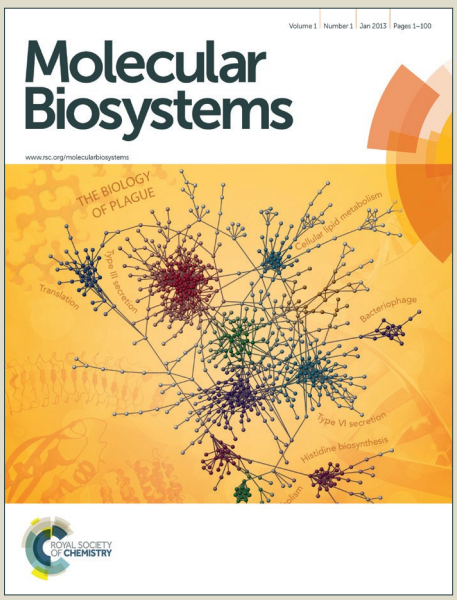
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Reconstructing and analysing cellular states, space and time from gene expression profiles of many cells and single cells

Mirko Francesconi[a,b] & Ben Lehner[a,b,c*]

Affiliations

[a] EMBL-CRG Systems Biology Unit, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain

[b] Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, 08003 Barcelona, Spain

[c] Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

[*]Corresponding author

E-mail: ben.lehner@crg.eu

Abstract

Genome-wide gene expression profiling is a fast, cheap and standardised analysis that provides a high dimensional measurement of the state of a biological sample. In this review we describe some of the computational methods that can be applied to identify and interpret sources of variance in gene expression in whole organisms, organs, tissues or single cells. This allows the identification of constituent cell types and states in complex mixtures, the reconstruction of temporal trajectories of development, differentiation and progression, and the reconstruction of spatial patterning. When applied to genetically variable samples, these methods allow the efficient investigation of how genetic variation influences gene expression in space and time.

1. Sources of variability in biological samples

Differences in gene expression measured by RNA sequencing or using DNA microarrays can be purely technical (this can be estimated[1]) or due to biological differences between samples. These biological differences may be part of the experimental design or they may be due to uncontrolled experimental variation in the state of each sample[2-7]. Uncontrolled experimental variation is usually regarded as a confounder but it might also be interesting when the sources of variation are correctly identified and understood (Fig1). In this review we provide an overview of the computational methods that can be used to identify and understand controlled and uncontrolled variance in gene expression datasets and highlight examples of how these have been used to make interesting discoveries.

In gene expression data from mixtures of cell types such as tissues, organs or tumours, an important source of variance derives from cell heterogeneity. This may come from the process of interest, for example a condition might change the proportion of different cell types within tissues[8]. Or different amounts of contaminating tissues could confound the analysis, for example with patient biopsies. In this case it is important to factor out this confounder to increase the power to detect differences in the tissue of interest.

Even when analysing populations of sorted pure cell types (or unicellular organisms) heterogeneity is still present. Cells are not static entities but they dynamically adjust their state in response to environmental stimuli. It might be hard to control all (micro-) environmental factors that can trigger some cell response. A typical example is growth rate: cells adjust their global gene expression according to growth rate - increasing for example ribosomes and translation-related genes and decreasing stress related genes - no matter what the growth-rate-limiting factor is[9, 10]. Any perturbation that changes growth rate will consequently impact gene expression.

Moreover cells go through oscillations, for example cell cycles, metabolic cycles and circadian rhythms. This heterogeneity will be present in unsynchronized single cells but also potentially in bulk measures of asynchronous populations where oscillatory dynamics are convoluted in the average expression data; population measures reflect the average of gene expression of different cell cycle or other oscillatory stages

weighted by the fractions of cells that are at each stage. Conditions, treatments (or confounders) that change the oscillatory dynamics will change these fractions and will leave a trace on gene expression.

Even more complex dynamics are present when studying systems undergoing development or differentiation. When studying fast developing model systems such as *Drosophila melanogaster* or *Caenorhabiditis elegans* that complete development, even a few hours' difference in precise staging can introduce substantial variation in gene expression[5, 11, 12]. This might be related to the experimental conditions of interest (that could cause a delay in development) or it could be a confounder if not properly controlled (experimental batches might be at slightly different developmental stages).

In summary, even in a simple perturbation and expression profiling experiment it is important to understand and correctly decompose gene expression variance into the corresponding sources. This will help to: (1) better understand and interpret the global effects of a treatment/condition/mutation, for example as developmental delays or changes in growth rate or cell composition; (2) tease apart specific effects and direct targets of treatment/condition/mutation beyond the developmental delays or changes in growth rate; and (3) control for experimental confounders and increase the power to detect the effects of interest. However, the real power of understanding and decomposing expression profiles is the application to large datasets in which new cell states, spatial patterning or temporal 'trajectories' of expression can be identified. In addition, if this is performed in combination with genetically varying samples, the impact of genetic variation on these states or spatial and temporal patterns can be determined.

2. Computational approaches for inferring cell states, sample compositions, time and space from gene expression

Global gene expression data is typically highly redundant because many genes (for example those involved in the same biological processes) share correlated expression profiles. Thus it is useful to represent (map) the high dimensional data listing the

expression of tens of thousands of genes onto a lower and more interpretable dimensional space, a task known as dimensionality reduction. Ideally after this step the data would be represented by a few dimensions that account for most of the variance in the data and where each dimension represents a distinct, interpretable biological process. At the same time it is desirable to filter out non-biological, uninteresting variance (technical noise).

Ab initio (unsupervised) methods for dimensionality reduction: principal components analysis

Depending on the goal of the study and on the nature of the biological process of interest, different computational approaches for reducing dimensionality can be used. If the goal is to discover new cell types, states or trajectories *ab initio* from the data, one of most widely used technique is principal component analysis[13] (PCA). PCA rotates the data into new orthogonal coordinate systems where the axes (components) are linear combinations of the original variables and represent the directions of maximal variance in the data. This means that the first component explains most of the variance, the second component explains most of the residual variance after subtracting the first one, and so on. Thus retaining only the first few principal components accounts for most of the variance present in the original variables and filters out noise. PCA can be performed by eigen decomposition[14, 1514, 1514, 1514, 1514, 1514, 1514, 1514, 15] of the covariance matrix of the data or, more efficiently, by singular value decomposition (SVD) of the data matrix[14, 15].

For each component PCA outputs singular values (or eigenvalues) that indicate the variance explained, sample scores (sample coordinates on the component) and gene loadings (the coefficients of the genes in the linear combination, i.e. how much each genes contribute to the component).

Visualizing the sample scores on the first few principal components helps to provide an overview of the global structure of the data, for example by highlighting clusters or trajectories (Fig2). The biological meaning of each component can be deduced by analysing the gene loadings, for example by traditional gene set enrichment analysis[16, 17]. To this purpose Chung et al. developed a method to systematically identify genes significantly associated to principal components avoiding over-fitting[18]. However

interpreting principal components is not always easy because the components might be enriched in many distinct biological processes and, vice versa, the same biological process can be enriched in different components. The reasons for this include that in PCA the components are defined maximizing the variance explained as a criterion,and they are also constrained to be orthogonal. These conditions and constraints mean that the principal components will not necessarily correspond to separate biological sources of variance. This is an important limitation of PCA when the aim is to clearly separate and remove unobserved confounders from interesting sources of variance (signal). Using PCA there is a risk is that interesting signal is also removed together with confounders.

Other unsupervised methods to decompose variance

Additional methods that relax some or all PCA constraints have been developed for better separating sources of variance and to increase interpretability, known as factor analysis methods. Similarly to PCA, these methods search for linear combinations of the data (factors) that best explain the correlations among the variables but improve interpretability by allowing for further rotations that better capture the underlining structure of the data. For example the varimax[19] method further rotates the data after PCA (preserving orthogonality) in a way such that the genes have high or low loadings only in one factor. Promax[20] also allows oblique rotations thus relaxing the orthogonality constraint. Other methods based on factor analysis have been recently proposed to better estimate sources of variance with the aim of correcting gene expression from hidden confounders, such as surrogate variable analysis (SVA)[21], probabilistic estimation of expression residuals (PEER)[22] and remove unwanted variation (RUV)[23] . The last two methods also allow estimating hidden factors in a semi-supervised manner only on selected gene sets (i.e. control genes) to minimize the risk of explaining away the signal together with confounders.

Among unsupervised method to deconvolve sources of signals, Independent Component Analysis[24] (ICA) is one of the most flexible. Its rationale stems from central limit theorem, which states that mixtures (convolution) of independent signals tend to be normally distributed. Thus an effective strategy to separate the hidden independent source signals from the measured mixed signal is to find linear combinations

(components) that maximize non-gaussianity (rather than variance as in PCA). Several measure of non-gaussianity have been proposed such as kurtosis, negentropy or mutual information[24]. Components obtained by ICA are linearly independent (a stronger condition than uncorrelated as in PCA) but they do not need to be orthogonal as in PCA. ICA has been applied to gene expression data[25, 26] and it can outperform PCA in teasing apart independent biological processes underlying expression differences[26](Fig3).

Using a reference expression dataset

If one wants to match data to predetermined states or types then a good approach is to compare the data to existing reference expression profiles. A simple approach is to use a subset of relevant genes from a reference dataset to build a model that predicts the corresponding state in the dataset of interest. For example, the expression of many genes in yeast correlates linearly with growth rate under many different conditions[9] and a simple linear model including these genes can infer the relative growth rates of new conditions from gene expression[27]. Similarly, the proportion of cells in a sample in each stage of the cell cycle can be inferred by comparison to reference datasets defining sets of genes activated at different phases during the cell cycle. The expression level in each of these genes in an asynchronous population (vector A) can then be expressed as the weighted average of their expression at each cell cycle stage (matrix R) where the weights (matrix W) are the unknown fractions of cells at each cell cycle stage in the population (A=WR). These fractions can be determined by solving the system for W[28]. The same modelling framework can be used to deconvolve cell type fractions from gene expression data of whole tissues when cell type-specific expression signature are known[8]. Similarly, cell type-specific expression profiles can be inferred in complex tissues if the fraction of each different cell type in the tissue is known[29].

More powerful methods to match expression data to a reference dataset include partial least squares (PLS) and canonical correlation analysis (CCA) [5]. These two related statistical techniques analyse the relationship between two datasets (covariance for PLS, correlation for CCA) of multiple dependent and multiple independent variables such as two gene expression datasets measuring the same genes in two different sets of conditions. They decompose the covariance (or correlation in CCA) between the two datasets by finding linear combinations of the reference dataset that best explain linear

combinations of the independent dataset, in a manner similarly to PCA except that in this case only the variance shared between the two datasets is taken into account. Using these methods as multiple advantages: first one can find multiple processes shared with a reference and quantify how much variance is explained by each. Further, the approach leaves the variance not explained by the reference untouched. This avoids that specific signals of interest are explained away together with global confounders when the aim is to correct gene expression before downstream analysis. This is in contrast to using reference genes whose expression in the data of interest might reflect a combination of underlining processes some of which might not be present in the reference.

Tackling non-linearity

Many biological processes such as the cell cycle, development and differentiation show complex non-linear dynamics such as oscillations or bifurcations. In these cases, linear methods (such as those described above) are a useful first step to reduce dimensionality, visualize the data and filter out noise, but they cannot directly be used to order the data along a non-linear dynamic process. Reconstructing non-linear dynamics from the data might be challenging because classical distance measures are not appropriate to define, for example, how close two data points are in a trajectory and hence their ordering (Fig4a).

In some cases, simple transformations can be used to infer the correct dynamics and to order the data. For example if PCA (or ICA) transformed data lie on (a portion of) a cycle in a low dimensional space, a simple transformation in polar coordinates can recover the correct order of the data points along the dynamics[5, 14, 15].

In the case of more complex dynamics, finding the geometry of the data and ordering or clustering data points might be harder. When studying development and the data points lie in a single trajectory, ordering them can be seen as an instance of the well known travelling salesman problem to find the shortest path connecting all the points, for which many algorithms have been developed[30].

However, often data points are arranged in a more complex way than one simple trajectory as in the case of lineage bifurcations during differentiation[31]. General

approaches to this problem start by building a graph that connects data points only to their nearest neighbours (with the aim to preserve only the local distances) and then finding the minimum spanning tree (MST) that connects all the data points (Fig4B). In the simplest case, the diameter of this graph represents the dynamic trajectory along which data points can then be sorted[32](Fig4C). This strategy has been successfully applied to uncover trajectories and bifurcations both for low and medium dimensional data such as flow cytometry coupled with mass spectrometry (cyto-mass) or single cell quantitative real time PCR (qRT-PCR) expression measurements[33, 34] and for high dimensional gene expression data after applying a linear dimensionality reduction step such as PCA[30, 32], ICA[35] or a clustering step[36].

Alternatively, several methods have been developed that start by building a nearest neighbour graph and use the shortest path (geodesic) distance between points instead of the euclidean distance to perform non-linear dimensionality reduction and clustering. Examples include Isomap[37], locally linear embedding (LLE)[38] and laplacian eigenmaps[39]. Other non-linear dimensionality reduction methods such as diffusion maps[40] or the t-distributed stochastic linear embedding (t-SNE)[41] are based on alternatives to classical distance metrics but again with the same objective of preserving local similarities rather than global ones. Isomaps[31], diffusion maps[42] and t-SNE[43] have been used in a biological context to discover trajectories, bifurcations and cell heterogeneity in medium or high dimensional data on differentiation, development and disease.

3. Applications

Interpreting functional genomics data

One of the first applications of expression deconvolution was in the interpretation of systematic functional genomic data such as analysing the consequences of gene deletion. In an early study in yeast, Lu et al showed that it is possible to deconvolve the fraction of cells in each cell cycle phase from bulk microarray expression data in

asynchronous populations by using reference genes that oscillate during the cell cycle[28]. This deconvolution made it possible to evaluate the effects of various environmental and genetic perturbations on the cell cycle dynamics from bulk gene expression, characterizing both the specific phase of the cell cycle and the severity of defects (Fig5). For example, based on the changes in gene expression the authors inferred that about half out of the 300 tested gene deletions affect cell cycle progression.

More recently O'Duibhir et al performed a similar analysis on 1485 gene expression profiles of yeast gene deletion strains[10]. They first found that 25% of the 700 mutants that differ from wild type share a common expression signature that is very similar to a 'slow growth' signature induced by nutrient limitation[9] or environmental stress. In yeast, the growth rate, the stress response and metabolic activity are tightly coordinated with the cell cycle[9], and the authors argued that a change in the fraction of cells at different cell cycle stages in a population can account for expression changes in many different experiments[10].

Analysis of complex tissues

Gene expression deconvolution is also useful for interpreting physiological changes in complex samples such as tissues. Tissues are a mixture of cell types so differential expression can be driven by: (1) changes in the relative abundance of different types, (2) changes that occur only in a subset of cell types, (3) changes common to every cell type, or (4) a combination of these three. Expression deconvolution can help discriminate among these scenarios [8, 29, 44-46]. Deconvolution of cell type fractions based on reference expression datasets showed that systemic lupus SLE patients have a specific increase in activated natural killer and T helper lymphocytes[8]. In contrast, deconvolution of cell-specific gene expression from whole blood samples in combination with cell-type frequency revealed that kidney transplant recipients experiencing rejection had hundreds of differentially expressed genes specifically in monocytes[29].

Discovery of new cell types and states.

Single cell RNA sequencing (RNA-seq) technologies[47-50]are opening up new possibilities for the analysis of complex heterogeneous samples. Whole tissues can be dissociated

into single cells that can be separately profiled[48, 51]. Whole genome single cell profiles are inherently stochastic which makes the analysis of biological variance more challenging[52], nonetheless they dramatically improve the ability to discover and characterize cell types and states. Cell types and states are classically defined by measuring a combination of a few selected markers using flow cytometry. Even in the most advanced configuration (flow cytometry coupled with mass spectrometry)[53], this approach allows about 40 markers to be measured in parallel for each cell, thus introducing selection bias. In contrast, RNA-seq allows the expression of all genes to be profiled across hundreds or thousands of cells, so providing an unbiased *ab initio* characterization of cell states and types including rare ones[48, 51]. Jaitin et al., for example, could decompose by hierarchical clustering the heterogeneous dendritic cell group into four functionally distinct subclasses and showed how the relative abundance of these cell types is remodelled after infection[48]. In another recent study, Zeisel et al. used single cell RNA-seq to identify 47 subclasses of cells in the mouse cortex and hippocampus[51].

Development and differentiation

Gene expression deconvolution is also useful for interpreting dynamical biological processes from simple responses to stimuli[54] to more complex dynamics such as development and differentiation in multicellular organisms. Cell state trajectories can be reconstructed and data points can be ordered along these with little or no *a priori* chronological information both from average[5, 30, 31, 36] and single cell expression data[35, 42, 54, 55]. Inferring the precise physiological time point of each sample from gene expression can also be important in experiments where the exact chronological time point at which each sample was collected was controlled and recorded, for example because of heterogeneity in the rates of development between genotypes, individual cells or experimental batches[5].

For example, Shalek et al showed by using PCA that the response of single dendritic cells to a pathogenic stimulus is variable in time and includes some precocious cells at early time points that are more advanced in the dynamic response[54] and more similar to cells at later time points. Studying early blood development in mouse embryos using PCA and diffusion maps Moignard et al. revealed heterogeneity along the differentiation

dynamics not only within embryos collected at the same chronological time but also between single cells within individual embryos[42]. These two examples highlight two important advantages of using single cell data and data from single individuals to analyse dynamical progressions: first, one can quantify how synchronous a process is; and second, while in average data the differences in physiological time within each time point are averaged out and decrease the signal, analysis of single cells or individuals ordered along a time-series allows analysis of the full dynamic response[42] even when the system is asynchronous. Another example is provided by the analysis of blood development[55], which provided more power to discover cascades of causal regulators of differentiation[35, 55].

Reconstructing 3D spatial gene expression

In multicellular organisms gene expression not only varies in time but also in space. Methods that retain full spatial information of genome-wide gene expression exist but they are still limited in throughput and are laborious[56]. Junker at al proposed a method similar to tomography where a sample is cryo-sectioned in different directions, each section is analysed by RNA-seq and spatial expression is mathematically reconstructed. Applying this method, they constructed an atlas of 3D expression patterns for zebrafish embryos. Although impressive, ambiguities remain in the atlas when genes are expressed in more than one contiguous region because the system is underdetermined[57]. To overcome this fundamental limitation, RNA-seq would have to be performed on slices at different angles across the sample, which would likely require averaging across different embryos. An alternative approach for reconstructing 3D expression patterns is to use the known spatial distributions of landmark genes, for example mapped by *in situ* hybridisation. This idea has been applied to reconstruct spatial gene expression in zebrafish embryos[58] and in the brains of annelid worms[59]. In these studies, samples were first dissociated into single cells that were RNA-seq profiled and spatial gene expression was computationally reconstructed by measuring the similarity of the expression of each gene to the marker genes with know spatial expression patterns. A similar approach was also used for the spatial reconstruction of a much smaller number of gene expression profiles in the mouse otocyst, the precursor of the inner ear[60].

Interpreting the effects of genetic variation

A central goal of many fields of biology such as human genetics and plant or animal breeding is to understand how natural genetic variation amongst individuals alters their characteristics. Here too the decomposition of expression profiles can be useful, either to remove non-genetic variation and improve the power when asking how genetic variation influences gene expression or to infer additional phenotypic traits or hidden environmental perturbations from gene expression components.

Expression quantitative trail loci (eQTLs) are genetic variants that alter gene expression. eQTLs are identified by performing genome-wide expression profiling on genetically heterogeneous populations. Gene expression is influenced by many non-genetic factors that can obscure subtle genetic effects. These non-genetic factors can be known covariates such as sex or age but they are often hidden uncontrolled experimental variables. eQTL studies are particularly sensitive to these because they usually include many experimental batches. Controlling for both known and hidden confounders greatly increase the power to detect significant eQTLs[7]. The simplest and most widely used approach to improve *cis*-eQTL detection is to remove the first few principal components, because this only removes broad variance components preserving local genetic effects. This approach however is not well suited for improving the detection of genetic loci that cause large-scale gene expression changes in *trans* because their signal might be might be removed together with confounders.

Beyond increasing statistical power, the analysis of hidden confounders can be useful for discovering genotype-environment interactions, i.e. genetic variants that change gene expression differently under different conditions, for example cell type-[61], tissue-[62] or environmental-specific eQTLs[63]. If a hidden confounder reflects a biological source of variance, it should be treated as a covariate instead of being corrected for when testing the genetic effects on gene expression. Parts et al., for example, used a sparse factor analysis model to infer different cellular states (defined by the activity of different molecular pathways) and showed that the effect of some genetic variants is highly dependent on the cellular state[6].

As another example, Curtis et al. used an integrative clustering approach to discover new molecular cancer subtypes from gene expression heterogeneity and to characterize the impact of genetic variation on these cancer subtypes. They further showed that integrative clustering combining gene expression and genomic information is predictive of survival[64].

If a dynamic biological process underlies the hidden source of variance in an eQTL dataset, then the data can be used to investigate how time-dependent processes such as development are influenced by genetic variation. For example, we recently investigated how natural genetic variation affects gene expression in both space and time during 12 hours of the development of *C. elegans* by identifying and exploiting small differences in the exact physiological stages at which each sample was expression profiled [65]. The physiological stage of each sample was inferred by comparing the expression to a reference gene expression time series using CCA (Fig6A), and the tissue-specificity of expression trends were inferred by comparison to expression profiles of sorted tissue samples. This allowed us to examine how sequence variation in the genome alters how genes are expressed in time (Fig6B), and also to ask whether these effects are tissue-specific or not. In this way we were able to identify hundreds of examples where genetic variation close to a gene increased the amplitude of oscillations, altered the rate of induction, or completely altered the dynamics[5].

Summary

We have highlighted in this review how both bulk and single cell gene expression data can be decomposed into the constituent cell types and states and used to reconstruct spatial and temporal patterns of expression. This allows biological processes to be studied at multiple levels using a single expression dataset (that may actually have been generated for a different purpose).

One important lesson to be learned is that 'hidden' confounders in expression data can be more than artefacts to correct for. Rather, they can identify important biological sources of variance that can be interpreted and used to make interesting discoveries.

Another important lesson is that, although it is tempting to apply a simplifying discrete view of a process, this often results in a loss of information and power because of the intrinsically continuous nature of many biological processes in both time and space.

As the cost of single cell RNA sequencing falls and the methods for sample preparation become more routine, these kinds of analyses will become increasingly important and widely used. In particular, we envisage that the application of single cell RNA sequencing to complex samples such as human tissue samples will facilitate the analysis of how genetic variation influences many different dynamic biological processes such as disease progression, development and tissue composition.

Acknowledgments

Research in our laboratory is supported by a European Research Council Consolidator grant IR-DC (616434), by the Spanish Ministry of Economy and Competitiveness (BFU2011-26206 and 'Centro de Excelencia Severo Ochoa 2013-2017' SEV-2012-0208), the AXA Research Fund, FP7 project 4DCellFate (277899), Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), and the EMBL-CRG Systems Biology Program.

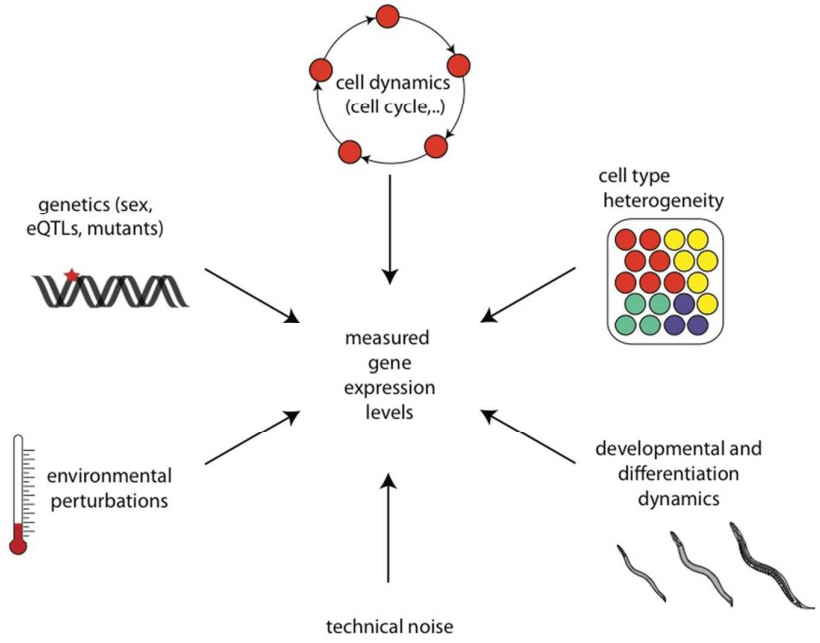


Fig 1 Sources of variance in gene expression data. Genome wide gene expression profiles can capture diverse intentional and unintentional influences.

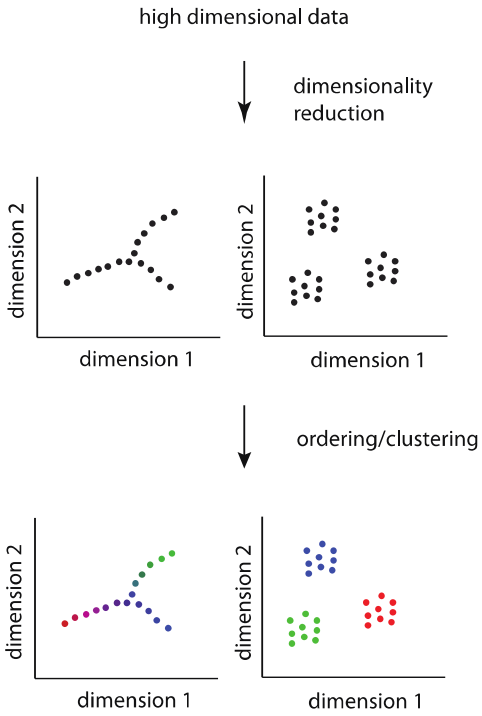


Fig 2. How to interpret high dimensional gene expression data. Projecting data in a low dimensional space (dimensionality reduction) is useful to filter out noise and helps to visualize the global structure of the data, for example highlighting clusters or trajectories. In the ideal case each dimension represents a distinct and interpretable biological process or function. Then clustering or methods to sort data points in trajectories can be applied.

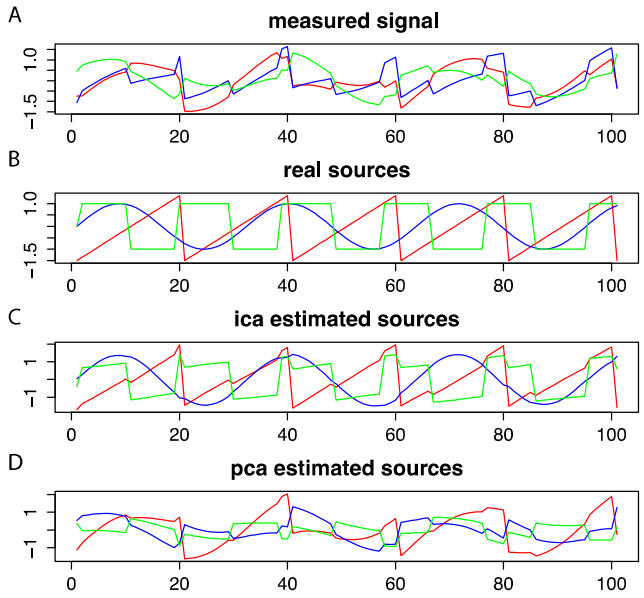


Fig 3. Finding the sources of variance in gene expression. Gene expression measures often include many mixed sources of signal. (A) Three measured signals constituted by a linear combination of (B) three original sources of signal along a time (or space) dimension. (C) ICA better estimates the original sources of signal than (D) PCA.

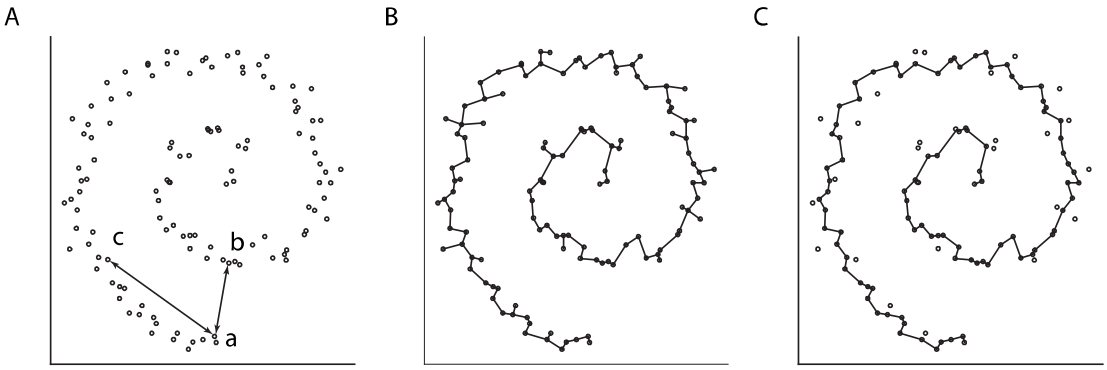


Fig 4. Identification of a non linear trajectory and the inference of the correct order of data points along this trajectory. A Synthetic dataset (jelly-roll) of points arranged in a spiral. Points a and b are closer than a and c in euclidean space despite being at opposite end of the spiral. B. The minimum spanning tree of the data – here

data points are connected according to their shortest path distance. C Data points are correctly sorted along a trajectory determined by the diameter of the minimum spanning tree (adapted from Magwene et al.[32]).

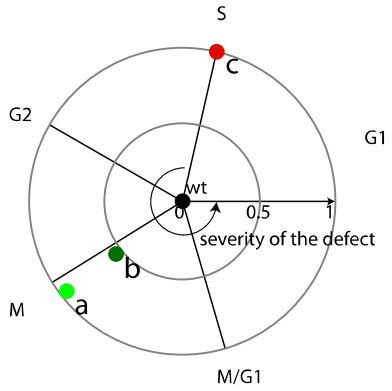


Fig 5. Interpreting the effect of genetic and environmental perturbations on cell cycle dynamics. The angular position indicates the phase of the cell cycle affected, while the distance from the centre (occupied by wild type) indicates the severity of the defect. Both mutants a and b affect mitosis but mutant a has a more severe effect. Mutant c affects S phase (adapted from Lu et al.[28]).

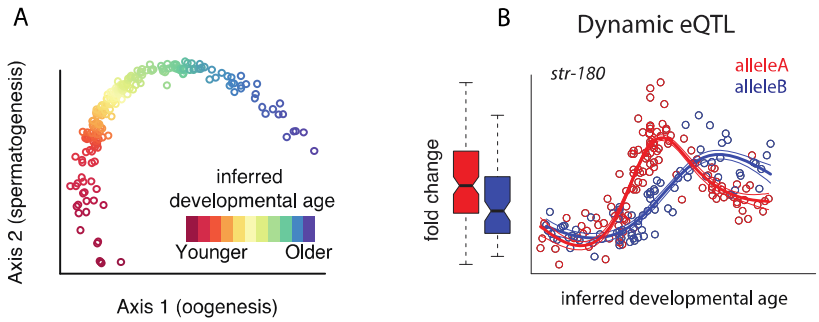


Fig 6. The impact of genetic variation on developmental dynamics. Worms switch from spermatogenesis to oogenesis when maturing. (A) the developmental age of each sample is inferred from the trajectory on the components related to spermatogenesis and oogenesis. (B) A dynamic eQTL analysis shows the complex effect of local genetic variation on the expression dynamics of the *str-180* gene (right) that could not be appreciated when developmental time is not included as a covariate in the analysis

(left). Analysis from Francesconi and Lehner [5] of data from Rockman et al.[65].

References

1. P. Brennecke, S. Anders, J. K. Kim, A. A. Kolodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni and M. G. Heisler, *Nat Methods*, 2013, **10**, 1093-1095.
2. N. Fusi, O. Stegle and N. D. Lawrence, *PLoS Comput Biol*, 2012, **8**, e1002330.
3. F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni and O. Stegle, *Nat Biotechnol*, 2015, **33**, 155-160.
4. N. Fusi, C. Lippert, K. Borgwardt, N. D. Lawrence and O. Stegle, *Bioinformatics*, 2013, **29**, 1382-1389.
5. M. Francesconi and B. Lehner, *Nature*, 2014, **505**, 208-211.
6. L. Parts, O. Stegle, J. Winn and R. Durbin, *PLoS Genet*, 2011, **7**, e1001276.
7. O. Stegle, L. Parts, R. Durbin and J. Winn, *PLoS Comput Biol*, 2010, **6**, e1000770.
8. A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan and H. F. Clark, *PLoS One*, 2009, **4**, e6098.
9. M. J. Brauer, C. Huttenhower, E. M. Airoldi, R. Rosenstein, J. C. Matese, D. Gresham, V. M. Boer, O. G. Troyanskaya and D. Botstein, *Mol Biol Cell*, 2008, **19**, 352-367.
10. E. O'Duibhir, P. Lijnzaad, J. J. Benschop, T. L. Lenstra, D. van Leenen, M. J. Groot Koerkamp, T. Margaritis, M. O. Brok, P. Kemmeren and F. C. Holstege, *Mol Syst Biol*, 2014, **10**, 732.
11. L. B. Snoek, M. G. Sterken, R. J. Volkers, M. Klatter, K. J. Bosman, R. P. Bevers, J. A. Riksen, G. Smant, A. R. Cossins and J. E. Kammenga, *Scientific reports*, 2014, **4**, 3912.
12. T. Hashimshony, M. Feder, M. Levin, B. K. Hall and I. Yanai, *Nature*, 2015, **519**, 219-222.
13. I. T. Jolliffe, *Principal Component Analysis*, Springer, Second edition edn., 2002.
14. O. Alter, P. O. Brown and D. Botstein, *Proc Natl Acad Sci U S A*, 2000, **97**, 10101-10106.
15. N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar and N. V. Fedoroff, *Proc Natl Acad Sci U S A*, 2000, **97**, 8409-8414.
16. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub and E. S. Lander, *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**, 15545-15550.
17. J. J. Goeman, S. A. Van De Geer, F. De Kort and H. C. Van Houwelingen, *Bioinformatics*, 2004, **20**, 93-99.
18. N. C. Chung and J. D. Storey, *Bioinformatics*, 2015, **31**, 545-554.
19. H. F. Kaiser, *Psychometrika*, 1958, **23**, 187-200.
20. A. E. Hendrickson and P. O. White, *British Journal of Statistical Psychology*, 1964, **17**, 65-70.
21. J. T. Leek and J. D. Storey, *PLoS Genet*, 2007, **3**, 1724-1735.
22. O. Stegle, L. Parts, M. Piipari, J. Winn and R. Durbin, *Nat Protoc*, 2012, **7**, 500-507.

23. J. A. Gagnon-Bartsch and T. P. Speed, *Biostatistics*, 2012, **13**, 539-552.
24. A. Hyvarinen and E. Oja, *Neural networks : the official journal of the International Neural Network Society*, 2000, **13**, 411-430.
25. W. Liebermeister, *Bioinformatics*, 2002, **18**, 51-60.
26. S. I. Lee and S. Batzoglou, *Genome Biol*, 2003, **4**, R76.
27. E. M. Airoldi, C. Huttenhower, D. Gresham, C. Lu, A. A. Caudy, M. J. Dunham, J. R. Broach, D. Botstein and O. G. Troyanskaya, *PLoS Comput Biol*, 2009, **5**, e1000257.
28. P. Lu, A. Nakorchevskiy and E. M. Marcotte, *Proc Natl Acad Sci U S A*, 2003, **100**, 10370-10375.
29. S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, T. Hastie, M. M. Sarwal, M. M. Davis and A. J. Butte, *Nat Methods*, 2010, **7**, 287-289.
30. L. Anavy, M. Levin, S. Khair, N. Nakanishi, S. L. Fernandez-Valverde, B. M. Degnan and I. Yanai, *Development*, 2014, **141**, 1161-1166.
31. S. Huang, Y. P. Guo, G. May and T. Enver, *Dev Biol*, 2007, **305**, 695-713.
32. P. M. Magwene, P. Lizardi and J. Kim, *Bioinformatics*, 2003, **19**, 842-850.
33. P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs, Jr., R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan and S. K. Plevritis, *Nat Biotechnol*, 2011, **29**, 886-891.
34. S. C. Bendall, K. L. Davis, A. D. Amir el, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan and D. Pe'er, *Cell*, 2014, **157**, 714-725.
35. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen and J. L. Rinn, *Nat Biotechnol*, 2014, **32**, 381-386.
36. P. Qiu, A. J. Gentles and S. K. Plevritis, *PLoS Comput Biol*, 2011, **7**, e1001123.
37. J. B. Tenenbaum, V. de Silva and J. C. Langford, *Science*, 2000, **290**, 2319-2323.
38. S. T. Roweis and L. K. Saul, *Science*, 2000, **290**, 2323-2326.
39. M. Belkin and P. Niyogi, *Neural computation*, 2003, **15**, 1373-1396.
40. R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner and S. W. Zucker, *Proc Natl Acad Sci U S A*, 2005, **102**, 7426-7431.
41. L. Van der Maaten and G. Hinton, *Journal of Machine Learning Research*, 2008, **9**, 85.
42. V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, S. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher and B. Gottgens, *Nat Biotechnol*, 2015, **33**, 269-276.
43. A. D. Amir el, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan and D. Pe'er, *Nat Biotechnol*, 2013, **31**, 545-552.
44. H. Lahdesmaki, L. Shmulevich, V. Dunmire, O. Yli-Harja and W. Zhang, *BMC Bioinformatics*, 2005, **6**, 54.
45. J. Clarke, P. Seo and B. Clarke, *Bioinformatics*, 2010, **26**, 1043-1049.
46. J. Ahn, Y. Yuan, G. Parmigiani, M. B. Suraokar, L. Diao, Wistuba, II and W. Wang, *Bioinformatics*, 2013, **29**, 1865-1871.
47. S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lonnerberg and S. Linnarsson, *Nat Methods*, 2014, **11**, 163-166.
48. D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay and I. Amit, *Science*, 2014, **343**, 776-779.
49. S. Picelli, A. K. Bjorklund, O. R. Faridani, S. Sagasser, G. Winberg and R. Sandberg, *Nat Methods*, 2013, **10**, 1096-1098.
50. T. Hashimshony, F. Wagner, N. Sher and I. Yanai, *Cell Rep*, 2012, **2**, 666-673.
51. A. Zeisel, A. B. M. Manchado, S. Codeluppi, P. Lönnerberg, G. L. Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-

Leffler and S. Linnarsson, *Science*, 2015.
52. O. Stegle, S. A. Teichmann and J. C. Marioni, *Nat Rev Genet*, 2015, **16**, 133-145.
53. S. C. Bendall, E. F. Simonds, P. Qiu, E. ad D Amir, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, R. S. Balderas, S. K. Plevritis, K. Sachs, D. Pe'er, S. D. Tanner and G. P. Nolan, *Science*, 2011, **332**, 687-696.
54. A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, S. Schwartz, B. Fowler, S. Weaver, J. Wang, X. Wang, R. Ding, R. Raychowdhury, N. Friedman, N. Hacohen, H. Park, A. P. May and A. Regev, *Nature*, 2014, **510**, 363-369.
55. S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan and D. Pe'er, *Cell*, 2014, **157**, 714-725.
56. J. H. Lee, E. R. Daugharthy, J. Scheiman, R. Kalhor, J. L. Yang, T. C. Ferrante, R. Terry, S. S. Jeanty, C. Li, R. Amamoto, D. T. Peters, B. M. Turczyk, A. H. Marblestone, S. A. Inverso, A. Bernard, P. Mali, X. Rios, J. Aach and G. M. Church, *Science*, 2014, **343**, 1360-1363.
57. J. P. Junker, E. S. Noel, V. Guryev, K. A. Peterson, G. Shah, J. Huisken, A. P. McMahon, E. Berezikov, J. Bakkers and A. van Oudenaarden, *Cell*, 2014, **159**, 662-675.
58. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier and A. Regev, *Nat Biotechnol*, 2015, DOI: 10.1038/nbt.3192.
59. K. Achim, J. B. Pettit, L. R. Saraiva, D. Gavriouchkina, T. Larsson, D. Arendt and J. C. Marioni, *Nat Biotechnol*, 2015, DOI: 10.1038/nbt.3209.
60. R. Durruthy-Durruthy, A. Gottlieb, B. H. Hartman, J. Waldhaus, R. D. Laske, R. Altman and S. Heller, *Cell*, 2014, **157**, 964-978.
61. M. Ackermann, W. Sikora-Wohlfeld and A. Beyer, *PLoS Genet*, 2013, **9**, e1003514.
62. T. Flutre, X. Wen, J. Pritchard and M. Stephens, *PLoS Genet*, 2013, **9**, e1003486.
63. J. Gagneur, O. Stegle, C. Zhu, P. Jakob, M. M. Tekkedil, R. S. Aiyar, A.-K. Schuon, D. Pe'er and L. M. Steinmetz, *PLoS Genet*, 2013, **9**, e1003803.
64. C. Curtis, S. P. Shah, S. F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Graf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerod, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A. L. Borresen-Dale, J. D. Brenton, S. Tavare, C. Caldas and S. Aparicio, *Nature*, 2012, **486**, 346-352.
65. M. V. Rockman, S. S. Skrovanek and L. Kruglyak, *Science*, 2010, **330**, 372-376.