Outlook for rest of semester

04.12.2017	Mark	single-cell	Removal of batch effects using distribution-matching residual networks (MH, SG)		
11.12.2017	Lukas	hands-on session #2: mass cytometry	X	X	**n.b.: Y11-J-05 for entire session**
18.12.2017	Mark	epigenomics, DNA methylation, ChIP data, gene set analysis	Linear models enable powerful differential activity analysis in massively parallel reporter assays (DP, ZY)		

Quick reminders

- Projects: due Friday 13th Jan 2018
  - (if needed) office hours Fridays 9.00-10.00 starting this week; otherwise, make appointment by email
  - somewhat unlikely to respond 25th Dec – 2nd Jan
- Exercises:
  - marks come from top 9
  - I hope to have another automatic update this week

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, pseudotime ordering, etc.

Mark D. Robinson, Institute of Molecular Life Sciences

SINGLE-CELL OMICS

Computational and analytical challenges in single-cell transcriptomics

Oliver Stegle[1], Sarah A. Teichmann[1,2] and John C. Marioni[1,2]

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].

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

Hypothetical situations

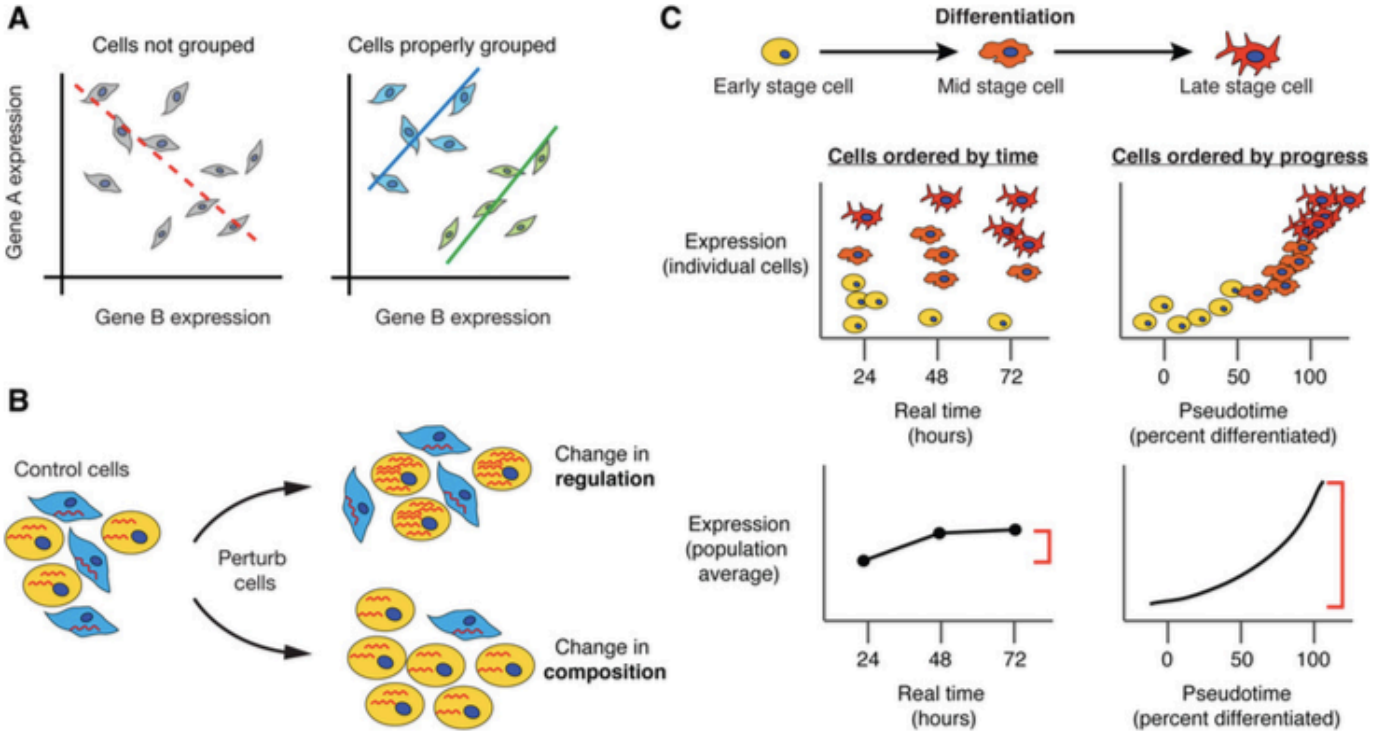
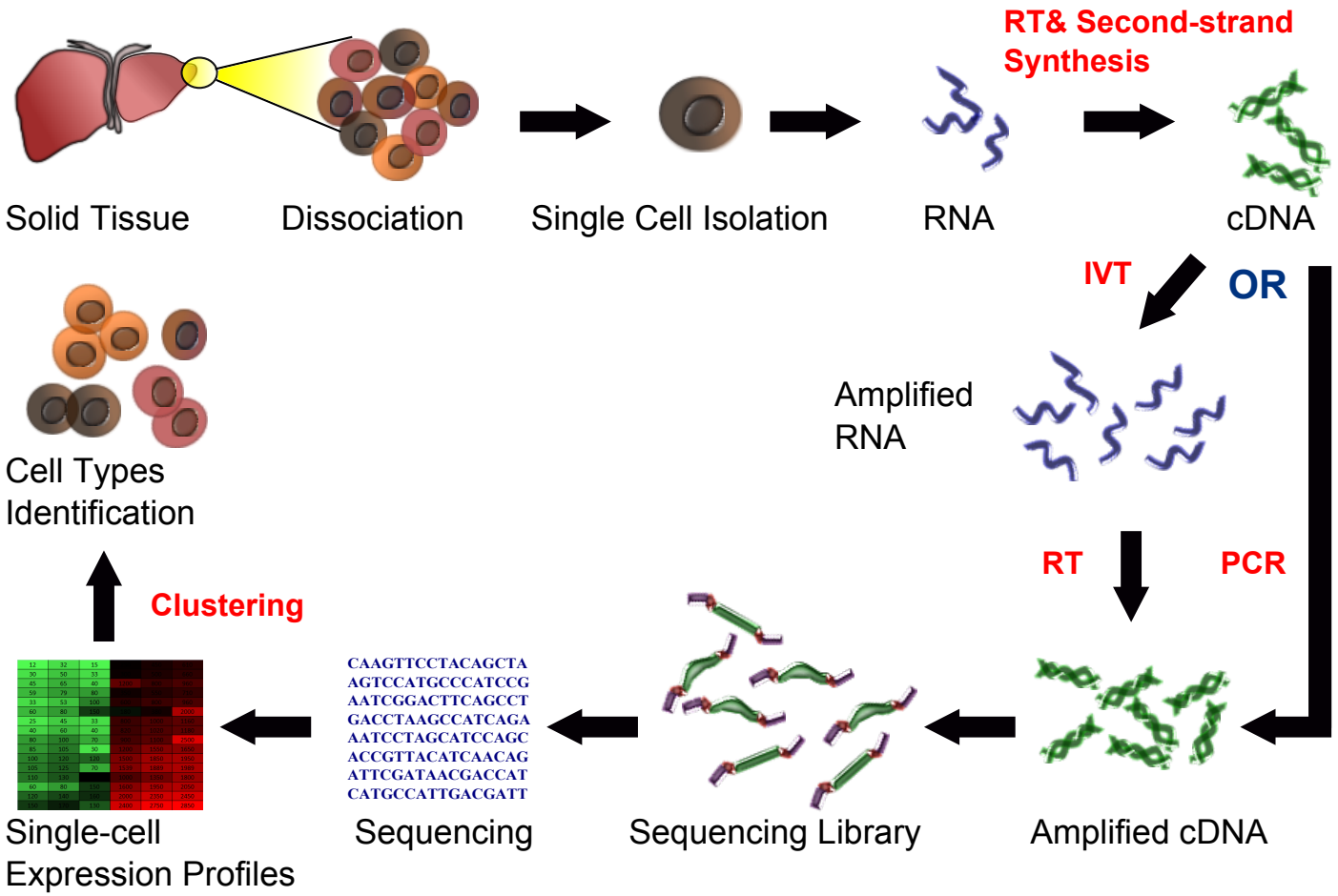


Figure 1. Single-cell measurements preserve crucial information that is lost by bulk genomics assays. (*A*) Simpson's Paradox describes the misleading effects that arise when averaging signals from multiple individuals. (*B*) Bulk measurements cannot distinguish changes due to gene regulation from those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

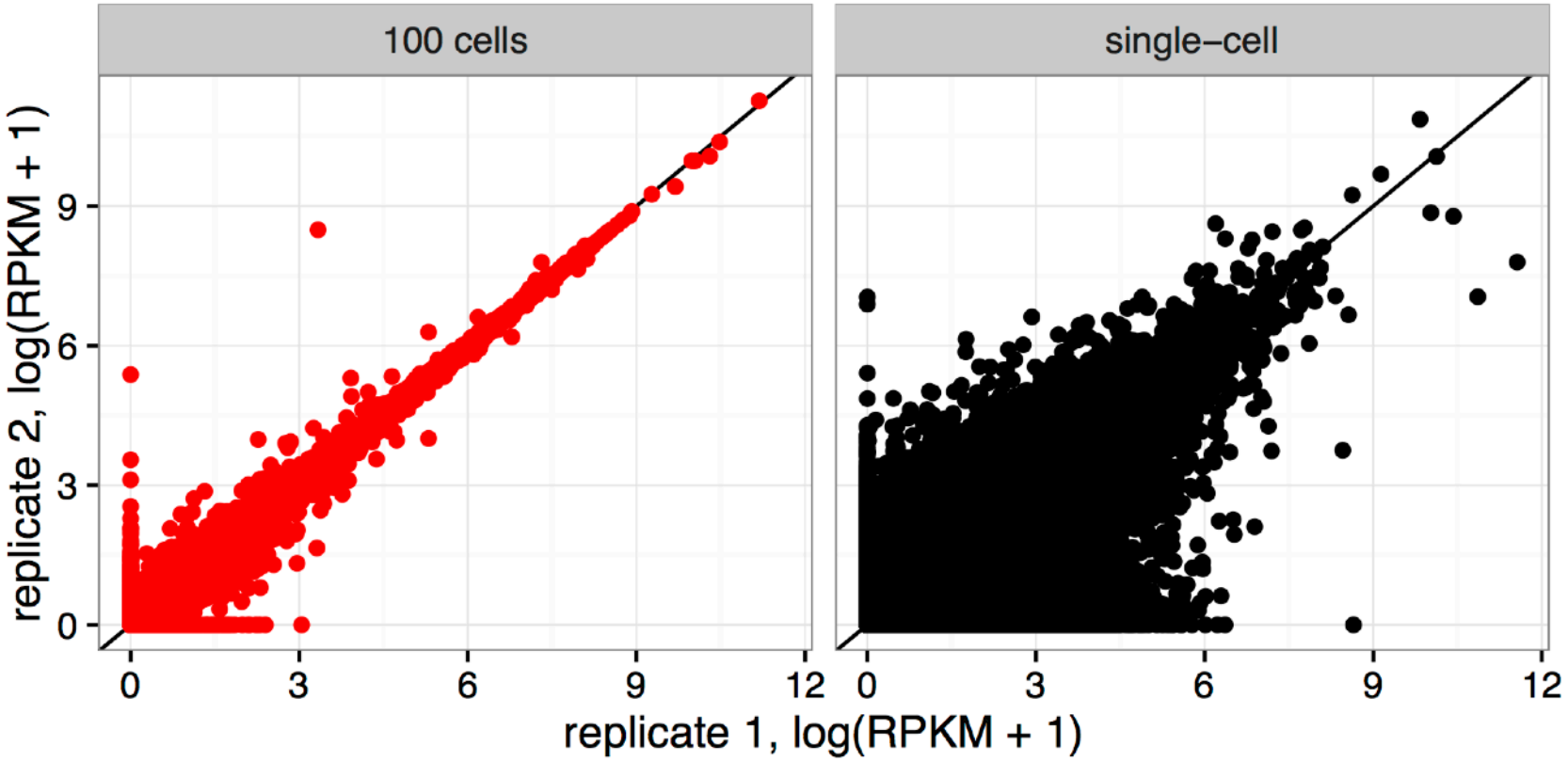
https://en.wikipedia.org/wiki/Single_cell_sequencing

Single Cell RNA Sequencing Workflow

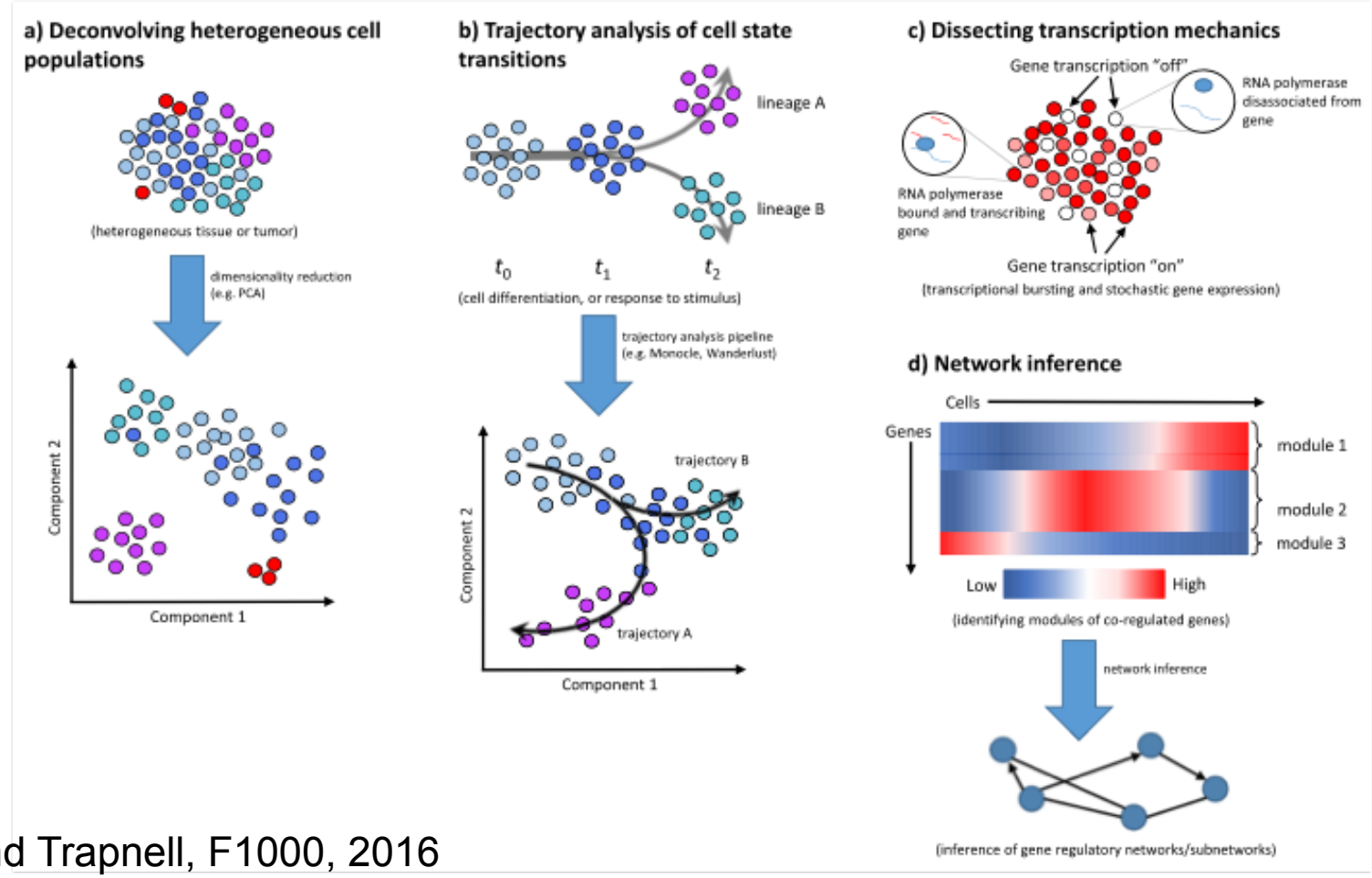


Using oil droplets loaded with reagents ..
Show 10x genomics video: https://10xgenomics.wistia.com/medias/f75ht43w1q

Variability levels



Tasks



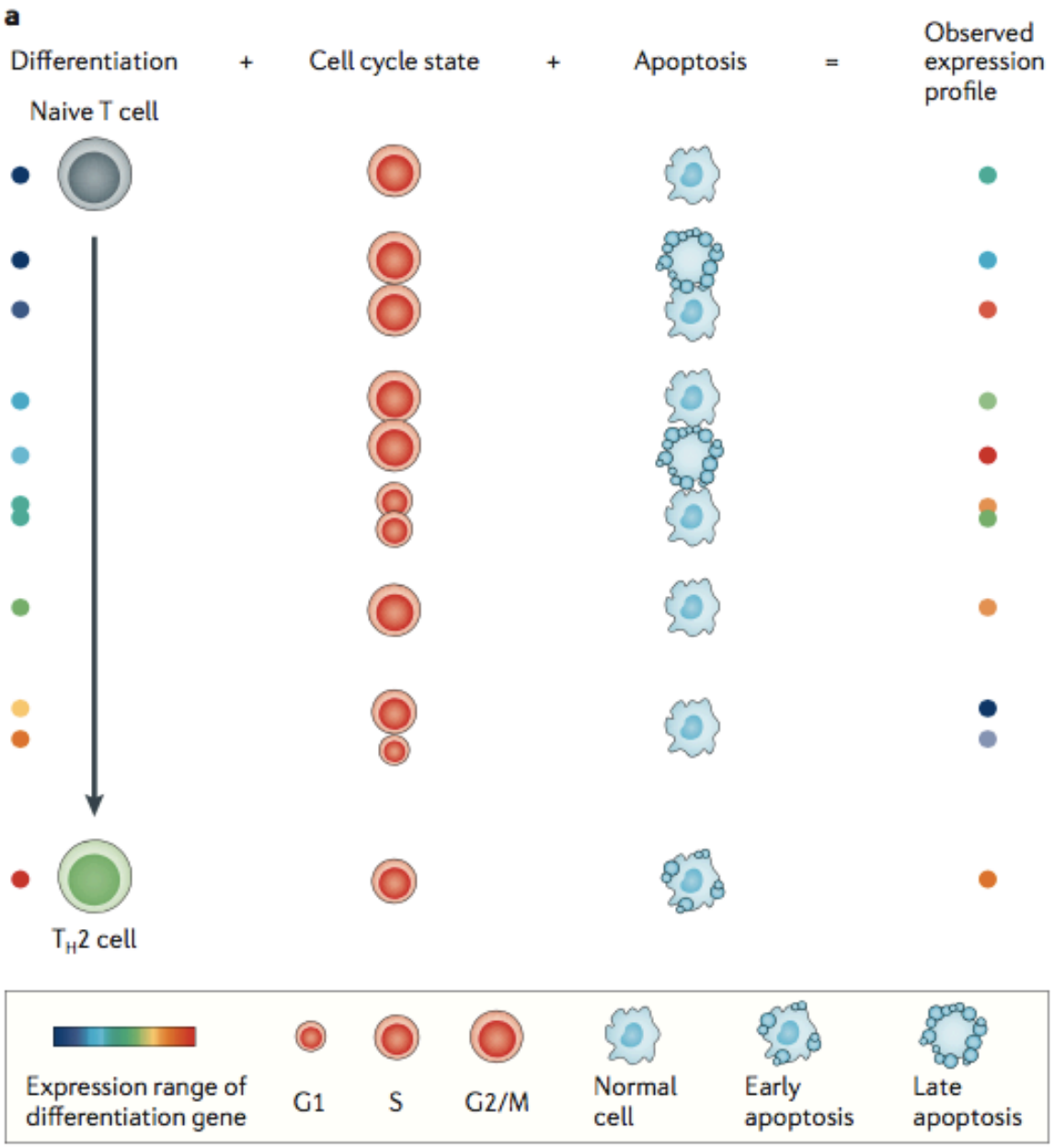
Liu and Trapnell, F1000, 2016

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

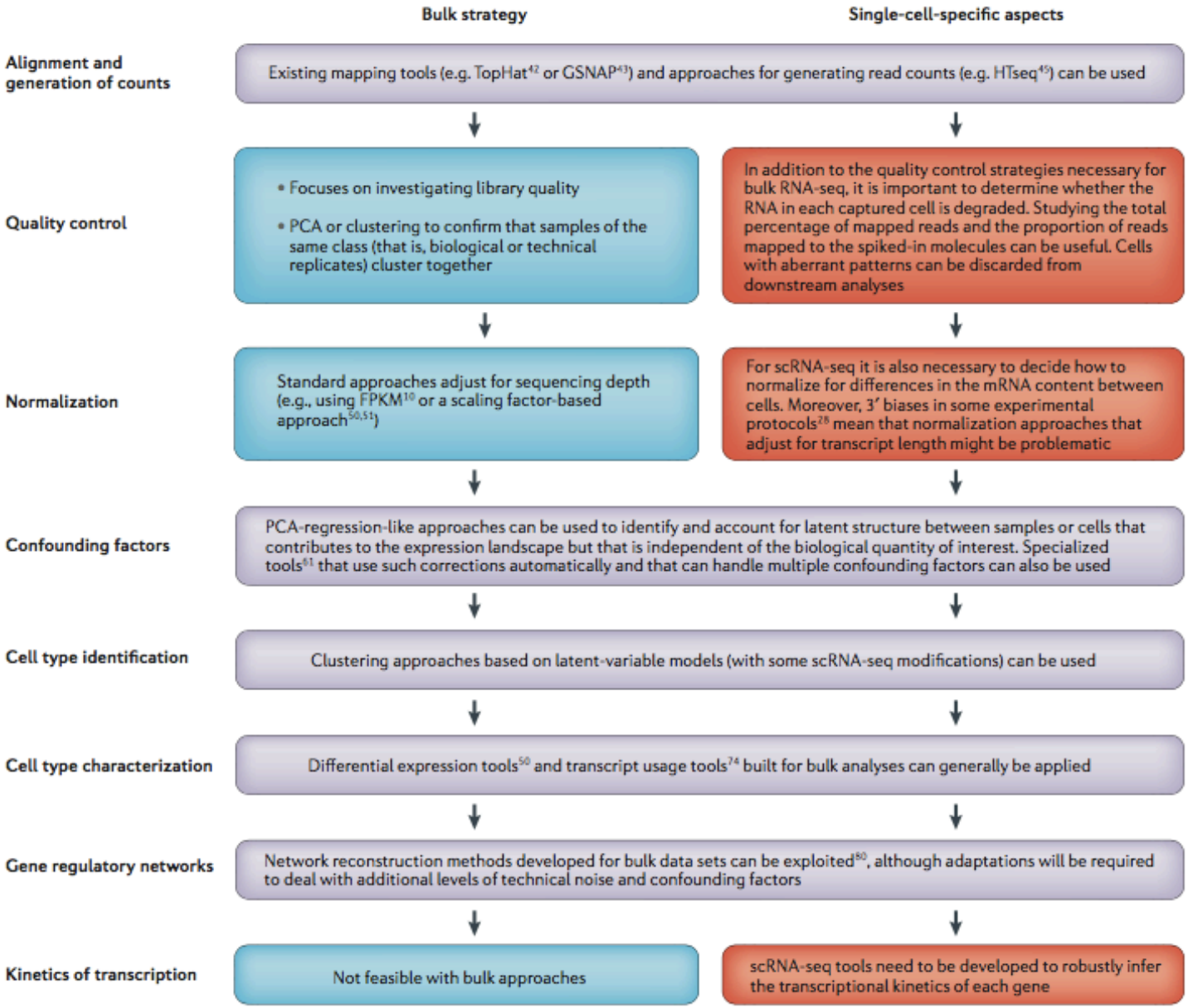
$$y_g \sim \mathcal{N}(\mu_g \mathbf{1}, XX^T + \sigma_v^2 CC^T + v_g^2 \mathrm{II}) \quad (1)$$

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 $\Sigma = XX^T$.



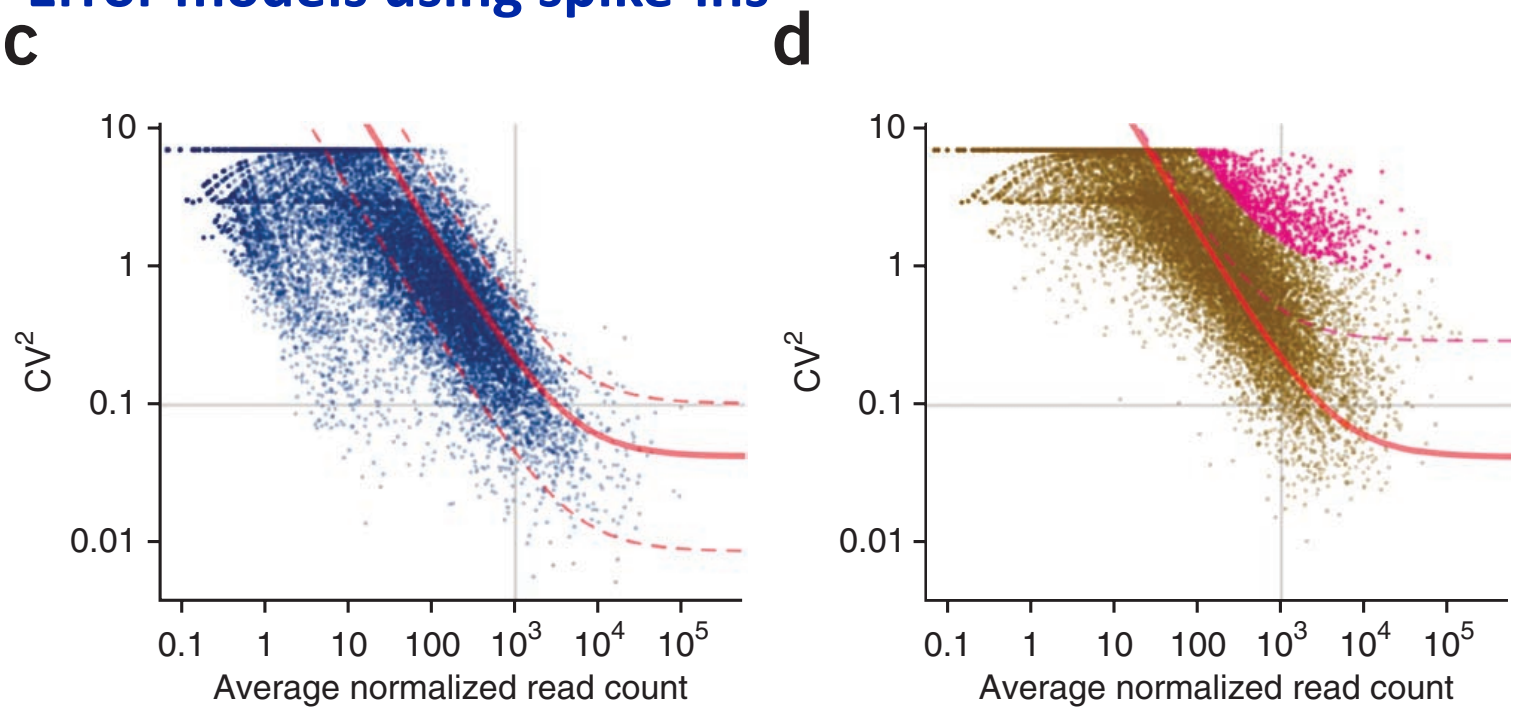
Stegle et al., NBT, 2015; Buettner et al., NBT, 2015

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



Stegle et al., NBT, 2015

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

Brennecke et al., Nature Methods, 2013

scRNA-seq Gotcha #1: batch effects

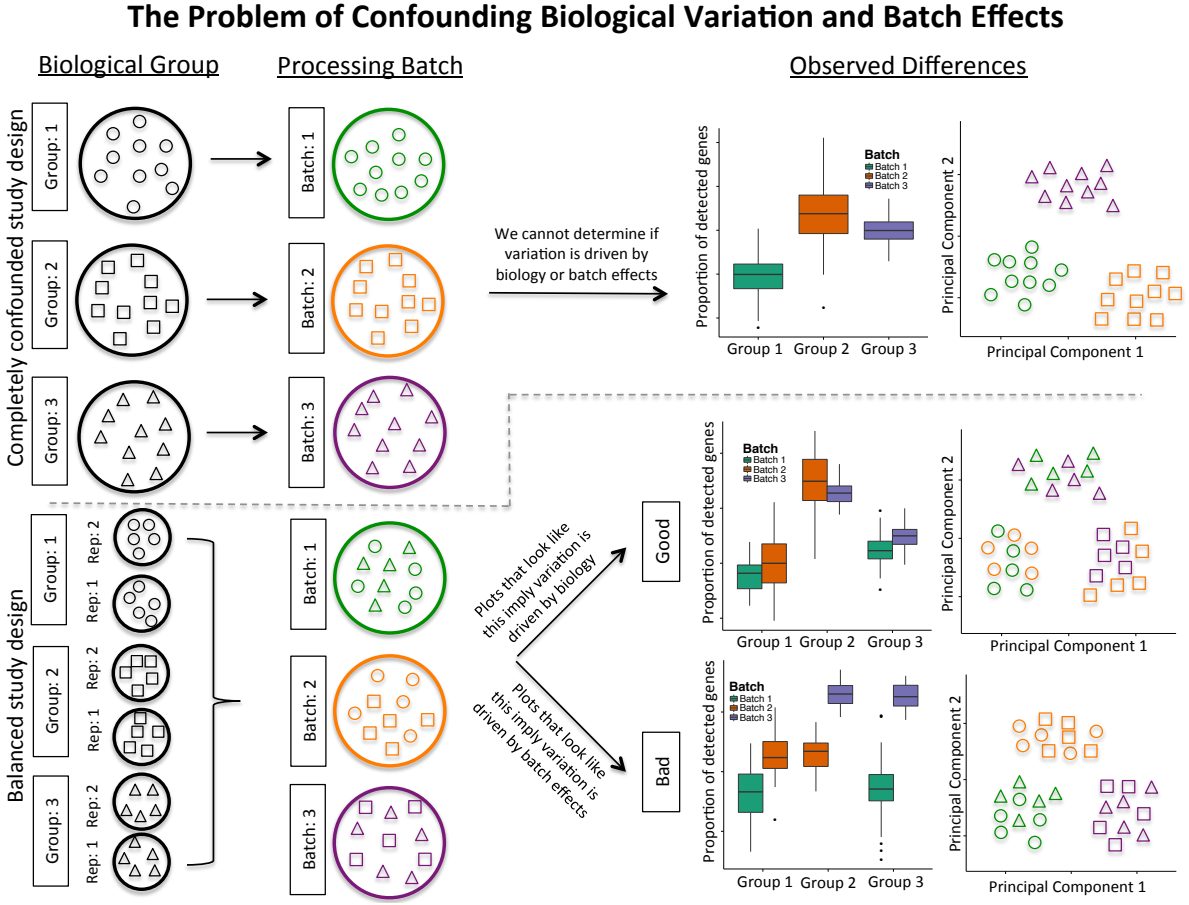
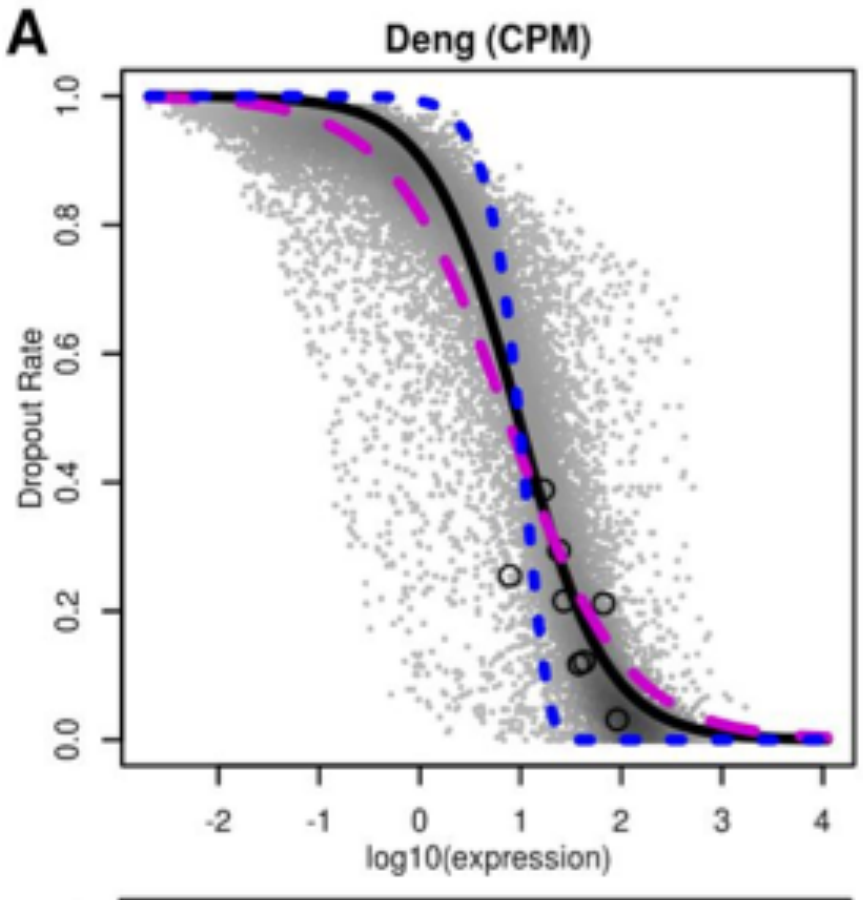


Figure 1: The problem of confounding biological variation and batch effects. The top sectio depicts a completely confounded study design of processing individual cells from three biological groups (represented by shapes) in three separate batches (represented by colors). In this case, we cannot determine 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).

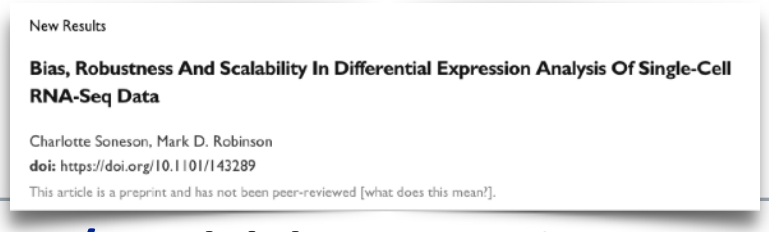
n.b.: some debate recently about the prominence of dropout in the newer droplet-based platforms.

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


New Results

Bias, Robustness And Scalability In Differential Expression Analysis Of Single-Cell RNA-Seq Data

Charlotte Soneson, Mark D. Robinson

doi: https://doi.org/10.1101/143289

This article is a preprint and has not been peer-reviewed [what does this mean?].

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

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

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

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

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

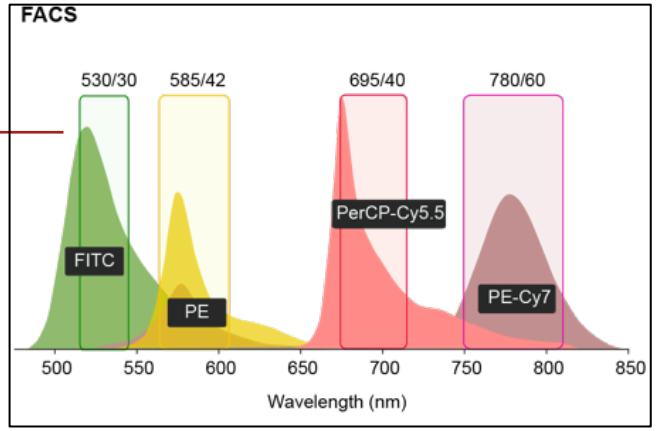
$$p(f) = \sum_{x \in X} p_S(x)p_G(fx)$$

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

Flow Cytometry: 6-12 markers measured

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



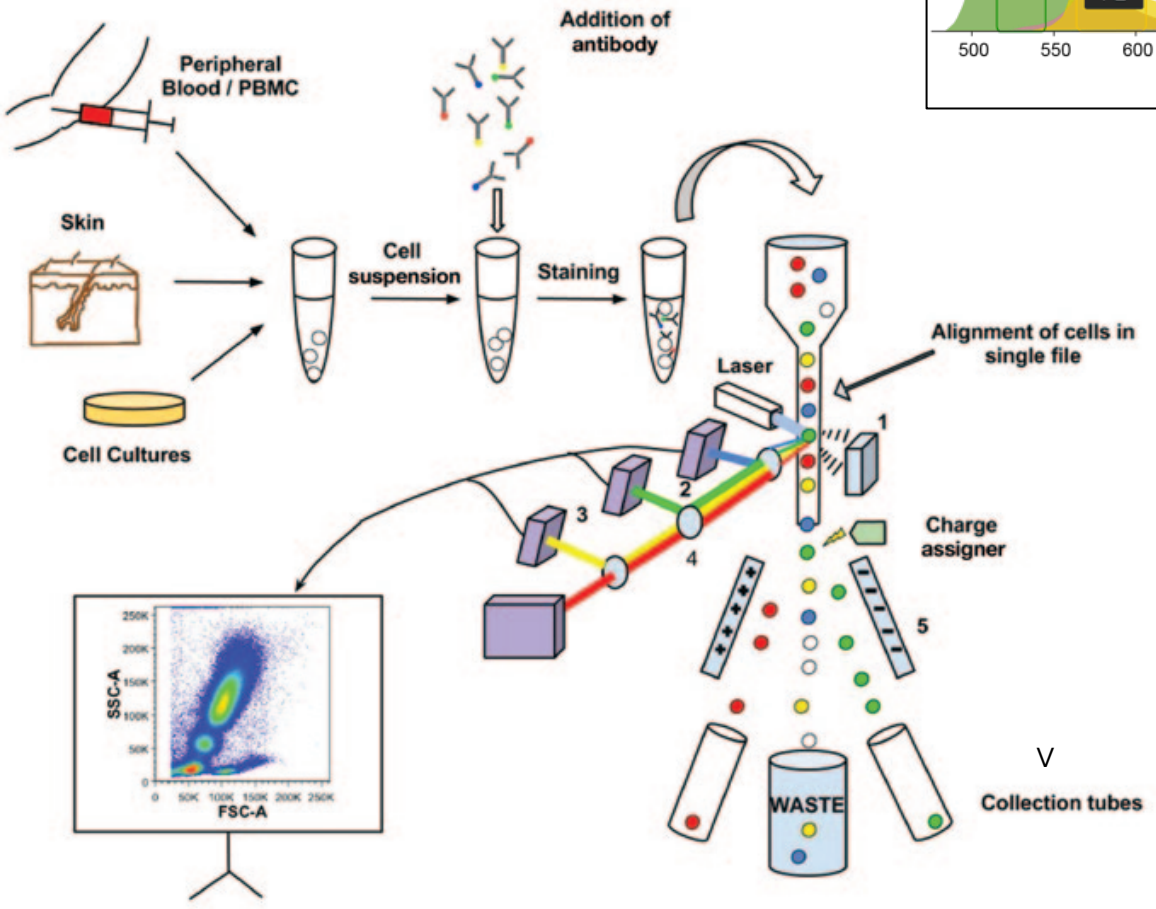
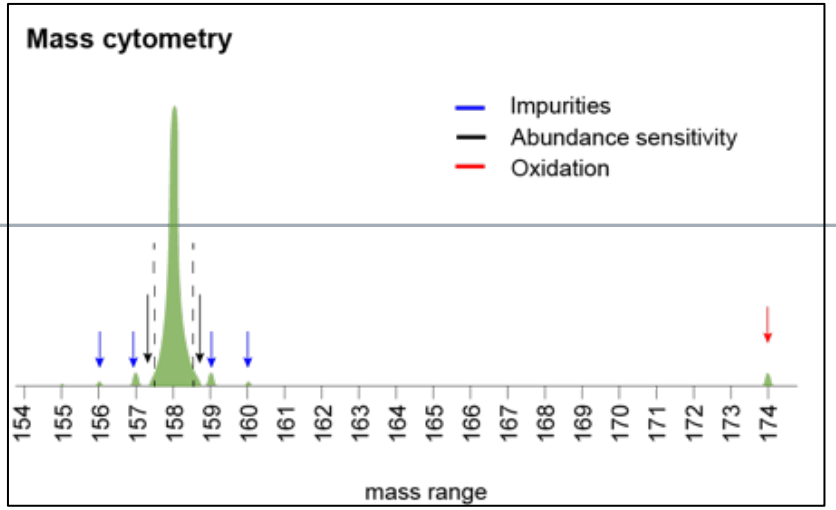
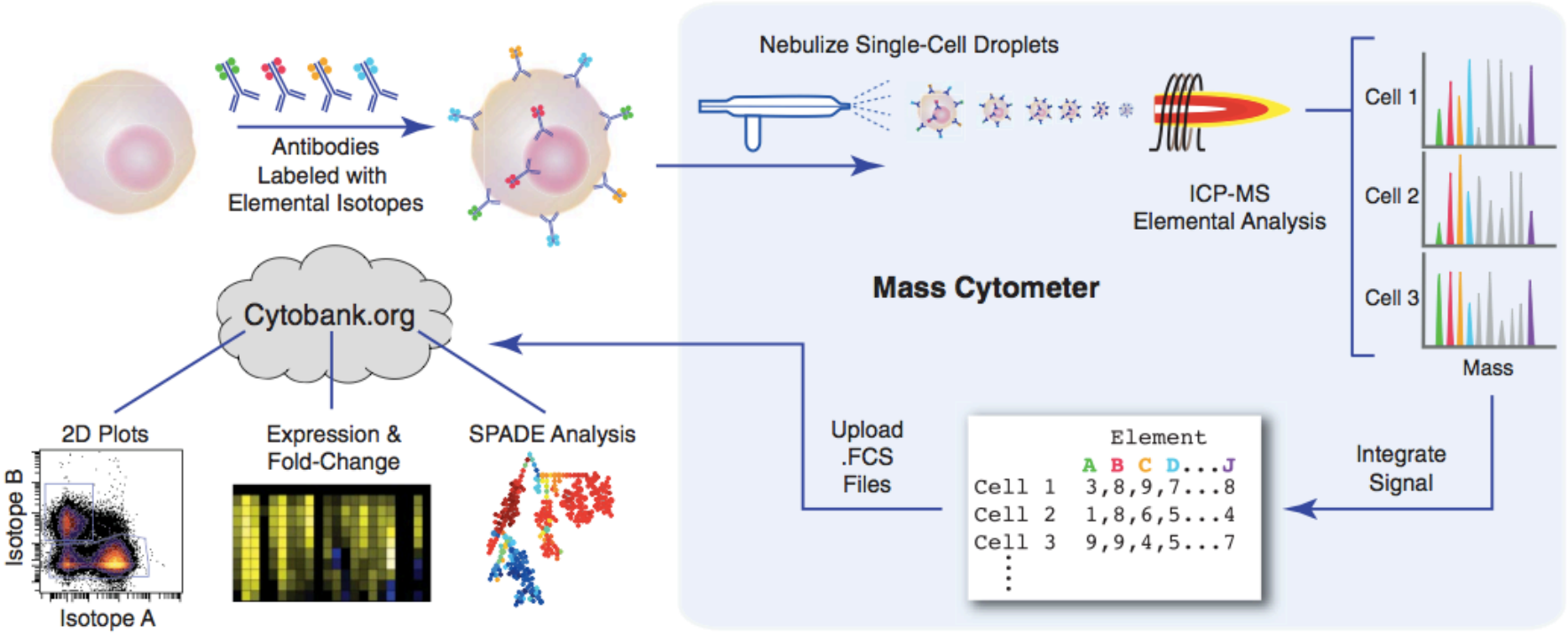


Figure 1. Schematic representation of a flow cytometer. For details please see text. (1) Forward-scatter detector, (2) side-scatter detector, (3) fluorescence detector, (4) filters and mirrors, and (5) charged deflection plates.

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

Mass cytometry (30-50 markers)

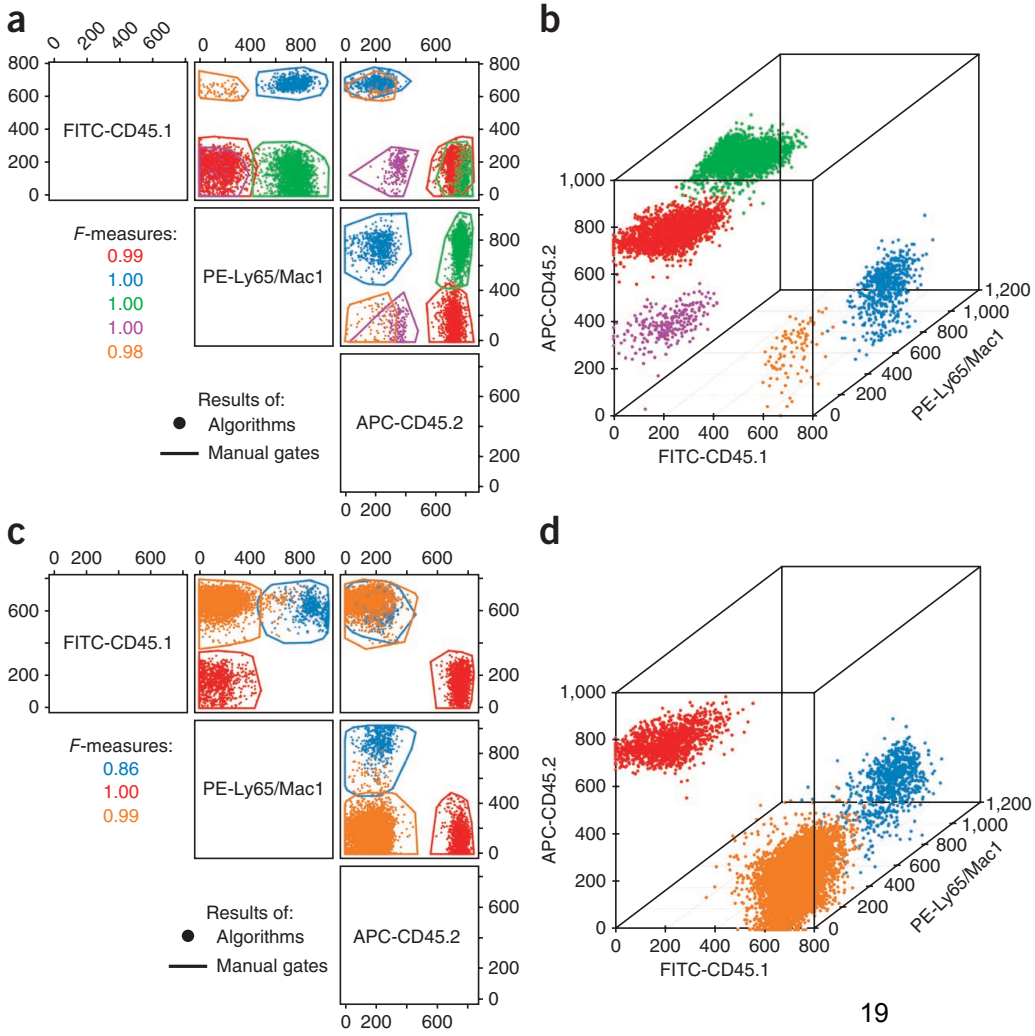




Aghaeepour et al., Nature Methods, 2013

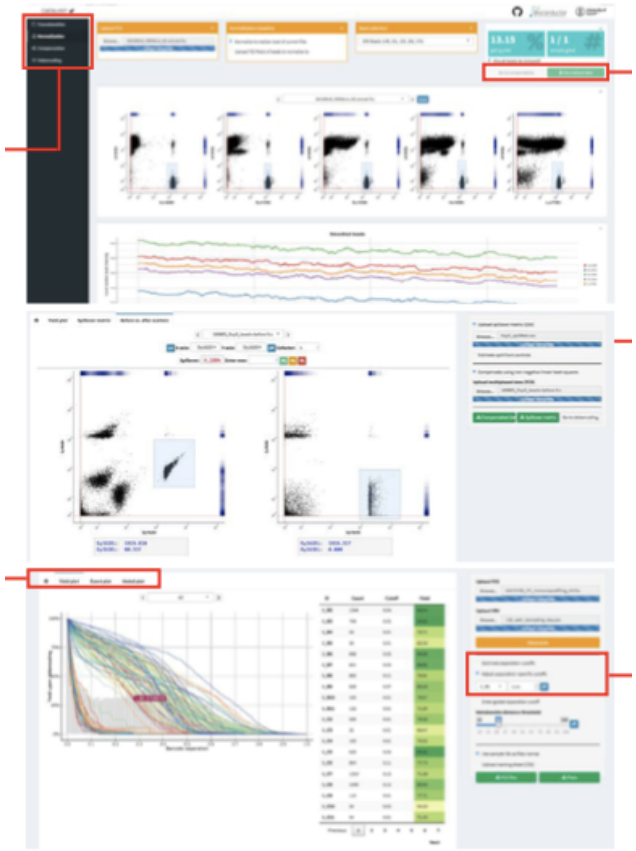
Manual gating versus clustering

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.



CATALYST package + **CATALYSTLite** Shiny app available: debarcoding, normalization, compensation

Helena



New Results

Channel crosstalk correction in suspension and imaging mass cytometry

Stephane Chevrier, Helena Crowell, Vito Riccardo Tomaso Zanotelli, Stefanie Engler, Mark D. Robinson, Bernd Bodenmiller

doi: https://doi.org/10.1101/185744

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract | **Info/History** | Metrics | Preview PDF

ARTICLE INFORMATION

doi https://doi.org/10.1101/185744

History September 7, 2017.

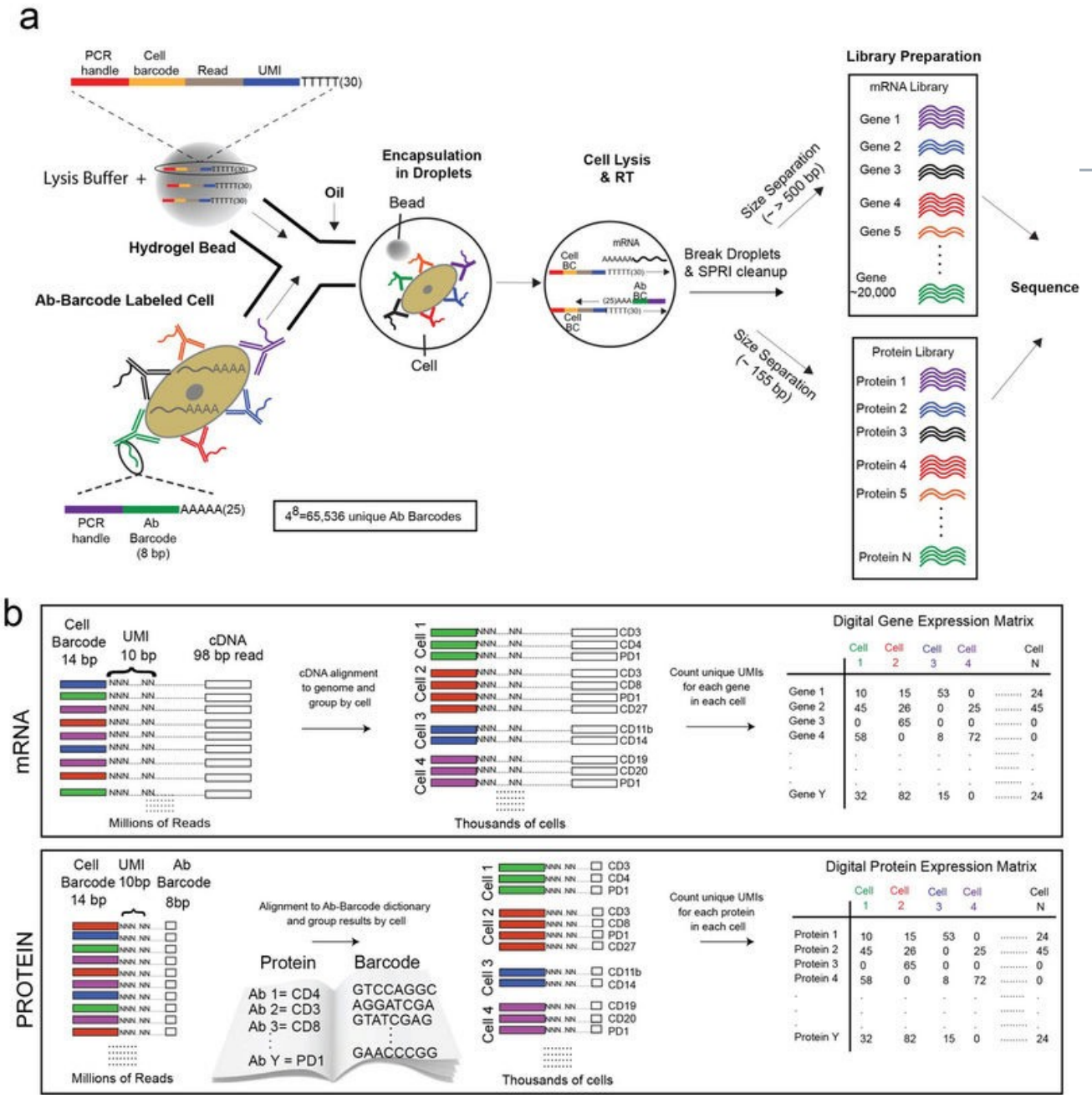
Copyright The copyright holder for this preprint is the author/funder. It is made available under a CC-BY 4.0 International license.

https://catalyst-project.github.io/

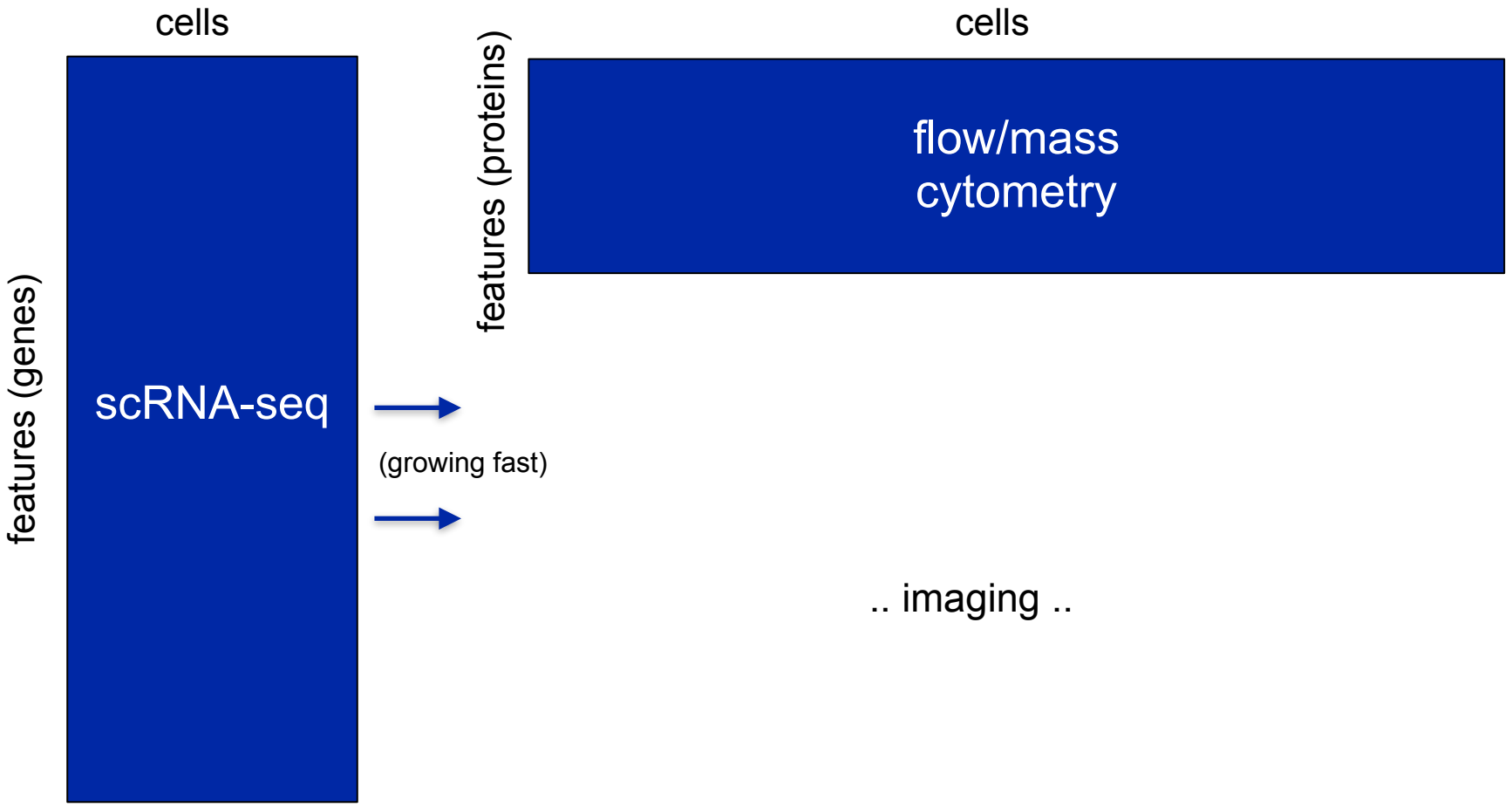
software on Bioconductor/github, preprint (Sept 7th 2017)

https://f1000research.com/posters/6-1662

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



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)

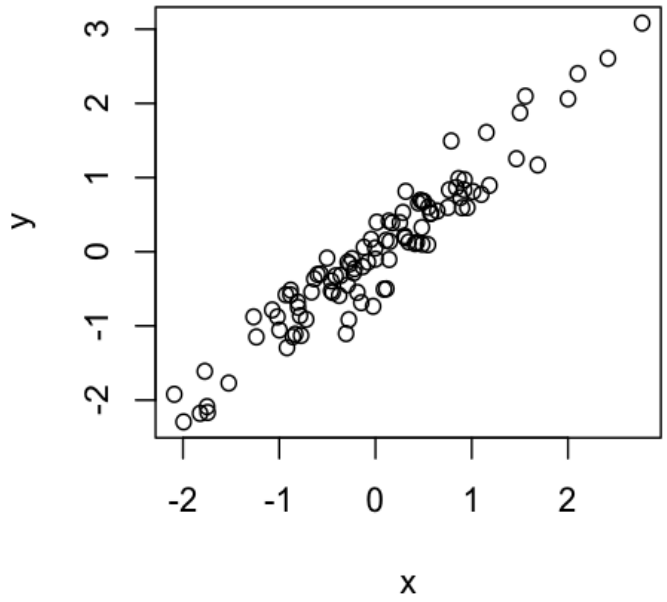
Data analytic techniques exist to **project** high-dimensional data (typical situation: 15k-20k gene expression measurements for each of N cells/samples) into a small number of dimensions (2 or 3, for humans)

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.

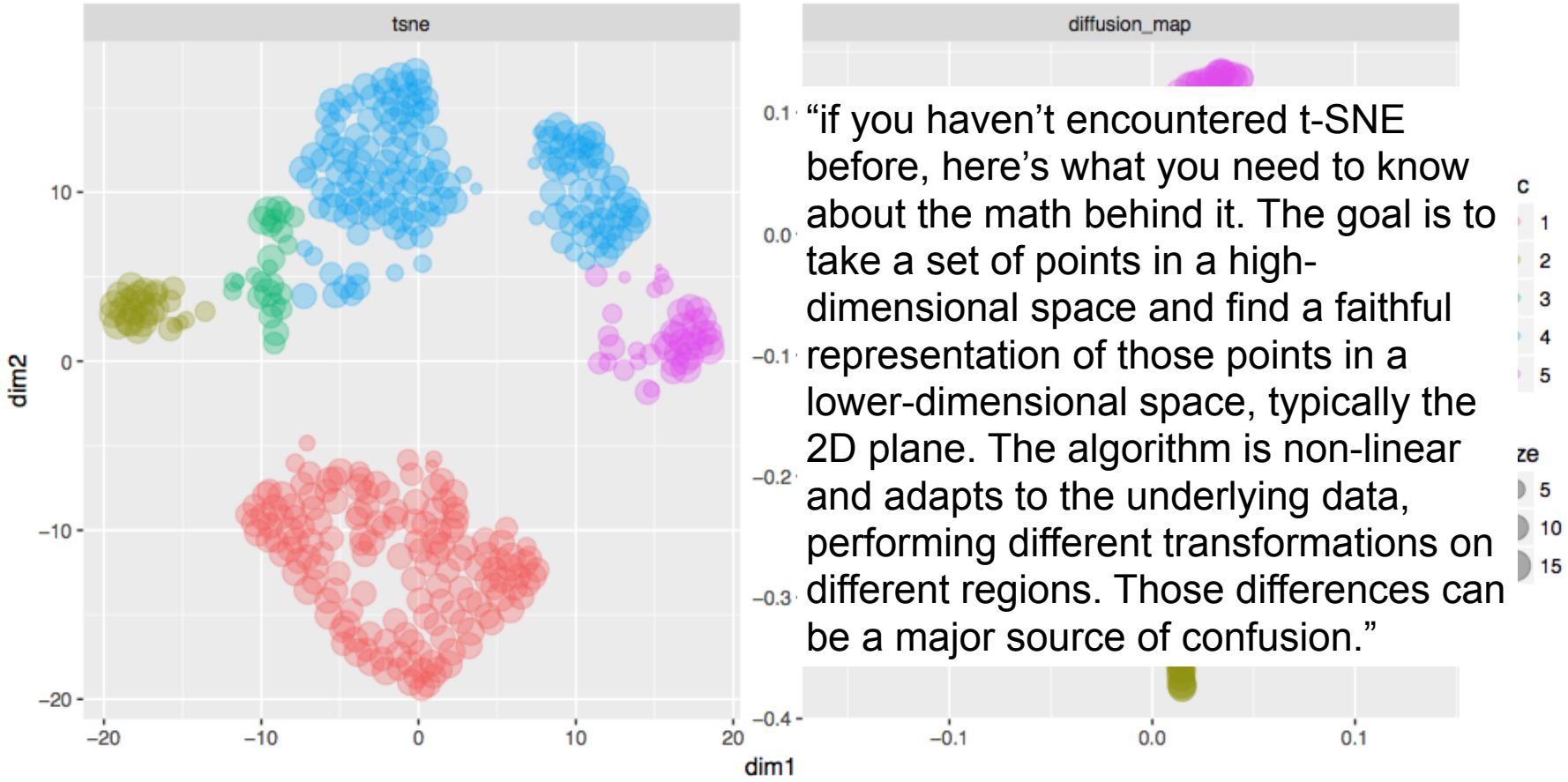
Visual explanation:

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



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

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

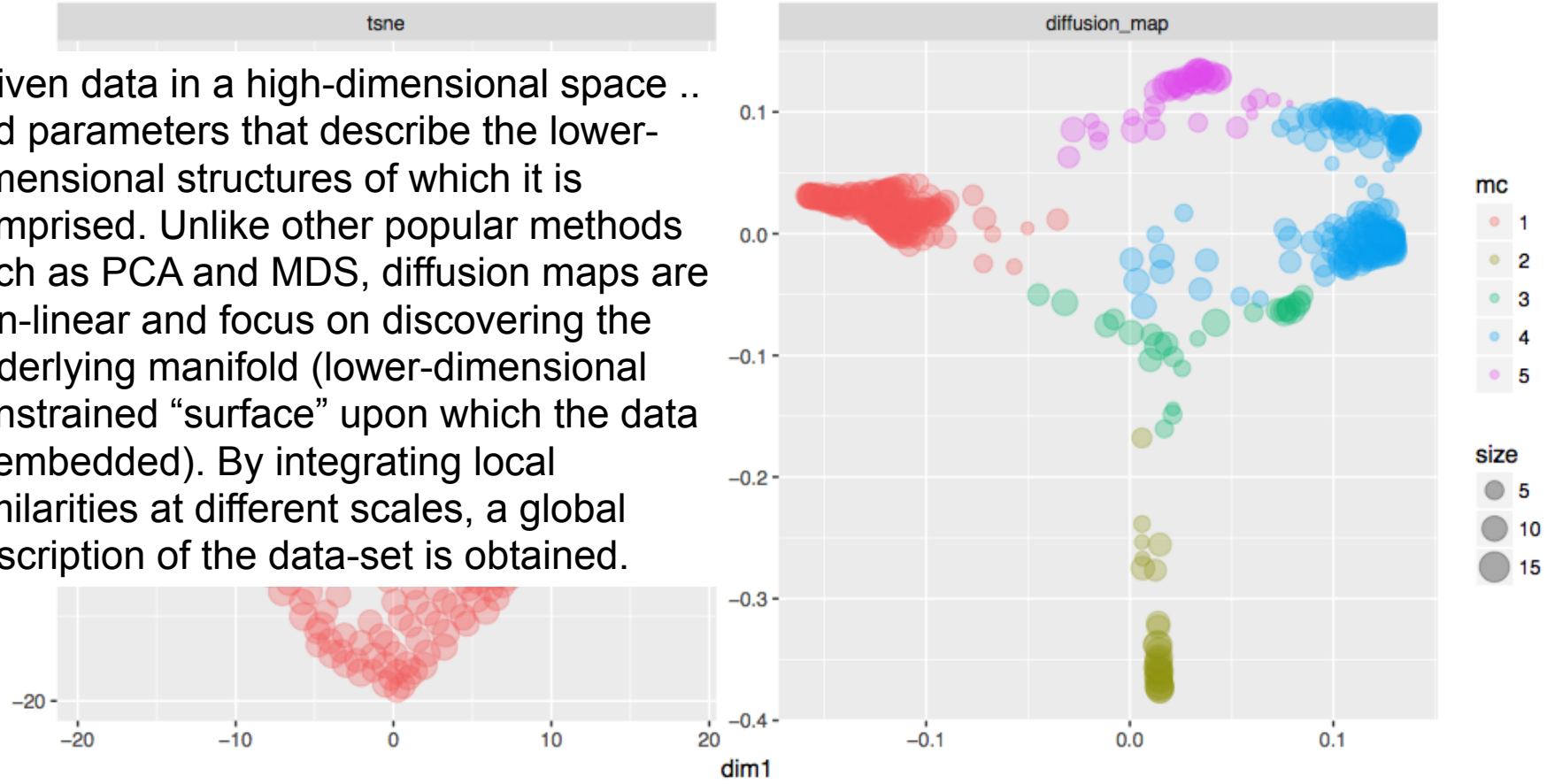


"if you haven't encountered t-SNE before, here's what you need to know about the math behind it. The goal is to take a set of points in a high-dimensional space and find a faithful representation of those points in a lower-dimensional space, typically the 2D plane. The algorithm is non-linear and adapts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

http://inside.mines.edu/~whereman/talks/delaPorte-Herbst-Hereman-vanderWalt-DiffusionMaps-PRASA2008.pdf

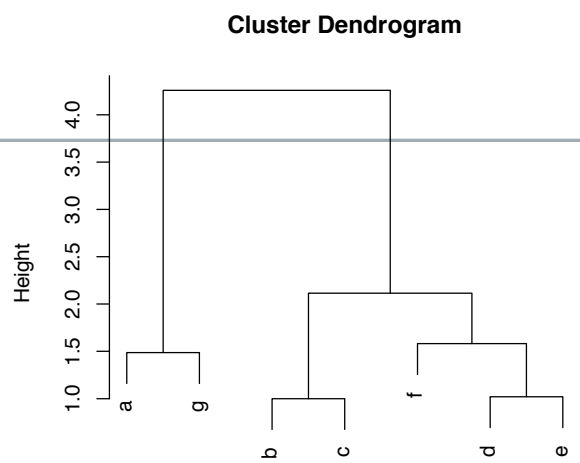
tSNE (t-dist'd stochastic neighbour embedding) + diffusion maps

"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

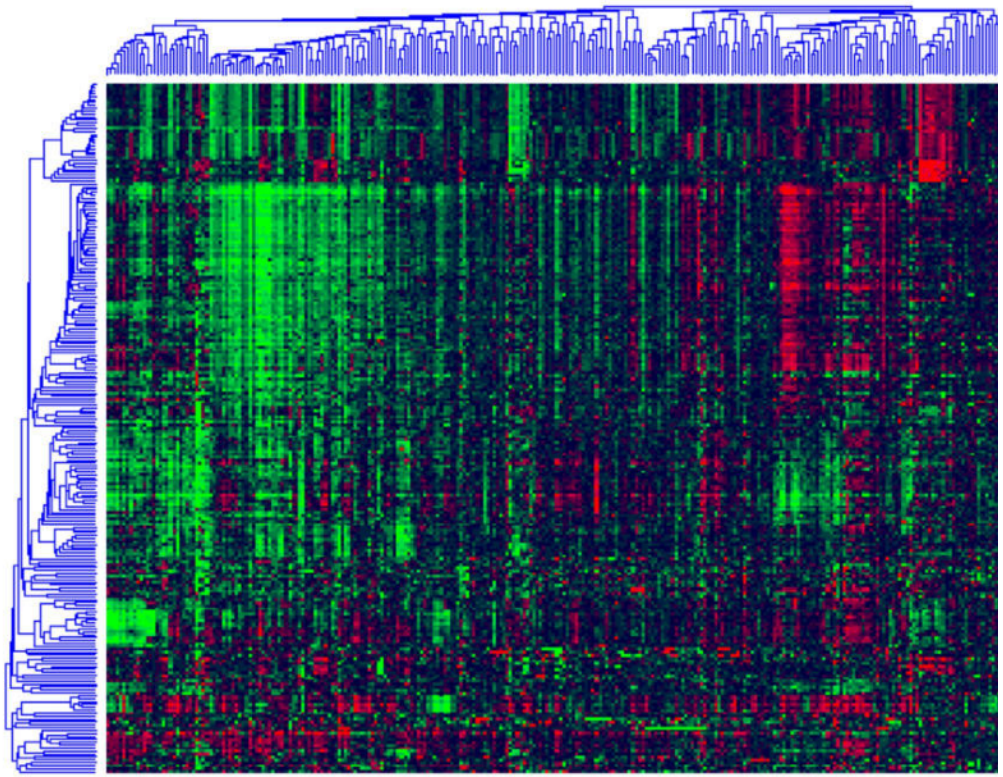
Hierarchical (Agglomerative) Clustering



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.

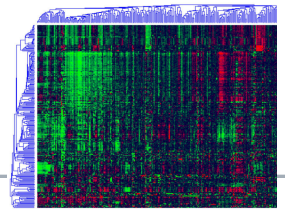
Linkage: determines how clusters are merged into a tree



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

Euclidean distance:

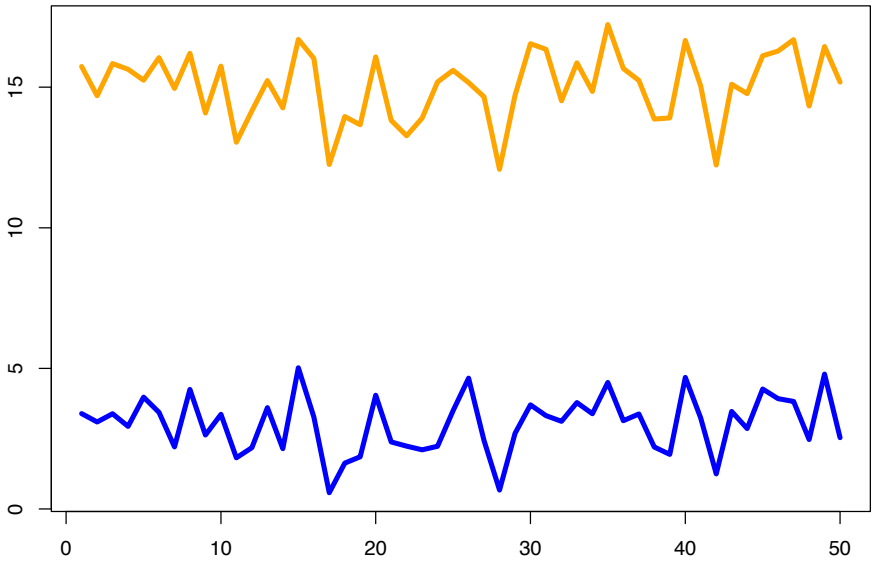
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

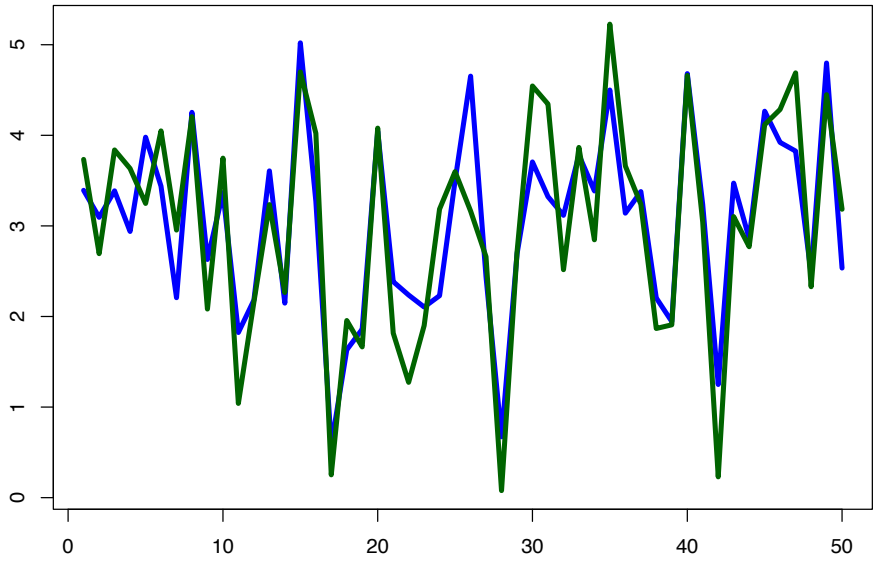


Are these “vectors” similar ?

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

It depends how you define similar.

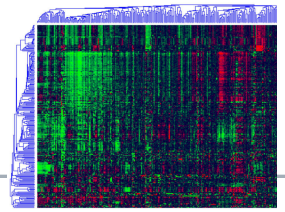




Euclidean distance: 84.84 3.92

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

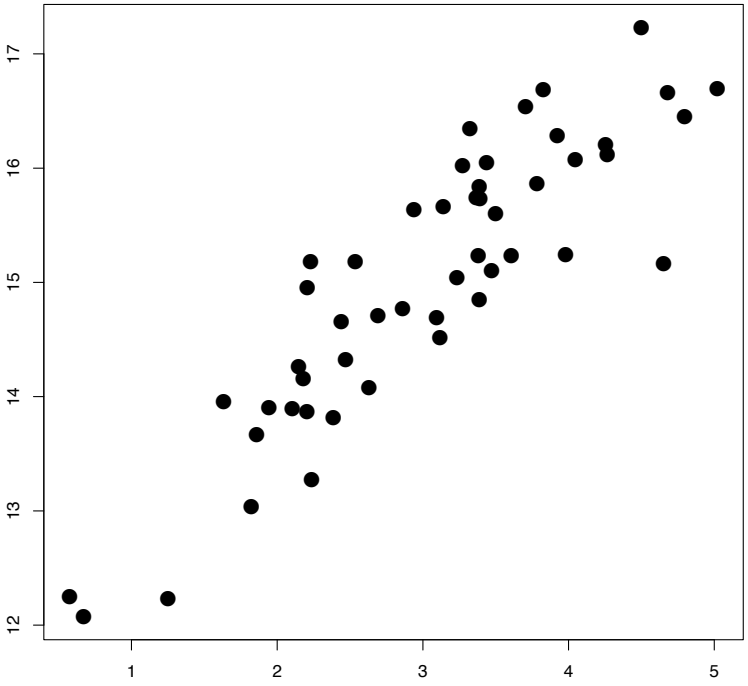
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

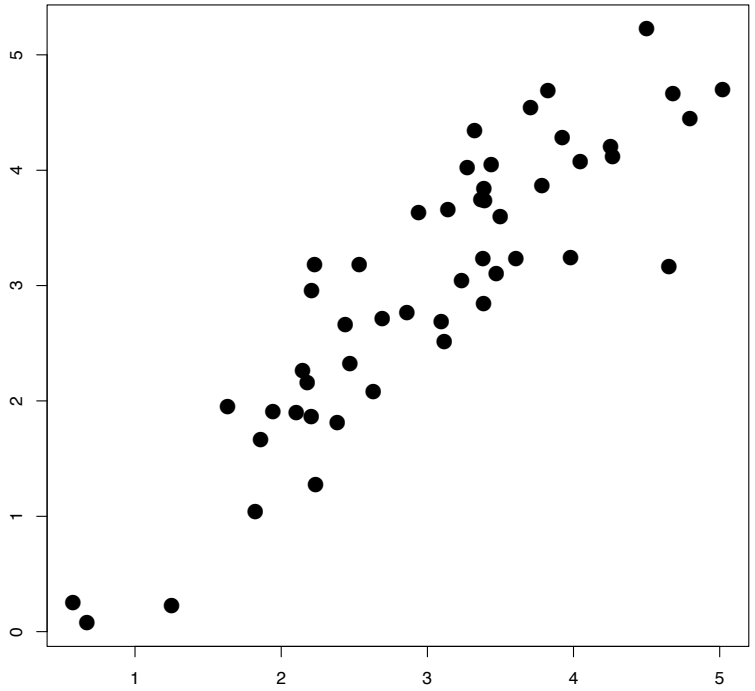


Are these "vectors" similar ?

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

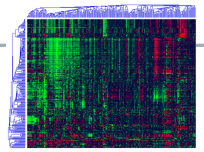
It depends how you define similar.





Correlation: 0.89 0.89

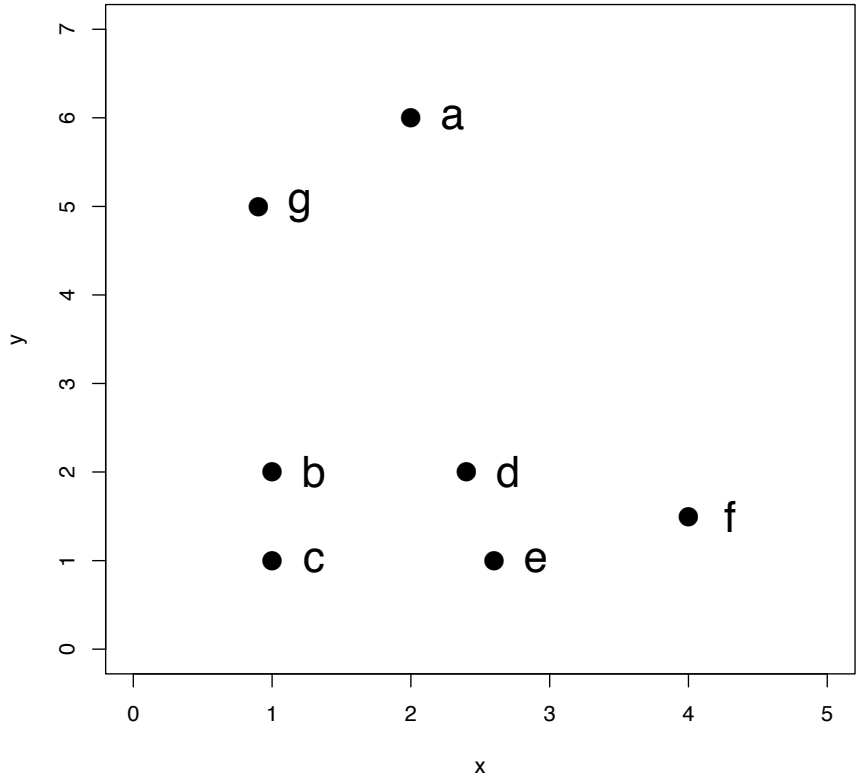
http://en.wikipedia.org/wiki/Hierarchical_clustering



Hierarchical (Agglomerative) Clustering

Start with distances.

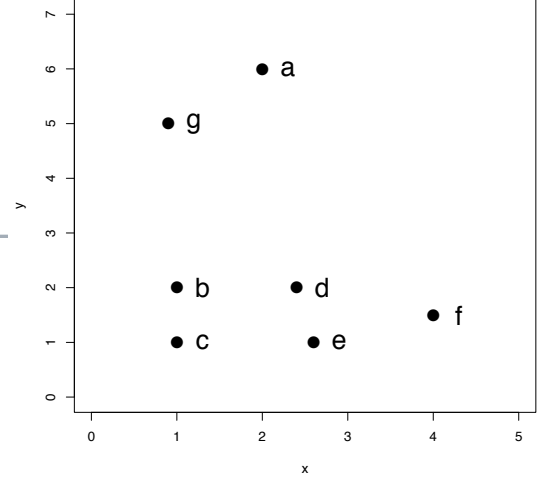
Linkage: determines how clusters are merged into a tree.

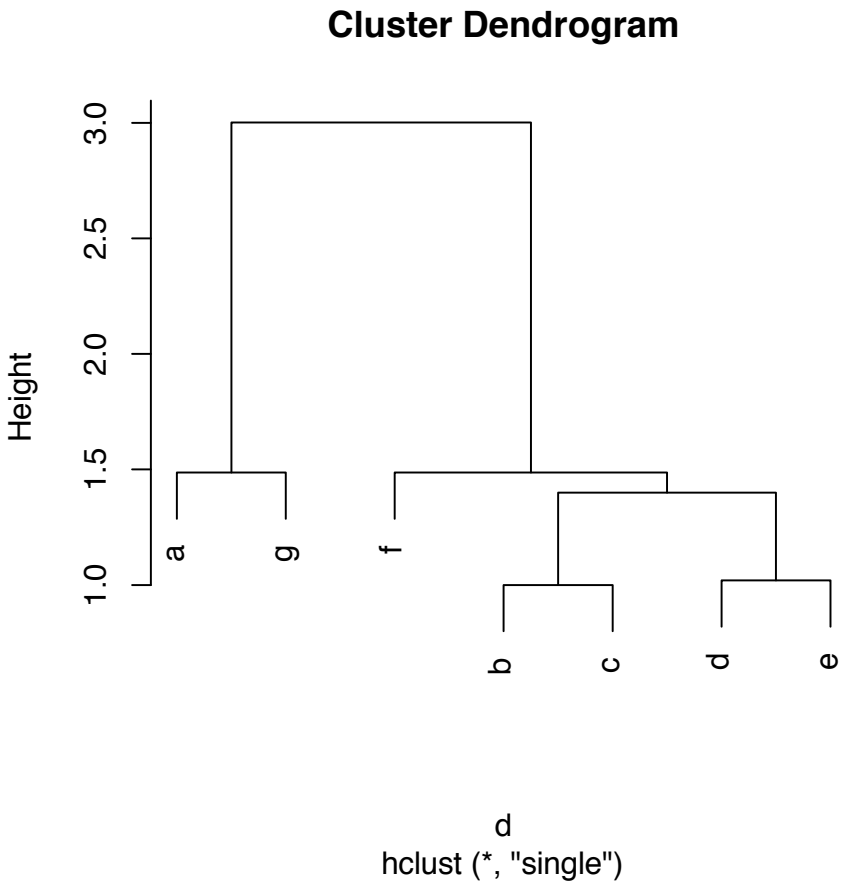


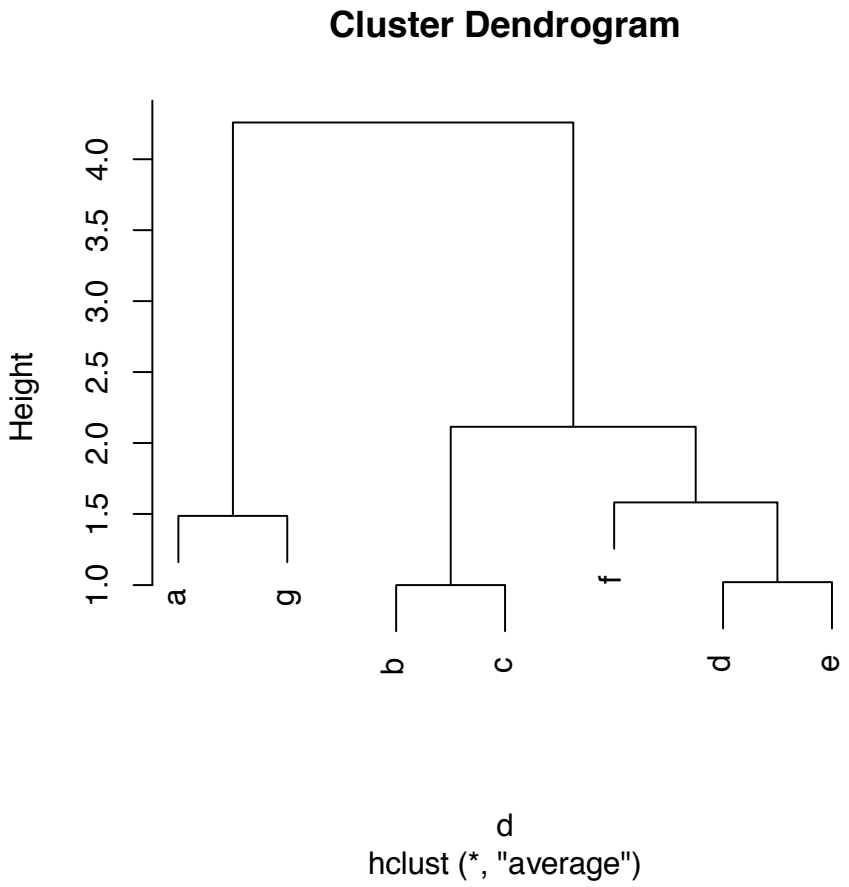
From eyeballing, here is a likely set of merges:

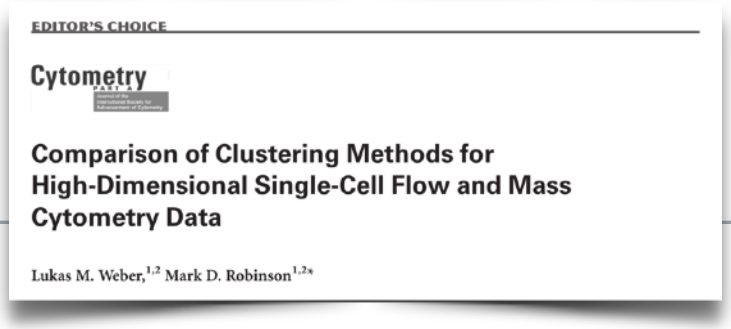
b,c
d,e
a,g,
(d,e),f
(b,c),((d,e),f)
ALL

Different linkages








EDITOR'S CHOICE

Cytometry

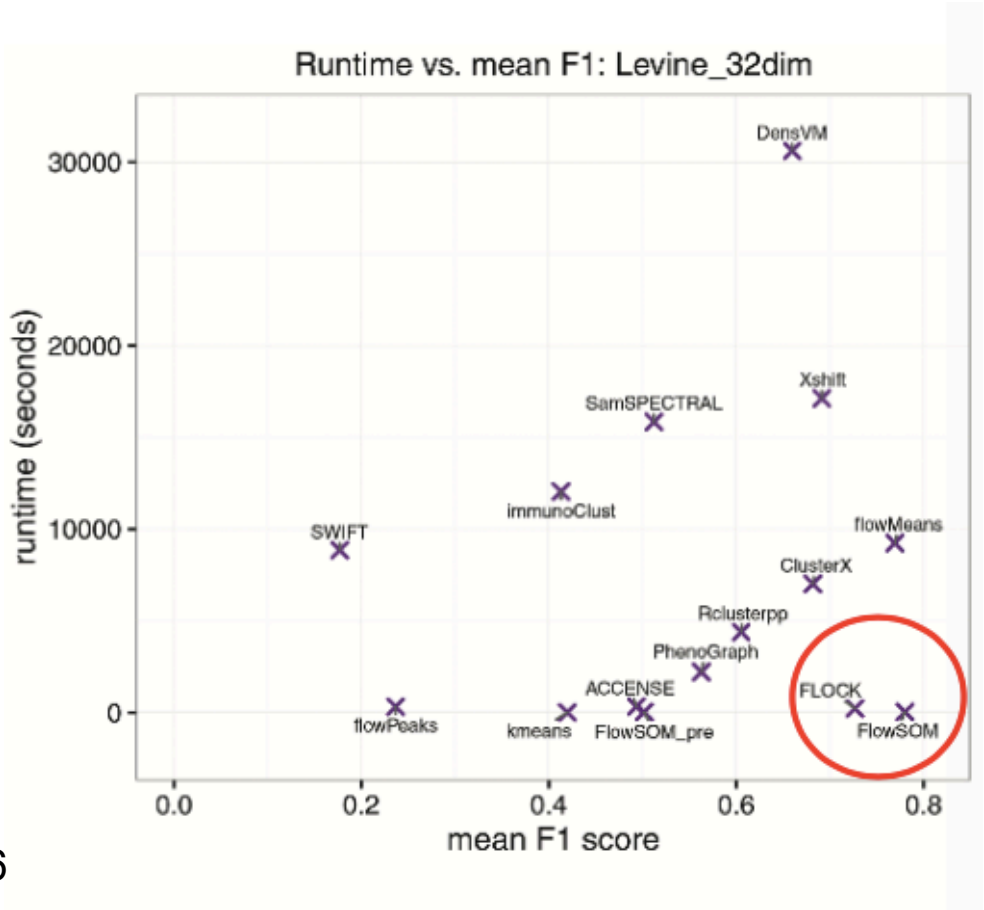
Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data

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

Many clustering algorithms (cytometry data): Phenograph + FlowSOM + ..

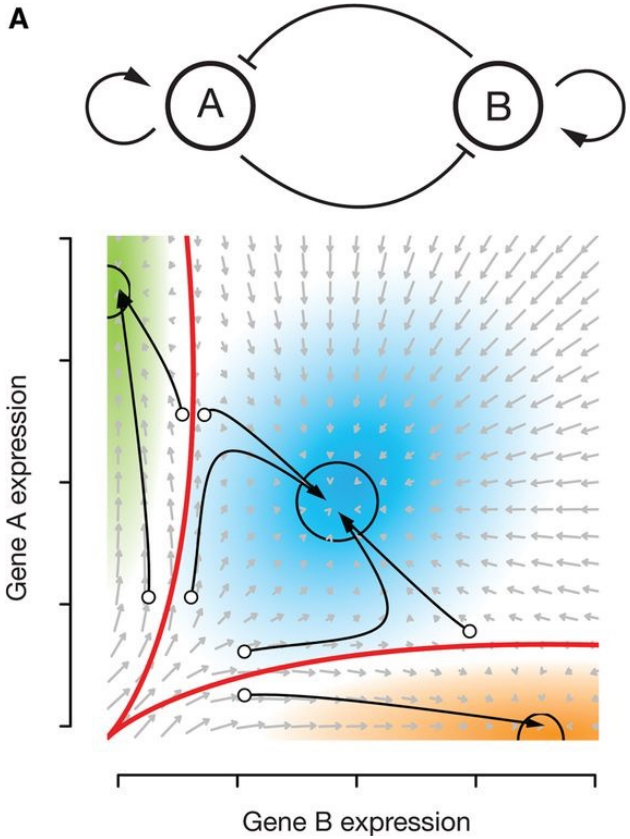
Using various "high" dimensional datasets with a manual gated truth.

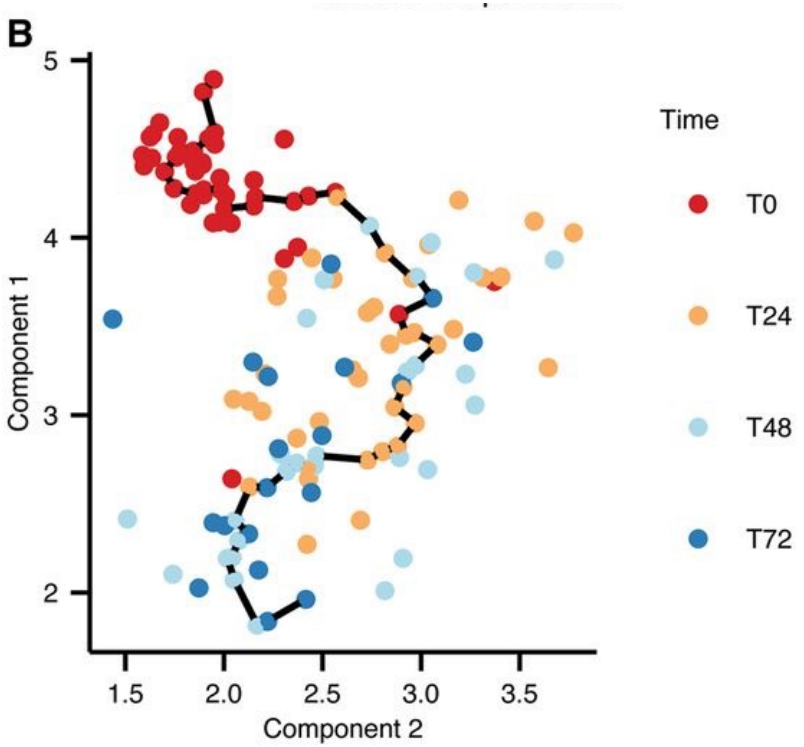
F1 score = geometric mean of precision and recall (mean = averaged over the known populations)



Weber and Robinson, Cytometry A, 2016

Trajectory analysis





Trapnell, 2015 Genome Research

Trajectory analysis

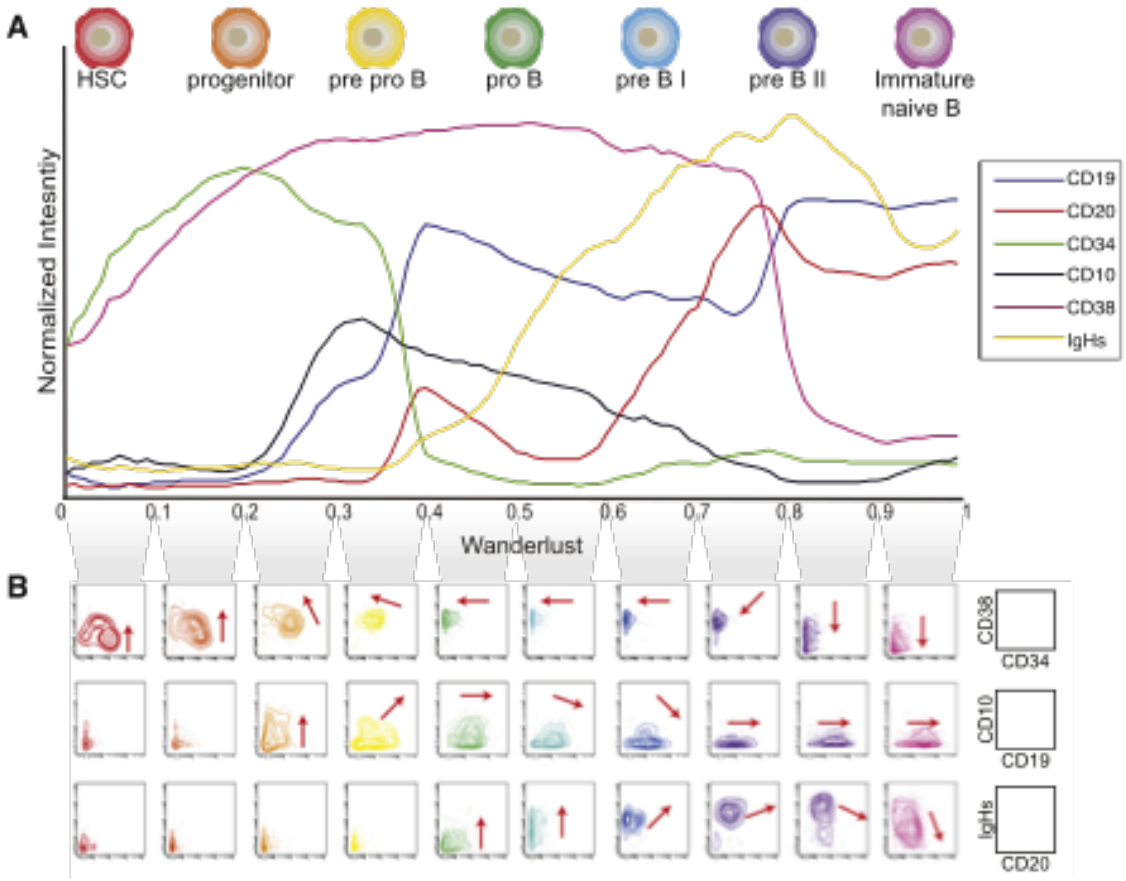


Figure 2. Wanderlust Confirms Known Hallmarks of Human B Cell Development and Is Consistent across Healthy Individuals

(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.

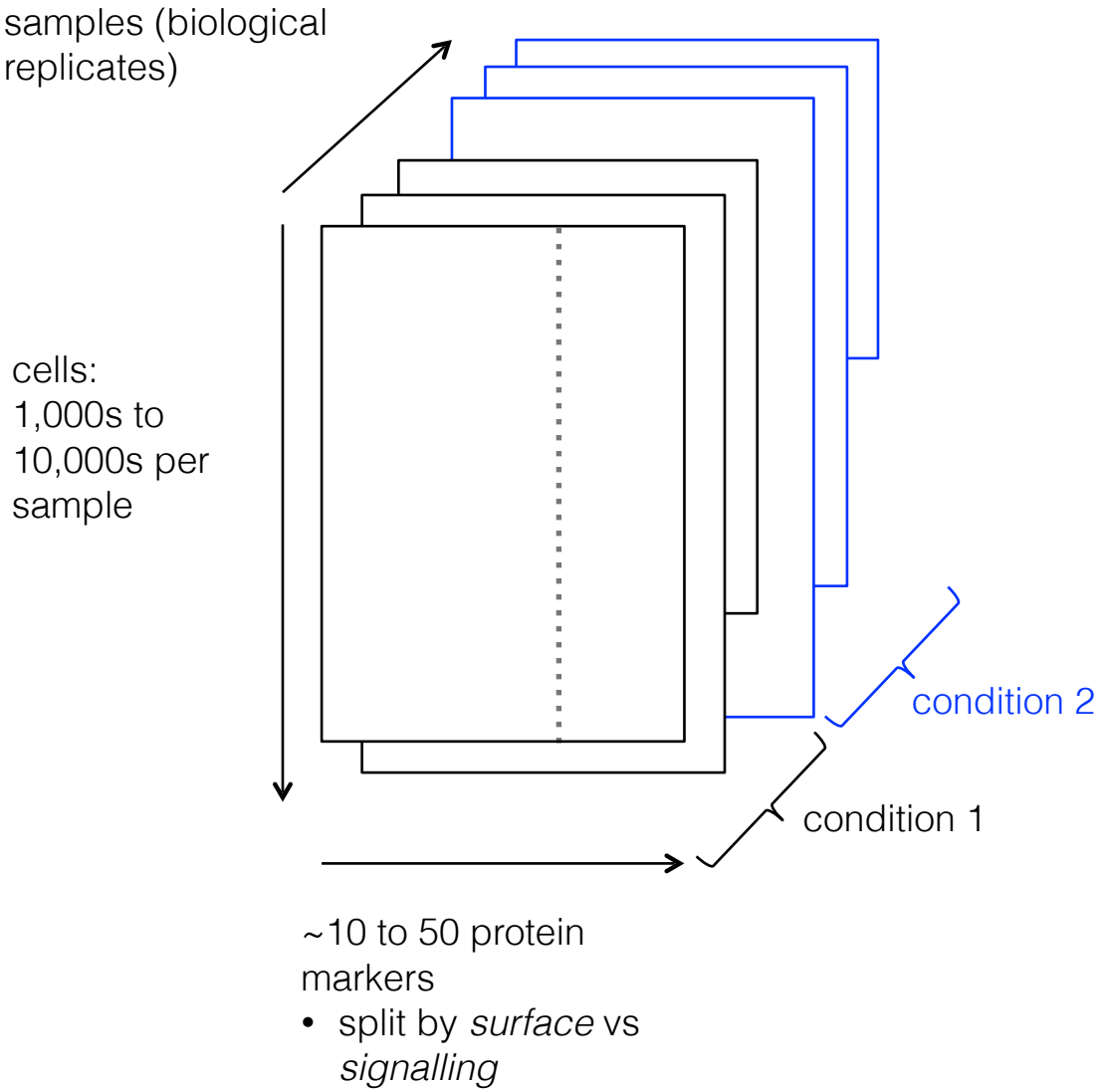
(B and C) Biaxial plots (B) demonstrate the two-dimensional progression of cellular marker expression (red arrow) across the Wanderlust trajectory taken in segments of 0.1. (C) Distribution of marker expression across the trajectory for CD24, TdT, and CD10. The green line indicates the relative standard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

Multi-sample differential analyses for single cell data



Aims:

- Part 1: **differential abundance** of cell populations (clusters)
- Part 2: within-cell-type **differential marker expression**

Lukas

Gosia