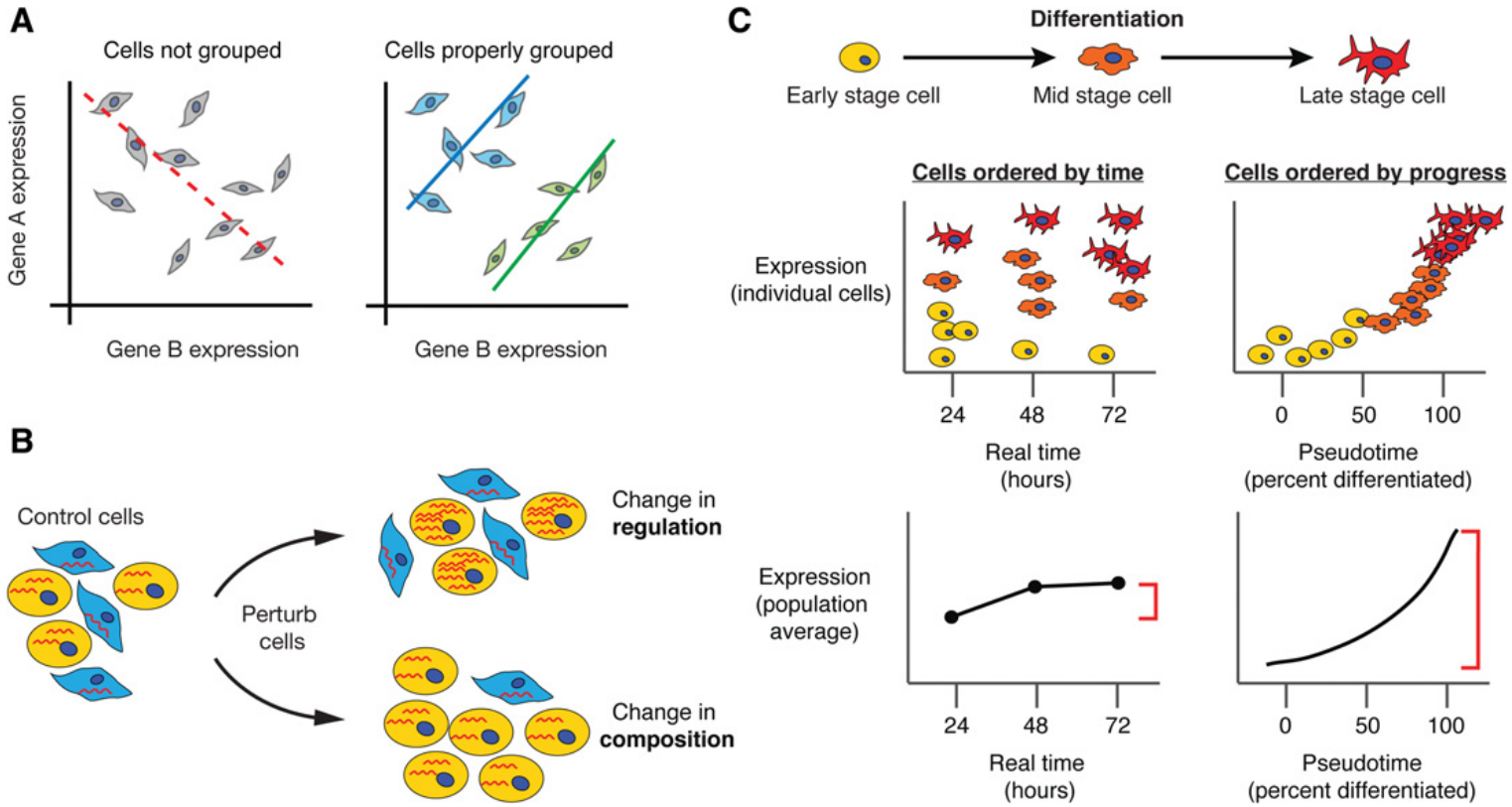
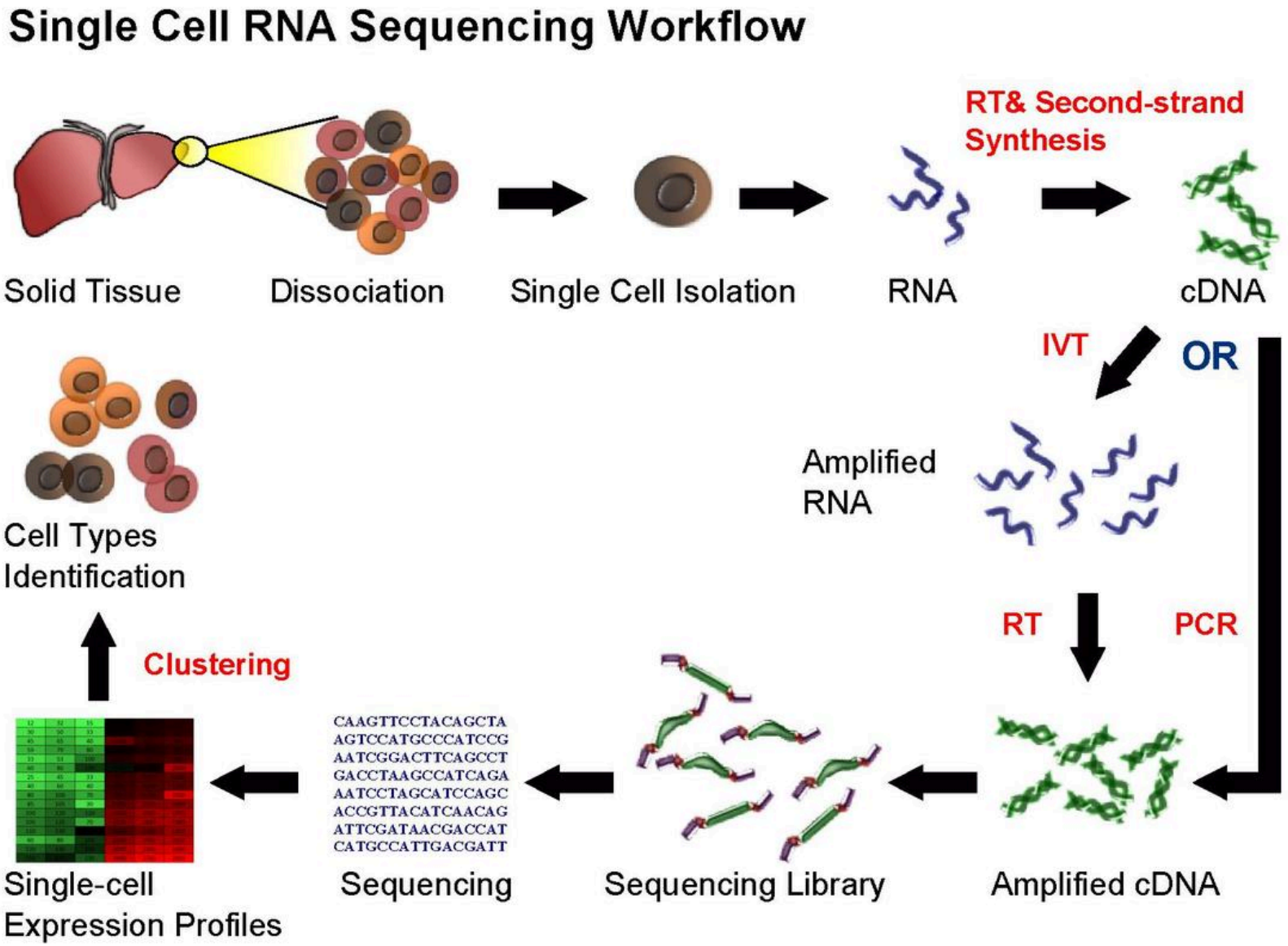


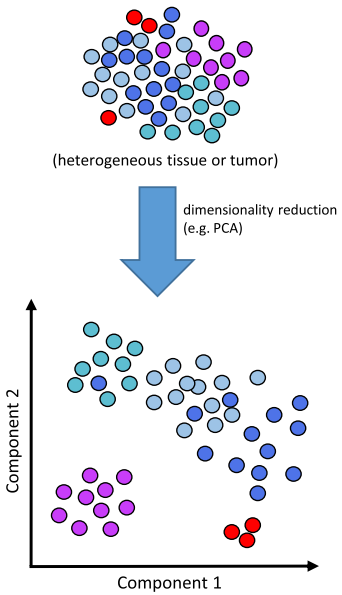

**Figure 1.** Single-cell measurements preserve crucial information that is lost by bulk genomics assays. (*A*) Simpson's Paradox describes the misleading effects that arise when averaging signals from multiple individuals. (*B*) Bulk measurements cannot distinguish changes due to gene regulation from 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).

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

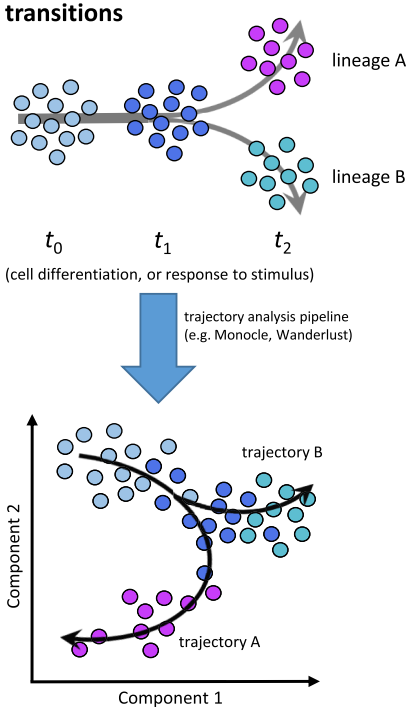


# Tasks

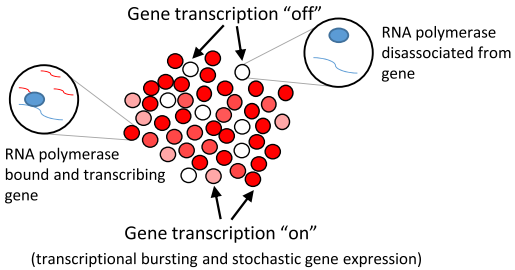
**a) Deconvolving heterogeneous cell populations**



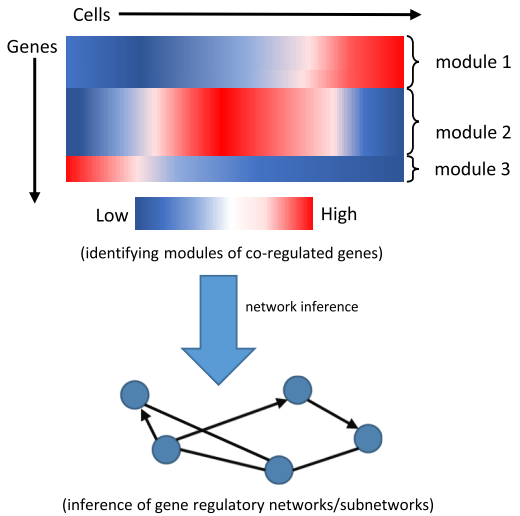

**b) Trajectory analysis of cell state transitions**




**c) Dissecting transcription mechanics**




**d) Network inference**




Liu and Trapnell, F1000, 2016

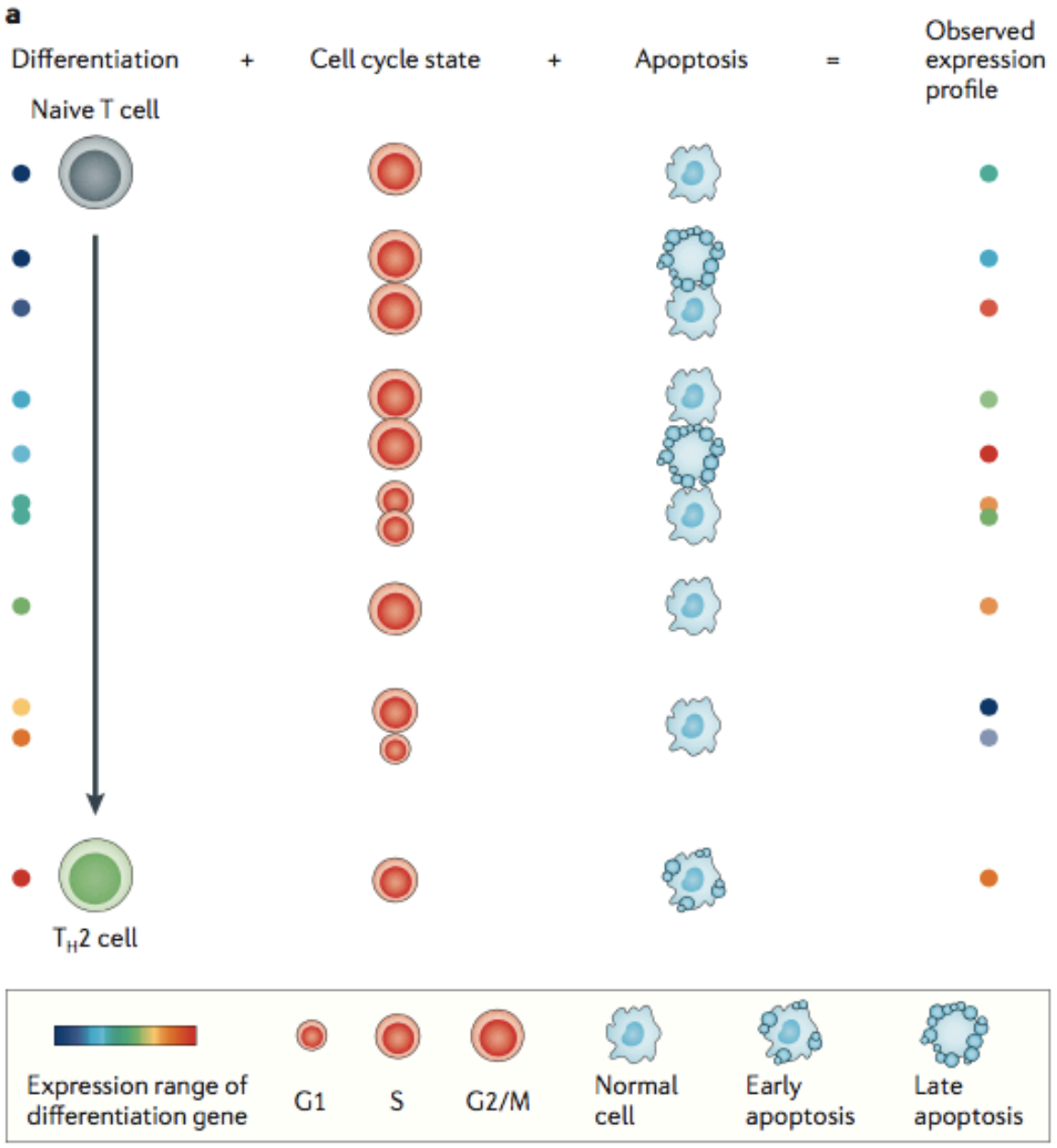
# Measurements are a convolution of other signals

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

$$y_g \sim \mathcal{N}(\mu_g \mathbf{1}, \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}} + \sigma_v^2 \boldsymbol{C}\boldsymbol{C}^{\boldsymbol{T}} + v_g^2 \mathbf{II}) \quad (1)$$

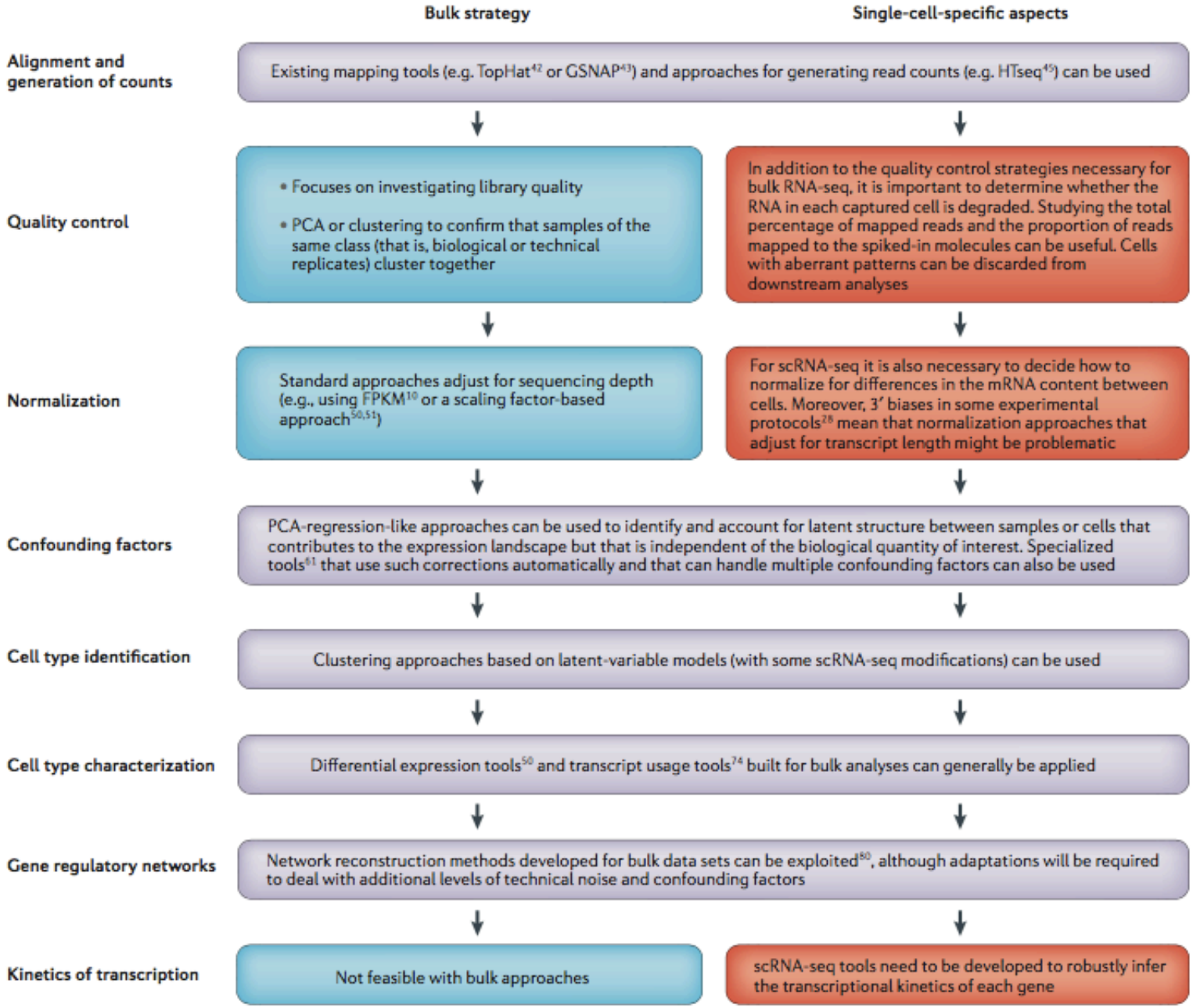
where $\boldsymbol{X}$ represents the hidden factor (such as cell cycle), $\boldsymbol{C}$ corresponds to additional observed covariates (if available) and $v_g^2$ denotes the residual variance. Because the same distributional assumptions are shared across a large set of genes in the annotated set, the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

An important choice when fitting the model is the dimensionality of the



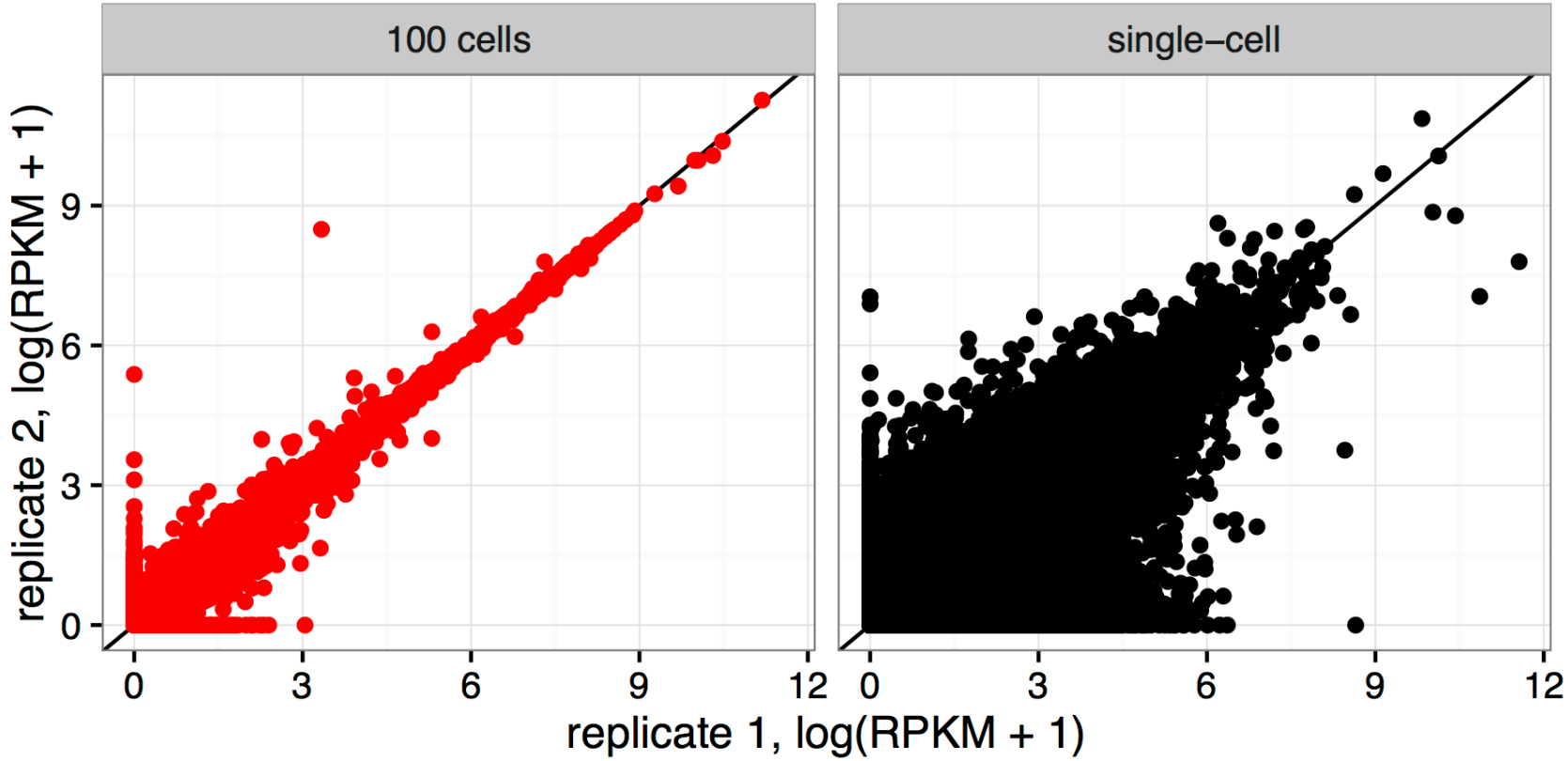

Stegle et al., NBT, 2015; Buettner et al., NBT, 2015

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

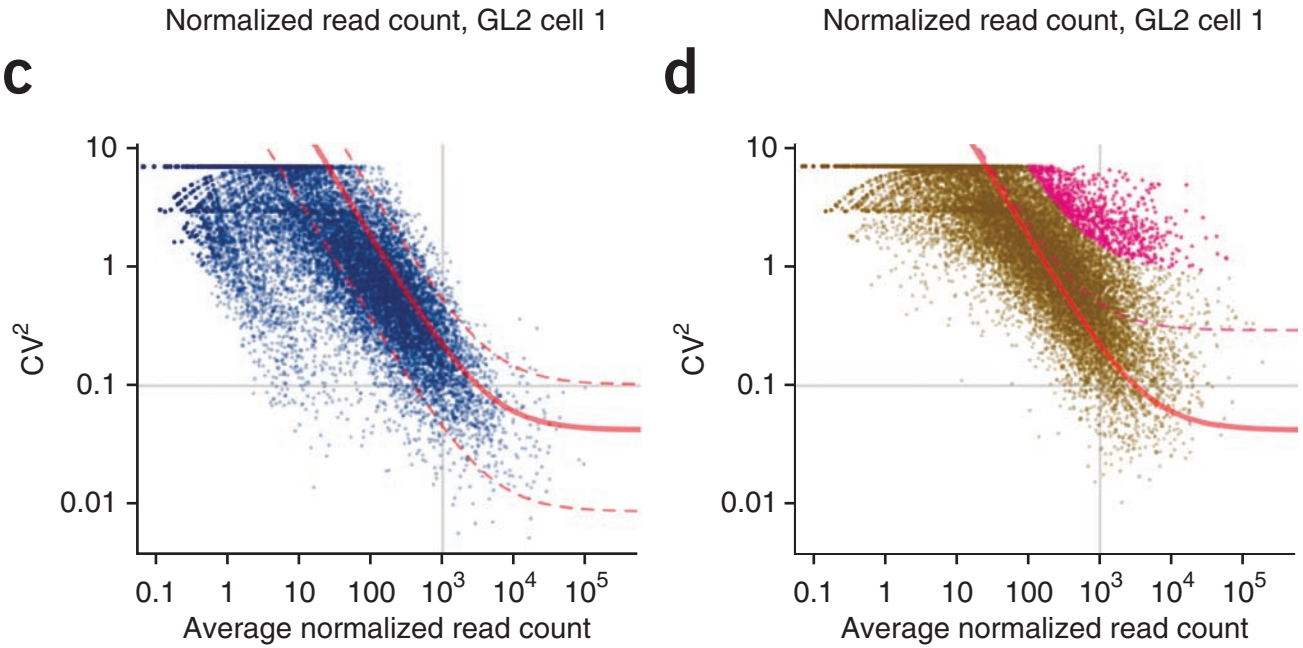



Stegle et al., NBT, 2015

# Variability levels



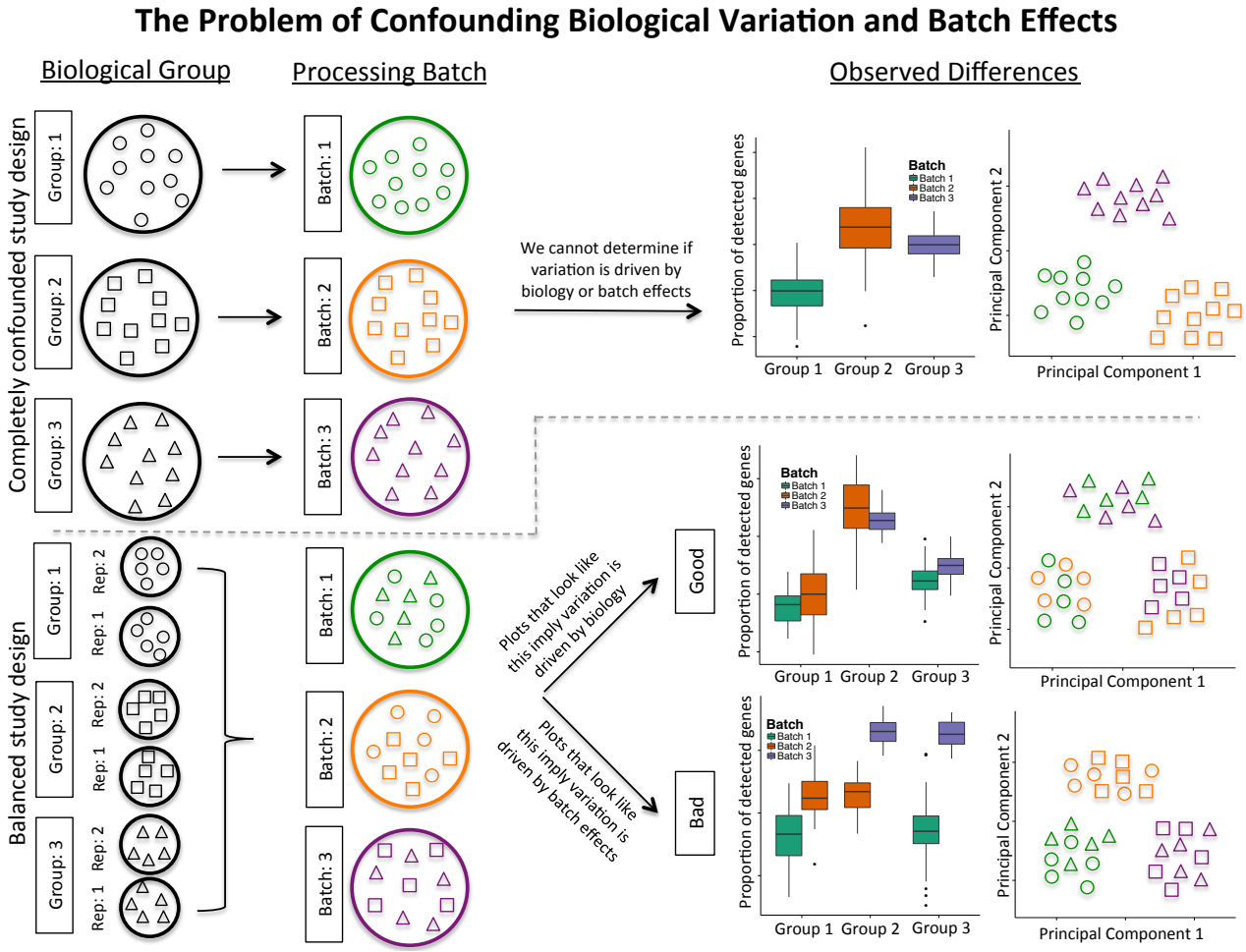
# Error models using spike-ins



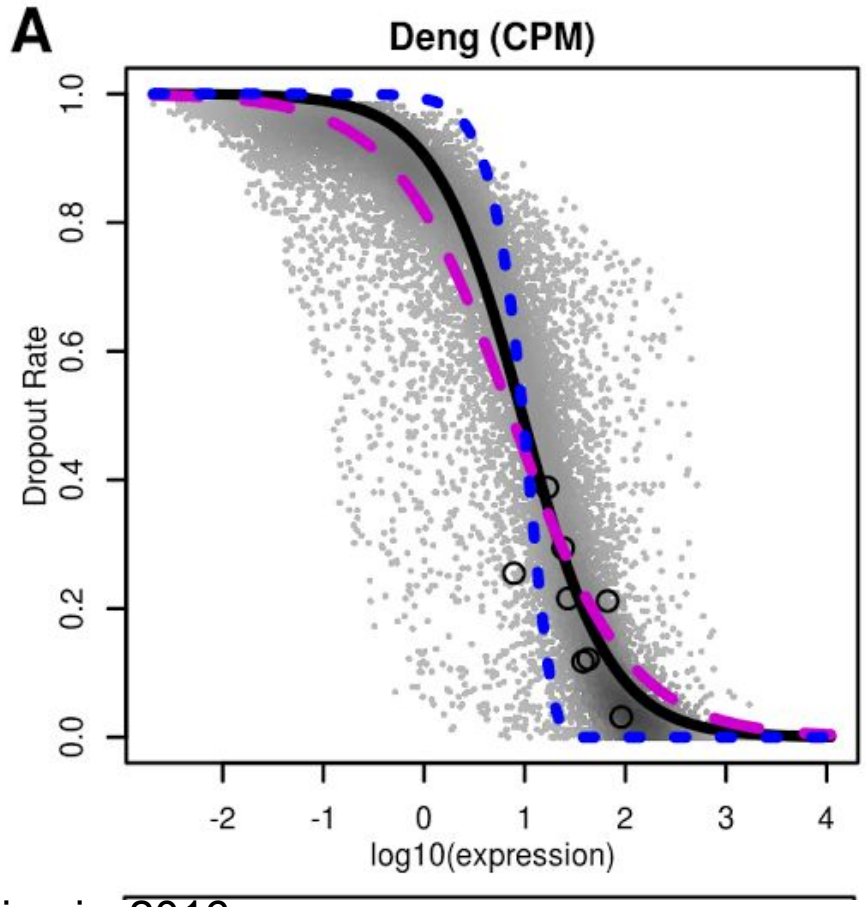

(**c**) Technical noise fit: squared coefficients of variation are plotted against the means of normalized read counts for each HeLa gene using data from all seven GL2 cells. The solid red curve represents the fitted variance-mean dependence; the dashed lines indicate a 95% interval for the expected residual distribution (Online Methods). (**d**) Identification of highly variable genes across all seven GL2 cells. For the genes highlighted in magenta, the coefficient of biological variation significantly exceeds 50% according to our test (with the false discovery rate controlled at 10%). The dashed line marks the expected position of genes with 50% biological CV; however, owing to the statistical uncertainty of CV estimation, statistical significance is achieved only for $CV^2$ values well above this line.

Brennecke et al., Nature Methods, 2013

# scRNA-seq Gotcha #1: batch effects



# scRNA-seq #2: dropout




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.

**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 $\Omega_c$ 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

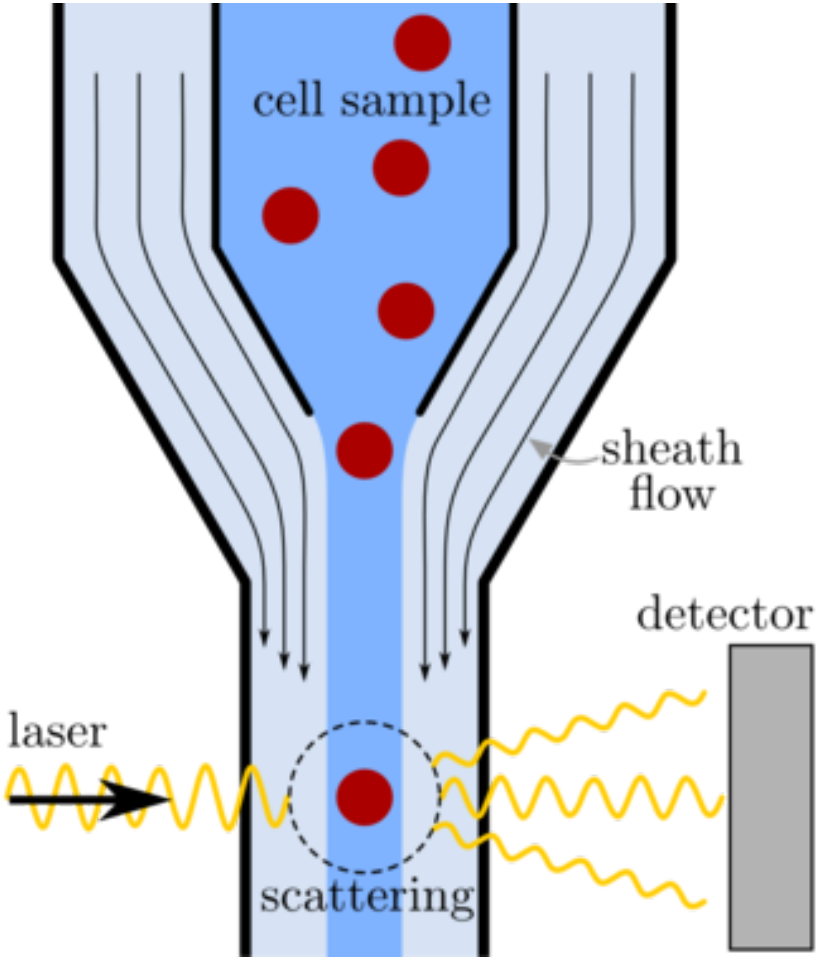
$$p(f)=\sum_{x\in X}p_S(x)p_G(fx)$$

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.

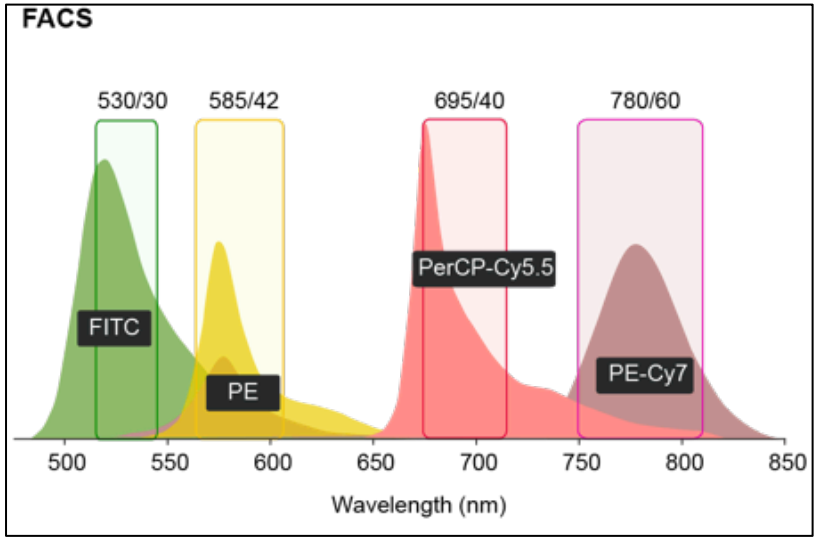
Finak et al., Genome Biology, 2015

Kharchenko et al., Nature Methods, 2015

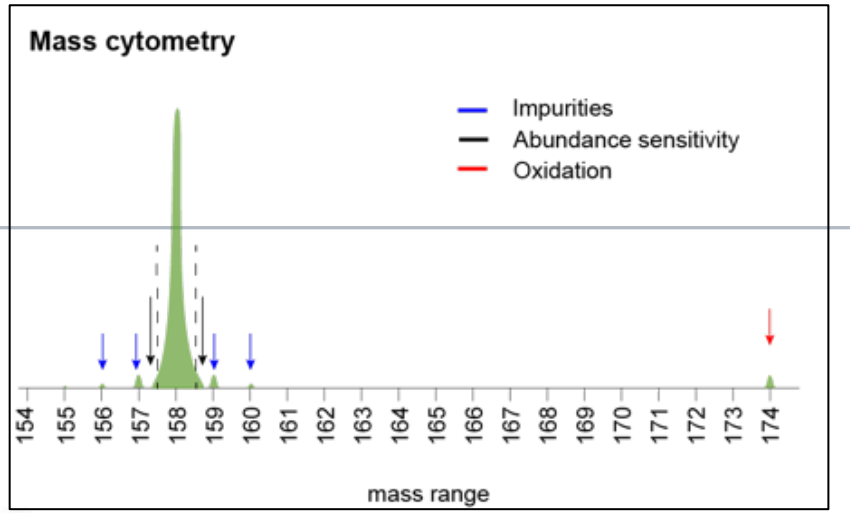
# Flow cytometry (5-10 markers)

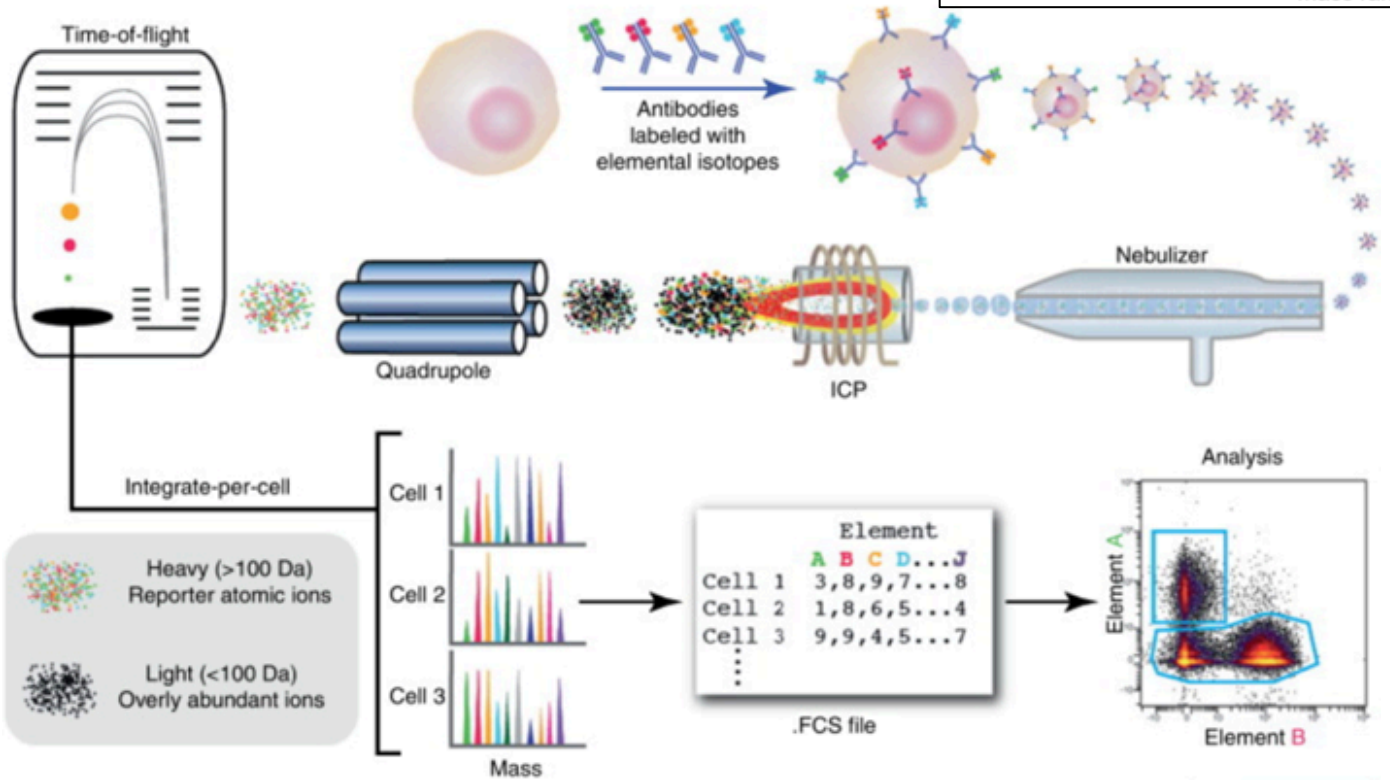



First, cells are stained with a panel of antibodies; these antibodies have a flourescent tag attached.



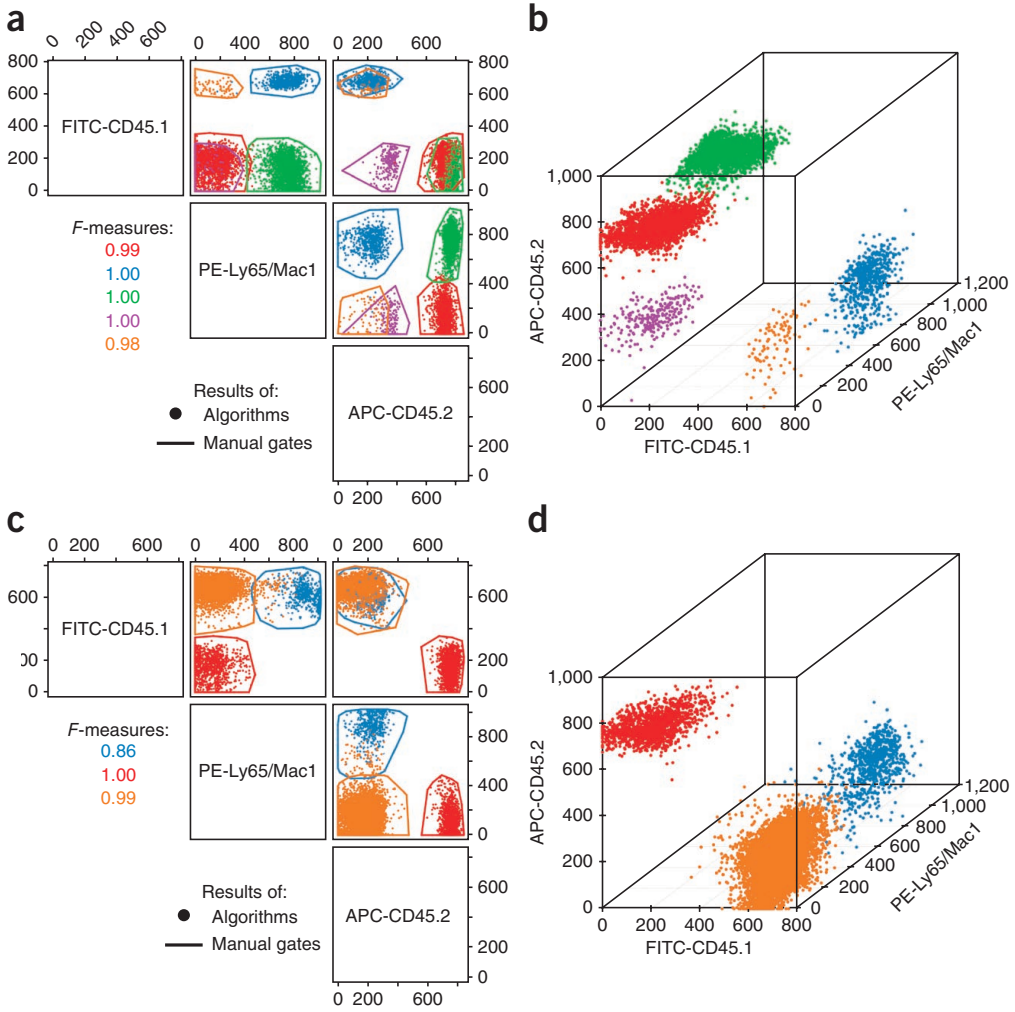
# Mass cytometry (30-50 markers)



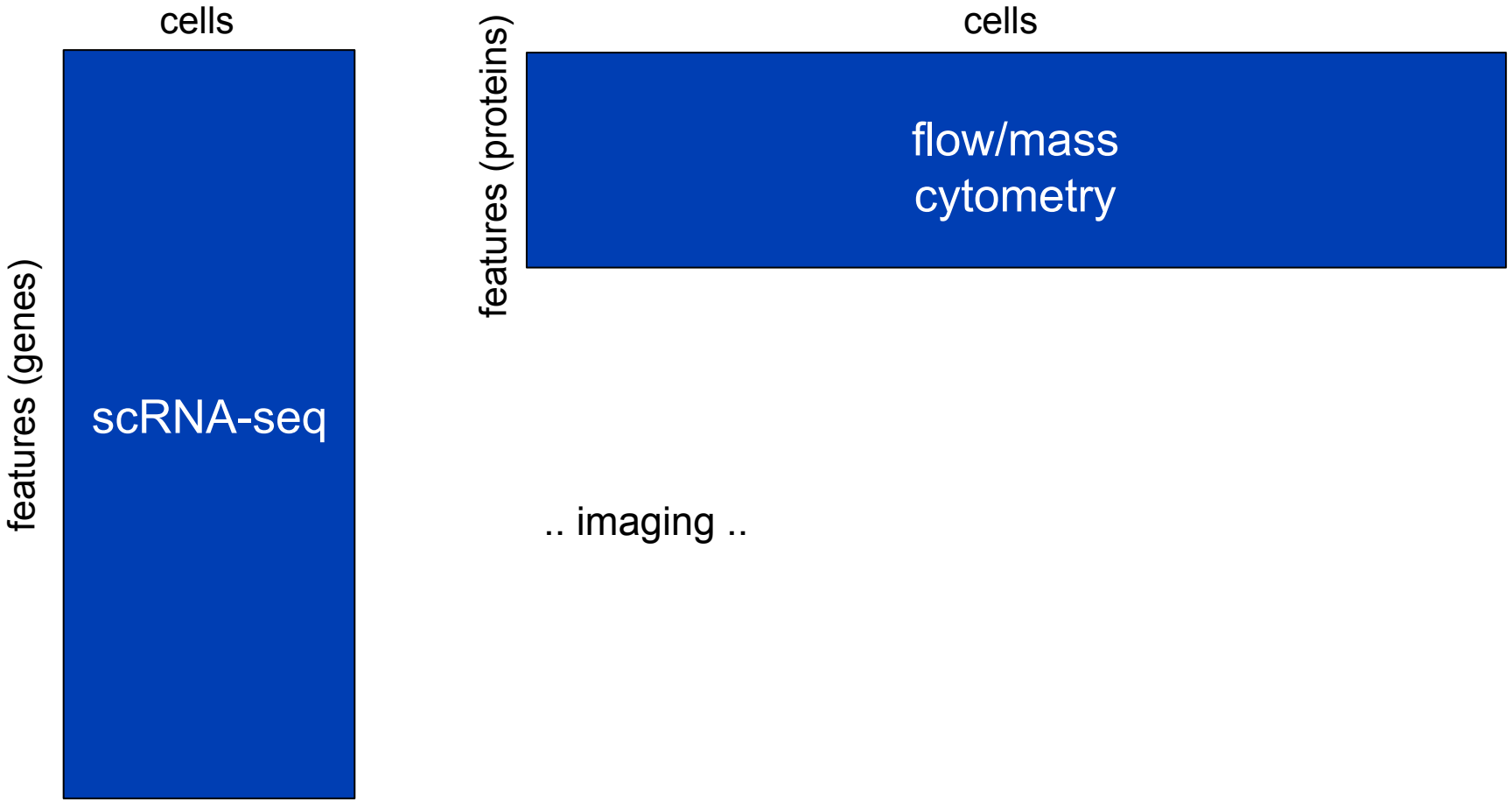



# Manual gating versus clustering

**Figure 3** | Comparison of manual-gate consensus and ensemble clustering results. Dots are color-coded by population membership as determined by ensemble clustering, with donor-derived (CD45.2$^+$) granulocytes/monocytes in green and donor-derived lymphocytes in red. Colored polygons enclose regions corresponding to the consensus clustering of manual gates. Fluorochromes used: FITC, fluorescein isothiocyanate; PE, phycoerythrin; APC, allophycocyanin. (**a**,**b**) Sample for which all of the cell populations have been accurately identified. (**c**,**d**) Sample in which the tail of the blue population has been misclassified as orange by the algorithms, resulting in a lower *F*-measure for the blue population. The red, blue, green, purple and orange cell populations match cell population 1–5 of **Figure 2**, respectively.



# Different shapes of single cell data



# Themes common to many single-cell techniques

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

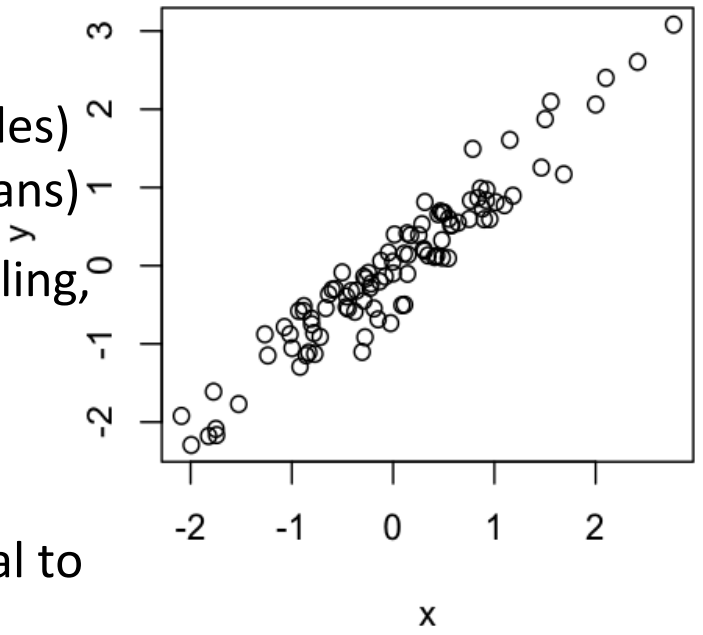
# Dimensionality reduction (generally)

Data analytic techniques exist to **project** high-dimensional data (our situation: 15'000 gene expression measurements for each of N cells/samples) into a small number of dimensions (2 or 3, for humans)

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

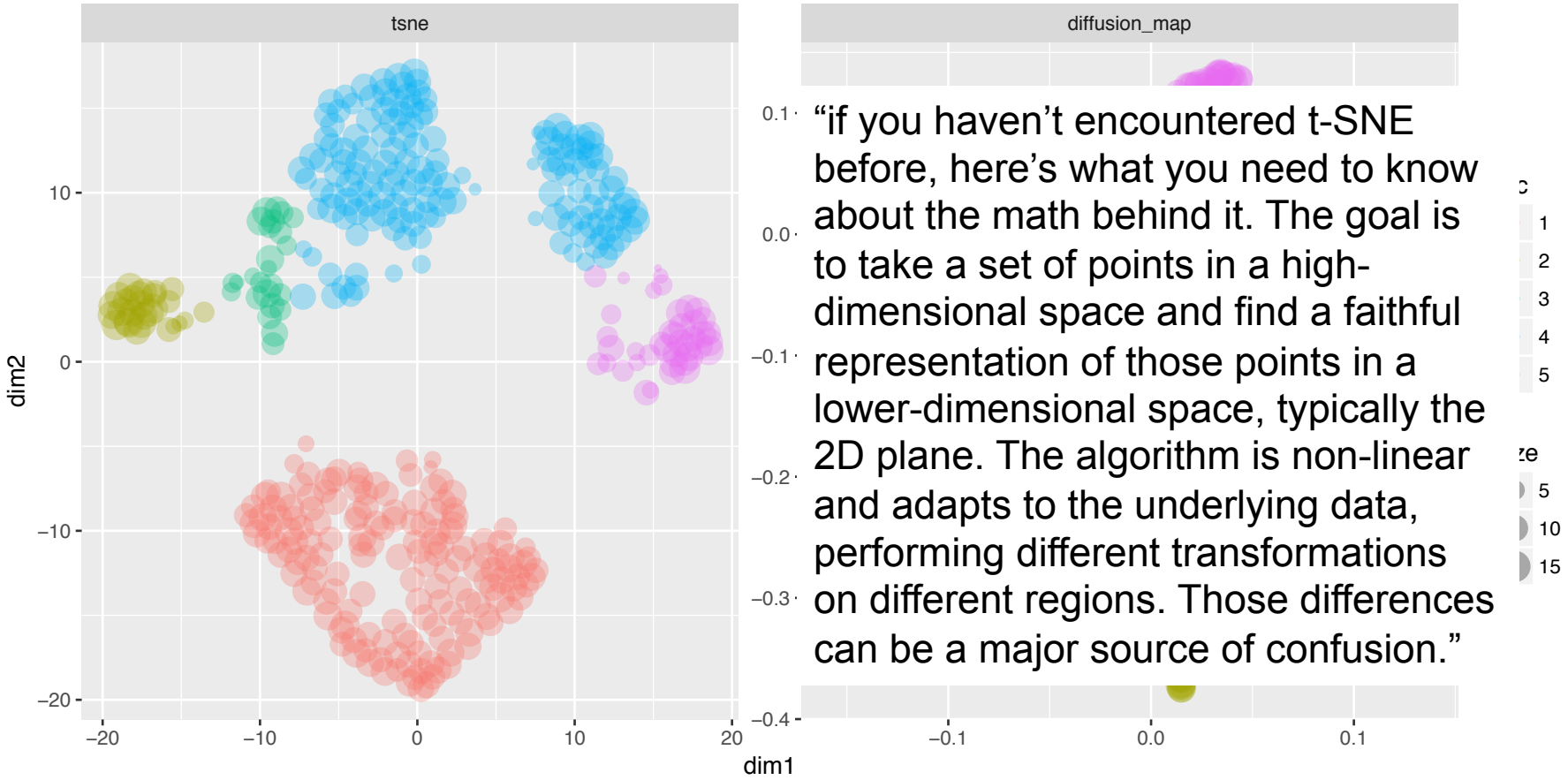
Linear PCA: uses a linear combination of original variables such that the components decrease in variability (highest variance first) and are orthogonal to previous dimensions. Often, first 2 or 3 are used.

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



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

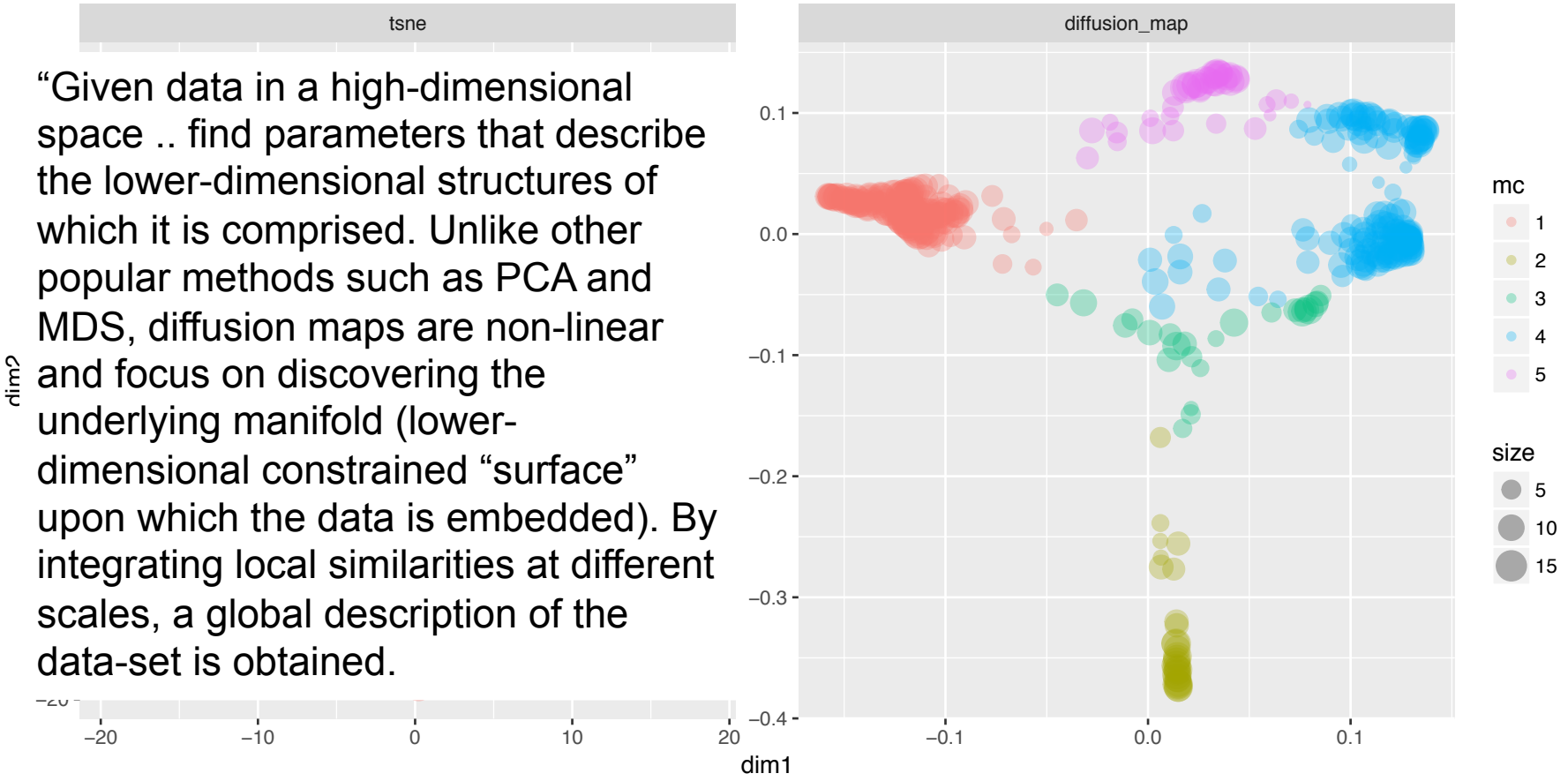
# tSNE (t-dist'd) stochastic neighbour embedding) + diffusion maps




"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."

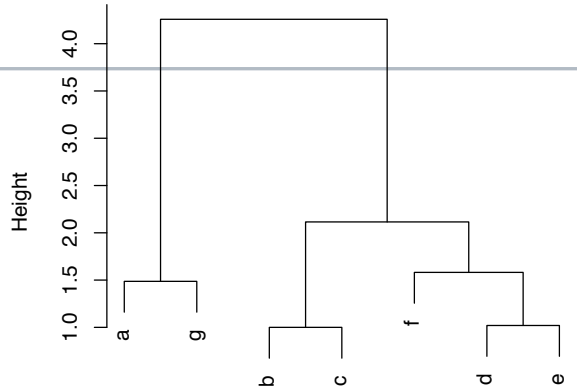
# tSNE (t-dist'd) stochastic neighbour embedding) + diffusion maps

"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

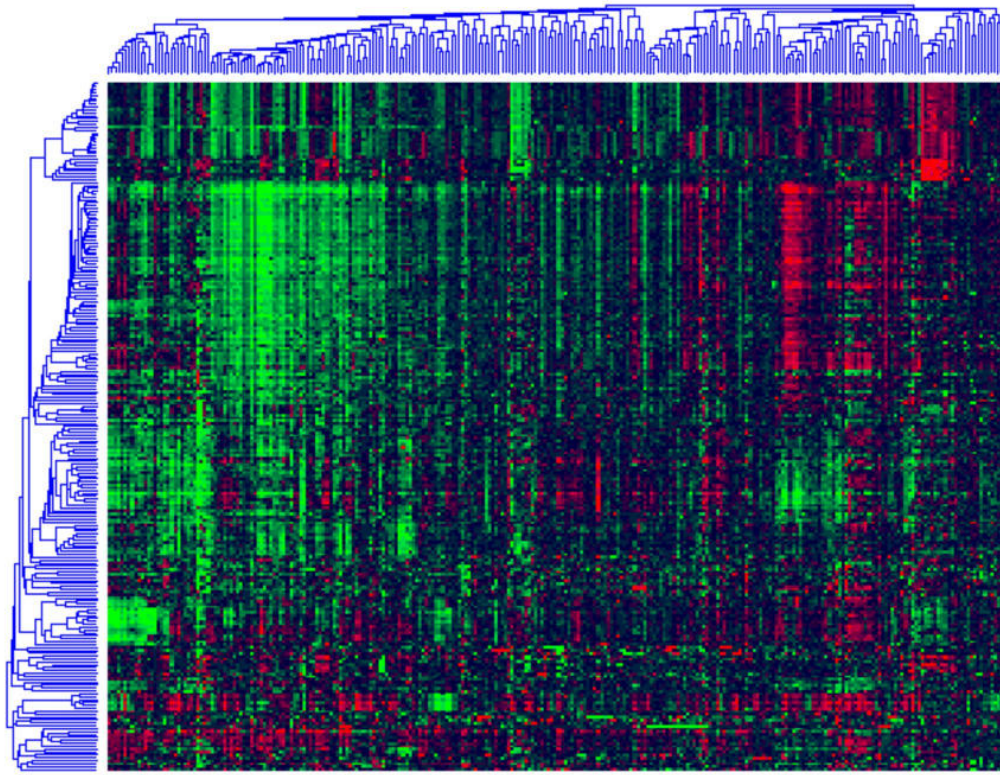
# Hierarchical (Agglomerative) Clustering




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.

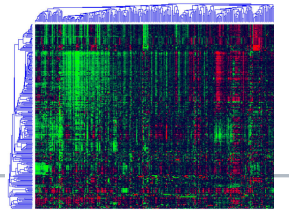
**Linkage**: determines how clusters are merged into a tree



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

Euclidean distance:

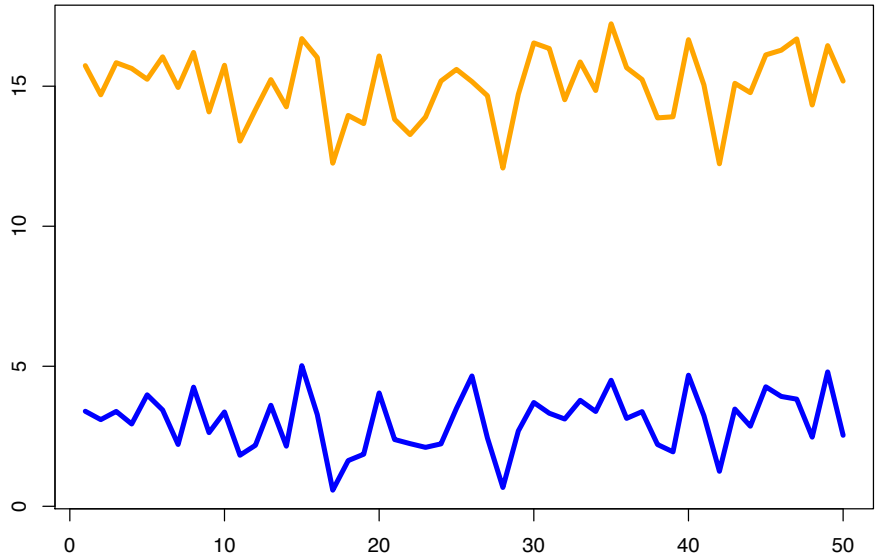
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

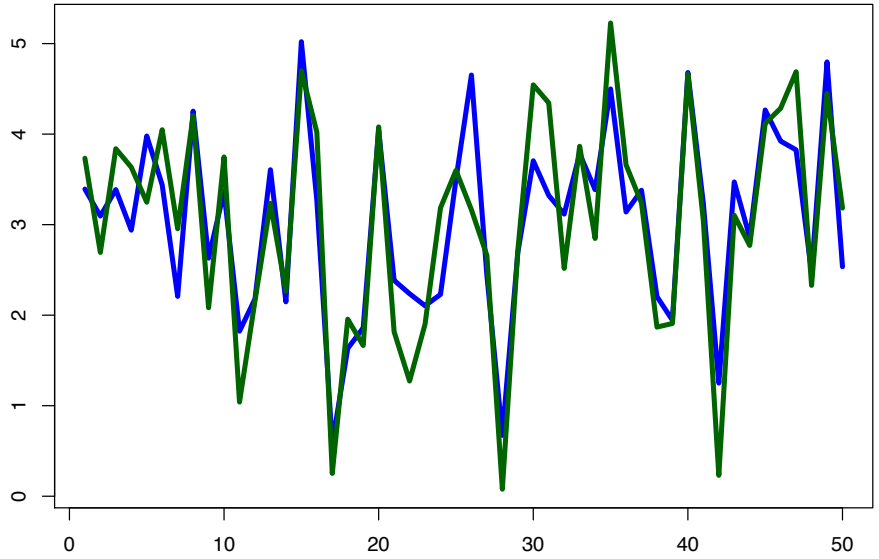


# Are these “vectors” similar ?

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

It depends how you define similar.

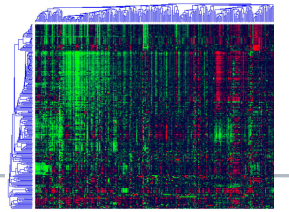



Euclidean distance: 84.84 3.92

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

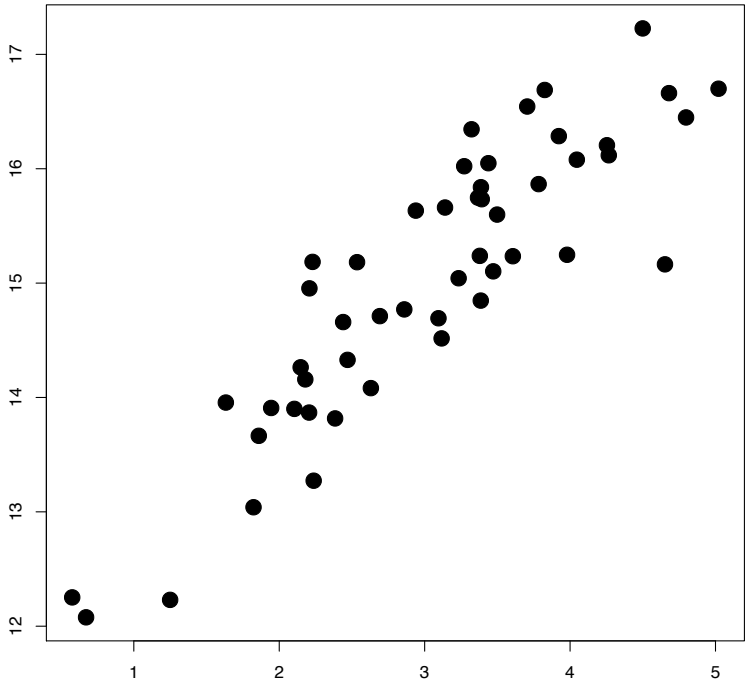
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

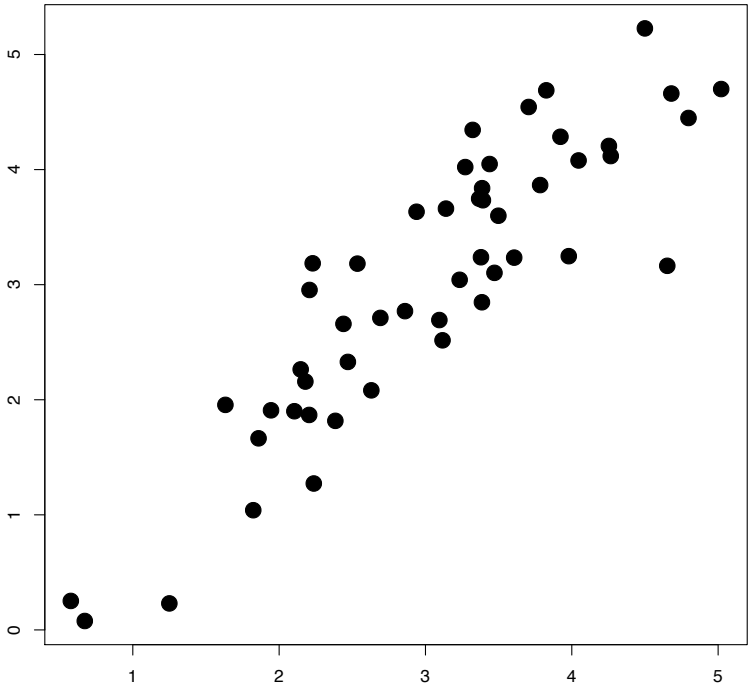


# Are these "vectors" similar ?

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

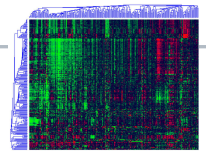
It depends how you define similar.




Correlation: 0.89 0.89

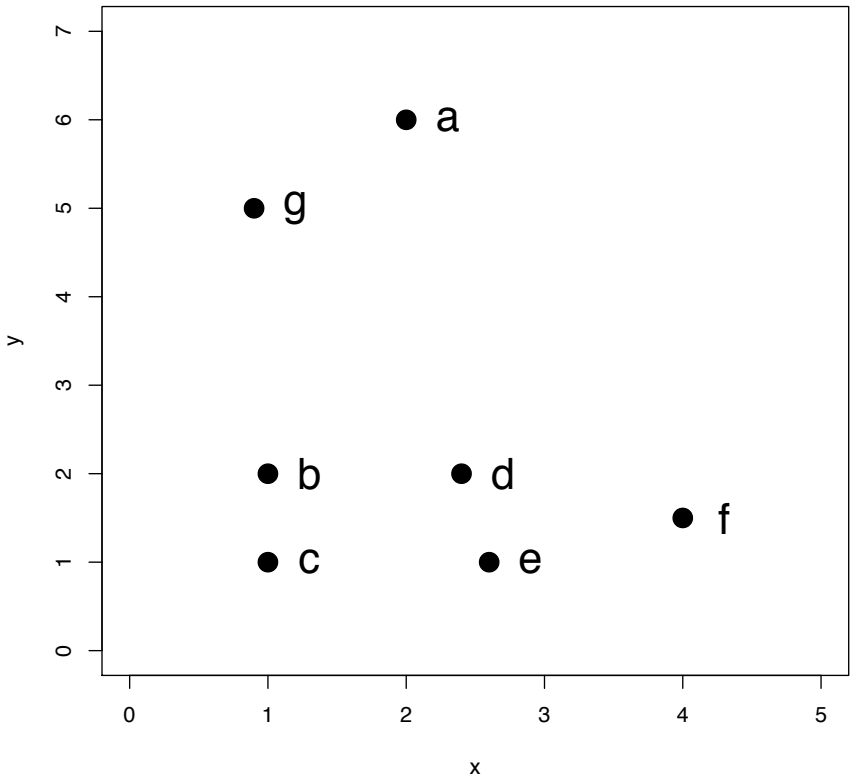
http://en.wikipedia.org/wiki/Hierarchical_clustering



# Hierarchical (Agglomerative) Clustering

Start with distances.

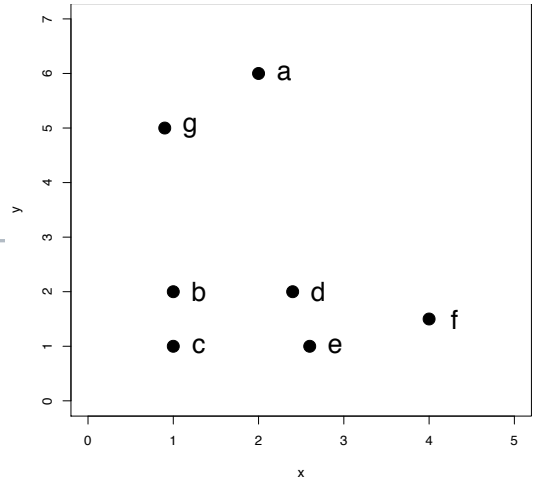
Linkage: determines how clusters are merged into a tree.

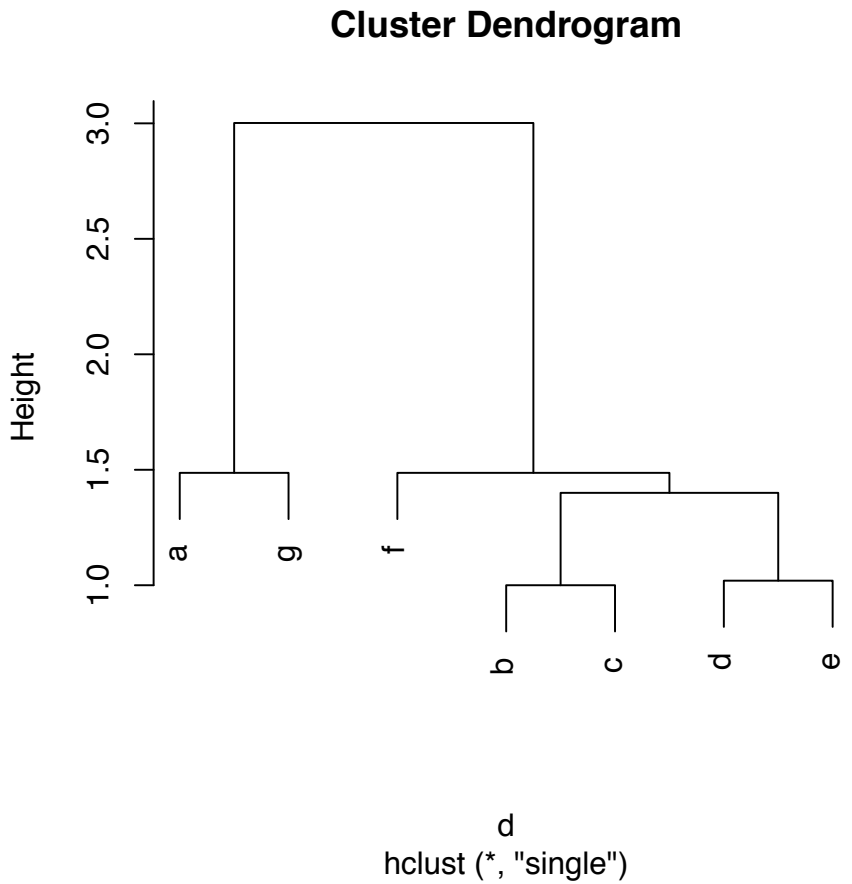


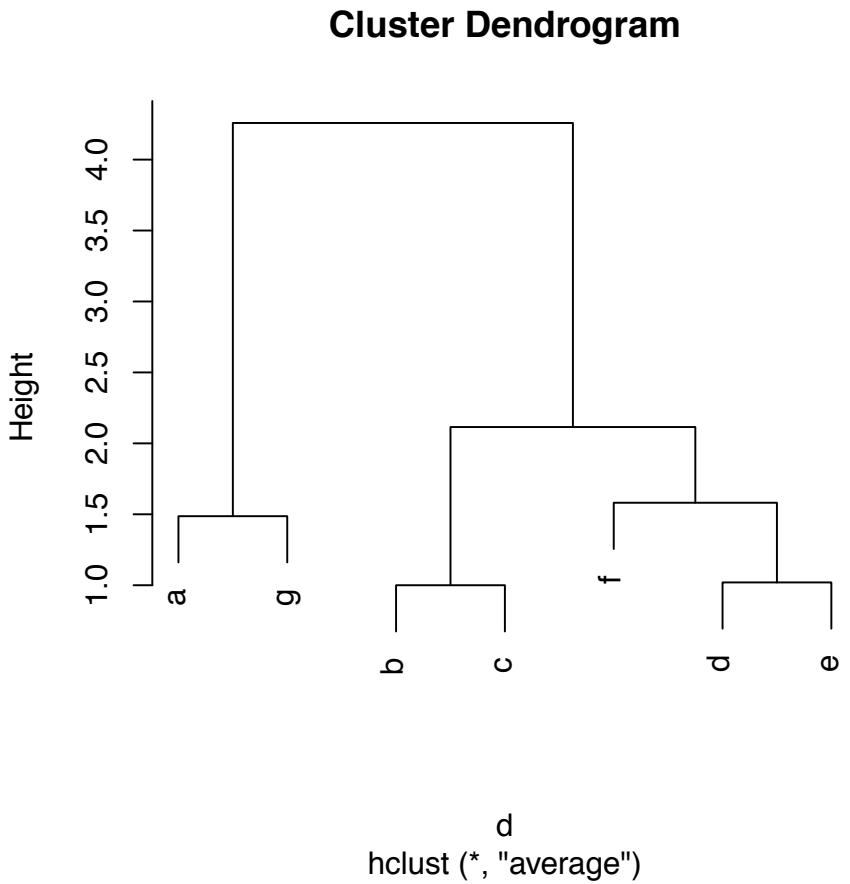

From eyeballing, here is a likely set of merges:

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

# Different linkages

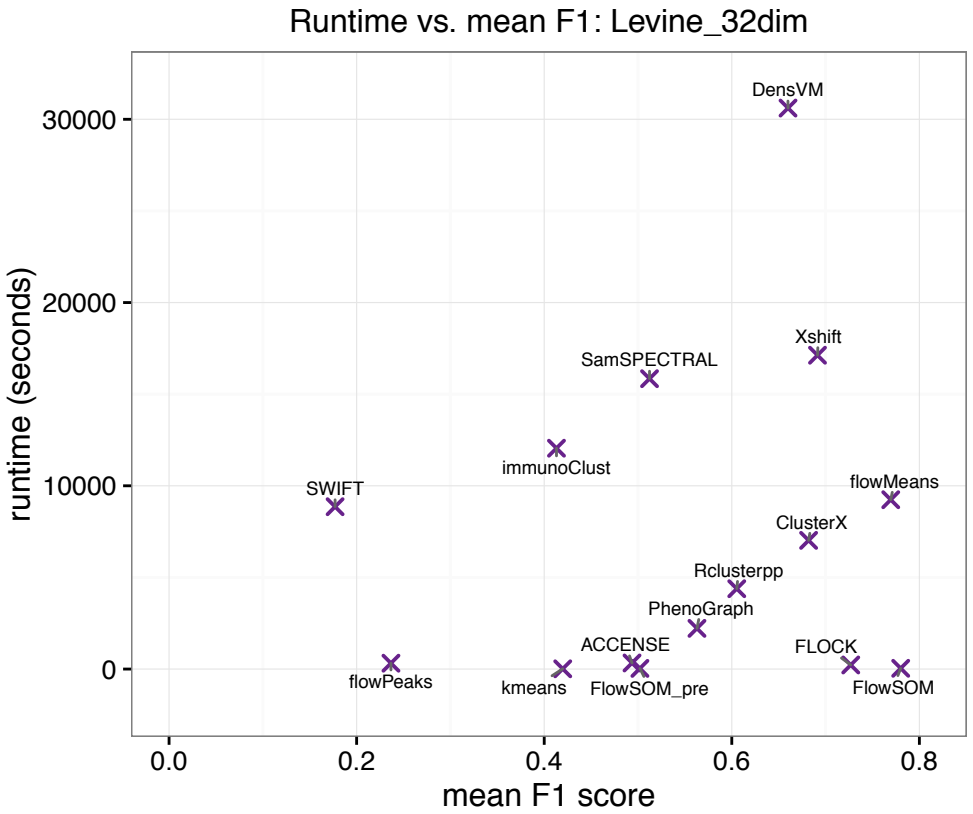


# Many clustering algorithms: Phenograph + FlowSOM + ..

**F**

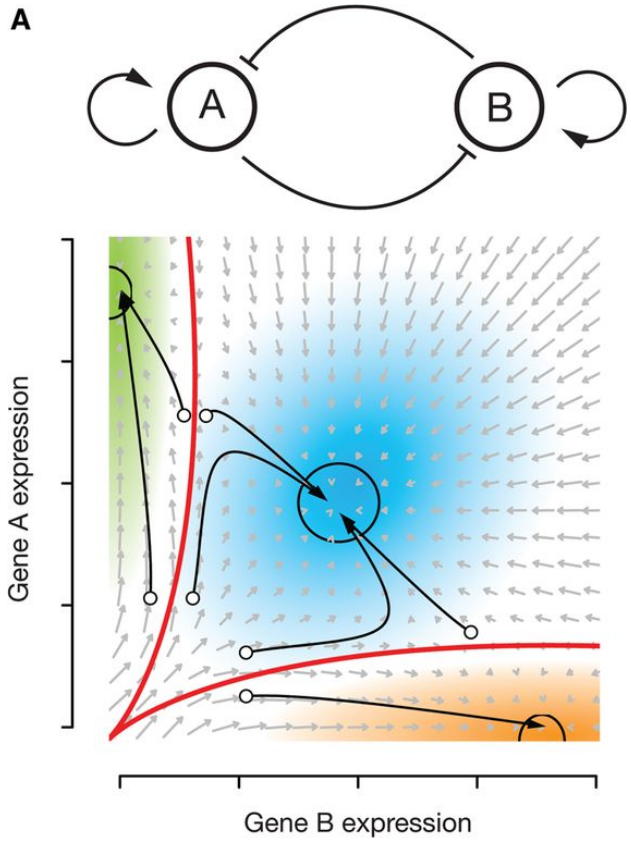


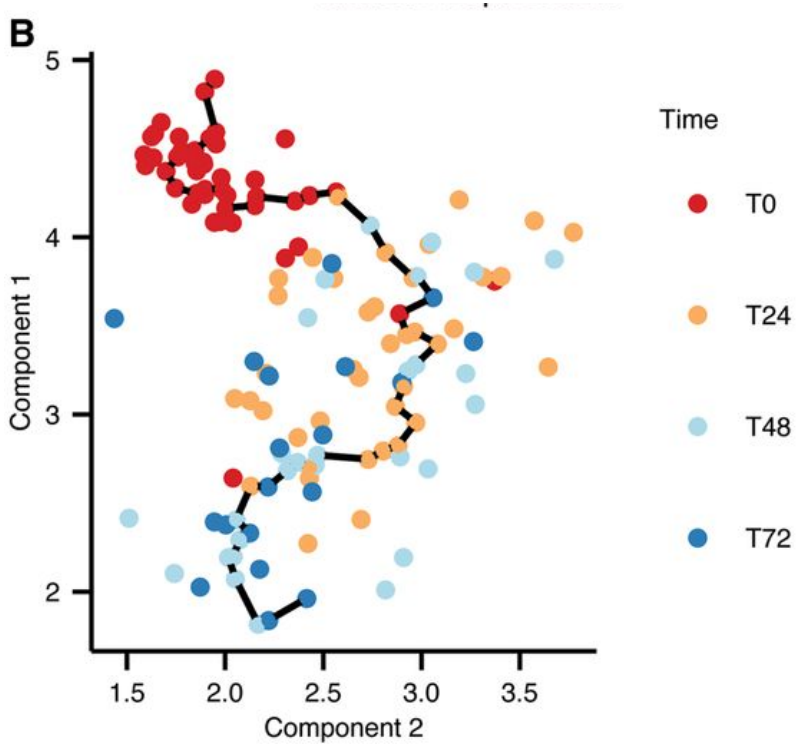

Using various “high” dimensional datasets with a manual gated truth.

F1 score = geometric mean of precision and recall
(mean = averaged over the known populations)

Weber and Robinson, Cytometry A, to appear

# Trajectory analysis




Trapnell, 2015 Genome Research

# Trajectory analysis

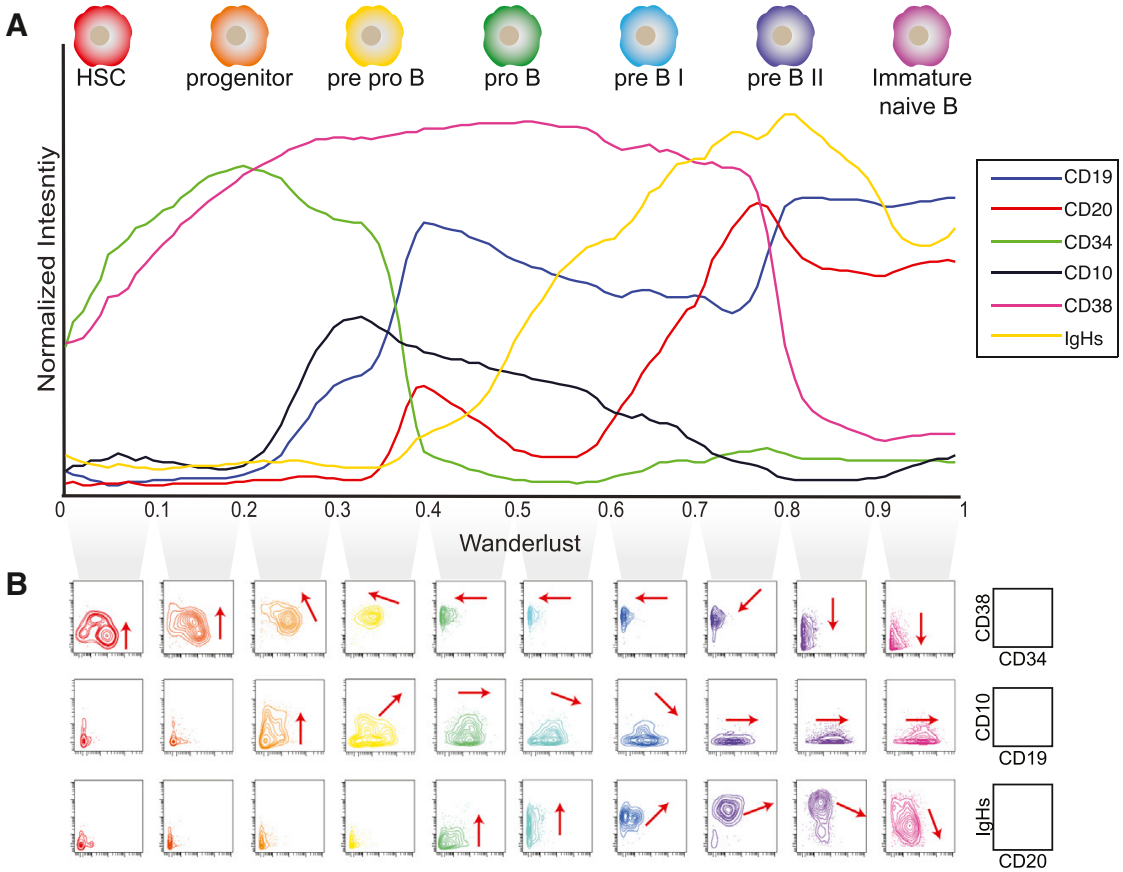



**Figure 2. Wanderlust Confirms Known Hallmarks of Human B Cell Development and Is Consistent across Healthy Individuals**

(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