**Spring 2021 – Epigenetics and Systems Biology**
**Discussion Session (Systems Biology)**
**Michael K. Skinner – Biol 476/576**
**Week 3 (February 4, 2021)**

## Systems Biology (Components)

Primary Papers

1. Kuster, et al. (2011) J Physiol 589.5 pp 1037-1045. (PMID: 21224228)
2. Garcia, et al. (2014) Systems Biol 8:34. (PMID: 24655443)
3. Griffiths, et al. (2018) Molecular Systems Biology 14:e8046. (PMID: 29661792)

## Discussion

Student 4 – Ref #1 above

- What was the systems model used to investigate cardiovascular disease?
- How were the components assessed?
- What networks and conclusions were obtained?

Student 5 – Ref # 2 above

- What combination of omics technology was used?
- What insight into respiratory disease was obtained?
- What do the networks indicate?

Student 6 – Ref #3 above

- What recent omics technology was used?
- How can single cell technology provide new omics insights?
- What are some of the advantages and disadvantages of single cell genomics?

TOPICAL REVIEW

# 'Integrative Physiology 2.0': integration of systems biology into physiology and its application to cardiovascular homeostasis

Diederik W. D. Kuster[1,2], Daphne Merkus[1], Jolanda van der Velden[3], Adrie J. M. Verhoeven[2] and Dirk J. Duncker[1]

[1] *Experimental Cardiology, Thoraxcenter, and* [2] *Dept. Biochemistry, Erasmus MC, University Medical Center Rotterdam, Rotterdam,* [3] *Laboratory for Physiology, Institute for Cardiovascular Research, VU University Medical Center, Amsterdam, The Netherlands*

**Abstract** Since the completion of the Human Genome Project and the advent of the large scaled unbiased '-omics' techniques, the field of systems biology has emerged. Systems biology aims to move away from the traditional reductionist molecular approach, which focused on understanding the role of single genes or proteins, towards a more holistic approach by studying networks and interactions between individual components of networks. From a conceptual standpoint, systems biology elicits a 'back to the future' experience for any integrative physiologist. However, many of the new techniques and modalities employed by systems biologists yield tremendous potential for integrative physiologists to expand their tool arsenal to (quantitatively) study complex biological processes, such as cardiac remodelling and heart failure, in a truly holistic fashion. We therefore advocate that systems biology should not become/stay a separate discipline with '-omics' as its playing field, but should be integrated into physiology to create 'Integrative Physiology 2.0'.

(Received 4 November 2010; accepted after revision 20 December 2010; first published online 4 January 2011)
**Corresponding author** D. J. Duncker: Experimental Cardiology, Thoraxcenter, Erasmus MC, University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands. Email: d.duncker@erasmusmc.nl

## Introduction

Maintenance of homeostasis is essential for survival of an organism. The cardiovascular system has therefore developed a high degree of plasticity to maintain circulatory homeostasis in a wide variety of circumstances. Defence mechanisms include acute adjustments, e.g. the cardiovascular adaptations to a sudden increase in physical activity, as well as chronic adjustments, e.g. cardiac remodelling to a chronic elevation in haemodynamic loading conditions following myocardial injury, volume or pressure overload. These adjustments require highly integrated and orchestrated responses involving a large number of controlled variables. In view of the importance of adequate circulatory responses for the survival of an organism, these processes are characterized by a high level of redundancy involving complex signalling pathways that display significant interactions at multiple levels. Integrative physiology has been able to decipher many aspects of cardiovascular homeostasis, including the regulation of coronary blood flow (Duncker & Bache, 2008) as well as the short- and long-term regulation of

**Dirk J. Duncker**, **Jolanda van der Velden**, **Diederik Kuster**, **Daphne Merkus** and **Adrie Verhoeven** work in the Cardiovascular Research Institutes of Rotterdam and Amsterdam and collaborate on research into the pathogenesis and therapies of cardiac remodelling and dysfunction following acute myocardial infarction. Their backgrounds are in biochemistry (D.K., J.vdV., A.V.), molecular biology (D.K., A.V.) and physiology (D.M., J.vdV., D.J.D.). Starting from *in vivo* observations in exercising pigs, they employ an integrative approach to unravel the cellular, biochemical and molecular basis of cardiac remodelling and dysfunction.


DOI: 10.1113/jphysiol.2010.201533

blood pressure and cardiac function (Guyton, 1992; Hester *et al*. 2011). In other areas of cardiovascular homeostasis, including cardiac hypertrophy, integrative physiology has provided tremendous insight into this process at the organ and cellular level, but only very limited insight into its molecular basis (Fig. 1). The emergence of the field of molecular biology has enabled cardiovascular researchers to obtain deeper insight into this complex process (Mudd & Kass, 2008).

Initial molecular studies in the cardiovascular field principally consisted of observational work, looking at gene and/or protein expression and changes therein in cardiovascular disease states (e.g. Katz, 1988; Brand *et al.* 1992). These studies were followed by more mechanistic approaches to test the involvement of identified (novel) genes and their products, mainly by virtue of knocking out and/or over-expressing a gene of interest (Frey & Olson, 2003; Heineke & Molkentin, 2006). This reductionist approach has significant value in monogenic diseases. However, the use of genetic models in studies of cardiovascular disease soon illustrated the complexity of cardiovascular diseases, as many gene knock-out animal models lacked a clear phenotype. These findings were initially interpreted to suggest that the gene was not important, while a more physiological interpretation is that other genes increased their activity and acted to compensate. These observations, in conjunction with the completion of the Human Genome Project and the advent of the '-omics' technologies, stimulated the emergence of the field of systems biology. As outlined elsewhere in this issue of *The Journal of Physiology*, systems biology aims to move beyond the traditional reductionist molecular approach (which focused on understanding the role of single genes or proteins), towards a more holistic approach by studying networks and interactions between individual components of networks. The strength of this integrative molecular approach is that, even when a perturbation in a molecular pathway does not result in clear phenotypic changes, the responsible compensatory adaptations will likely be mirrored in adaptations in the transcriptome, proteome and/or metabolome. Until now, systems biology has been mainly considered a research field in its own right. However, to date systems biology has been applied to relatively simple systems, including cultured cells and bacteria, but has not been applied to studies of homeostasis in complex organisms, including mammals, a field that has traditionally been the domain of integrative physiology (Fig. 1). We believe that integration of the complementary disciplines of systems biology and integrative physiology is essential to advance our understanding of complex biological processes.

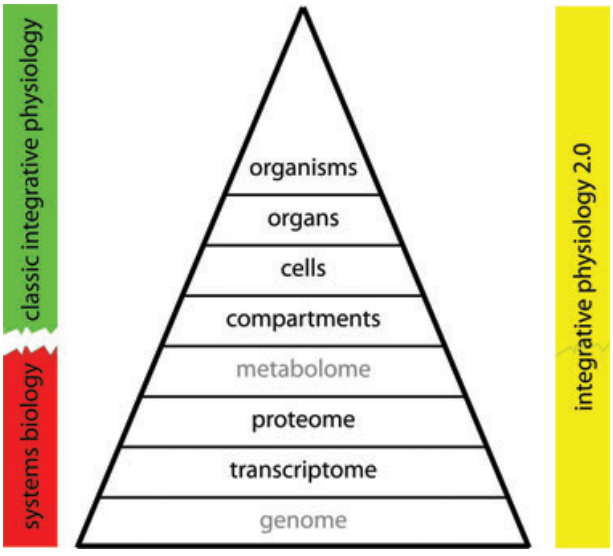


**Figure 1. From systems biology and classical integrative physiology towards Integrative Physiology 2.0**
A process such as cardiac remodelling should be studied at different levels and the findings integrated. The bars on the left illustrate the dichotomy between classic integrative physiology and systems biology. The bar on the right illustrates the 'Integrative physiology 2.0' approach, which integrates the large scale unbiased '-omics' studies of systems biology with integrative physiology. Levels shown with a grey font have not been studied by our group, to date.

In this article we will present studies on the adjustments of the myocardium to acute and chronic increases in loading conditions, in order to highlight the established strengths of classical integrative physiology and the promise of integrating systems biology and physiology. We begin to review our studies using classical *in vivo* physiology approaches to study regulation of cardiac function and coronary blood flow in response to acute exercise. We will then discuss how we have implemented biochemistry, molecular biology, and more recently bioinformatics to study biological processes in a more holistic rather than reductionistic fashion to understand complex processes such as cardiac remodelling and hypertrophy.

## Plasticity of the cardiovascular system: acute responses to exercise

One of the most dramatic challenges for the cardiovascular system is represented by sudden heavy physical exercise, requiring both central and regional haemodynamic adjustments in order to meet increases in metabolic needs of skeletal and cardiac muscle. A fivefold increase in cardiac output together with a redistribution of flow away from visceral organs and tissues is needed to accommodate sufficient increases in skeletal muscle and myocardial blood flow. The increases in muscle blood flow are facilitated by a small increase in aortic blood pressure but are opposed by the compressive forces generated by the contracting muscle, acting on the intramuscular vasculature. Consequently, the increases in flow are principally due to vasodilatation of the resistance vessels within the skeletal and cardiac muscle.

A large number of vascular control mechanisms have been identified that can contribute to metabolic regulation of resistance vessel tone in the heart and skeletal muscle (Fig. 2), including blood-derived, endothelial, metabolic and sympathetic influences. However, unravelling of the exact mechanism that mediates the exercise-induced vasodilatation has proven to be difficult (Laughlin *et al.* 1996; Rowell, 2004; Tune *et al.* 2004; Duncker & Merkus, 2007; Duncker & Bache, 2008). Since maintenance of tissue perfusion is essential for adequate cardiac and skeletal muscle function and organismal survival, it is not surprising that regulation of tissue blood flow is characterized by a high number of redundant control mechanisms (Rowell, 2004; Duncker & Bache, 2008). A consequence of this non-linear redundancy design is that pharmacological blockade of a single vasodilator mechanism may have little or no effect (and may thus not reveal the actual contribution of that mechanism), as other vasodilator pathways will increase their activity and act to compensate. Only when multiple pathways are blocked will an effect become apparent, which is then greater than the sum of the effects of blocking the individual pathways. Indeed, studies in cardiac and skeletal muscle have demonstrated that simultaneous blockade of various vasodilator substances was required to attenuate the increase in skeletal muscle flow (Murrant & Sarelius, 2002; Boushel, 2003) or coronary blood flow (Duncker & Bache, 2008) during exercise. These observations demonstrate the importance of an integrative approach looking at the whole system and the interaction between the individual components.

## Plasticity of the cardiovascular system: cardiac remodelling after myocardial infarction

The cardiovascular system is not only able to respond quickly to acute challenges, but also has the plasticity to respond to chronic changes in haemodynamic loading conditions, for example as occurs following an acute myocardial infarction (MI). Loss of a significant portion of myocardial tissue results in an immediate decrease in cardiac pump function, leading to neurohumoral activation that is aimed at restoring pump function.

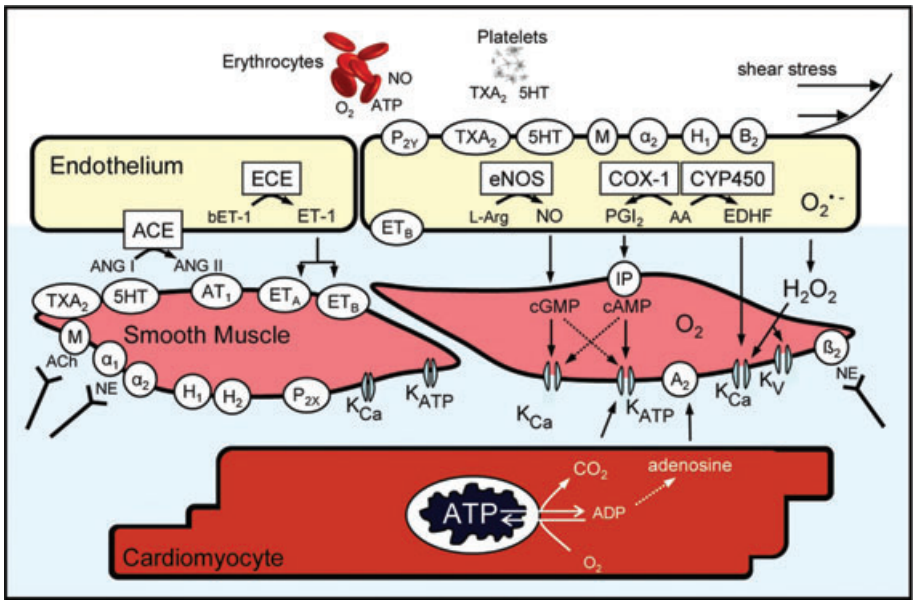


**Figure 2. Schematic drawing of the various influences that determine coronary vasomotor tone and diameter**

Influences include autonomic nervous system activity, metabolic factors from cardiomyocytes and endothelial factors. The latter are modified by physical forces (shear stress), as well as erythrocyte and platelet-derived products acting on the endothelium. $TxA_2$, thromboxane $A_2$ (receptor); 5HT, serotonin or 5-hydroxytryptamine (receptor); $P_{2X}$ and $P_{2Y}$, purinergic receptor subtypes 2X and 2Y that mediate ATP-induced vasoconstriction and vasodilatation, respectively; ACh, acetylcholine; M, muscarinic receptor; $H_1$ and $H_2$, histamine receptors type 1 and 2; $B_2$, bradykinin receptor subtype 2; ANG I and ANG II, angiotensin I and II; $AT_1$, angiotensin II receptor subtype 1; ET, endothelin; $ET_A$ and $ET_B$, endothelin receptor subtypes A and B; $A_2$, adenosine receptor subtype 2; $\beta_2$, $\beta_2$-adrenergic receptor; $\alpha_1$ and $\alpha_2$, $\alpha$-adrenergic receptors; NO, nitric oxide; eNOS, endothelial NO synthase; $PGI_2$, prostacyclin; IP, prostacyclin receptor; COX-1, cyclooxygenase-1; EDHF, endothelium-derived hyperpolarizing factor; $CYP_{450}$, cytochrome $P_{450}$ 2C9; $K_{Ca}$, calcium-sensitive $K^+$ channel; $K_{ATP}$, ATP-sensitive $K^+$ channel; $K_V$, voltage-sensitive $K^+$ channel; AA, arachidonic acid; L-Arg, L-arginine; $O_2^-$, superoxide. Receptors and enzymes are indicated by an oval and rectangle, respectively. From Duncker & Bache (2008), modified with permission from the American Physiological Society.

The neurohumoral activation results in a wide array of responses varying from the immediate (seconds–minutes) positive chronotropic, inotropic and lusitropic cardiac effects and sub-acute (hours–days) volume retention, to the chronic (days–months) cardiac remodelling, characterized by hypertrophy of the cardiac muscle (Katz, 2003). All these responses aim to maintain pump function of the injured heart. However, despite the apparent appropriateness of the hypertrophic remodelling response to maintain cardiac pump function early after MI (van Kats *et al.* 2000), hypertrophic remodelling constitutes an independent risk factor for the long-term development of congestive heart failure (Levy *et al.* 1990; Vakili *et al.* 2001). The mechanism underlying progressive deterioration of left ventricular (LV) function towards overt heart failure remains incompletely understood, but may involve (i) continuous loss of cardiomyocytes through apoptosis (Narula *et al.* 2006), (ii) a primary reduction in contractile function of the surviving myocardium (van der Velden *et al.* 2004), (iii) alterations in extracellular matrix leading to progressive LV dilatation (Spinale, 2007), and/or (iv) myocardial blood flow abnormalities, resulting in impaired myocardial $O_2$ delivery to the non-infarcted region (van Veldhuisen *et al.* 1998). Blood flow to the remodelled myocardium can become impeded as the coronary vasculature does not grow commensurate with the increase in LV mass and because extravascular compression of the coronary vasculature increases with increased LV filling pressures (Haitsma *et al.* 2001). In addition, an increase in coronary resistance vessel tone, secondary to neurohumoral activation and endothelial dysfunction, could also contribute to the impaired myocardial oxygen supply.

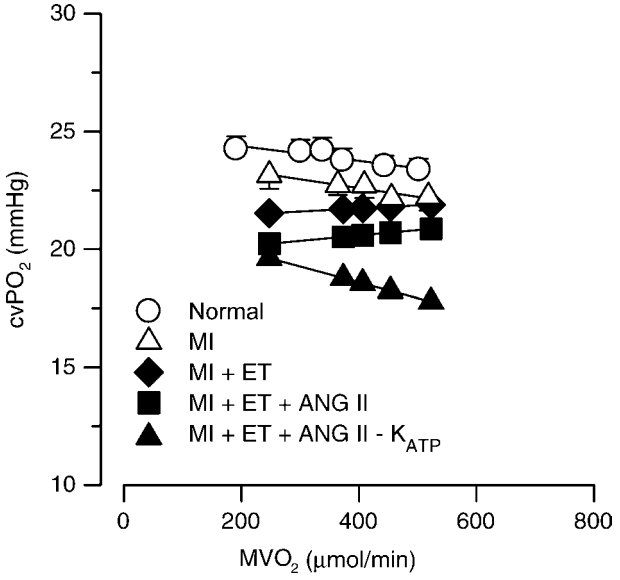


**Figure 3. Myocardial oxygen balance in normal and MI swine**
Shown are the relations between myocardial oxygen consumption ($M\dot{V}_{O_2}$) and coronary venous oxygen tension ($cvP_{O_2}$) in 30 normal swine (open circles) and 20 MI swine (open triangles) under control conditions. Data were obtained at rest and during increases in $M\dot{V}_{O_2}$ produced by graded treadmill exercise (1–5 km $h^{-1}$ in normal swine and 1–4 km $h^{-1}$ in MI swine). In addition, we have depicted the computed relations in MI swine if the ET (filled diamonds) and ANG II (filled squares) vasoconstrictor influences (which were both attenuated in MI swine) and the $K_{ATP}$ (filled triangles) vasodilator influences (which were enhanced in MI swine) would have been identical to those in normal swine. The graph clearly illustrates that the adaptations in coronary vasomotor control act to blunt perturbations in oxygen balance in remodelled myocardium of swine with a recent MI. Modified from Duncker *et al*. (2008) with permission from Springer Science+Business Media.

Consequently, we explored in a series of studies the alterations in regulation of coronary resistance vessel tone in post-MI remodelled myocardium. For this purpose we employed a porcine model of MI produced by permanent ligation of the left circumflex coronary artery, which results in transmural infarction of 20–25% of the LV free wall, and studied swine at 2–3 weeks after induction of MI. Swine were not only studied at rest but also during graded treadmill exercise to further stress the remodelled hearts and recruit the cardiac and coronary functional reserve capacity, to facilitate elucidation of compensatory mechanisms that become activated to maintain cardiovascular homeostasis. These studies indicate that myocardial oxygen balance is mildly perturbed in remodelled myocardium. Thus at a similar level of cardiac work and hence oxygen consumption, coronary blood flow and hence myocardial oxygen supply are lower in MI compared to normal swine, forcing the myocardium to increase its oxygen extraction leading to a lower coronary venous oxygen content (Fig. 3). That the relatively small degree of perturbation in the oxygen balance was associated with myocardial metabolic distress was also reflected in the increased vasodilator influence through opening of $K_{ATP}$ channels, particularly during exercise (Merkus *et al.* 2005b). Unexpectedly, we observed that despite increased circulating levels of noradrenalin, angiotensin II and endothelin-1, the coronary influences of $\alpha$-adrenergic tone were not increased (Duncker *et al.* 2005), while the coronary vasoconstrictor influences of endogenous endothelin (Merkus *et al.* 2005*a*) and angiotensin II (Merkus *et al.* 2006) were virtually abolished. Thus, early after myocardial infarction, small perturbations in myocardial oxygen balance were observed in remodelled myocardium. However, adaptations in coronary resistance vessel control, consisting of increased vasodilator influences in conjunction with blunted vasoconstrictor influences, acted to minimize the impairments of myocardial oxygen balance (Fig. 3). These studies not only highlight the plasticity of the post-MI remodelled heart and coronary circulation, to minimize perturbations in myocardial oxygenation in the face of increased compressive forces and reduced capillary densities, but also illustrate the necessity

to study these phenomena in an integrative manner in an intact animal model.

Neurohumoral activation following MI initially contributes to circulatory homeostasis, but will eventually contribute to the progressive deterioration in LV function. This concept is supported by studies showing detrimental effects of amplification of neurohumoral activity by phosphodiesterase-3 (PDE3) inhibitors in patients with heart failure (Packer *et al.* 1991), while on the other hand $\beta$-adrenergic receptor blockade (CIBIS Investigators and Committees, 1994; CIBIS-II Investigators and Committees, 1999; MERIT-HF Study Group, 1999) and inhibitors of the RAAS system (Pfeffer *et al.* 1992) have clearly shown long-term benefits in large cohorts of patients with heart failure. Starting from these observations in patients with heart failure, we took an integrative approach to study the cellular and molecular mechanisms underlying LV dysfunction observed in our swine model ~3 weeks after acute MI. In a first series of studies, we demonstrated the presence of LV remodelling (van Kats *et al.* 2000) and dysfunction (Duncker *et al.* 2001; Haitsma *et al.* 2001), necessitating an increased oxygen extraction by the peripheral tissues (Fig. 4*A*) and causing an increase in neurohumoral activation (Fig. 4*B*) (Haitsma *et al.* 2001). Despite the increased neurohumoral activation, $\beta$-adrenergic inotropic (Fig. 4*C*) and lusitropic (Fig. 4*D*) influences on the left ventricle were markedly blunted, particularly during treadmill exercise (van der Velden *et al.* 2004; Duncker *et al.* 2005). A loss of $\beta$-adrenergic signalling was also suggested by an attenuated response to PDE3 inhibition (Duncker *et al.* 2001). To further investigate the cellular mechanisms underlying the global LV dysfunction, we performed studies in isolated permeabilized individual cardiomyocytes (van der Velden *et al.* 2004). In myocytes from the remote LV zone in MI hearts, we observed abnormalities in myofilament force development, which correlated well with the degree of LV remodelling, and an increase in myofilament $Ca^{2+}$ sensitivity (Fig. 5*A*) (van der Velden *et al.* 2004). These alterations in myofilament function are likely to contribute to the systolic (Fig. 4*C*) and diastolic (Fig. 4*D*) LV dysfunction observed in swine during $\beta$-adrenergic receptor activation produced by treadmill exercise. The abnormalities in myofilament function could be prevented, at least in part, by treatment with chronic $\beta_1$-adrenergic receptor blockade during the post-MI period (Duncker *et al.* 2009). Analysis of myofilament proteins with one- and two-dimensional gel-electrophoresis failed to demonstrate significant alterations in phosphorylation status under basal conditions, including to our surprise the $\beta$-adrenergic target proteins cardiac myosin binding protein C (Fig. 5*C*) and troponin I (Fig. 5*D*) (Duncker *et al.* 2009). When the heart was stimulated with the $\beta$-adrenergic receptor agonist dobutamine, the increase in troponin I phosphorylation was blunted in remodelled myocardium (Fig. 5*D*) (Boontje *et al.* 2010). The increased $Ca^{2+}$ sensitivity of force development of post-MI myocytes could be restored to normal (sham) values by incubation with the catalytic subunit of protein kinase A (PKA), the downstream kinase of the $\beta_1$-adrenergic receptor (Fig. 5*B*). Taken together, these observations suggest that PKA specific phosphorylation sites may be selectively altered in post-MI hearts, which are the subject of ongoing studies within our laboratory.

To complement the top-down approach (from organism towards proteome) outlined above and to further investigate the mechanisms underlying the LV dysfunction following MI, we recently set out to investigate transcriptional control of LV remodelling and dysfunction. For this purpose, we performed microarray analysis to find genes that are differentially expressed in post-MI *versus* control hearts (Kuster *et al.* 2010). Relations between the differentially expressed genes were assessed by Ingenuity Pathway Analysis. This program

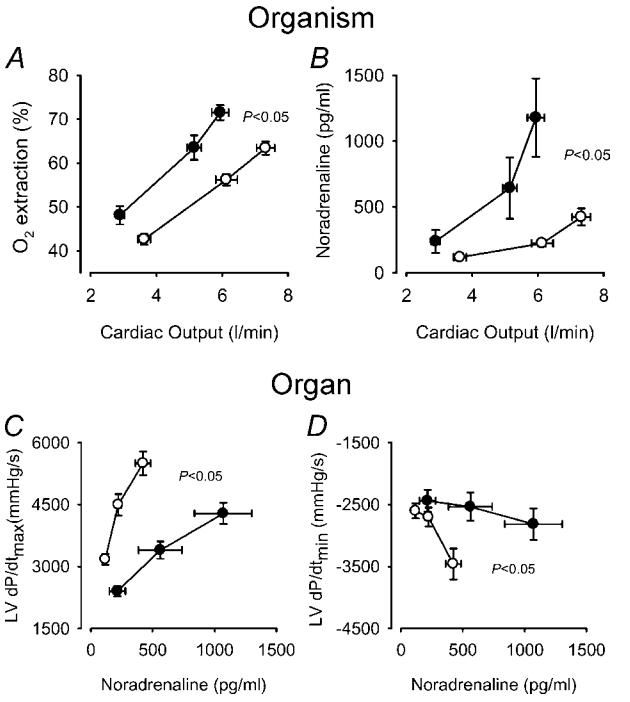


**Figure 4. Functional changes at the whole-body and cardiac level**
Whole-body oxygen extraction (*A*) and circulating noradrenaline levels (*B*) in resting and exercising swine with cardiac dysfunction 3 weeks after MI (filled circles) or sham surgery (open circles), and maximum rates of rise (*C*) and fall (*D*) of left ventricular pressure were plotted as a function of circulating noradrenaline levels. In each group, data are shown during resting conditions and during treadmill running at 2 and 4 km $h^{-1}$. The data show relatively little functional deficit in resting conditions, but functional deficits at higher endogenous sympathetic activation increasing with exercise intensity. Based on Haitsma *et al.* (2001) with permission from the European Society of Cardiology and van der Velden *et al.* (2004).

builds networks of interacting molecules by connecting as many differentially expressed genes as possible, and allowing for hub molecules of which the expression remains unchanged. Taken a non-supervised approach (Fig. 6*A*), an important network was identified that contained several genes encoding proteins involved in $\beta$-adrenergic signalling, including the regulatory subunit of PKA (PRKAR2B), A-kinase anchoring protein 5 (AKAP5), calmodulin and calmodulin kinase (CaMK), of which the expression was altered. In addition, subsequent analysis of the $\beta$-adrenergic signalling network revealed increased expression of PDE4 (Fig. 6*B*). If confirmed at the protein level, the increased expression could contribute to the observed blunted PKA influence on myofilament $Ca^{2+}$ sensitivity via (i) reduced cAMP production through increased CaMK-mediated inhibition of adenylyl-cyclase and increased cAMP breakdown by PDE4, and (ii) inactivation of the catalytic subunit of PKA by increased binding to the regulatory subunit of PKA.

## Integrative Physiology 2.0

Systems biology approaches have not yet been applied to the study of cardiac remodelling, largely because of its tremendous complexity. Starting from observations in

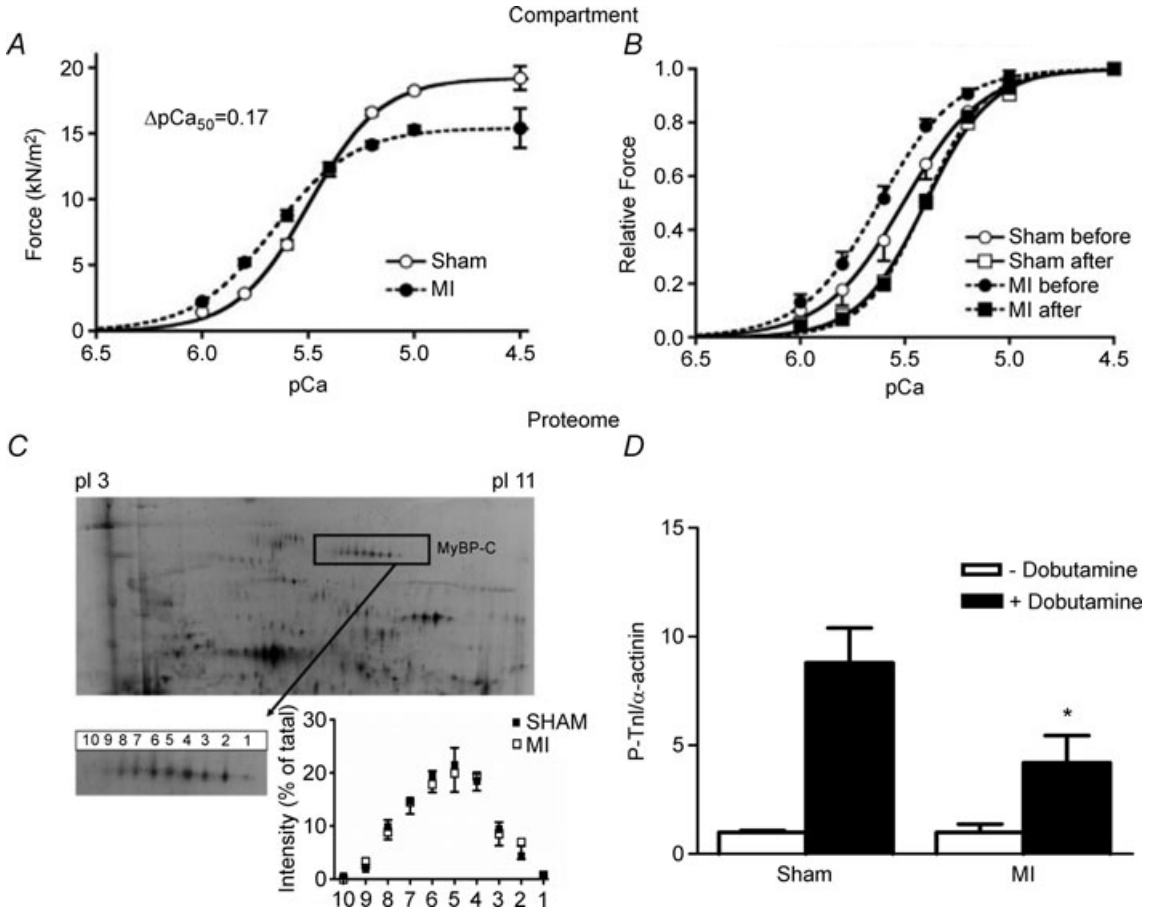


**Figure 5. Myofilament function and protein phosphorylation**
*A*, determination of force development by skinned cardiomyocytes isolated from sham and post-MI pig hearts at different exogenous $Ca^{2+}$ concentrations showed reduced maximal force and increased $Ca^{2+}$ sensitivity in post-MI remodelled myocardium. *B*, $Ca^{2+}$ sensitivity in the MI hearts was normalized to control (sham) values by pre-incubation of skinned cardiomyocytes with exogenous protein kinase A (PKA). Force development was measured before and after incubation with PKA. Force at maximal [$Ca^{2+}$] was set to 1. The observation that PKA abolished the difference in $Ca^{2+}$ sensitivity between sham and post-MI cardiomyocytes suggests that the increase in myofilament $Ca^{2+}$ sensitivity is caused by lower levels of PKA-mediated phosphorylation of sarcomeric proteins. *C*, two-dimensional gel electrophoresis showed no difference in the phosphorylation pattern of the PKA target protein cardiac myosin binding protein (cMyBP-C) between sham and MI hearts. *D*, troponin I (TnI) phosphorylation did not differ under baseline conditions between sham and MI heart. However, following intravenous infusion of dobutamine the increase in TnI phosphorylation was attenuated in post-MI myocardium. Panels *A* and *B* were adapted from van der Velden *et al.* (2004), *C* was adapted from Duncker *et al.* (2009) and *D* shows data from Boontje *et al.* (2010).

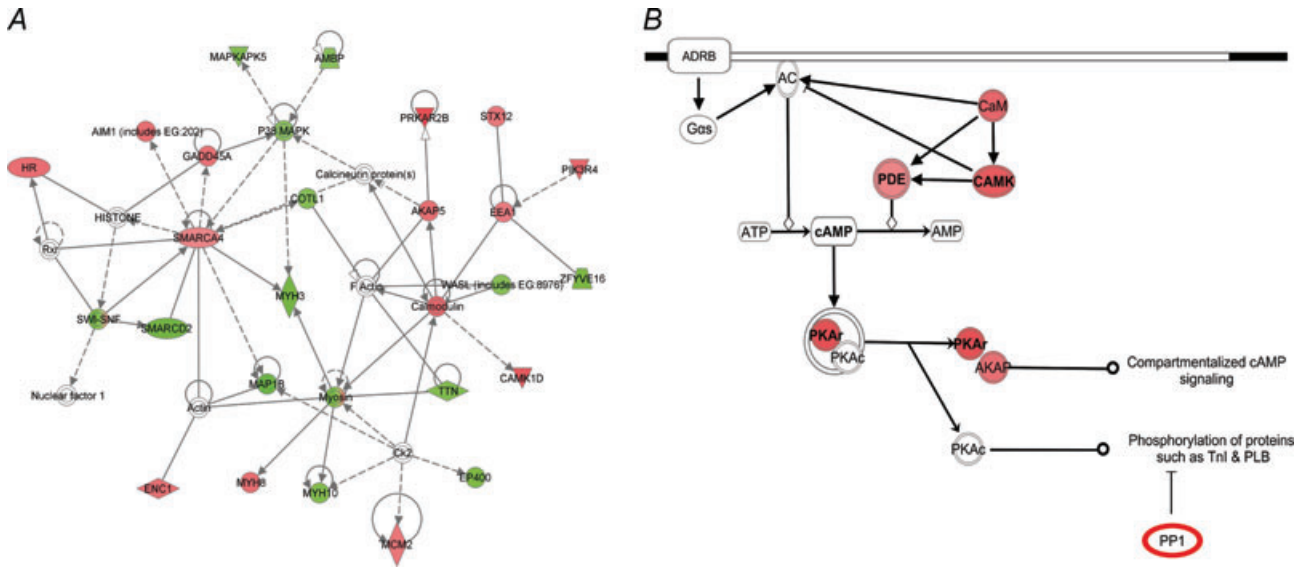


**Figure 6. Network identification by Ingenuity Pathway Analysis**
*A*, one of the major networks identified by unsupervised analysis of genes differentially expressed in post-MI *vs.* sham myocardium. Genes in red and green are up- and downregulated after MI, respectively. Genes in white, such as calcineurin, are not changed in expression but represent hubs between a large number of differentially expressed genes. The data show that a number of genes of the $\beta$-adrenergic pathway are changed in expression. *B*, simplified $\beta$-adrenergic signalling pathway identified by supervised data analysis, with upregulated genes in red and downregulated genes in green. Colour intensities correspond to the degree of change, with a deeper colour indicating a greater change. PP1 has been depicted with a red outline to indicate that we previously found an increase in PP1 protein level. Data are from Kuster *et al.* (unpublished).

patients showing detrimental effects of PDE3 inhibitors and beneficial effects of $\beta$-blockers, we have taken an integrative approach to studying the mechanisms underlying LV dysfunction after MI (Fig. 7). We began by narrowing our experimental focus to the well-defined clinical phenotype of post-MI LV remodelling and took a top-down approach, starting in the awake pig and ending with specific and generalized molecular investigations

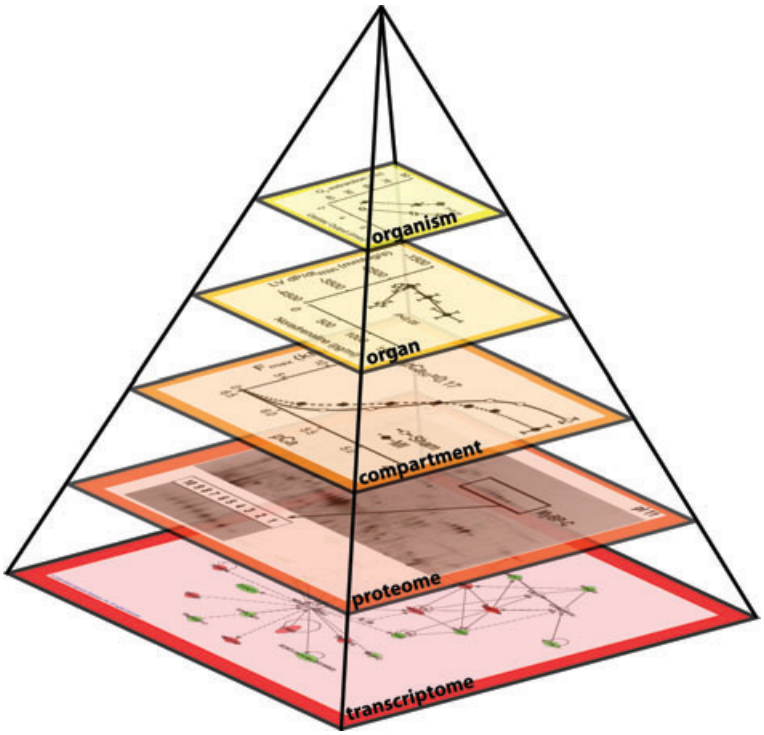


**Figure 7. Illustration of our 'Integrative Physiology 2.0' approach**
Complex physiological processes such as cardiac remodelling must be studied in detail at different levels ranging from the transcriptome of cells all the way up to the intact organism, and possibly even further to population-based functional responses to pharmacons (not shown). At each level, data should be integrated with 'higher' and 'lower' levels, to build a multidimensional picture of the ongoing processes.

centred on transcriptomic and proteomic correlations (Fig. 7) based on current knowledge (Adams, 2010). Using a porcine model of post-MI remodelling, we first demonstrated the presence of LV remodelling and pump dysfunction in swine, necessitating increased oxygen extraction by the peripheral tissues and causing an increase in neurohumoral activation (*organism*). Despite the increased neurohumoral activation, $\beta$-adrenergic receptor mediated increases of LV function (*organ*) were blunted (Duncker *et al.* 2005), which coincided with attenuated LV inotropic responses to PDE3 inhibition (Duncker *et al.* 2001). Further studies at the cardiomyocyte level revealed abnormalities of myofilament force development that correlated well with the degree of LV remodelling (*cellular compartment*) (van der Velden *et al.* 2004). The alterations in myofilament $Ca^{2+}$ sensitivity appeared to be mediated by loss of PKA catalytic activity (*proteome*), and could be prevented by simultaneous treatment with $\beta_1$-adrenergic receptor blockade, coinciding with an improvement in LV pump function (Duncker *et al.* 2009). Non-supervised as well as supervised network analysis of microarray data (*transcriptome*) revealed significant alterations in expression of genes encoding proteins involved in $\beta$-adrenergic receptor signalling (Fig. 7). These preliminary findings will be followed up by further studies into translational and post-translation modifications.

Since the completion of the Human Genome Project and the advent of the large scaled unbiased '-omics' techniques, the field of systems biology has emerged. Systems biology aims to move away from the traditional reductionist molecular approach, which focused on understanding the role of single genes or proteins, towards a more holistic approach by studying networks and interactions between individual components of networks. From a conceptual standpoint, systems biology elicits a 'back to the future' experience for any integrative physiologist, and we feel that systems biology can benefit from the knowledge and existing models of interaction between systems available in physiology. Conversely, many of the new techniques and modalities employed by systems biologists yield tremendous potential for integrative physiologists to expand their tool arsenal to (quantitatively) study complex biological processes, such as cardiac remodelling and heart failure, in a truly holistic fashion. Such an approach may generate new hypotheses, concepts and eventually novel treatments for the process of cardiac remodelling and heart failure, which should subsequently be tested in a physiological setting. We therefore advocate that systems biology should not become/stay a separate discipline with '-omics' as its playing field, but should be integrated into physiology to create 'Integrative Physiology 2.0', allowing interconnection and integration of processes at the various levels of complexity and organization within the pyramid of life.

## References


Adams KF (2010). Systems biology and heart failure: concepts, methods, and potential research applications. *Heart Fail Rev* **15**, 371–398.

Boontje NM, Merkus D, Zaremba R, Versteilen A, de Waard MC, Mearini G, de Beer VJ, Carrier L, Walker LA, Niessen HW, Dobrev D, Stienen GJ, Duncker DJ & van der Velden J (2010). Enhanced myofilament responsiveness upon beta-adrenergic stimulation in post-infarct remodeled myocardium. *J Mol Cell Cardiol* **Dec 13**, Epub ahead of print.

Boushel R (2003). Metabolic control of muscle blood flow during exercise in humans. *Can J Appl Physiol* **28**, 754–773.

Brand T, Sharma HS, Fleischmann KE, Duncker DJ, McFalls EO, Verdouw PD & Schaper W (1992). Proto-oncogene expression in porcine myocardium subjected to ischemia and reperfusion. *Circ Res* **71**, 1351–1360.

CIBIS Investigators and Committees (1994). A randomized trial of beta-blockade in heart failure. The Cardiac Insufficiency Bisoprolol Study (CIBIS). *Circulation* **90**, 1765–1773.

CIBIS-II Investigators and Committees (1999). The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial. *Lancet* **353**, 9–13.

Duncker DJ & Bache RJ (2008). Regulation of coronary blood flow during exercise. *Physiol Rev* **88**, 1009–1086.

Duncker DJ, Boontje NM, Merkus D, Versteilen A, Krysiak J, Mearini G, El Armouche A, de Beer VJ, Lamers JM, Carrier L, Walker LA, Linke WA, Stienen GJ & van der Velden J (2009). Prevention of myofilament dysfunction by $\beta$-blocker therapy in postinfarct remodeling. *Circ Heart Fail* **2**, 233–242.

Duncker DJ, de Beer VJ & Merkus D (2008). Alterations in vasomotor control of coronary resistance vessels in remodelled myocardium of swine with a recent myocardial infarction. *Med Biol Eng Comput* **46**, 485–497.

Duncker DJ, Haitsma DB, Liem DA, Heins N, Stubenitsky R & Verdouw PD (2001). Beneficial effects of the $Ca^{2+}$ sensitizer EMD 57033 in exercising pigs with infarction-induced chronic left ventricular dysfunction. *Br J Pharmacol* **134**, 553–562.

Duncker DJ, Haitsma DB, Liem DA, Verdouw PD & Merkus D (2005). Exercise unmasks autonomic dysfunction in swine with a recent myocardial infarction. *Cardiovasc Res* **65**, 889–896.

Duncker DJ & Merkus D (2007). Exercise hyperaemia in the heart: the search for the dilator mechanism. *J Physiol* **583**, 847–854.

Frey N & Olson EN (2003). Cardiac hypertrophy: the good, the bad, and the ugly. *Annu Rev Physiol* **65**, 45–79.

Guyton AC (1992). Kidneys and fluids in pressure regulation. Small volume but large pressure changes. *Hypertension* **19**, I2–I8.

Haitsma DB, Bac D, Raja N, Boomsma F, Verdouw PD & Duncker DJ (2001). Minimal impairment of myocardial blood flow responses to exercise in the remodeled left ventricle early after myocardial infarction, despite significant hemodynamic and neurohumoral alterations. *Cardiovasc Res* **52**, 417–428.

Heineke J & Molkentin JD (2006). Regulation of cardiac hypertrophy by intracellular signalling pathways. *Nat Rev Mol Cell Biol* **7**, 589–600.

Hester RL, Iliescu R, Summers R and Coleman TG (2011). Systems biology and integrative physiological modelling. *J Physiol* **589**, 1053–1060.

Katz AW (1988). Molecular biology in cardiology, a paradigmatic shift. *J Mol Cell Cardiol* **20**, 355–366.

Katz AW (2003). The "modern" view of heart failure: how did we get here? *Circ Heart Failure* **1**, 63–71.

Kuster DW, Merkus D, Verhoeven AJ & Duncker DJ (2010). Integrative approach to study the molecular basis of post-myocardial infarction remodeling in porcine heart. *FASEB J* **24**, 110.8.

Laughlin MH, McAllister RM, Jasperse JL, Crader SE, Williams DA & Huxley VH (1996). Endothelium-medicated control of the coronary circulation. Exercise training-induced vascular adaptations. *Sports Med* **22**, 228–250.

Levy D, Garrison RJ, Savage DD, Kannel WB & Castelli WP (1990). Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. *N Engl J Med* **322**, 1561–1566.

MERIT-HF Study Group (1999). Effect of metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL Randomised Intervention Trial in Congestive Heart Failure (MERIT-HF). *Lancet* **353**, 2001–2007.

Merkus D, Haitsma DB, Sorop O, Boomsma F, de Beer VJ, Lamers JM, Verdouw PD & Duncker DJ (2006). Coronary vasoconstrictor influence of angiotensin II is reduced in remodeled myocardium after myocardial infarction. *Am J Physiol Heart Circ Physiol* **291**, H2082–H2089.

Merkus D, Houweling B, van den Meiracker AH, Boomsma F & Duncker DJ (2005*a*). Contribution of endothelin to coronary vasomotor tone is abolished after myocardial infarction. *Am J Physiol Heart Circ Physiol* **288**, H871–H880.

Merkus D, Houweling B, van Vliet M & Duncker DJ (2005*b*). Contribution of $K^+_{ATP}$ channels to coronary vasomotor tone regulation is enhanced in exercising swine with a recent myocardial infarction. *Am J Physiol Heart Circ Physiol* **288**, H1306–H1313.

Mudd JO & Kass DA (2008). Tackling heart failure in the twenty-first century. *Nature* **451**, 919–928.

Murrant CL & Sarelius IH (2002). Multiple dilator pathways in skeletal muscle contraction-induced arteriolar dilations. *Am J Physiol Regul Integr Comp Physiol* **282**, R969–R978.

Narula J, Haider N, Arbustini E & Chandrashekhar Y (2006). Mechanisms of disease: apoptosis in heart failure—seeing hope in death. *Nat Clin Pract Cardiovasc Med* **3**, 681–688.

Packer M, Carver JR, Rodeheffer RJ, Ivanhoe RJ, DiBianco R, Zeldis SM, Hendrix GH, Bommer WJ, Elkayam U, Kukin ML, *et al*. (1991). Effect of oral milrinone on mortality in severe chronic heart failure. The PROMISE Study Research Group. *N Engl J Med* **325**, 1468–1475.

Pfeffer MA, Braunwald E, Moye LA, Basta L, Brown EJ Jr, Cuddy TE, Davis BR, Geltman EM, Goldman S, Flaker GC, *et al*. (1992). Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. Results of the survival and ventricular enlargement trial. The SAVE Investigators. *N Engl J Med* **327**, 669–677.

Rowell LB (2004). Ideas about control of skeletal and cardiac muscle blood flow (1876–2003): cycles of revision and new vision. *J Appl Physiol* **97**, 384–392.

Spinale FG (2007). Myocardial matrix remodeling and the matrix metalloproteinases: influence on cardiac form and function. *Physiol Rev* **87**, 1285–1342.

Tune JD, Gorman MW & Feigl EO (2004). Matching coronary blood flow to myocardial oxygen consumption. *J Appl Physiol* **97**, 404–415.

Vakili BA, Okin PM & Devereux RB (2001). Prognostic implications of left ventricular hypertrophy. *Am Heart J* **141**, 334–341.

van der Velden J, Merkus D, Klarenbeek BR, James AT, Boontje NM, Dekkers DH, Stienen GJ, Lamers JM & Duncker DJ (2004). Alterations in myofilament function contribute to left ventricular dysfunction in pigs early after myocardial infarction. *Circ Res* **95**, e85–e95.

van Kats JP, Duncker DJ, Haitsma DB, Schuijt MP, Niebuur R, Stubenitsky R, Boomsma F, Schalekamp MA, Verdouw PD & Danser AH (2000). Angiotensin-converting enzyme inhibition and angiotensin II type 1 receptor blockade prevent cardiac remodeling in pigs after myocardial infarction: role of tissue angiotensin II. *Circulation* **102**, 1556–1563.

van Veldhuisen DJ, van den Heuvel AF, Blanksma PK & Crijns HJ (1998). Ischemia and left ventricular dysfunction: a reciprocal relation? *J Cardiovasc Pharmacol* **32**(Suppl 1), S46–S51.

## Acknowledgements

The present study was supported by grants from the Netherlands Heart Foundation (2000T042 (D.M.), NHS2005B234 (D.K.) and 2005B220 (J.vdV.)).

Garcia et al. BMC Systems Biology 2014, 8:34
http://www.biomedcentral.com/1752-0509/8/34

BMC Systems Biology

RESEARCH ARTICLE Open Access

# Network and matrix analysis of the respiratory disease interactome

Benjamin Garcia[1,2,3], Gargi Datta[1,2], Gregory P Cosgrove[4] and Michael Strong[1,2*]

## Abstract

**Background:** Although respiratory diseases exhibit in a wide array of clinical manifestations, certain respiratory diseases may share related genetic mechanisms or may be influenced by similar chemical stimuli. Here we explore and infer relationships among genes, diseases, and chemicals using network and matrix based clustering methods.

**Results:** In order to better understand and elucidate these shared genetic mechanisms and chemical relationships we analyzed a comprehensive collection of gene, disease, and chemical relationships pertinent to respiratory disease, using network and matrix based analysis approaches. Our methods enabled us to analyze relationships and make biological inferences among over 200 different respiratory and related diseases, involving thousands of gene-chemical-disease relationships.

**Conclusions:** The resulting networks provided insight into shared mechanisms of respiratory disease and in some cases suggest novel targets or repurposed drug strategies.

**Keywords:** Interactome, Networks, Respiratory diseases, Lung disease

## Background

The capability to catalog interactions among diseases, chemicals, and genes into well-curated databases offers a collective knowledge of experimental results that has great potential for the generation of hypotheses and meta-analyses. To date, many biological databases have been established to catalog relationships among genes [1], diseases [2], and chemicals [3]. Many of these databases focus on one particular type of relational interaction, ranging from protein-protein interaction databases [1], gene-chemical databases [4], and disease-gene databases [2], and are often constructed using data mining methods complemented by manual curation. The described databases, in many instances, serve as the foundation for a wide array of predictive and analytical methods to examine interactions. They can also be extended to analyze interactions among overarching themes, including analyzing gene-chemical interactions within the context of a given set of diseases or protein-protein interactions within the context of peptide recognition [5,6]. Integration of multiple sources and types of relational data remains an important and challenging research area with great potential toward the development of furthering our understanding complex diseases and interactions.

Each year over 400,000 deaths occur in the United States as a result of respiratory and related diseases (RRD) [7]. Given the high prevalence and importance of lung and respiratory diseases, we hypothesized that a better understanding of the respiratory gene-chemical-disease interactome would lead to better understanding of the molecular mechanisms of lung disease, including the environmental and drug influences, and more importantly, may lead to new treatment or intervention strategies. In this study, we focus our efforts on the analysis of gene-disease-chemical relationships, in order to elucidate and infer novel interactions and to understand biology pertinent to respiratory diseases using network and matrix-based methods.

Current network and matrix-based analyses of disease relationships has relied heavily on gene or protein-centric examinations [8-11], neglecting chemical features that may also influence disease. Likewise, network analysis techniques have often been developed and utilized to examine gene or protein relationships among diseases [12], but often

* Correspondence: StrongM@NJHealth.org
[1]Integrated Center for Genes, Environment, and Health, National Jewish Health, Denver, CO 80206, USA
[2]Computational Bioscience Program, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO 80045, USA
Full list of author information is available at the end of the article

neglect environmental or chemical factors that may influence disease. In cases where genes, diseases, and chemicals have been analyzed, often the networks are decoupled to allow for the analysis of a single entity or relationship type, such as the effect of a drug on a gene network or the elucidation of molecular mechanisms in disease [13-15]. Host-pathogen studies have also largely focused on a single relational type, predominantly protein-protein interaction relationships [16]. Here we apply methods to investigate gene-chemical-disease networks, in order to better understand the genetic and chemical contributors of diseases, elucidating novel biology and helping to further understand shared disease pathology.

## Results and discussion

### Network construction

In order to compile a comprehensive dataset to examine gene, disease, and chemical relationships pertinent to respiratory disease, we extracted information from the Comparative Toxicogenomics Database (CTD) [4] and the Human Protein Reference Database (HPRD). CTD houses manually curated information pertinent to gene-disease-chemical relationships for a wide variety of diseases, and HPRD houses information focusing on protein-protein interactions from a wide array of experiments in humans and other model organisms. CTD offers a conservative and expert curated source of interactions to form networks, and HPRD uses the same normalized gene names as CTD.

We compiled and filtered our in-house database in two ways. The first database, we refer to as the whole respiratory network (Additional file 1: Table S1), and the second database we refer to as the therapeutic network (Additional file 2: Table S2). The whole respiratory network represents disease-gene, disease-chemical, chemical-gene, and gene-gene interactions associated with respiratory diseases. The therapeutic network, in contrast, consists of a subset of the respiratory network, containing only chemicals with curated therapeutic interactions with diseases and the genes that interact with those chemicals. These curated therapeutic interactions are established using the "DirectEvidence" field from CTD. This network was called the therapeutic network as a reference to this inclusion criterion. In addition to the therapeutic inclusion criteria, chemical-chemical interactions were also included based upon curated chemical relationships derived from chemical gene-interaction information. Gene-gene interactions were established using the HPRD database [1].

To assess the directionality of chemical-gene interactions, the uniqueness of chemical-gene and gene-chemical interactions were assessed. First, chemicals with disease interactions were batched queried using CTD, with an output of curated chemical-gene interactions. Second, genes with disease interactions were batched queried using CTD, with an output of curated gene-chemical interactions. The intersection between these two sets was then calculated. In the whole respiratory network, there were 27075 total chemical-gene and gene-chemical linkages with 13543 remaining after accounting for bi-directionality of interactions. Given the small percentage of directional linkages (~0.05%), all links were treated as bi-directional.

The type of interaction was established for disease-chemical, disease-gene, and chemical-gene interactions. For disease-chemical and disease-gene interactions, there were three types of interactions based upon CTD curation: therapeutic, marker/mechanism, and both therapeutic and marker/mechanism. Chemical-gene interactions had three major effects and one minor effect based upon CTD curation. The major effects are increasing, decreasing, and affecting expression or activity. The minor effect is based upon the type of protein modification imparted by the chemical onto the protein. The list of protein modification includes: ubiquitination, phosphorylation, oxidation, cleavage, methylation, hydrolysis, hydroxylation, glycosylation, glucuronidation, acetylation, nitrosation, ribosylation.

To establish chemical-chemical linkages and the type of gene-chemical linkage, CTD was used [4]. Chemical-gene interactions were extracted with a query specifying interaction type. Co-interactions between multiple chemicals and a gene were extracted from this list and chemical-chemical linkages were established if two chemicals had a curated co-occurrence with a gene. A co-occurrence was determined when a secondary chemical appeared in the interaction characteristics between a chemical and a gene. The type of linkage between the two chemicals was classified using the same type of link used to classify chemical-gene interaction in which the co-occurrence appeared. As there is often discordance between the naming of chemicals, especially those with pharmaceutical implications, a chemical reaction database and drug interaction database were not utilized for establishing chemical-chemical interactions.

After construction of the network, Jaccard similarity coefficients were generated between all nodes. Each coefficient was then classified based upon whether the two nodes were connected and the type of nodes being connected. Figure 1 represents the three node interaction types of interest: disease-gene interactions, disease-chemical interactions, and chemical-gene interactions. To test the alternative hypothesis that linked nodes are more similar than unlinked nodes based upon a Jaccard coefficient, Mann–Whitney U tests were performed on each of the three sets with a null hypothesis that the similarity between linked nodes and unlinked nodes is the same. In all three cases, Mann–Whitney U tests showed with greater than 99.9% confidence that linked nodes were more similar than unlinked nodes ($p < 0.01$). This suggests that the greater the

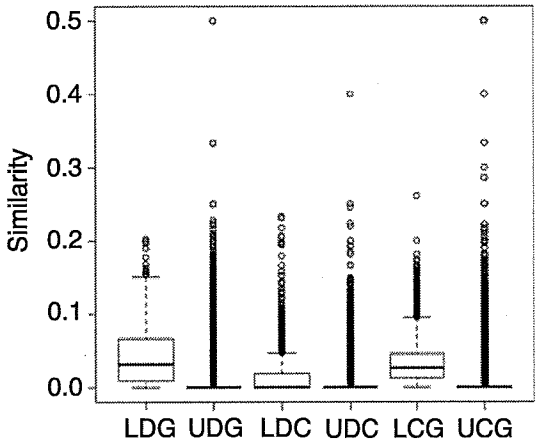


**Figure 1 Similarity of linked and unlinked nodes.** Jaccard similarity coefficients between linked and unlinked nodes in the therapeutic network. LDG – linked disease-gene, UDG – unlinked disease-gene, LDC – linked disease-chemical, UDC – unlinked disease-chemical, LCG – linked chemical-gene, UCG – unlinked chemical-gene. In all three sets, the linked nodes were more similar than unlinked nodes determined by Mann Whitney U tests ($p < 0.01$). The median for all unlinked node pairs is 0, with the medians for the linked pairs: DG – 0.026, DC – 0, CG – 0.032.

similarity between nodes, the more likely they are to interact. To assess the stability of the Jaccard coefficient, single edge additions were added to sub-networks. Kolmogorov-Smirnov tests were then run on the Jaccard coefficient distributions of the individual sub-network against perturbations within that sub-network. The result is that no perturbation caused a significant shift in distribution (average p-value ~ 0.99), with smaller sub-networks being more affected by perturbations (minimum p-value ~ 0.10). This lack of significant change is due to an addition of one edge having only small impacts on network topology, validating the Jaccard similarity as a stable measure of similarity for small amounts of missing data.

### Clustering methods

Evaluation of protein-protein interaction network clustering methods is generally performed through the comparison of gold standard regulatory networks or pathways. Since an analogous gold standard gene-chemical-disease network does not exist, for us to evaluate clustering methods, we selected high performing methods used for clustering protein-protein interaction networks, with the added stipulation that their output is scalable to a more sparse and dissimilar network. MCODE and MCL, two widely accepted and utilized clustering methods were tested for scalability when adding additional node types [11,17-19]. The gene-gene portion of the network was used as a baseline for the types and sizes of sub-networks that can be expected in an ideal situation. In the gene-gene network, both algorithms performed similarly with median sub-network sizes of 4 for MCODE and 3 for MCL. In the larger sub-networks both methods displayed highly interconnected clusters. In the therapeutic network, however, the clustering methods performed much differently. MCODE had a median sub-network size of 18, while still maintaining the highly interconnected networks, and MCL had a median cluster size of 3, and no longer exhibited an interconnected feature. We also applied hierarchical clustering, utilizing a Pearson's correlation coefficient. Pearson's has been shown to be a highly robust unsupervised correlation that performs well under a multitude of protein-protein interaction analyses, from identifying regulatory networks to identifying groups of proteins with shared functions [20,21]. A lack of a gold standard gene-disease-chemical network is also why no semi-supervised or supervised methods were chosen.

### Node-edge analysis

For the whole respiratory network, nodes were input based upon type (disease, chemical, gene) and edges based upon types of nodes involved (disease-gene, gene-gene, chemical-gene, disease-chemical) into Cytoscape [22], creating a network of 1,830 nodes and 17,275 edges. This network became a test-bed for methods to improve subsequent analyses including constructing networks with only one type of edge, and networks with filtered diseases, chemicals, and genes of interest. These tests led to the creation of both a gene-disease sub-network and the therapeutic chemical network.

The gene-disease sub-network was visualized by Cytoscape to determine clusters of similar genes not seen in the overall network. Figure 2 shows one such cluster of shared

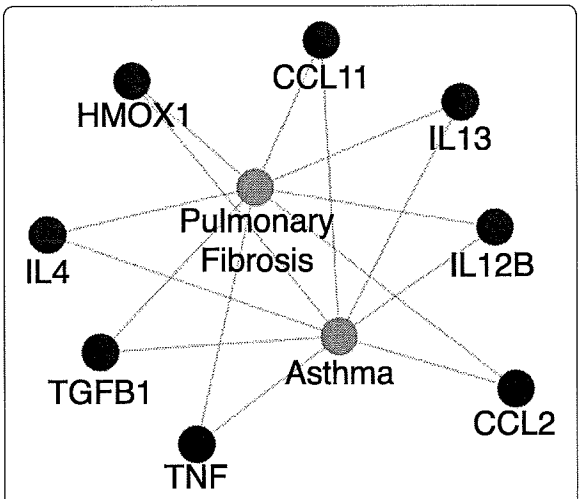


**Figure 2 Gene-disease sub-network.** Shared genes between Pulmonary Fibrosis and Asthma. Many of these genes represent similar pathophysiology in both diseases, such as activation of Th2 cell inflammation by IL13 and IL4. Red links represent marker/mechanism links.

genes between asthma and pulmonary fibrosis. Among the genes we observe linked to pulmonary fibrosis and asthma, we see the IL4 and IL13 cytokines. Both IL4 and IL13 are involved in activating Th2 cell inflammation, involved in asthma. Both IL4 and IL13 antagonists have also been shown to be effective in asthma therapy through the dampening of inflammation associated with asthma. In addition to being involved in asthma, IL13 has also been linked to pulmonary fibrosis, stimulated by the activation of Th2 cell inflammation, leading to tissue fibrosis. TGFB1 also induces inflammation, apoptosis, and fibrosis in mouse models [23,24], and has been associated with asthma. Networks such as these may be used to identify shared genetic mechanisms or molecular pathways of disease, and can also be used to identify novel drug targets or repurposed drug strategies to combat diseases that may be clinically very different, but that may share common genetic or molecular relationships.

For the therapeutic network, full information about the interaction between nodes was input into Cytoscape and visualized using an organic graph layout [22]. Nodes were colored by disease, chemical, or gene. Edges were colored by positive interactions (therapeutic or increases), negative interactions (marker/mechanism or decreases), mixed interactions (affects or therapeutic with marker/mechanism), and color intensity weighted by any protein modifications. Based upon database inclusion criteria, there were 388 genes, 227 diseases, and 578 chemicals. There were 10,679 linkages between these nodes, with each linkage having a characteristic path length of 3 and each node having an average of 18 neighbors. These numbers are about half that of the whole respiratory network, both decreasing the size of the network and making the network more directed towards finding positive interactions between chemicals and diseases. Linkage statistics from both networks can be seen in Table 1. A schematic of the overall process of creating and analyzing the therapeutic network can be seen in Figure 3.

To elucidate clusters of interest, the Cytoscape plugin MCODE was run on the network using a degree cutoff of 2, a node score cutoff of 0.2, a K-Core of 2, and a max depth of 100 [17]. This resulted in 18 highly interconnected clusters with a diverse set of node types (Additional file 3: Table S3), allowing the therapeutic network to be investigated and parsed into manageable sub-networks. These sub-networks offer a more manageable network to elucidate and identify novel and relevant interactions. Figure 4 demonstrates two of these sub-networks. Non-connected nodes that occur in highly interconnected sub-networks, particularly those with shared neighbors, offer a refined starting point for inferring novel interactions. Connections of interest were investigated by randomly choosing 23 unlinked node-pairs from the resulting sub-networks. These 23 inferred links were then analyzed by manually mining literature and databases for evidence that the two nodes might be linked by methods beyond those we used to establish our networks. In the absence of a gold standard, manual literature mining is often used to for validating inferences [25]. Supporting evidence for these inferred links can be seen in Table 2.

**Table 1 Network nodes and links**

| | Whole respiratory | Therapeutic |
|---|---|---|
| Nodes | | |
| Genes | 426 | 388 |
| Chemicals | 1177 | 578 |
| Diseases | 227 | 227 |
| Total | 1830 | 1193 |
| Links | | |
| Gene-chemical | 13543 | 7587 |
| Gene-gene | 438 | 433 |
| Chemical-chemical | 0 | 435 |
| Disease-gene | 577 | 536 |
| Disease-chemical | 2717 | 1688 |
| Total | 17275 | 10679 |

Counts of each type of node and linkage for both the whole respiratory network and the therapeutic network.

One of these sub-networks, shown in Figure 4A, contains non-small-cell lung carcinoma and closely interacting genes and chemicals. From this sub-network, three links were analyzed in greater detail: aspirin - EGFR, acetaminophen - non-small-cell carcinoma, and piroxicam - non-small-cell carcinoma. Aspirin - EGFR is an inferred link in this sub-network that was added as a direct link to an update of CTD that occurred after the creation of this network [4]. There was strong support in literature for aspirin promoting EGFR inhibitors, enough for a curated interaction between these two elements [26,27]. This link represents a verified prediction both by literature and by CTD, representing the effectiveness of using sub-networks to find novel links. Acetaminophen - Non-small-cell lung carcinoma is a link that has negative support in literature [28]. In studies involving testing multiple anti-inflammatory drugs for change in non-small-cell lung carcinoma outcome, they found no correlation between Acetaminophen and change in prognosis [28]. The negative support for this link shows that while sub-networks offer a starting point for testing inferred interactions, not all of the nodes will have a direct link. Lastly, Piroxicam - non-small-cell carcinoma had direct and indirect literature support for this link [29,30]. There was increased immune function in lung cancer patients that had piroxicam added to their drug regimens [29]. Also, piroxicam showed decreased tumorigenesis in mice with colon

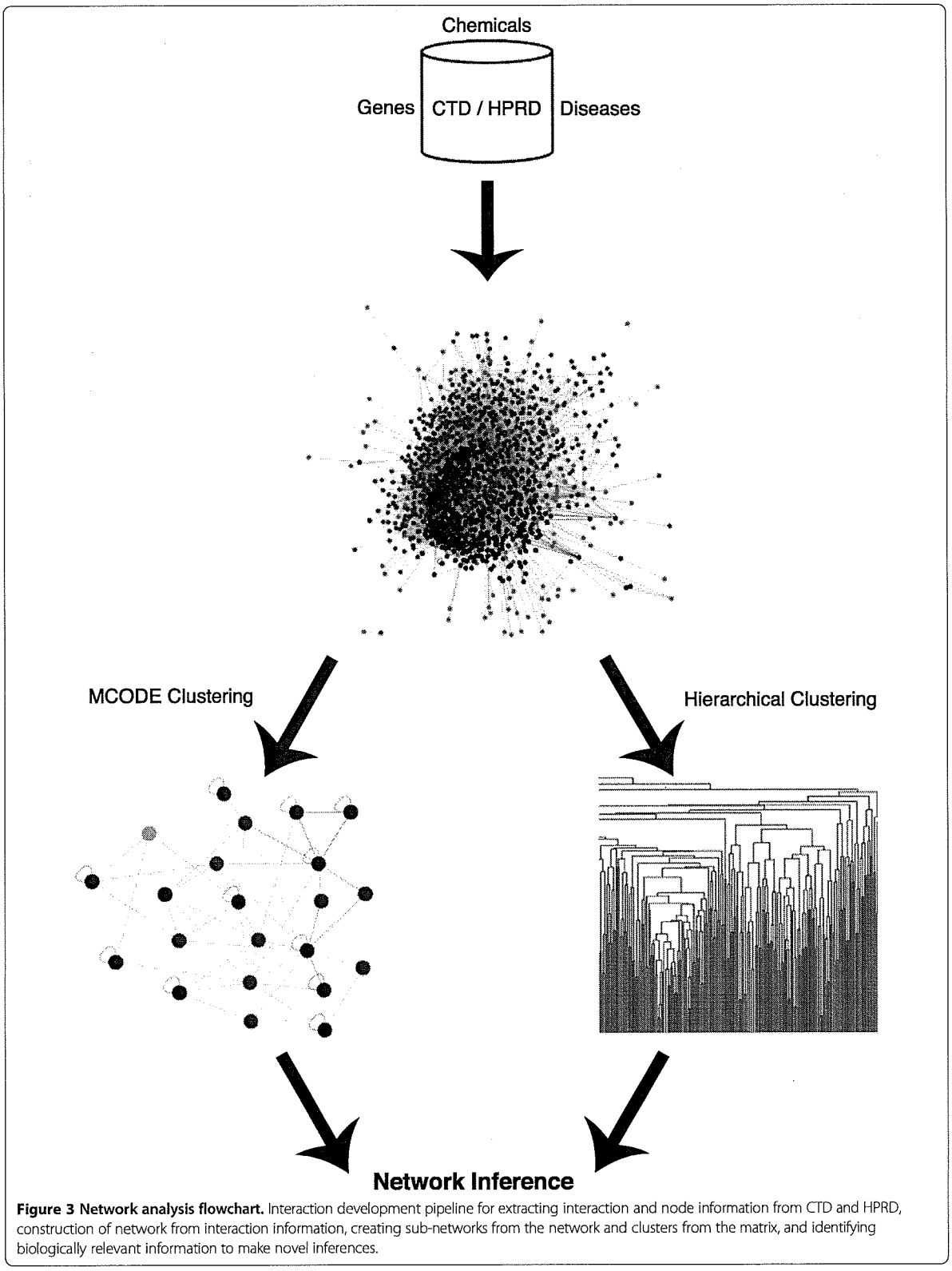


**Figure 3 Network analysis flowchart.** Interaction development pipeline for extracting interaction and node information from CTD and HPRD, construction of network from interaction information, creating sub-networks from the network and clusters from the matrix, and identifying biologically relevant information to make novel inferences.

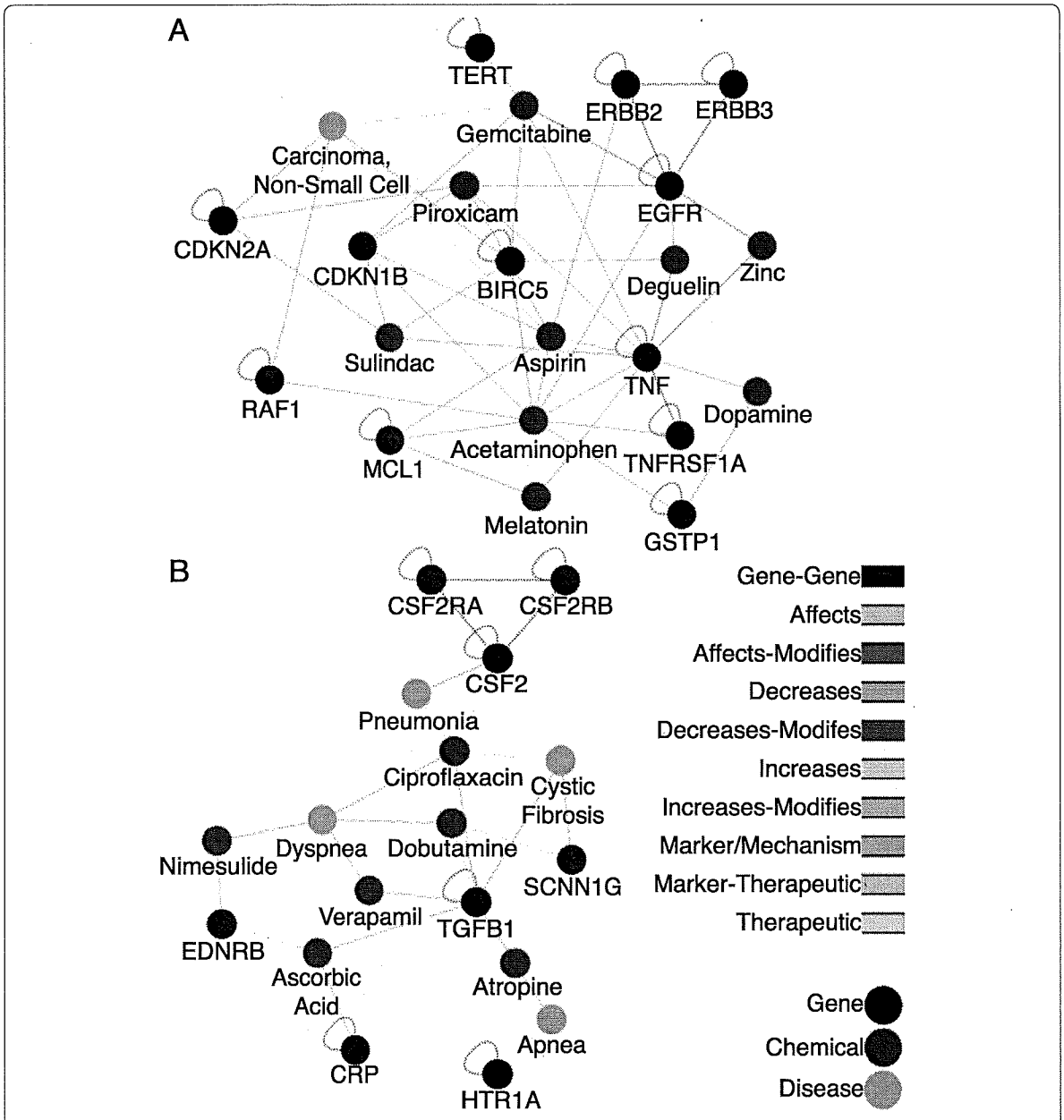


**Figure 4 Visualization of sub-network interactions. A)** Clustered sub-network containing non-small-cell lung carcinoma and a restricted view of closely interacting chemicals and genes for visualization purposes. **B)** Clustered sub-network containing Cystic Fibrosis and closely interacting diseases, chemicals, and genes.

cancer, suggesting this link might be present in other cancers as well [30]. This link represents a possible avenue for further research. There is evidence to support that there are beneficial effects of piroxicam on non-small-cell carcinoma prognosis; however, the full effects of this interaction are not well understood.

Analysis of sub-networks also presents the ability to find links for similar or comorbid diseases. In the cystic fibrosis sub-network, Figure 4B, dobutamine interacts with both of cystic fibrosis' genes in the sub-network, suggesting a link between dobutamine and cystic fibrosis. Upon searching the literature, dobutamine, especially in combination with nitric oxide, improves pulmonary hypertension in cystic fibrosis patients, a common comorbidity [31]. CTD neither contains a link between dobutamine and cystic fibrosis nor dobutamine and pulmonary hypertension.

**Table 2 Inferred interaction summary**

| Link | Inferred links | Literature support | Expression support | Database support | Anti-support | No support |
|---|---|---|---|---|---|---|
| Gene-disease | 10 | 9 | 0 | 3 | 0 | 1 |
| Gene-chemical | 9 | 3 | 5 | 1 | 1 | 4 |
| Chemical-disease | 4 | 2 | 0 | 1 | 1 | 0 |

Type of inferred link and support for each link. Inferred link is number of currently non-linked node pairs analyzed in each category. Literature support means a PubMed search resulted in a published article that supports the link. Expression support means there is literature support for gene expression changes. Database support means that there is support for a link due to curation methodology or the link was added in later version of CTD. Anti-support means that literature specifically says this link is not real, and no support means that no evidence could be found for or against the link.

Jaccard similarity coefficients were generated for each sub-network. These coefficients measured similarity using only nodes and links present within the sub-network. Similarities were then averaged for each node, representing how similar a given node is to the sub-network as a whole. The same 23 unlinked node pairs from the previous analysis were used to determine the relationship between similarity and literature evidence. Similarity between the two nodes was ranked against the similarity of all other pairwise Jaccard coefficients within the sub-network, with the similarity being broken into one of three sets: upper 25$^{th}$ percentile, middle percentile, and the lower 25$^{th}$ percentile. These comparisons represent how similar the two nodes are to each other, relative to the sub-network as a whole. Evidence for a possible interaction was then manually mined from published articles, and then compared to their similarity classifications. Table 3 represents mined literature support against similarity classification. With increasing similarity between the two nodes, relative to their ranked similarities within the sub-network, there was increasing evidence in literature to support connection between the two nodes. In addition to having a greater likelihood of evidence based upon similarity, just being in the same sub-network increased the likelihood of two nodes having a connection over the 0.015 probability of any two random nodes being linked together in the databases used for constructing the network. This shows a complimentary relationship between clustering and similarity when trying to determine if there is evidence to support two nodes being linked.

A more systematic evaluation of the relationship between Jaccard similarity and identifying novel links was performed on a human signaling network [32]. Protein-Protein interactions from the human signaling network were selected based upon both the interacting genes being present in the therapeutic network while their interaction was not present in the network. While self-interacting genes were utilized in generating Jaccard similarity values, they were excluded from both the background and the human signaling network during the analysis. This is due to the fact there is no way to distinguish between likely self-interactions and unlikely self-interactions using a similarity measure that will always be 1.0 in the case of a self-interaction. This selection resulted in 1057 additional interactions for use in validation.

A Mann–Whitney U test was performed on the human signaling network gene-gene interactions with the null hypothesis that there is no similarity difference from the background of possible gene-gene interactions. The alternative hypothesis is that the novel interactions from the human signaling interactions are more similar than the background. This test resulted in a $p < 0.01$, showing that these novel interactions are more similar than the background. Just as the literature study, rank of the Jaccard coefficient was also important to whether or not an interaction was found. There was an exponential relationship between the rank and inclusion into the human signaling network with roughly 40% of the additional interactions being in the 90$^{th}$ percentile or greater (Figure 5).

### Matrix analysis

A binary interaction matrix was created using the network interaction triples for both the whole respiratory and therapeutic networks, Values of 1 represent an interaction; whereas, values of 0 represent a lack of interactions. These matrixes were then used as input to Cluster 3.0, an open source clustering tool [33]. An uncentered similarity matrix with average linkage was used to calculate hierarchical clustering. Output of the dendrogram

**Table 3 Jaccard similarity assessment**

| Jaccard percentile | Support | Anti-support | No support | Percent support |
|---|---|---|---|---|
| 75-100 | 7 | 1 | 1 | 77.8% |
| 25-75 | 5 | 0 | 2 | 71.4% |
| 0-25 | 4 | 1 | 2 | 57.1% |

Supporting evidence for and inferred linkage utilizing the Jaccard coefficient between two nodes. The rank of the pairwise Jaccard coefficient within the sub-network was compared to the ability to find evidence supporting the pairwise connection. A rank of 100 represents the highest Jaccard coefficient within the sub-network and a rank of 0 represents the lowest Jaccard coefficient within the sub-network.

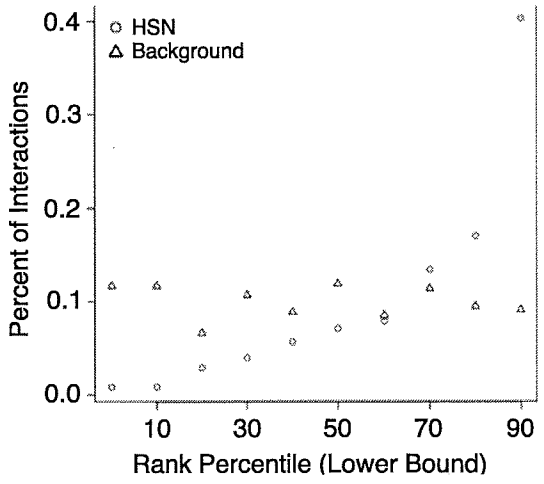


**Figure 5 Jaccard rank of human signaling network interactions.** Percent of gene-gene interactions that fall into a given percentile range (broken into ranges of 10 percent) for both the background of possible interactions and the novel interactions represented by the human signaling network (HSN). A lower bound of the 10$^{th}$ percentile represents the range of greater than the 10$^{th}$ percentile to less than or equal to the 20$^{th}$ percentile. The discrepancy in that the background percentiles are not exactly equal to 10% is due to the fact that duplicates of Jaccard coefficients at the boundary percentiles were treated the same as the boundary.

was viewed in TreeView [34]. Clustered interactions from the therapeutic matrix are shown in Figure 6.

Individual clusters from the therapeutic matrix were established using a 0.7 and 0.4 similarity threshold. Both of these thresholds were chosen as they represent inflection points in the node count versus similarity graph, as shown in Figure 7. Inflection points represent possible changes in cluster characteristics, such as separating high similarity clusters with medium similarity clusters. The 0.7 threshold resulted in 71 clusters. The smallest cluster had 2 nodes and the largest with 13 nodes. The 0.4 threshold resulted in 211 clusters (Additional file 4: Table S4). The smallest cluster had 2 nodes and the largest with 45 nodes. The 0.7 threshold offers the highest similarity between nodes; however, it often results in the inclusion of nodes that only have a few total number of interactions. The ERBB gene family was found in the 0.4 threshold but not in the 0.7 threshold. Also, the 0.4 threshold included both expansions and additions of clusters, such as the expansion of and anti-histamine cluster to include additional anti-histamines, and the addition of a tumorigenesis gene cluster. This expanded set of clusters supports the idea that the 0.4 threshold is more useful for finding clusters of similar function, while still maintaining a similar specificity as the clusters found in the 0.7 threshold.

Unlike sub-networks, clustering of the matrix elucidates families of chemicals, genes, and diseases with similar phenotypes and chemical characteristics. Figure 8 shows clusters in each of these three node categories from a similarity cutoff of 0.4. These clusters contain a group of beta2-agonists (Figure 8A), ERBB family proteins (Figure 8B), and a group of fungal lung diseases (Figure 8C). For the matrix clusters, genes had a tendency to cluster with other genes, chemicals with other chemicals, and diseases with other diseases. Almost all of the clusters were made up of elements of the same type, supporting the idea that this matrix clustering approach is suitable for finding nodes with similar properties versus the more diverse interactomes in the traditional sub-networks.

The assertion that subclusters can identify nodes with similar properties can be used for predicting interactions by analyzing overlap between cluster nodes and their shared interactions. In a cluster containing SEPP1, GJB1, SELENBP1, SLC22A18, A2M, and PDFGA, five out of the six genes in this cluster have an association with lung neoplasms. PDGFA, the gene not linked with lung neoplasms, has associations with breast, prostate, head and neck, and pancreas cancers. In addition, PDGFA increases with asbestos exposure, a chemical linked to mesothelioma [35]. This increase is also associated with tumorigenicity, supporting the assertion that PDFGA is also a marker for lung neoplasms [35].

Ebastine, levocabastine, hydroxyzine, SUN1334H, azelastine, olopatadine, cetirizine, desloratadine, sho-seiryu-to, epinastine, and tripolidine are a group of anti-histamine drugs that target HRH1, all of which also have interactions with rhinitis. These anti-histamine drugs also have anti-inflammatory properties, revealed by seven drugs having links to IL4, four having links to IL5, and four having links to IL8. This is supported by a study that shows various anti-histamines having anti-inflammatory properties in rhinitis pathology [36].

MT2, MT1, CCL9, CCL8, ECM1, and SLC39A4 represent a diverse cluster of two metallothionein proteins, two macrophage proteins, one extracellular matrix protein, and one zinc transporter protein. Many of these genes regulate metal concentrations within cells and are linked to respiratory hypersensitivity. Out of the five shared chemicals, only acetaminophen is linked to respiratory hypersensitivity. However, four out of these five chemicals have links to asthma, suggesting they may play a greater role in respiratory hypersensitivity in general. This hypothesis is supported by the fact that zinc deficiency alters respiratory epithelium in allergic response of mice [37].

Ofloxacin, amoxicillin clavulanate, clarithromycin, and azizthromycin are a group of antibiotics that treat respiratory infections. The interactome of these antibiotics

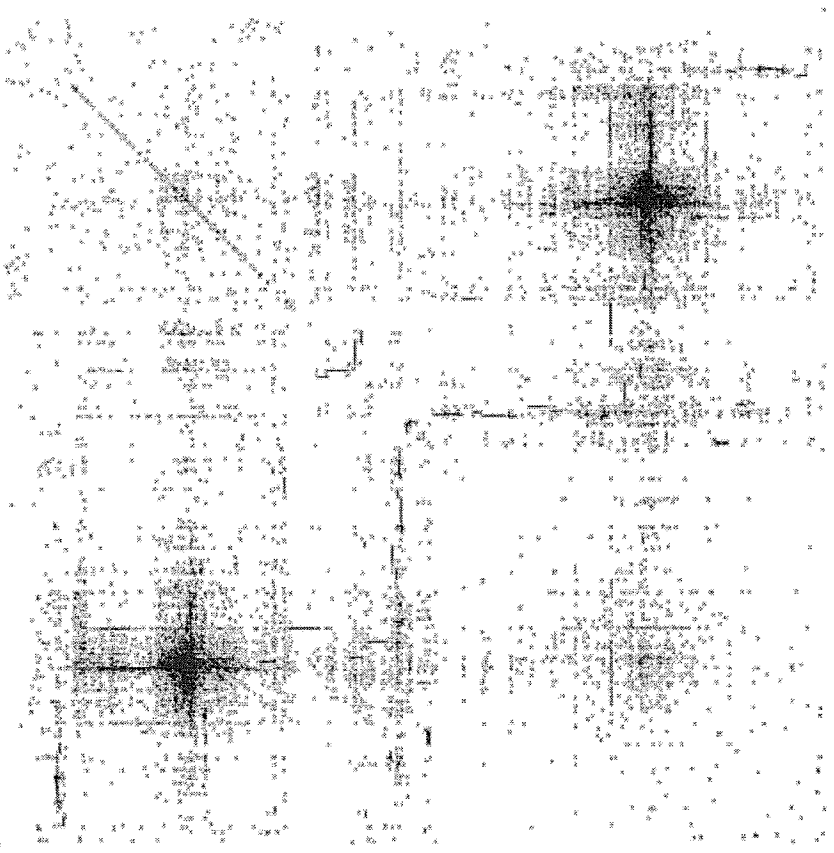

**Figure 6 Therapeutic diffusion matrix.** Therapeutic matrix clustered based upon uncentered Pearson's correlation coefficient with average linkages and then hierarchical clustering. Each red dot represents an interaction pair with the rows and columns representing nodes. The green represents diffusion to aid in visualizing the sparse network. Node ordering is based upon similarity to adjacent nodes.

is shown in Figure 9. Of these antibiotics, only clarithromycin and ofloxacin have direct links to tuberculosis in CTD. The drug combination amoxicillin-clavulanate has literature support that it is effective in treating tuberculosis, whereas; amoxicillin alone is ineffective [38,39]. This increase in effectiveness with clavulanate is due to the fact clavulanate inhibits an enzyme that makes *Mycobacterium tuberculosis* resistant to amoxicillin [38,39]. While literature shows that azithromycin alone is also ineffective in treating tuberculosis isolates, literature shows that azithromycin in combination with capreomycon, pyrazinamide, ethambutol, and isoniazid improves outcomes in multi-drug resistant patients over streptomycin, ethambutol, pyrazinamide, and isoniazid [40,41]. Given the fact that tuberculosis is often treated with a combination of drugs, further evaluation of amoxicillin-clavulanate and azithromycin within the context of a drug regimen would offer a more practical approach to evaluating the effectiveness of treating tuberculosis patients with these antibiotics. Also of note are the links from azithromycin and clarithromycin to IL6 and IL4 respectively. It is thought that even though azithromycin does not directly kill *M. tuberculosis* in cell culture, it may have a pro-immune effects that improves outcomes of tuberculosis patients, or may play a role as an anti-inflammatory. BCL2L1 is affected by clarithromycin, a known tuberculosis drug, and azithromycin, an inferred TB drug. This coupled with a shared interaction of CCL2 between tuberculosis and azithromycin promotes that idea that azithromycin may have a therapeutic effect on tuberculosis through an anti-inflammatory response. Through the analysis of gene-disease-chemical networks we may gain better insight into both the direct target and off target activities of certain drugs, useful in the identification of drug repurposing strategies.

### Node-edge versus matrix

While these two approaches take the same input, clustering produces two distinct results. Only eight of the

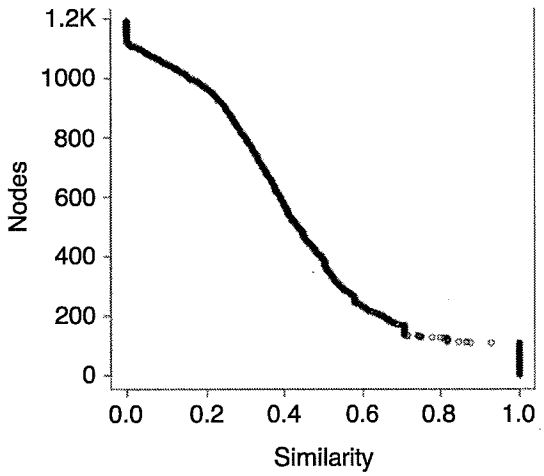


**Figure 7 Matrix similarity graph.** Number of hierarchical clustered nodes represented by a static cutoff for a given uncentered Pearson's correlation coefficient. The inflection point at ~ 0.4 similarity was used to generate clusters for analysis.

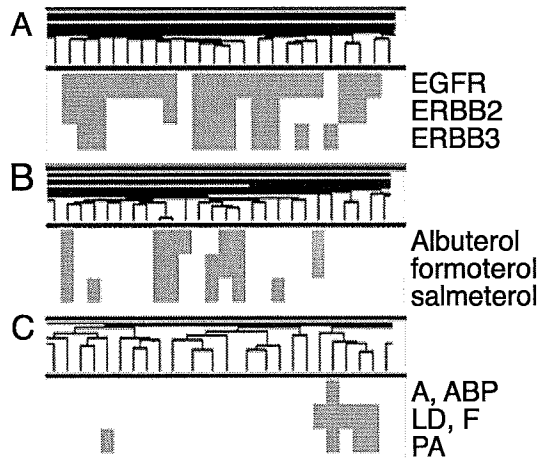


**Figure 8 Clusters of elements from 0.4 similarity cutoff. A)** Beta2-agonists **B)** ERBB family proteins. **C)** Fungal lung diseases; A, ABP – Aspergillosis, Allergic Bronchopulmonary; LD, F – Lung Disease, Fungal; PA – Pulmonary Aspergillosis.

eighteen sub-networks contained a cluster from the matrix where at least 50% of the nodes present in the matrix cluster were also present in the sub-network. Most of the matrix clusters that overlapped with the sub-networks contained only two or three nodes. However, one sub-network contained 11 of the 28 nodes in one matrix subcluster, making it the most nodes shared between a sub-network and a matrix cluster. These differences can be attributed to both network construction and the types of interactions that are obtained from each approach. Given the sparsity of the network, especially in chemical-chemical interactions, and the lack of disease-disease interactions, clustering coefficients and pairwise comparisons produce non-overlapping results. Clustering coefficients from node-edge based approaches represent closely interacting genes, chemicals, and diseases. These closely interacting nodes offer avenues of exploration for finding novel interactions. Pairwise comparisons from matrixes represent nodes that share the same interaction profile. This interaction profile can then be used for determining both biological meaning and novel interactions for any pairs between the cluster nodes and the interaction profile nodes. Thus, these two approaches offer a complimentary analysis strategy for sparse networks, enabling elucidation of both novel interactions and increasing our biological understanding of node clusters.

The second distinction these two approaches offer is in the visualization of interactions. Node-edge network approaches illustrate which nodes form a sub-network, which nodes interact within these sub-networks, and the types of interactions between each node, giving an all encompassing view of the sub-network. Matrix-based approaches provide a broader view of interactions, offering a tool for visualizing not only how similar nodes and clusters are to each other, but also the interactions nodes share outside of their individual clusters.

## Conclusion

Current network analyses of disease are still highly focused on gene and protein-based networks, neglecting environmental and drug effects that contribute to the pathophysiology of a disease or sets of diseases. Our proposed methods integrate both the chemical and disease entities into network and matrix-based analyses, allowing for a more complete systems understanding of the underlying biology. With this addition of multiple different entity types comes the lack of a gold standard for identifying specific genes, chemicals, and diseases that should cluster together, providing a similar role as the curated regulatory and pathway networks used to establish accuracy in protein-protein and gene-gene network analyses.

In order to better investigate complex and sparse networks, such as the respiratory disease interactome, a multi-method approach utilizing methods proven effective in gene-gene and protein-protein network-based analyses has proven useful to elucidate and investigate different network properties and the underlying biological context. In this case we have used two approaches: a node-edge-based clustering coefficient with Jaccard similarity comparison approach applied to traditional networks, and a matrix-based Pearson's correlation coefficient with hierarchical clustering approach. This allows identification of closely interacting diseases, chemicals, and genes, as well as similar interaction profiles either within or between these same elements of interest. These two approaches help facilitate investigations

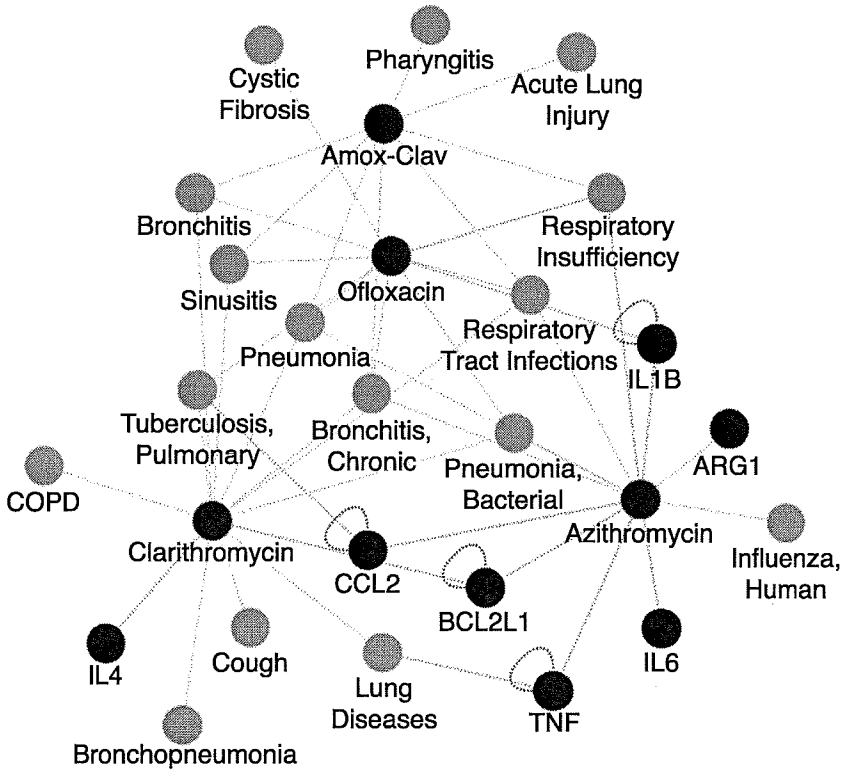


**Figure 9 Matrix cluster interactome.** Cluster of oflaxacin, amoxicillin-clavulanate (Amox-Clav), azithromycin, and clarithromycin with closely interacting genes and diseases.

on the underlying biology for a given disease, pathophysiology similarities across diseases, and chemicals that may have a therapeutic indication outside of their original use.

The shared interactome of four therapeutic antibiotics (ofloxacin, amoxicillin clavulanate, clarithromycin, and azizthromycin, (Figure 9) allows for an inference of interaction between azithromycin and tuberculosis based upon the interaction profile of a cluster generated by hierarchically clustering a Pearson's correlation coefficient matrix. This profile represents the layering of diseases, chemicals, genes and the interactions between them, showing that while azithromycin has no known anti-Mycobacterium tuberculosis properties, it does have pro-host immune properties that may have therapeutic merit for tuberculosis treatment.

These methods are also useful for finding drug targets. The shared interactome of pulmonary fibrosis and asthma (Figure 1) demonstrates that Th2 cell inflammation is an important factor in both of these diseases, where a drug that improves the outcomes in one of these diseases may also be useful for the other disease. Looking at these interactomes provides a broader context for drug discovery and drug repurposing.

Chemical, gene, and disease interactomes offer a novel approach to not only identify shared biology among diseases, but also offer a method for identifying possible new drug targets and repurposed drug strategies. Layering additional interaction information, additional databases, and additional analysis techniques will allow for a more complete systems-based analysis that will extend to any complex disease interactome.

## Methods

### Network generation

Respiratory diseases and the curated chemical and genes interactions with these diseases were extracted from CTD using the January 9, 2012 database version [4]. Curated chemical-gene interactions were extracted from batch queries using the chemicals and genes associated with respiratory diseases. Genes, chemicals, and their associated links that did not contain a link to a respiratory disease were removed from the list. Duplicates of gene-chemical and chemical-gene links were also removed from this list. Gene-gene interactions were established using the April 13, 2010 version of the HPRD database [1]. Genes and their associated interactions were excluded from the list if they did not

contain a direct link to a respiratory disease. These interactions were further specialized by including only chemicals with therapeutic interactions to respiratory diseases in a therapeutic network, with the "therapeutic" name stemming from including only chemicals with at least one therapeutic indication. The therapeutic indication for a chemical is determined from the by the direct evidence field from CTD. Genes were then excluded if they did not contain a link to one of these therapeutic chemicals. Chemical-chemical links and chemical-gene interaction characteristics for the therapeutic network were established using the February 10, 2012 version of CTD [4]. Chemical-chemical links were established using co-occurrence of chemicals in chemical-gene interactions. A chemical was established as co-occurring when a secondary chemical appeared in the interaction characteristics of chemical-gene interactions. A triple was stored for each interaction, including both interacting nodes and the type of interaction between them.

### Network and matrix visualization

A file containing the triples of interactions and a file containing the type of node (chemical, gene, disease) were loaded into Cytoscape [22]. Nodes were colored based upon their type, with chemicals represented as blue, genes as black, and diseases and orange. Interactions were colored based upon interaction characteristics, with positive interactions as green, negative interactions as red, mixed interactions as purple, and additional characteristics as increasing intensity.

A binary interaction matrix between nodes was created using the network construction file containing interaction triples. A value of 1 was used for any interaction type between nodes and a value of 0 was used for a lack of interaction between nodes. This binary interaction matrix was visualized by creating a bitmap of clustered interactions and the resulting dendrograms by using TreeView [34].

### Network and matrix clusters

MCODE, a Cytoscape plugin, was used to generate each of the sub-networks [17,22]. A degree cutoff of 2, and node score cutoff of 0.2, a k-core of 2, and a max depth of 100 were used as the MCODE parameters for generating clusters.

Cluster 3.0 was used to generate clusters for this matrix [33]. An uncentered similarity with average linkage was used to calculate the hierarchical clustering. Similarity scores of 0.4 and 0.7 were used for creating clusters, based upon inflection points Figure 7.

### Jaccard similarity

Jaccard similarity coefficients were generated for both the therapeutic network and for sub-networks using the following formula: $\frac{Node1 \cap Node2}{Node1+Node2-(Node1 \cap Node2)}$. This formula calculates the intersection of the two sets divided by their union. A set, in all cases, is all the nodes that interact with a given node, including any self-interactions. The intersection of two nodes is all shared interactions between the two nodes, with the union of the two nodes being all the nodes that interact with at least one of the nodes of interest. For the entire therapeutic network, a Mann–Whitney U test was run with the alternative hypothesis that linked nodes are more similar than unlinked nodes. For sub-networks, ranks of Jaccard coefficients were calculated using the individual sub-network that a node pair come from and then compared to the evidence of there being an interaction.

### Network stability

Sub-networks were used to assess the stability of the network in respect to changes in Jaccard coefficient. For a given sub-network, an additional network was generated for each missing edge. In each of these networks one additional edge was added between two existing unlinked nodes, creating a unique set of networks. Jaccard coefficients were then generated for each additional network. Two-sample Kolmogorov-Smirnov tests were used to assess whether or not the distribution of the original sub-network and the altered sub-networks was shifted. This was done for each of the sub-networks and their corresponding altered networks. The null hypothesis was that the Jaccard coefficient distribution of the network with an additional edge is the same as the unaltered sub-network, with the alternative hypothesis being that the distribution is shifted.

### Programming

Original network parsing to establish interactions between nodes was done using perl version 5.12.4 on Mac OSX 10.7. This includes parsing interactions between genes, chemicals, and diseases, finding which chemicals have co-interactions with genes, finding unique interactions and directional interactions between chemicals and genes, finding interaction characteristics for disease-gene and disease-chemical interactions, and selecting inclusion criteria for interactions of interest to develop each network.

Further network parsing, matrix construction, and dendrogram parsing was done using C#/.NET 4.0 on a Windows 7 machine. This includes finding specific interaction characteristics for chemical-gene and chemical-chemical interactions, construction of the interaction matrix, visualization of the interaction matrix, and extracting clusters based upon a threshold from the output from Cluster 3.0.

## Additional files

**Additional file 1: Table S1.** Whole Respiratory Network.
**Additional file 2: Table S2.** Therapeutic Network.
**Additional file 3: Table S3.** MCODE Clusters.
**Additional file 4: Table S4.** Hierarchical Clusters.

### Competing interests
The authors declare that they have no competing interests.

### Authors' contribution
BG – Co-conceived the Project, Methods development, coding, paper writing. GD – Methods development, coding, paper editing. GC – Methods development. MS – Conceived the Project, Supervised the Project, Methods development, paper editing. All authors read and approved the final manuscript.

### Acknowledgements
BG acknowledges support from a NLM Institutional Training Grant, NIH 5T15LM009451, and from NICTA, which is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. All authors acknowledge support from a Translational Research Inititative Grant from National Jewish Health, MS acknowledges support from the Boettcher Foundation Webb-Waring Award, and BG thanks Cheng Soon Ong and Melissa Davis of NICTA in Melbourne and Sonia Leach of National Jewish Health for our discussions. BG also thanks Joseph Cursons of NICTA in Melbourne for his help with matrix visualization.

### Author details
[1]Integrated Center for Genes, Environment, and Health, National Jewish Health, Denver, CO 80206, USA. [2]Computational Bioscience Program, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO 80045, USA. [3]NICTA, Victoria Research Lab, Melbourne, Victoria 3010, Australia. [4]Department of Medicine, National Jewish Health, Denver, CO 80206, USA.

Received: 25 November 2013 Accepted: 10 March 2014
Published: 22 March 2014

### References
1. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A: **Human protein reference database–2009 update.** *Nucleic Acids Res* 2009, **37**:D767–D772.
2. Baxevanis AD: **Searching Online Mendelian Inheritance in Man (OMIM) for information for genetic loci involved in human disease.** *Current protocols in bioinformatics.* 2003, **35**(9.13):1–9. 13.15.
3. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M: **LIGAND: database of chemical compounds and reactions in biological pathways.** *Nucleic Acids Res* 2002, **30**:402–404.
4. Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, Rosenstein M, Wiegers T, Mattingly CJ: **The comparative toxicogenomics database: update 2011.** *Nucleic Acids Res* 2011, **39**:D1067–D1072.
5. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR: **The connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.** *Science* 2006, **313**:1929–1935.
6. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, Quondam M, Zucconi A, Hogue CW, Fields S, Boone C, Cesareni G: **A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules.** *Science* 2002, **295**:321–324.
7. American Lung Association: *State of lung disease in diverse communities.* 2010.
8. Kaneko Y, Yatagai Y, Yamada H, Iijima H, Masuko H, Sakamoto T, Hizawa N: **The search for common pathways underlying asthma and COPD.** *Int J Chron Obstruct Pulmon Dis* 2013, **8**:65–78.
9. Dockstader K, Nunley K, Karimpour-Fard A, Medway A, Nelson P, Port JD, Liggett SB, Bristow MR, Sucharov CC: **Temporal analysis of mRNA and miRNA expression in transgenic mice overexpressing Arg- and Gly389 polymorphic variants of the beta1-adrenergic receptor.** *Physiol Genomics* 2011, **43**:1294–1306.
10. Janjic V, Przulj N: **Biological function through network topology: a survey of the human diseasome.** *Brief Funct Genomics* 2012, **11**:522–532.
11. Islam MF, Hoque MM, Banik RS, Roy S, Sumi SS, Hassan FM, Tomal MT, Ullah A, Rahman KM: **Comparative analysis of differential network modularity in tissue specific normal and cancer protein interaction networks.** *J Clin Bioinformatics* 2013, **3**:19.
12. Barzel B, Barabasi AL: **Network link prediction by global silencing of indirect correlations.** *Nat Biotechnol* 2013, **31**:720–725.
13. Iorio F, Saez-Rodriguez J, Bernardo D: **Network based elucidation of drug response: from modulators to targets.** *BMC Syst Biol* 2013, **7**:139.
14. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H: **Three-dimensional reconstruction of protein networks provides insight into human genetic disease.** *Nat Biotechnol* 2012, **30**:159–164.
15. Yeh SH, Yeh HY, Soo VW: **A network flow approach to predict drug targets from microarray data, disease genes and interactome network - case study on prostate cancer.** *J Clin Bioinformatics* 2012, **2**:1.
16. Gulbahce N, Yan H, Dricot A, Padi M, Byrdsong D, Franchi R, Lee DS, Rozenblatt-Rosen O, Mar JC, Calderwood MA, Baldwin A, Zhao B, Santhanam B, Braun P, Simonis N, Huh KW, Hellner K, Grace M, Chen A, Rubio R, Marto JA, Christakis NA, Kieff E, Roth FP, Roecklein-Canfield J, Decaprio JA, Cusick ME, Quackenbush J, Hill DE, Münger K, Vidal M, Barabási AL: **Viral perturbations of host networks reflect disease etiology.** *PLoS Comput Biol* 2012, **8**:e1002531.
17. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinforma* 2003, **4**:2.
18. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30**:1575–1584.
19. Nguyen P, Srihari S, Leong H: **Identifying conserved protein complexes between species by constructing interolog networks.** *BMC Bioinforma* 2013, **14**:S8.
20. Maetschke SR, Madhamshettiwar PB, Davis MJ, Ragan MA: **Supervised, semi-supervised and unsupervised inference of gene regulatory networks.** *Brief Bioinform* 2013, **15**:195–211.
21. Trinidad JC, Thalhammer A, Burlingame AL, Schoepfer R: **Activity-dependent protein dynamics define interconnected cores of co-regulated postsynaptic proteins.** *Mol Cell Proteomics: MCP* 2013, **12**:29–41.
22. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization.** *Bioinformatics* 2011, **27**:431–432.
23. Lee CG, Cho SJ, Kang MJ, Chapoval SP, Lee PJ, Noble PW, Yehualaeshet T, Lu B, Flavell RA, Milbrandt J, Homer RJ, Elias JA: **Early growth response gene 1-mediated apoptosis is essential for transforming growth factor beta1-induced pulmonary fibrosis.** *J Exp Med* 2004, **200**:377–389.
24. Wills-Karp M: **Interleukin-13 in asthma pathogenesis.** *Immunol Rev* 2004, **202**:175–190.
25. Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA: **Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets.** *Genome Med* 2012, **4**:41.
26. Becker JC, Muller-Tidow C, Stolte M, Fujimori T, Tidow N, Ilea AM, Brandts C, Tickenbrock L, Serve H, Berdel WE, Domschke W, Pohle T: **Acetylsalicylic acid enhances antiproliferative effects of the EGFR inhibitor gefitinib in the absence of activating mutations in gastric cancer.** *Int J Oncol* 2006, **29**:615–623.
27. Selvendiran K, Bratasz A, Tong L, Ignarro LJ, Kuppusamy P: **NCX-4016, a nitro-derivative of aspirin, inhibits EGFR and STAT3 signaling and modulates Bcl-2 proteins in cisplatin-resistant human ovarian cancer cells and xenografts.** *Cell Cycle* 2008, **7**:81–88.
28. Van Dyke AL, Cote ML, Prysak G, Claeys GB, Wenzlaff AS, Schwartz AG: **Regular adult aspirin use decreases the risk of non-small cell lung cancer among women.** *Canc Epidemiol Biomarkers Prev: a Pub of the Am Assoc for Cancer Res, cosponsored by the Am Soc of Preventive Oncolgy* 2008, **17**:148–157.
29. Braun DP, Bonomi PD, Taylor SG, Harris JE: **Modification of the effects of cytotoxic chemotherapy on the immune responses of cancer patients with a nonsteroidal, antiinflammatory drug, piroxicam. A pilot study of the Eastern Cooperative Oncology Group.** *J Biol Response Modif* 1987, **6**:331–345.

30. Palmerini E, Fan K, Yang K, Risio M, Edelmann W, Lipkin M, Biasco G: **Piroxicam increases colon tumorigenesis and promotes apoptosis in Mlh1 +/-- /Apc1638(N/+) mice.** *Anticancer Res* 2007, **27**:3807–3812.
31. Vizza CD, Rocca GD, Roma AD, Iacoboni C, Pierconti F, Venuta F, Rendina E, Schmid G, Pietropaoli P, Fedele F: **Acute hemodynamic effects of inhaled nitric oxide, dobutamine and a combination of the two in patients with mild to moderate secondary pulmonary hypertension.** *Critical Care* 2001, **5**:355–361.
32. Zaman N, Li L, Jaramillo ML, Sun Z, Tibiche C, Banville M, Collins C, Trifiro M, Paliouras M, Nantel A, O'Connor-McCourt M, Wang E: **Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets.** *Cell Rep* 2013, **5**:216–223.
33. de Hoon MJ, Imoto S, Nolan J, Miyano S: **Open source clustering software.** *Bioinformatics* 2004, **20**:1453–1454.
34. Saldanha AJ: **Java Treeview–extensible visualization of microarray data.** *Bioinformatics* 2004, **20**:3246–3248.
35. Metheny-Barlow LJ, Flynn B, van Gijssel HE, Marrogi A, Gerwin BI: **Paradoxical effects of platelet-derived growth factor-A overexpression in malignant mesothelioma. Antiproliferative effects in vitro and tumorigenic stimulation in vivo.** *Am J Respir Cell Mol Biol* 2001, **24**:694–702.
36. Mandhane SN, Shah JH, Bahekar PC, Mehetre SV, Pawar CA, Bagad AS, Chidrewar GU, Rao CT, Rajamannar T: **Characterization of anti-inflammatory properties and evidence for no sedation liability for the novel antihistamine SUN-1334H.** *Int Arch Allergy Immunol* 2010, **151**:56–69.
37. Truong-Tran AQ, Ruffin RE, Foster PS, Koskinen AM, Coyle P, Philcox JC, Rofe AM, Zalewski PD: **Altered zinc homeostasis and caspase-3 activity in murine allergic airway inflammation.** *Am J Respir Cell Mol Biol* 2002, **27**:286–296.
38. Chambers HF, Kocagoz T, Sipit T, Turner J, Hopewell PC: **Activity of amoxicillin/clavulanate in patients with tuberculosis.** *Clin Infect Dis: an official publication of the Infectious Dis Soc of America* 1998, **26**:874–877.
39. Nadler JP, Berger J, Nord JA, Cofsky R, Saxena M: **Amoxicillin-clavulanic acid for treating drug-resistant Mycobacterium tuberculosis.** *Chest* 1991, **99**:1025–1026.
40. Agarwal S: **To assess the clinical efficacy of azithromycin and capreomycin in the threatment of multi-drug resistant pulmonary tuberculosis.** *Chest* 2004, **126**:752S.
41. Watt B, Rayner A, Harris G: **Comparative activity of azithromycin against clinical isolates of mycobacteria.** *J Antimicrob Chemother* 1996, **38**:539–542.

doi:10.1186/1752-0509-8-34
**Cite this article as:** Garcia *et al.*: Network and matrix analysis of the respiratory disease interactome. *BMC Systems Biology* 2014 8:34.

Review

OPEN ACCESS

molecular systems biology

# Using single-cell genomics to understand developmental processes and cell fate decisions

Jonathan A Griffiths[1], Antonio Scialdone[2,3,4,5] & John C Marioni[1,2,6,*]

## Abstract

**High-throughput *-omics* techniques have revolutionised biology, allowing for thorough and unbiased characterisation of the molecular states of biological systems. However, cellular decision-making is inherently a unicellular process to which "bulk" -omics techniques are poorly suited, as they capture ensemble averages of cell states. Recently developed single-cell methods bridge this gap, allowing high-throughput molecular surveys of individual cells. In this review, we cover core concepts of analysis of single-cell gene expression data and highlight areas of developmental biology where single-cell techniques have made important contributions. These include understanding of cell-to-cell heterogeneity, the tracing of differentiation pathways, quantification of gene expression from specific alleles, and the future directions of cell lineage tracing and spatial gene expression analysis.**

**Keywords** cell fate; development; differentiation; single-cell RNA-seq; transcriptome

**DOI** 10.15252/msb.20178046 | Received 11 October 2017 | Revised 20 December 2017 | Accepted 19 January 2018

**Mol Syst Biol. (2018) 14: e8046**

## Introduction

High-throughput *-omics* techniques have revolutionised molecular biology, providing insight at every step of the central dogma. At the level of DNA, we now know the genome sequences for many species and how these vary between individuals of these species (The 1000 Genomes Project Consortium, 2015). Differences in gene expression between organisms, tissues and disease states have been extensively quantified by microarrays and RNA-seq (for both coding and non-coding transcripts), while mass spectrometry and other approaches have begun to yield a high-throughput overview of protein expression. Other techniques reveal how each level of the dogma affects the other: where protein binds DNA (Aparicio *et al*, 2004; Johnson *et al*, 2007), how DNA conformation affects gene expression (Belton *et al*, 2012) and which RNA molecules are being translated (Ingolia *et al*, 2009).

However, these approaches typically require as input hundreds to millions of cells, revealing only an average reading across cell populations. For developmental biology, where individual cells make decisions about their fate, these ensemble measures provide only limited information, as individual cellular measurements are lost. Nonetheless, procedures such as fluorescence-activated cell sorting enable isolation of specifically labelled cell populations. Isolation of specific cell types or subpopulations allows for meaningful bulk genomic analysis and has contributed a great deal to our understanding of developmental biology (Spitz & Furlong, 2006), albeit large numbers of input cells are required.

Recently developed single-cell *-omics* techniques (Tang *et al*, 2009; Smallwood *et al*, 2014; Buenrostro *et al*, 2015b; Heath *et al*, 2016), by contrast, are particularly apposite for developmental biology, transferring high-throughput molecular techniques onto the correct scale for understanding cellular decision-making. In particular, knowledge of the set of genes that different cells express allows characterisation of cell state, thus providing a direct read-out of how dynamic decisions are made. Transcriptional information can be supplemented with the results of other assays, such as chromatin accessibility (Buenrostro *et al*, 2015a), allowing even deeper insight into the mechanisms by which cell fate is regulated.

This review focusses on transcriptomic assays, which make up the large majority of single-cell genomic research published to date. We first summarise the processes involved in generating and analysing single-cell expression data. We then identify areas of developmental biology where these assays have provided unique insight, as well as outlining future challenges and opportunities.

## Generating single-cell transcriptomic data

Quantifying gene expression via microscopy is familiar in contemporary biology, whether using hybridisation techniques or artificially created fusion proteins. Flow cytometry scales up optical approaches to hundreds of thousands of cell measurements without

1 Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
2 EMBL-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK
3 Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, München, Germany
4 Institute of Functional Epigenetics, Helmholtz Zentrum München, München, Germany
5 Institute of Computational Biology, Helmholtz Zentrum München, München, Germany
6 Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK
*Corresponding author. Tel: +44 1223 494583; E-mail: marioni@ebi.ac.uk

compromising cellular resolution (Fulwyler, 1965). Historically, these methods have not been suitable for assaying many genes simultaneously, due to constraints imposed by fluorophore emission spectra. Nucleotide-focussed methods pushed beyond this limitation: real-time PCR (Van Gelder *et al*, 1990) can quantify hundreds of genes, with cellular throughput improved using microfluidic systems (White *et al*, 2011; Sanchez-Freire *et al*, 2012). The recent development of sequencing-by-hybridisation (described later in this review) has addressed the gene-throughput problems of optical approaches, allowing the quantification of thousands of transcripts in the same cell.

To achieve truly transcriptome-wide expression coverage, however, RNA-sequencing-based methods are best suited. Shortly after the first application of RNA-seq to bulk populations of cells (Bainbridge *et al*, 2006), the feasibility of applying RNA-seq to individual cells was demonstrated (Tang *et al*, 2009). Over the past 5 years, single-cell RNA-seq (scRNA-seq) has become the most commonly used approach for assaying single-cell gene expression profiles. There are two broad sets of methods for applying single-cell RNA-seq—"plate-based" and "droplet-based" (Fig 1).

Initially, most studies used plate-based assays, where library preparation is performed manually on cells sorted into and lysed in individual wells of a microwell plate (Jaitin *et al*, 2014; Picelli *et al*, 2014). Robotic and microfluidic systems (e.g. Fluidigm C1) have been developed to automate some of these processes.

Droplet-based methods employ microfluidics to capture individual cells in nanolitre-sized droplets, each loaded with reagents and unique labels: reverse transcription and transcript labelling take place within these small volumes. The droplet suspension is later broken down for pooling of cell libraries prior to sequencing. These methods have been developed by academic groups (Klein *et al*, 2015; Macosko *et al*, 2015) and commercially, by 10X Genomics (Zheng *et al*, 2017).

Each approach has its own advantages and disadvantages. Plate-based methods tend to provide higher-quality libraries at the cost of lower cellular throughput, processing hundreds or thousands of cells compared to the hundreds of thousands that droplet methods can process. More subtle differences also differentiate the two sets of methods. To capture rare cell types with known cell-surface markers, it is generally more efficient to flow-sort and prepare plates of single-cell libraries rather than to capture more cells using a droplet method. Additionally, current droplet methods capture gene information exclusively from the 3′ or 5′ end of each transcript, while plate approaches can generate reads from across entire transcripts; the latter allows splice-variant and allele-specific transcriptional information to be retrieved. Finally, droplet methods are more likely to produce "multiplet" cell transcriptomes, where multiple different

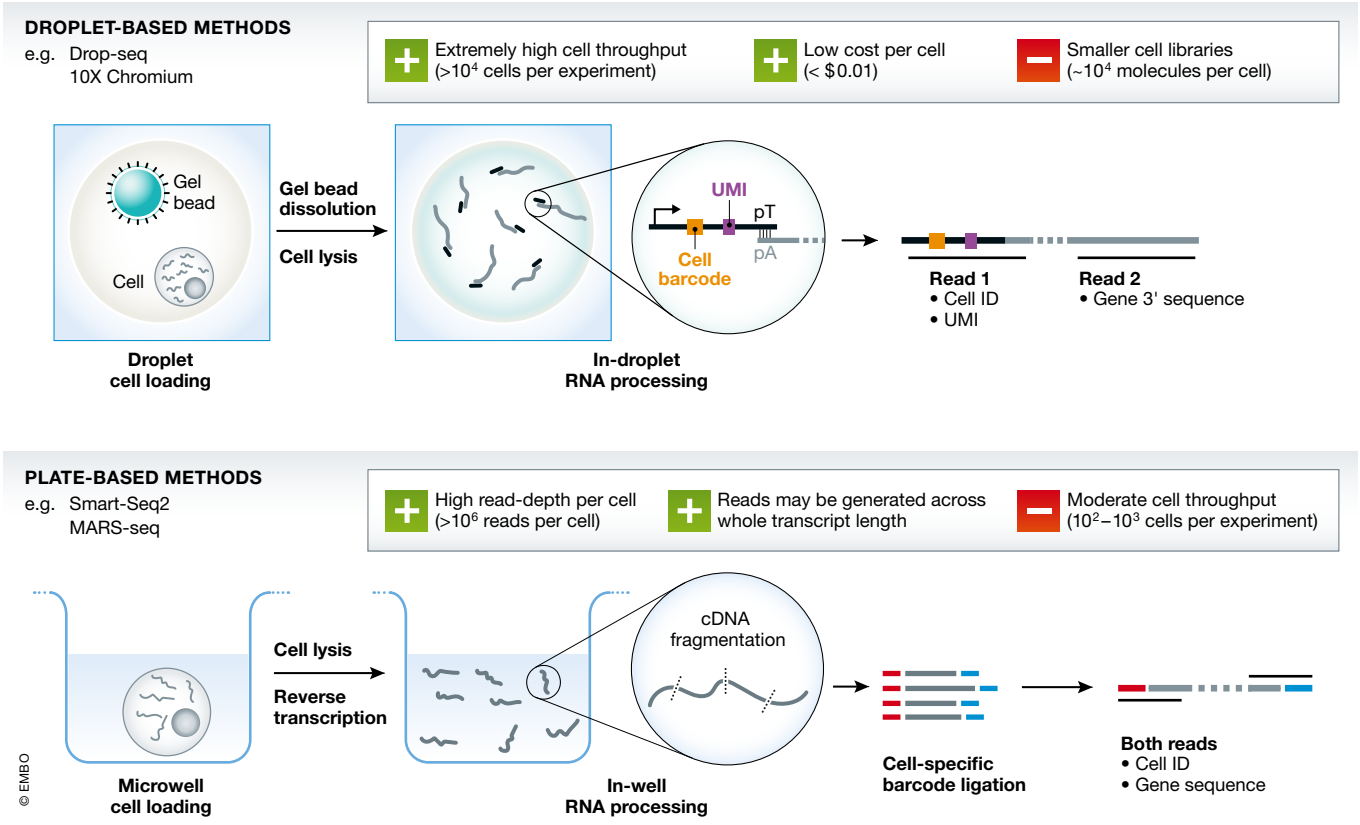


**Figure 1. Single-cell library preparation summary.**
There are two primary methods for generating single-cell transcriptomics data: plate-based and droplet-based methods, shown above. In summary, droplet-based approaches offer high cell throughput, while plate-based approaches provide higher resolution in each individual cell. Note that different implementations of these methods provide slightly different outputs and that some steps are excluded for clarity (e.g. cDNA amplification).

 

cells become labelled with the same barcode. This is largely due to the lack of user oversight (e.g. it is more difficult to identify attached pairs of cells) and the possible reuse of cell barcodes from the labelling beads. The doublet rate in droplet experiments is proportional to the number of loaded cells (Zheng *et al*, 2017).

For a researcher, the decision about which method to use is typically driven by the nature of the biological system under consideration—whether the quality or quantity of cells is important. For example, plate-based methods may be more suitable for young embryos, given the small number of cells present. For later stages of development, where there are tens of thousands of cells and a higher level of heterogeneity in each embryo, a droplet method is better suited because it is relatively easy to capture a greater number of cells, which facilitates a more complete sampling and allows unbiased capture of rare cell types. Additionally, droplet methods may be preferable for studying continuous systems, as the higher number of cells sampled can be used to better approximate the continuous process that is being studied.

Both methods exploit cell-specific DNA barcodes to allow the pooling of libraries from different cells prior to sequencing. These barcodes allow different transcriptomic reads to be assigned to individual cells. Both can also exploit unique molecular identifiers (UMIs): small, randomly generated nucleotide sequences that allow PCR duplicate reads to be collapsed, providing a more precise estimate of the actual number of RNA molecules present in a sample. For an in-depth discussion of existing approaches, see Svensson *et al* (2017).

A new method of library preparation holds much promise for combining the benefits of both plate and droplet approaches. Here, pools of cells are repeatedly split and randomly allocated to different sets of barcodes, combinatorially building up a large diversity of possible barcode labels. The method's utility has been demonstrated for DNA sequencing (Vitak *et al*, 2017), RNA-seq (Cao *et al*, 2017) and chromatin accessability assays (Cusanovich *et al*, 2015).

#### *Multi-omic assays*

The vast majority of single-cell genomics research has focussed on capturing only RNA. However, several protocols exist that allow integration of genomic, epigenomic and transcriptional information from the same cells. For example, G&T-seq (Macaulay *et al*, 2015) combines DNA sequencing with RNA-seq and is adept at identifying how copy-number changes may impact transcription. M&T-seq (Angermueller *et al*, 2016) captures DNA methylation and transcriptome data, with NMT-seq (preprint: Clark *et al*, 2018) further adding chromatin-accessibility information using a GpC methyltransferase (Kelly *et al*, 2012). While these assays offer unique advantages, they are typically experimentally challenging to run, and handle many fewer cells than scRNA-seq.

## State-of-the-art analysis techniques

#### *Quality control*

After demultiplexing barcodes and alignment of suitably trimmed reads to the appropriate reference genome, the resulting data from an scRNA-seq experiment can be represented as an integer matrix of gene expression levels, with each entry representing the number of sequenced reads (or molecules, if UMIs were used) assigned to a particular gene in a specific cell. Notably, barcode decomposition is not trivial—particularly for the random sequences of UMIs—as sequencing errors can alter their observed sequences. Methods have been developed to account for this by predicting which barcodes have arisen by error and which truly existed within the sample (Smith *et al*, 2017).

Subsequently, it is important to assess the quality of the transcriptome for each cell: incomplete cell lysis or failures during library preparation can provide output that confounds analyses. There are many parameters that quality control (QC) tests may focus on, but there are three attributes that may be easily assessed in all single-cell data sets: the total number of transcripts detected; the total number of genes found to be expressed; and the fraction of expression contributed by mitochondrial genes. Cells that show aberrant behaviour for these characteristics are typically removed from further analysis, albeit care must be taken when studying a heterogeneous population of cells as total mRNA content and other features can vary substantially (Ilicic *et al*, 2016).

Drop-out is a phenomenon observed in scRNA-seq whereby cells that are expected to express a certain gene show an observed count of zero. This is most commonly understood to be driven by stochastic failures of transcripts to be reverse-transcribed or amplified, and therefore never sequenced. This is of particular importance for data generated by droplet assays, where capture efficiency varies considerably across cells. In order to recover expression values from dropped-out genes, it is possible to impute expression values from other cells that show similar expression patterns (preprint: Dijk *et al*, 2017). However, the user should make sure that weak signals are not being artificially inflated. A researcher must also be aware of the possibility that doublets can drive technical signal in a data set, particularly for droplet-based methods. While there are no published methods for doublet detection at the time of writing, a number of papers have implemented heuristic approaches for excluding multiplet libraries. These include rejecting cells expressing sets of biologically mutually exclusive markers (e.g. *Xist* and Y chromosome genes; Ibarra-Soria *et al*, 2017), and by identifying small clusters composed of cells with large library size whose expression profiles correlate strongly with at least two other clusters in the data set (Bach *et al*, 2017).

#### *Confounding factors*

Single-cell RNA-seq experiments are sensitive to confounding factors. For example, as in any -omics experiment, systematic differences between experimental batches must be removed before the expression profiles of cells can be compared, emphasising the importance of good experimental design (Lun & Marioni, 2017). Even when controlling for these effects, true biological differences may produce signals orthogonal to the experiment's aim. In particular, cell size (as reflected by total mRNA content) often manifests itself in the number of detected genes in each cell (McDavid *et al*, 2016; Hicks *et al*, 2017), which can lead to structure in the high-dimensional expression space. Cell library size differences are controlled by the critical step of normalisation (reviewed in Vallejos *et al* (2017)), which aims to remove differences due to sequencing depth and total RNA content. The addition of precisely quantified exogenous RNA species ("spike-in" genes) to each cell's lysate allows the estimation of absolute amounts of RNA (Brennecke *et al*, 2013). However, their use is rare in droplet-based assays: spike-in

RNA will be present in every droplet, not only those containing cells. Consequentially, spike-in genes may consume a large amount of the sequencing read space and would be confounded by repeated use of the same cell barcode in multiple droplets (resulting in a variable amount of spike per barcode). Other biological factors such as cell-cycle stage can also lead to structure that can mask the signal of interest; computational strategies exist to identify and remove these effects (Buettner *et al*, 2015).

*Cell type identification*

A common first step in the analysis of scRNA-seq data is to classify cells into a number of groups. By identifying these subgroups of cells, the degree of heterogeneity within the population of interest can be assessed and comparisons can be performed, even between potentially small or rare groups of cells (e.g. primordial germ cells).

Cell-type clustering performance can be improved by using only genes that vary more between cells than would be expected by chance (Brennecke *et al*, 2013), or by using "eigengenes" that explain variability in the data (e.g. derived via principal components analysis). For additional discussion of these features, see Trapnell (2015).

*Developmental trajectories and pseudotime*

In many systems, cells display a continuous spectrum of states that is considered to represent the differentiation process. In these cases, a discrete classification of cells is not appropriate, and a researcher may prefer to use a method that summarises the continuity of cell states in the data.

Such methods are typically referred to as *pseudotime* methods, a term first introduced by the software package Monocle (Trapnell *et al*, 2014). Pseudotime describes an ordering of cells according to some characteristic in the data; this may represent developmental processes occurring over time, or the effects of continuous spatial heterogeneity in a system. Because pseudotime is an ordering of cells, it allows identification of the cell types at the beginning and end states of the trajectory, as well as those cells in intermediate stages (Fig 2). From the ordering of cells, it is possible to identify the transcriptional changes that accompany developmental processes, which can also permit the reconstruction of gene regulatory networks (Moignard *et al*, 2015). Additionally, recent developments allow detection of branching points in trajectories (Haghverdi *et al*, 2016), which serve to identify critical points of cellular decision-making. Note that caution must be exercised when applying classification and pseudotime methods, as they are guaranteed to generate output irrespective of the quality of data supplied. There is rarely any quantification of uncertainty, and results typically depend on specific parameter choices. For proper interpretation, it is important to ensure that input data are of high quality and not confounded by, for example, batch effects. Moreover, it should be stressed that scRNA-seq's static "snapshot" data possess intrinsic limitations for the study of dynamic processes, which are common throughout developmental biology (preprint: Weinreb *et al*, 2017).

## The contribution of single-cell expression data to developmental biology

In this section, we highlight examples from developmental biology where the application of single-cell gene expression assays has played a key role in providing new biological insights.

*Understanding cellular heterogeneity*

There are two ways to look at scRNA-seq data: how the expression profiles of individual cells differ from each other, and what structure in the data drives this; or how different genes behave across the population of cells and with respect to other genes' expression. In this section, we describe how cultured mouse embryonic stem cells have been used as a model for understanding the role of dynamic gene expression patterns, before discussing how expression variability observed between cells in mouse embryos defines cell fate choices in early development.

**Observing heterogeneity in cultured cells** Embryonic stem cells are a foundational tool of developmental biology research, offering a platform to investigate specific cell fate choices by signal-induced differentiation. Early work on mouse embryonic stem cells (mESCs) identified archetypal gene expression patterns across cells, highlighting bimodal and lognormally expressed genes (which were typically pluripotency regulators) as well as sporadically expressed transcripts (mostly differentiation markers; Kumar *et al*, 2014). It is difficult to address the dynamics of cellular gene expression from scRNA-seq data alone, as it captures only snapshots of cells' gene expression (preprint: Weinreb *et al*, 2017). To address this, Kumar *et al* allowed individual cells to grow into colonies over 3 days and quantified the expression levels of key pluripotency genes in individual cells of each colony. A higher level of inter-colony variance than intra-colony variance was observed, demonstrating that the initial gene expression differences that existed within the originating cells had not been overcome by gene expression pattern changes over the course of several cell cycles. The rate of change of pluripotency markers was therefore shown to be relatively slow.

Further work in mESCs focussed on identifying differences between cell culture conditions: a foetal calf serum + LIF environment promotes self-renewal in stem cells, while adding additional inhibitors ("2i") further prevents differentiation. Cells treated in each of these conditions were profiled using scRNA-seq (Kołodziejczyk *et al*, 2015). Although global levels of gene expression variability were equivalent between environments, specific functional groups of genes were more or less variable in each condition. Gene ontology terms such as "organ development" were more variably expressed in the serum condition, where differentiation is less repressed, while 2i-treated cells showed greater variability in the expression of cell-cycle genes. Whole-transcriptome comparisons additionally revealed that the different treatments produce distinct transcriptome profiles, suggesting no overlap between subpopulations of serum-treated and 2i-treated cells, as was previously thought to be the case.

**Heterogeneity *in vivo*** To form the axes that define embryonic structure, an embryo must break the initial symmetry of the zygote. The degree to which stochastic fluctuations in gene expression bias cell fate in symmetry breaking is controversial (Hadjantonakis & Arias, 2016), so application of single-cell approaches is particularly appropriate.

An analysis of mouse embryonic cells (from the zygote to 16-cell stage) explored expression heterogeneity between cells in each embryo. Cell expression profiles become increasingly diverse immediately following the first zygotic division, driven by both transcript partitioning error during mitosis and stochastic gene expression (Shi

 

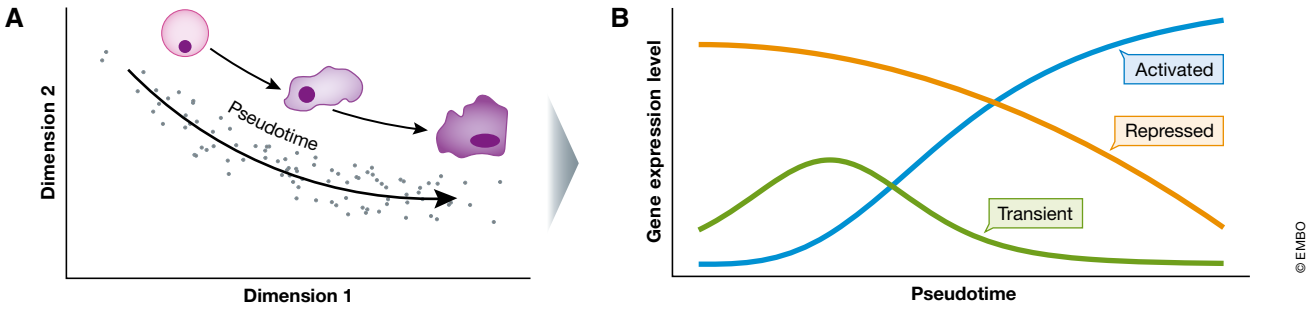


**Figure 2. Pseudotime recapitulates developmental trajectories.**
(A) By observing similarities between the expression profiles of cells, it is possible to order cells along an axis of pseudotime that recapitulates developmental processes. (B) Having established this ordering, genes that show significant changes in expression along the developmental pathway may be identified.

*et al*, 2015). Different groups of genes showed different behaviours, with some showing transiently or progressively increased variability. Few already variable genes become more variable after the 8-cell stage: it is possible that transcriptional differences between cells in an embryo begin to become fixed at this time. Finally, the authors highlighted how the ratio of two genes' expression may display particularly large amounts of heterogeneity due to asymmetric RNA distribution at mitosis, particularly if one or both of the initial transcripts is expressed at a low level. Given that many developmental decisions are specified by opposing lineage specifiers, stochastically driven heterogeneity in the expression of lineage specifiers seems a reasonable explanation for how symmetry can be broken.

Another study applied scRNA-seq to mouse embryonic cells from the 2-cell to 16-cell stage of development (Goolam *et al*, 2016), identifying highly heterogeneous expression of *Sox2* and *Oct4* (master pluripotency regulators) gene targets at the 4-cell stage. *Sox21* was identified as a gene of potential importance due to particularly heterogeneous expression across cells within an embryo and its joint regulation by *Sox2* and *Oct4*. Moreover, *Sox21* knockdown was shown to subtly bias cells towards an extraembryonic fate. Coupling the observed heterogeneity in *Sox21* expression with its fate-biasing effect, it was suggested that this heterogeneity may be responsible for pushing cells towards specific lineages during early development. However, definitively identifying the origin of these heterogeneities remains a challenge.

As development proceeds, cells become specialised into differentiated cell types through processes that are often summarised as a set of binary decisions. Single-cell approaches are especially useful in this context, because they capture cells before, during and after lineage commitment, unlike the discrete population averages of bulk sequencing (Fig 3).

One study has analysed gastrulation in the mouse, capturing epiblast cells at embryonic day (E) 6.5 along with mesodermal cells (marked using the cell-surface marker *Flk1*) at E7.0, E7.5 and E7.75 (Scialdone *et al*, 2016). Different cell types were readily identified, with pseudotime constructed over the blood precursor lineage recapitulating known gene expression changes and facilitating identification of new marker genes.

Using these data as an "atlas" of normal embryonic development allowed the authors to investigate how perturbations to developmental mechanisms affect cells' expression patterns and the cell types that they can differentiate into. A common hypothesis, driven by work in embryonic stem cell systems, states that cell fate commitment follows a path of binary choices. In the mesodermal lineage analysed here, *Tal1* is a transcription factor essential for specification of the blood lineage through an unknown mechanism of action. Under a binary decision model, *Tal1*$^{-/-}$ cells would necessarily differentiate to a cardiac lineage in the absence of this key transcription factor, as supported by *in vitro* studies (Org *et al*, 2015).

The authors generated *Tal*1 knockout embryos, applied scRNA-seq to the mesodermal lineage and computationally mapped cells from the *Tal1*$^{-/-}$ embryos on to the clusters identified from wild-type cells. This allowed proper comparison between similar cell types between the two sets of embryos while controlling for compositional changes.

Cells from the mutant embryos did not map to the blood progenitor or erythroid clusters, consistent with the absence of *Tal1*. However, cardiac markers were not upregulated in the *Tal1*$^{-/-}$ cells, unlike observations *in vitro* (Van Handel *et al*, 2012). Because the cells were not committing towards the cardiac fate, the findings called into question whether binary cell fate choices previously reported from *Tal1* knockout cells are an *in vitro* artefact, or instead occur at a later stage *in vivo* (Van Handel *et al*, 2012).

### *Developmental trajectories*

A particular advantage of single-cell methods is the ability to capture cells at various developmental stages in a single experiment. It is possible to reconstruct developmental pathways using the variety of cell states assayed using techniques motivated by the concept of pseudotime (see Fig 2 and section "State-of-the-art analysis techniques", above). Using this cell ordering, it is possible to inspect how cells change over the course of development, and which genes are critical for driving progression. This approach has been applied very widely and here we discuss some examples of how it has provided insight from different gene expression measurement technologies.

Cultured embryonic stem cells offer a versatile platform for following developmental pathways, as different morphogens can guide their development into a number of different tissues. One example is a study of the development of human definitive endoderm cells (Chu *et al*, 2016). In this study, cells were ordered along the developmental pathway, successfully reconstructing the behaviour of known markers. This ordering allowed the discovery of novel candidate regulators; for example, a driver of definitive

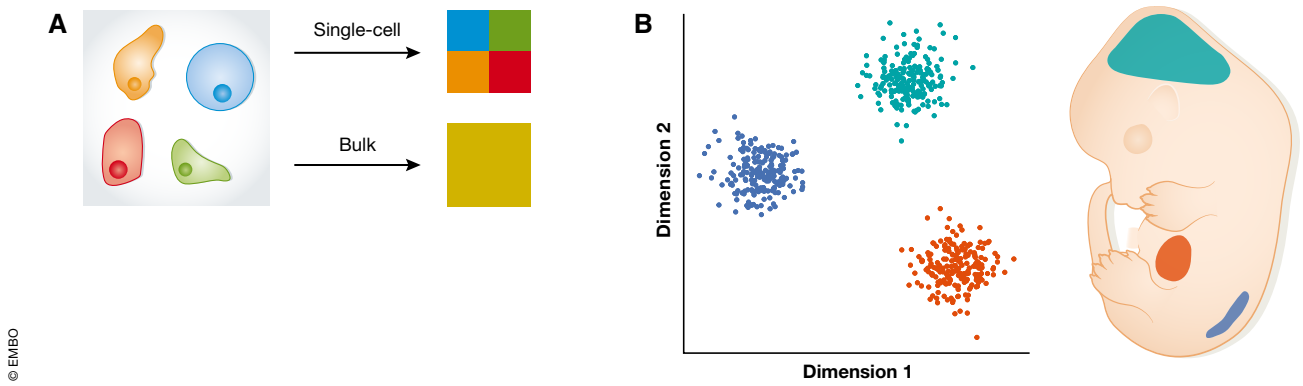


**Figure 3. scRNA-seq resolves cellular heterogeneity.**
(A) While bulk gene expression assays provide an average read-out of transcription over many cells, single-cell RNA-seq allows the assaying of gene expression in individual cells. (B) Single-cell approaches facilitate working with complex systems such as embryos, where groups of cells with radically different expression profiles can be analysed without contamination from neighbouring tissues.

endoderm differentiation (KLF8) was identified and validated by testing for changes in the fraction of differentiated cells post-KLF8 knockdown.

Trajectory inference is not limited to transcriptome data. Single-cell protein expression data (acquired by mass cytometry) have been used to identify the developmental progression of B cells in human bone marrow (Bendall *et al*, 2014). In addition to identifying a developmental progression consistent with known marker proteins, rapid changes in protein expression along pseudotime were used to identify points of cellular coordination—these correspond to the checkpoints that define progression between developmental stages. Additionally, changes in the structure of the regulatory network of STAT5 along B-cell development were noted.

Mesodermal cells from 7- to 8-day-old mouse embryos were analysed using single-cell qPCR to understand the early development of blood lineages (Moignard *et al*, 2015). Here, diffusion maps were used to identify developmental pseudotime trajectories (Haghverdi *et al*, 2015), recovering correctly the ordering of known markers. Cell states were defined via binarisation of the expression data, and a network was constructed that linked cells through changes in a single gene's expression state. This facilitated a mechanistic interpretation of the data, where predicted gene regulators were supported by motif searches and, for *Erg1*, validated in reporter systems.

Finally, it has also been shown that developmental trajectories inferred from chromatin-accessibility assays correspond closely to those inferred from expression information (preprint: Pliner *et al*, 2017).

Coupling information from different expression modalities along developmental trajectories offers potential for improved experimental design. For example, rare but important cell types could be identified using very high-throughput proteomic or flow cytometry techniques, before using identified markers to sort cells for transcriptome-wide analysis with scRNA-seq.

### *Allele-specific expression*

Biases of expression of different alleles is a difficult problem to dissect in bulk populations: Is it driven by subpopulations of cells that express only one allele at a time, or by a consistent but small bias across all cells? How much does the noisy process of transcription affect the way individual alleles are expressed?

To assay allele-specific expression (ASE) at the single-cell level, experiments must be designed carefully. Library preparation should ideally follow a protocol that allows reads to be generated across the whole length of the transcript [e.g. Smart-Seq2 (Picelli *et al*, 2014)], to maximise the number of inter-allele polymorphisms that can be assayed. Additionally, a system with the greatest possible number of allelic sequence differences is preferred. A frequently used system is the F1 hybrid mouse, that is the offspring of two different inbred lines.

The first single-cell RNA-seq study of ASE used early-stage mouse embryos (up to the blastocyst stage) and adult tissues (Deng *et al*, 2014), observing a high rate of monoallelic expression (12–25%) for even highly expressed autosomal genes. Cells in the same embryo expressed different genes monoallelically, implicating chance in deciding which alleles are expressed in individual cells. Similar behaviour has been observed in primary human fibroblasts (Borel *et al*, 2015), suggesting that stochastic monoallelic expression is common across many species and cell types.

While certain genes are known to produce predictable allele-specific expression patterns (i.e. imprinted and sex-biased genes), many genes display expression from a specific allele chosen apparently at random. This type of allele-specific expression is referred to as autosomal random monoallelic expression (aRME). aRME describes a heritable attribute of gene expression, for which single-cell analysis provides a particularly useful experimental tool. Reinius *et al* (2016) applied single-cell RNA-seq to clonal cell populations, showing that less than 1% of genes demonstrating aRME had conserved behaviour; this is in contrast to previous bulk RNA-seq work that observed aRME for over 7% of assayed genes (Gimelbrant *et al*, 2007). This single-cell work hints at the very dynamic nature of transcription (as expressed alleles change at least as fast as the cell cycle) and a lack of coordination between expression of different alleles.

Allele-specific expression is a useful tool for studying X chromosome inactivation (XCI), the process by which the dosage of X chromosome genes is controlled between sexes in mammals (Fig 4).

 

Experiments in both mice (Chen *et al*, 2016) and humans (Petropoulos *et al*, 2016) showed that the process is asynchronous across cells and that gene expression from the silenced X chromosome is gradually and uniformly reduced. One interesting difference between the two is that *Xist* is biallelically expressed during XCI in humans and monoallelically expressed in mice.

### *Lineage tracing*

Nearly all measurements of gene expression kill the cell, providing a snapshot of cellular development but losing information about a cell's lineage. As a cell's lineage represents a history of the decisions that cells have made during development, it is closely intertwined with cell fate choice. Assays have now been developed to reconstruct cell lineage alongside the capturing of expression data.

The most direct approach for identifying lineage relationships between cells using sequencing technologies lies in the genome. The pattern of mutations that individual cells acquire over time is passed on to their daughter cells upon division—a lineage tree can therefore be constructed from the distributions of these mutations across cells. However, single-cell whole-genome sequencing is expensive and presents many technical challenges (Gawad *et al*, 2016).

In particular, the relative infrequency of neutral mutations per cell cycle makes lineage determination over short timescales difficult. Given this, two techniques have been designed to implement CRISPR/Cas9 genome editing via a synthetic construct within a cell, which can accumulate mutations in a rapid manner. One of these methods provides output via imaging (Frieda *et al*, 2017) and the other via transcriptome or genome sequencing (McKenna *et al*, 2016). Both rely on the editing of a DNA-inserted barcode: endogenously expressed Cas9 (with an appropriate guide RNA) progressively and randomly alters this barcode, leaving permanent sequence changes that are inherited by daughter cells. The cell may transcribe the barcode, amplifying its presence within the cell, from where the sequence can be read out by probe labelling (Frieda *et al*, 2017), by RNA-seq (preprint: Raj *et al*, 2017) or simply by DNA sequencing (McKenna *et al*, 2016). The similarities and differences between cells' barcodes catalogue the mutational history of the assayed cells, and therefore the lineage relationship between them (Fig 5).

The sequencing approach was applied to zebrafish embryos by McKenna *et al* (2016), showing that adult organs were derived from only a small number of progenitor cells and that individual ancestral progenitor cells contributed to multiple organs and germ layers. The imaging approach has been demonstrated by a proof-of-concept study in mouse embryonic stem cells (Frieda *et al*, 2017).

Such a scarring system may be made inducible by some signal provided experimentally or naturally within a biological system. This adaptation allows for improved lineage resolution at particularly important time points.

### *Spatial transcriptomics*

Cellular decision-making is heavily influenced by a cell's environment and the signals it receives from its neighbours. However, existing scRNA-seq techniques require tissue dissociation, thereby discarding spatial information. Recovering this information has been the subject of several computational investigations.

Several groups have utilised gene expression atlases onto which cellular expression profiles can be remapped (Fig 6B). One approach used existing *in situ* hybridisation maps of spatially restricted genes as a "barcode" to which the complete expression profiles of individual cells can be matched. This was applied by two groups to reconstruct expression patterns in zebrafish embryos (Satija *et al*, 2015), and to the brain of the marine annelid

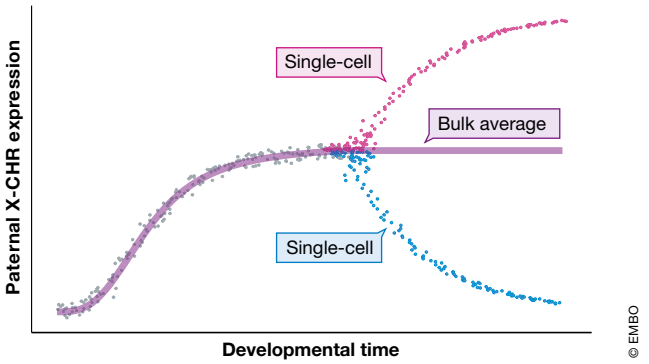


**Figure 4. Allele-specific expression at single-cell resolution.**
By exploiting single nucleotide polymorphisms in single-cell RNA-seq reads, it is possible to quantify how much individual alleles contribute to a gene's total expression. For developmental biology, this can be applied to study, for example, when monoallelic expression patterns become set during embryonic development and how they relate to fate decision, as in the case of X chromosome inactivation (Chen *et al*, 2016).

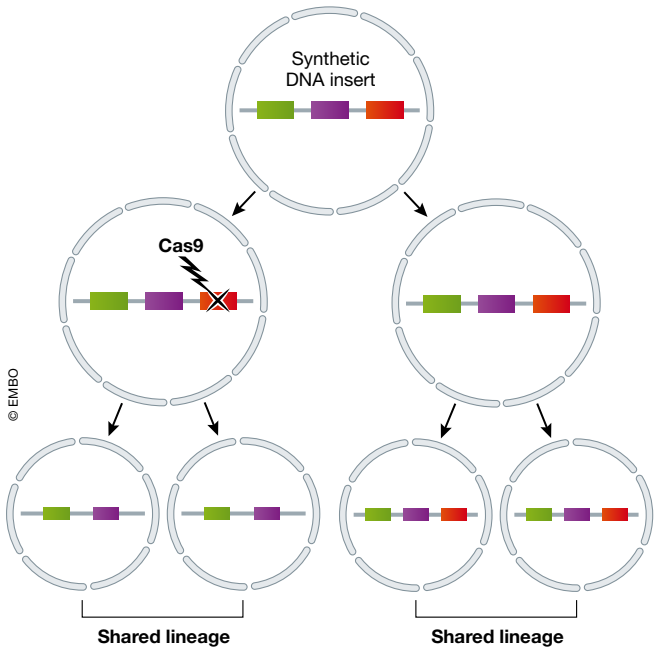


**Figure 5. Lineage tracing.**
Understanding how cells are related to each other is central to understanding how developmental processes work. However, comparison of transcriptomic profiles does not allow the reconstruction of these lineage relationships. Recent approaches used CRISPR/Cas9 to mutate a synthetic DNA construct, providing a genomic or transcriptional read-out containing cell lineage information.

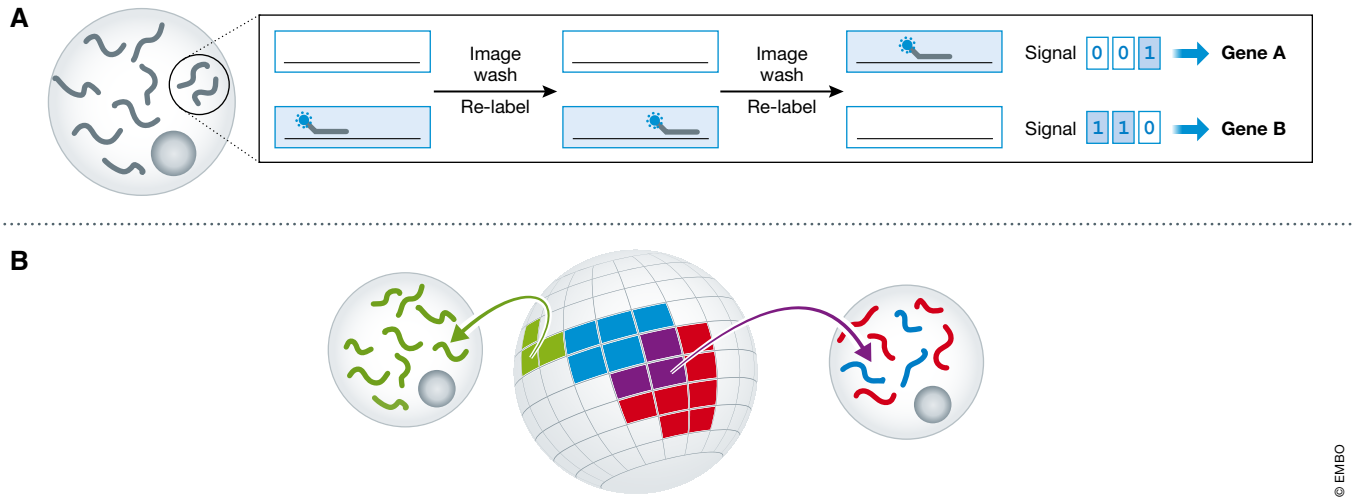


**Figure 6. Spatial gene expression data.**
(A) Most single-cell gene expression assays require dissociation of tissues, destroying locational information. New *in situ* hybridisation methods, however, offer high-throughput transcriptomic quantification captured alongside intra- and inter-cellular localisation. (B) In the absence of such techniques, others have used reference "atlases" to map back sequenced cells onto structures with known expression patterns.

*Platynereis dumerilii* (Achim *et al*, 2015). This type of approach is particularly useful where the biological structure is robust between samples, or where many high-quality reference data sets exist.

Where the system considered is known to have a robust or invariant structure, it is possible to reconstruct pseudospatial information from scRNA-seq expression data alone. Scialdone *et al* (2016) used an unsupervised approach to position cells along the anterior–posterior axis of the primitive streak during gastrulation, identifying genes expressed posteriorly (biasing cells towards, e.g., blood fate) and those expressed anteriorly (biasing cells towards, e.g., endoderm). Despite successes with post hoc reconstruction, methods that preserve spatial information experimentally will likely prove more accurate and generalisable, particularly to tissues with complex structure. Consequently, several groups have worked to develop such techniques.

The recently developed methods of merFISH (Chen *et al*, 2015) and seqFISH (Shah *et al*, 2016b) use sequencing-by-hybridisation techniques for transcriptomic quantification. In these assays, fixed cells are subject to repeated washes of fluorescently labelled DNA probes coupled with matched rounds of imaging; careful design of the probes allows individual RNA species to be identified by different sequences of fluorescence across washes, building up a unique barcode for each transcript (Fig 6A). The accuracy and resolution of these techniques have been improved by sample background clearing (Moffitt *et al*, 2016; Shah *et al*, 2016a), but the number of genes that can be reliably assayed has remained much lower than can be achieved with scRNA-seq (e.g. 249 genes in Shah *et al*, 2016b). However, recent efforts have reported the quantification of over 10,000 different transcripts in the same cells (Eng *et al*, 2017).

Locational information in these FISH assays is encoded at the individual transcript level, allowing the examination of intra-cellular effects (e.g. organelle localisation) as well as inter-cellular influences. These imaging techniques offer vast potential in developmental biology, particularly with regard to understanding signalling processes in complex systems such as embryos.

## The importance of perturbations in single-cell analyses

High-throughput -omics techniques have found their forte in hypothesis generation: because they quantify vast amounts of information, they offer considerable scope for identifying differences between samples that can form the basis of future targeted studies. However, in and of themselves, changes in gene expression levels do not provide conclusive evidence for hypotheses: Are cellular phenomena driving or being driven by the expression change? Is the expression change a function of some orthogonal effect? Have apparently significant changes arisen by chance? Follow-up experiments are therefore critical—by inducing over- or underexpression of a gene, strong signals should be detectable from further -omic assays, or through cellular behaviour alone. An appealing alternative exists for single-cell transcriptomics: natural variation in expression levels. As cells stochastically express more or less of individual genes than other cells in a population, differences in overall gene expression should propagate through gene regulatory networks, forming a large set of "micro-perturbations". However, such small differences can be readily confounded by technical artefacts (e.g. batch effects), and inference of gene regulatory networks from scRNA-seq data has been challenging to date. For instance, the SCENIC package utilises cis-regulatory information to reinforce transcriptional gene network learning (Aibar *et al*, 2017).

One possible solution to this problem is the combination of single-cell RNA-seq with targeted CRISPR screens to produce more impactful perturbations at high throughput (Adamson *et al*, 2016; Dixit *et al*, 2016; Jaitin *et al*, 2016; Datlinger *et al*, 2017). Implementations of this approach are Perturb-seq and CROP-seq. Specifically, these methods infect pools of cells with viral constructs containing

 

guide RNAs, which together with endogenously expressed Cas9 protein can target specific areas of the genome. Single-cell RNA-seq can then be applied to profile the transcriptome of each cell in addition to the specific guide RNAs that were transduced, linking a holistic view of gene expression with the knowledge of which perturbations have caused these transcriptional changes. Because of the pooled nature of such experiments and the ability to tune the multiplicity of infection, it is possible to load a large assortment of guide RNAs into a single experiment, allowing the investigation of a complex set of interacting perturbations without needing to massively increase the experiment's scale.

### The future of single-cell transcriptomics in developmental biology

Already single-cell transcriptomics has had a transformative effect in developmental biology: the ability to assay individual cells has facilitated the study of highly heterogeneous but small cell populations from the earliest stages of development. Moving forward, there are several areas where new developments will lead to even deeper insights than have already been obtained.

Perhaps most obviously, the vast majority of single-cell experiments performed to date divorce the spatial location of a cell from its transcriptional profile. Especially in early development, where spatial location affects the signals that a cell receives and thus its eventual fate, marrying these two sources of information will be extremely powerful. New approaches that increase the throughput of multiplexed RNA FISH, and other *in situ* sequencing technologies, promise to make this a reality. One important challenge will be to computationally record the location of individual cells within the embryo using a common coordinate framework—this will facilitate cross-sample comparisons. Interestingly, such a framework has already begun to be developed within the context of the Allen Brain Atlas (Sunkin *et al*, 2013) and will be an important challenge for the nascent Human Cell Atlas project (Regev *et al*, 2017). Extending this to early development will be critical, with effective work in the fly having already begun (Karaiskos *et al*, 2017).

Once generated, these spatially resolved maps of expression within the embryo will facilitate computational inference of signalling gradients, enabling both known and novel morphogen patterns to be found. This will play a key role in understanding how cells incorporate signalling information to make decisions about their downstream fate. While interesting, such new hypotheses will have to be complemented by additional experiments, for example involving the use of conditional knockout models.

Another key area where technology is driving biological discovery is the ability to assay multiple molecular layers within the same cell. Recent advances have allowed the epigenome, transcriptome and chromatin accessibility of the same cell to be profiled (preprint: Clark *et al*, 2018), therefore allowing insight into the mechanisms driving changes in gene expression. When coupled with information about a cell's location in the embryo (and the associated signalling gradients introduced above), we will begin to move towards a holistic model of cell fate choice and, indeed, of embryogenesis itself.

Underpinning all of these advances will be developments in computational methods. It is critically important that computational methods are developed in parallel with new technologies and that computational biologists work in close partnership with the experimental laboratories generating the data. Together, the potential for transforming our understanding of development is tremendous.

### Acknowledgements

J.A.G. was supported by Wellcome Trust Grant "Systematic Identification of Lineage Specification in Murine Gastrulation" (109081/Z/15/A). A.S. was supported by Wellcome Trust Grant "Tracing early mammalian lineage decisions by single cell genomics" (105031/B/14/Z). J.C.M. was supported by core funding from Cancer Research UK (award no. A17197) and EMBL.

### Conflict of interest

The authors declare that they have no conflict of interest.

## References

Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC (2015) High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol* 33: 503 – 509

Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, Villalta JE, Gilbert LA, Horlbeck MA, Hein MY, Pak RA, Gray AN, Gross CA, Dixit A, Parnas O, Regev A, Weissman JS (2016) A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* 167: 1867 – 1882.e21

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, van den Oord J, Atak ZK, Wouters J, Aerts S (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 14: 1083

Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood S, Ponting CP, Voet T, Kelsey G, Stegle O, Reik W (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* 13: 229 – 232

Aparicio O, Geisberg JV, Struhl K (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences *in vivo*. *Curr Protoc Cell Biol* Chapter 17: Unit 17.7

Bach K, Pensa S, Grzelak M, Hadfield J, Adams DJ, Marioni JC, Khaled WT (2017) Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun* 8: 2128

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, Mardis ER, Sadar MD, Siddiqui AS, Marra MA, Jones SJ (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genom* 7: 246

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58: 268 – 276

Bendall SC, Davis KL, Amir ED, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157: 714 – 725

Borel C, Ferreira PG, Santoni F, Delaneau O, Fort A, Popadin KY, Garieri M, Falconnet E, Ribaux P, Guipponi M, Padioleau I, Carninci P, Dermitzakis ET, Antonarakis SE (2015) Biased allelic expression in human primary fibroblast single cells. *Am J Hum Genet* 96: 70 – 80

Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG (2013) Accounting for

technical noise in single-cell RNA-seq experiments. *Nat Methods* 10: 1093 – 1095

Buenrostro J, Wu B, Chang H, Greenleaf W (2015a) ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 109: 21.29.1 – 21.29.9

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ (2015b) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523: 486 – 490

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* 33: 155 – 160

Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, Adey A, Waterston RH, Trapnell C, Shendure J (2017) Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357: 661 – 667

Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348: aaa6090 .

Chen G, Schell JP, Benitez JA, Petropoulos S, Yilmaz M, Reinius B, Alekseenko Z, Shi L, Hedlund E, Lanner F, Sandberg R, Deng Q (2016) Single-cell analyses of X Chromosome inactivation dynamics and pluripotency during differentiation. *Genome Res* 26: 1342 – 1354

Chu LF, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, Choi J, Kendziorski C, Stewart R, Thomson JA (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol* 17: 173

Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, Stegle O, Reik W (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Comm* 9: 781

Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J (2015) Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348: 910 – 914

Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* 14: 297 – 301

Deng Q, Ramsköld D, Reinius B, Sandberg R (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343: 193 – 196

Dijk Dv, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D (2017) MAGIC: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv* https://doi.org/10.1101/111591 [PREPRINT]

Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, Adamson B, Norman TM, Lander ES, Weissman JS, Friedman N, Regev A (2016) Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167: 1853 – 1866.e17

Eng CHL, Shah S, Thomassie J, Cai L (2017) Profiling the transcriptome with RNA SPOTs. *Nat Methods* 14: 1153 – 1155

Frieda KL, Linton JM, Hormoz S, Choi J, Chow KHK, Singer ZS, Budde MW, Elowitz MB, Cai L (2017) Synthetic recording and *in situ* readout of lineage information in single cells. *Nature* 541: 107 – 111

Fulwyler MJ (1965) Electronic separation of biological cells by volume. *Science* 150: 910 – 911

Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 17: 175 – 188

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A (2007) Widespread monoallelic expression on human autosomes. *Science* 318: 1136 – 1140

Goolam M, Scialdone A, Graham SJL, Macaulay IC, Jedrusik A, Hupalowska A, Voet T, Marioni JC, Zernicka-Goetz M (2016) Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* 165: 61 – 74

Hadjantonakis AK, Arias AM (2016) Single-cell approaches: Pandora's box of developmental mechanisms. *Dev Cell* 38: 574 – 578

Haghverdi L, Buettner F, Theis FJ (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 31: 2989 – 2998

Haghverdi L, Buettner M, Wolf FA, Buettner F, Theis FJ (2016) Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* 13: 845 – 848

Heath JR, Ribas A, Mischel PS (2016) Single-cell analysis tools for drug discovery and development. *Nat Rev Drug Discovery* 15: 204 – 216

Hicks SC, Townes FW, Teng M, Irizarry RA (2017) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* https://doi.org/10.1093/biostatistics/kxx053

Ibarra-Soria X, Jawaid W, Pijuan-Sala B, Ladopoulos V, Scialdone A, Jörg DJ, Tyser R, Calero-Nieto FJ, Mulas C, Nichols J, Vallier L, Srinivas S, Simons BD, Göttgens B, Marioni JC (2017) Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat Cell Biol* 20: 127 – 134

Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* 17: 29

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* 324: 218 – 223

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343: 776 – 779

Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, Salame TM, Tanay A, van Oudenaarden A, Amit I (2016) Dissecting immune circuits by linking CRISPR-pooled screens with single-Cell RNA-seq. *Cell* 167: 1883 – 1896.e15

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 316: 1497 – 1502

Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, Kocks C, Rajewsky N, Zinzen RP (2017) The *Drosophila* embryo at single-cell transcriptome resolution. *Science* 358: 194 – 199

Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* 22: 2497 – 2506

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161: 1187 – 1201

Kołodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, Marioni JC, Teichmann SA (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 17: 471 – 485

Kumar RM, Cahan P, Shalek AK, Satija R, Jay DaleyKeyser A, Li H, Zhang J, Pardee K, Gennert D, Trombetta JJ, Ferrante TC, Regev A, Daley GQ, Collins JJ (2014) Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516: 56 – 61

 

Lun ATL, Marioni JC (2017) Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* 18: 451 – 464

Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, Smith M, Van der Aa N, Banerjee R, Ellis PD, Quail MA, Swerdlow HP, Zernicka-Goetz M, Livesey FJ, Ponting CP, Voet T (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* 12: 519 – 522

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161: 1202 – 1214

McDavid A, Finak G, Gottardo R (2016) The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nat Biotechnol* 34: 591 – 593

McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J (2016) Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* 353: aaf7907

Moffitt JR, Hao J, Bambah-Mukku D, Lu T, Dulac C, Zhuang X (2016) High-performance multiplexed fluorescence *in situ* hybridization in culture and tissue with matrix imprinting and clearing. *Proc Natl Acad Sci USA* 113: 14456 – 14461

Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, Nishikawa SI, Piterman N, Kouskoff V, Theis FJ, Fisher J, Göttgens B (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol* 33: 269 – 276

Org T, Duan D, Ferrari R, Montel-Hagen A, Van Handel B, Kerényi MA, Sasidharan R, Rubbi L, Fujiwara Y, Pellegrini M, Orkin SH, Kurdistani SK, Mikkola HK (2015) Scl binds to primed enhancers in mesoderm to regulate hematopoietic and cardiac fate divergence. *EMBO J* 34: 759 – 777

Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, Plaza Reyes A, Linnarsson S, Sandberg R, Lanner F (2016) Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* 165: 1012 – 1026

Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* 9: 171 – 181

Pliner H, Packer J, McFaline-Figueroa J, Cusanovich D, Daza R, Srivatsan S, Qiu X, Jackson D, Minkina A, Adey A, Steemers F, Shendure J, Trapnell C (2017) Chromatin accessibility dynamics of myogenesis at single cell resolution. *bioRxiv* https://doi.org/10.1101/155473 [PREPRINT]

Raj B, Wagner DE, McKenna A, Pandey S, Klein AM, Shendure J, Gagnon JA, Schier AF (2018) Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat Biotechnol* https://doi.org/10.1038/nbt.4103

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell PJ, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N *et al* (2017) Science forum: the human cell atlas. *eLife* 6: e27041

Reinius B, Mold JE, Ramsköld D, Deng Q, Johnsson P, Michaëlsson J, Frisén J, Sandberg R (2016) Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat Genet* 48: 1430 – 1435

Sanchez-Freire V, Ebert AD, Kalisky T, Quake SR, Wu JC (2012) Microfluidic single cell real-time PCR for comparative analysis of gene expression patterns. *Nat Protoc* 7: 829 – 838

Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 33: 495 – 502

Scialdone A, Tanaka Y, Jawaid W, Moignard V, Wilson NK, Macaulay IC, Marioni JC, Göttgens B (2016) Resolving early mesoderm diversification through single-cell expression profiling. *Nature* 535: 289 – 293

Shah S, Lubeck E, Schwarzkopf M, He TF, Greenbaum A, Sohn CH, Lignell A, Choi HMT, Gradinaru V, Pierce NA, Cai L (2016a) Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing. *Development* 143: 2862 – 2867

Shah S, Lubeck E, Zhou W, Cai L (2016b) *In situ* transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* 92: 342 – 357

Shi J, Chen Q, Li X, Zheng X, Zhang Y, Qiao J, Tang F, Tao Y, Zhou Q, Duan E (2015) Dynamic transcriptional symmetry-breaking in pre-implantation mammalian embryo development revealed by single-cell RNA-seq. *Development* 142: 3468 – 3477

Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 11: 817 – 820

Smith TS, Heger A, Sudbery I (2017) UMI-tools: modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* 27: 491 – 499

Spitz F, Furlong EEM (2006) Genomics and development: taking developmental biology to new heights. *Dev Cell* 11: 451 – 457

Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, Hawrylycz M, Dang C (2013) Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* 41: D996 – D1008

Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* 14: 381 – 387

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6: 377 – 382

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526: 68 – 74

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32: 381 – 386

Trapnell C (2015) Defining cell types and states with single-cell genomics. *Genome Res* 25: 1491 – 1498

Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* 14: 565 – 571

Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc Natl Acad Sci USA* 87: 1663 – 1667

Van Handel B, Montel-Hagen A, Sasidharan R, Nakano H, Ferrari R, Boogerd CJ, Schredelseker J, Wang Y, Hunter S, Org T, Zhou J, Li X, Pellegrini M, Chen JN, Orkin SH, Kurdistani SK, Evans SM, Nakano A, Mikkola HKA (2012) Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium. *Cell* 150: 590 – 605

Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, Carbone L, Steemers FJ, Adey A (2017) Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* 14: 302 – 308

Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM (2017) Fundamental limits on dynamic inference from single cell snapshots. *bioRxiv* https://doi.org/10.1101/170118 [PREPRINT]

White AK, VanInsberghe M, Petriv OI, Hamidi M, Sikorski D, Marra MA, Piret J, Aparicio S, Hansen CL (2011) High-throughput microfluidic single-cell RT-qPCR. *Proc Natl Acad Sci USA* 108: 13999 – 14004

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R *et al* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8: 14049