Advancing haematopoietic stem and progenitor cell biology through single cell profiling

Fiona K. Hamey[1*], Sonia Nestorowa[1*], Nicola K. Wilson[1#], Berthold Göttgens[1#]

1: Department of Haematology and Wellcome Trust - MRC Cambridge Stem Cell Institute, University of Cambridge, Hills Road, Cambridge, CB2 0XY, UK

*: These authors contributed equally to this work

#: corresponding authors:

N. Wilson: E-mail: nkw22@cam.ac.uk , phone +44-1223-336822

B. Göttgens: E-mail: bg200@cam.ac.uk , phone +44-1223-336829

Abstract

Haematopoietic stem and progenitor cells (HSPCs) sit at the top of the haematopoietic hierarchy, and their fate choices need to be carefully controlled to ensure balanced production of all mature blood cell types. As cell fate decisions are made at the level of the individual cells, recent technological advances in measuring gene and protein expression in increasingly large numbers of single cells have been rapidly adopted to study both normal and pathological HSPC function. In this review we emphasise the importance of combining the correct computational models with single-cell experimental techniques, and illustrate how such integrated approaches have been used to resolve heterogeneities in populations, reconstruct lineage differentiation, identify regulatory relationships and link molecular profiling to cellular function.

Keywords

Single-cell, Haematopoietic stem/progenitor cells, computational models, heterogeneity, fate choice, lineage differentiation

Introduction

Haematopoietic stem/progenitor cells (HSPCs) lie at the apex of the haematopoietic tree, characterised by their ability to differentiate and give rise to all the mature blood cell types of the haematopoietic system. The HSPC compartment includes functionally distinct cells, defined by their differing abilities to self-renew and contribute to different haematopoietic lineages. At the top of the hierarchy sit the long-term haematopoietic stem cells (LT-HSCs), which can self-renew and are capable of reconstituting the whole blood system upon serial transplantation into lethally irradiated recipients. As LT-HSCs differentiate, they give rise to more specialised cell types, lose the ability to self-renew and become restricted in terms of lineage potential. At steady state, HSPC fate choices are balanced to ensure maintenance of the haematopoietic system, while system-wide dysregulation can lead to serious blood disorders such as leukaemia [1]. ~~Accordingly, extensive research has been carried out with the aim of characterising HSPC populations and understanding their biology in both normal and disease-state haematopoiesis.~~

~~The haematopoietic system benefits from being well-defined, which allows for the prospective isolation of cells with defined differentiation potential based on specific cell surface marker combinations.~~ While functional properties of HSPC populations have been assessed both at the population and single-cell level, expression profiling has historically been limited to measurements representing population averages [2]. Gene expression analysis was either restricted to measuring a handful of genes in single cells, with techniques such as fluorescence in situ hybridisation, or generating population-average (bulk) data by quantitative real-time PCR (qRT-PCR), microarray and RNA sequencing (RNA-seq). Whilst population-average data is undeniably informative, it represents an average state of gene expression, which assumes a population is homogeneous and so may fail to capture important

information on the heterogeneity of decision-making processes in individual cells. Lineage tracing studies in several adult stem cell systems have shown that equilibrium at the population level is achieved by what may be stochastic decisions of single cells [3,4]. Bulk measurements lack the resolution to uncover differences within a population that could be influencing fate decisions of individual cells. Recent studies have therefore explored single-cell technologies to profile cells from the HSPC compartment, highlighting the variation within defined cell types ~~and expanding the understanding of HSPC function in haematopoiesis~~.

The growing interest in studying single cells is accompanied by rapid technological innovation. An essential first step in many of these techniques is the isolation of individual cells. Many approaches take advantage of fluorescence-activated cell sorting (FACS), where cells are stained with fluorophore-conjugated cell surface marker antibodies and subsequently separated based on multiple parameters, including size, granularity, and fluorescent properties linked to surface marker expression [5]. The cells are then available for various applications, such as functional or gene expression analysis. Collected cells can often only be used for one type of experiment; therefore, FACS isolates cells that are a representative snapshot of a particular cell population at the point of collection, which, if collected at a single-cell level, is likely to reveal heterogeneity in many factors. Index sorting is an important advance in FACS, which collects data for all parameters measured, including well position, for each single cell sorted into 96- or 384-well plates, thus making an index of all cells on the plate. As such, it is possible to obtain the complete FACS phenotype of every cell for retrospective review [6,7]. These data can then be paired with other methods of analysis, such as gene expression analyses, to compare how population subsets vary in their gene and

surface marker expression [8]. Importantly, FACS-based index sorting is widely applicable and can be used to understand the characteristics of many cellular systems [7,9].

Another technique for studying single cells is mass cytometry (Cytometry by Time of Flight/CyTOF). In contrast to traditional flow cytometry, which labels antibodies with fluorochromes, the antibodies used in mass cytometry are labelled with transition element isotopes and quantified by concentrations of metal-tagged antibody [10,11]. Although FACS can now measure up to 30 parameters [12], CyTOF can measure in the region of 40 or more parameters simultaneously, allowing for in-depth study of cell phenotype. This technology can be applied to study signalling states of single cells in a variety of experimental conditions, limited only by the antibodies with which the cells are tagged [11] ~~which can be antibodies to cell surface markers as well as intracellular signalling proteins~~. However, unlike with FACS-based index sorting, after mass cytometry the cells are not available for further gene expression and functional analysis, but the results may provide insights for designing FACS sorting strategies for further analysis [11].

To study gene expression at the single-cell level, techniques such as qRT-PCR or RNA-seq can be applied. Fluidigm Biomark™ is a dynamic array integrated microfluidics circuit that enables the study of gene expression for up to 96 selected genes in 96 cells. As the genes of interest are chosen by the investigator this technique lends itself to looking at specific questions, targets or systems [13]. High-throughput qPCR analysis of thousands of single cells is possible using bioinformatic methods to discover trends in gene expression for the selected genes. In contrast, single-cell RNA-seq (scRNA-seq) offers a transcriptome-wide approach for measuring gene expression [14,15]. It can be used to profile gene expression in individual cells within a population of interest, and provide insight into the regulatory

programs governing these populations. Similarly, massively parallel single-cell RNA sequencing (MARS-seq) is an automated method of RNA sequencing, ideally designed to process thousands of multiplexed cells which are barcoded at multiple stages (at the molecular, cellular and plate level) [16]. An advantage of this method is that it allows the processing of thousands of cells and enables characterisation of multiple heterogeneous cell populations within a single data set, creating both an in-depth and broad picture of variability and heterogeneity.

As well as quantifying gene and protein levels in single cells, techniques exist for assessing the functional properties of individual cells. One such technology is single-cell barcoding. Barcoding involves tagging individual cells with unique barcodes, which are semi-random, noncoding stretches of DNA [17]. A barcode library is created and cells are labelled with these barcodes, commonly by retroviral transductions allowing the cells to be tracked in vivo [18]. Overall, multiple techniques currently exist to isolate and study single cells for gene expression and functional analysis, opening up many avenues for further characterising haematopoiesis. This review discusses several ways in which single-cell profiling of haematopoietic cells has enhanced our understanding of HSPC biology. Population heterogeneity, lineage differentiation, transcriptional regulatory relationships and the link between molecular profiles and cellular function will be explored. These topics highlight the diverse application of single-cell technologies and the variety of information gained from coupling such techniques with computational approaches and functional assays.

Resolving heterogeneous populations

Haematopoietic populations are known to be heterogeneous, both in terms of functionality in transplantation or colony forming assays, and in their molecular profiles obtained by single-

cell technologies. Such heterogeneities have been observed in many HSPC populations such as common myeloid progenitors (CMPs) [18] and present challenges when investigating the properties of a population. Haematopoietic cell types isolated based on surface marker expression are up to 98% pure for the surface markers, but may actually contain functionally different cells [19]. These cells and subpopulations would not be picked up by bulk analysis, which assumes homogeneity. Single-cell profiling provides a powerful technique for resolving heterogeneity without relying on the isolation of functionally pure populations (Figure 1A).

Dimensionality reduction

Single-cell profiling using techniques such as scRNA-seq produces complex multi-dimensional data sets. Large numbers of individual cells can be profiled, giving tens to thousands of gene expression measurements per cell. Due to the high number of dimensions, direct interpretation of such data is not straightforward. These challenges are not unique to single-cell data: a concept from machine-learning and statistics, known as dimensionality reduction, has been widely applied to population expression data to discover differences between samples of cells (Figure 1B). Dimensionality reduction methods enable complex high-dimensional data to be visualised in a low-dimensional space, most frequently two or three dimensions, allowing differences between groups of cells to be observed. Widely-used linear dimensionality reduction methods include principal component analysis (PCA) and independent component analysis (ICA), which are popular methods for interpreting high-dimensional single-cell data [20,21].

More recently, dimensionality reduction techniques such as t-distributed stochastic neighbour embedding (t-SNE) [22] and diffusion maps [23] have been applied to single-cell data as

these non-linear methods are able to uncover more complex relationships in the data. t-SNE finds a low-dimensional embedding of the data that aims to conserve the distribution of distances in the high-dimensional space so that cells with similar expression profiles are nearby on the dimensionality reduction plot. viSNE is an algorithm, based on t-SNE, specifically developed for visualising single-cell expression data. Amir et al. applied viSNE to single cell mass cytometry data to explore heterogeneities within leukaemic bone marrow, and showed phenotypic differences between wild-type and cancerous bone marrow, as well as the ability to detect the rare minimal residual disease phenotype [24]. Diffusion maps use a different approach, considering lengths of diffusion-like random walks between cells in the high-dimensional space and from these distances determine a projection of the cells. Some of these dimensionality-reduction methods have been specifically adapted for use with single-cell data [25]. A recent study by Moignard et al. describes the use of diffusion maps to visualise progression of cells during early blood development based on single-cell gene expression measurements, in which diffusion maps were able to successfully separate cell populations from early and late time points within the data [26]. They further illustrated a progression through differentiation, where heterogeneity within and between populations was visible at each time point.

Clustering single-cell profiles

Dimensionality reduction not only allows visualisation of cellular heterogeneity, but is also often useful in assigning cells to groups to query differences between populations. Assigning cells to subpopulations within a sample using prior knowledge is not always possible, or even desirable. Instead, clustering methods can be used to separate cells into populations in an unbiased way, based only on information such as expression profiles (Figure 1C). The expression of specific genes within each cluster can then be used to identify cell types or find

novel marker genes for populations. Well-established methods such as hierarchical clustering have been extensively applied to single-cell data to identify subgroups of cells within samples [8,13,27]. These methods calculate distance measurements between cell expression profiles and assign similar cells, i.e. those with small distances, to clusters. Recently, additional clustering methods have been developed specifically for partitioning single-cell expression profiles. Jaitin et al. apply a probabilistic mixture model to scRNA-seq data in order to cluster cells into groups with distinct gene expression profiles [16]. An alternative method which has been applied to single-cell protein expression data is PhenoGraph, an unbiased graph-based clustering algorithm that searches for highly connected groups of nodes to identify clusters of cell types [28]. An advantage of using a graph-based approach is that it is easily scaled for use with large numbers of cells and high-dimensional data.

Levine et al. [28] investigated intra-tumour heterogeneity in acute myeloid leukaemia (AML) by obtaining single-cell mass cytometry data for 16 surface markers and 14 antibodies against intracellular protein phosphorylation, in order to measure protein expression and activation in samples from AML patients and healthy bone marrow donors. They then applied PhenoGraph to reveal differences in the distribution of cell types in the bone marrow between AML and healthy bone marrow samples. When leukaemic and healthy cells were mapped together, signalling and surface phenotypes were tightly coupled in healthy cells, with CD34 levels distinguishing between primitive and mature phenotypes. In contrast, each leukaemia had distinct surface phenotypes, and signalling and surface phenotypes were decoupled, demonstrating that surface phenotypes alone are not enough to characterise a leukaemic population. They further showed that leukaemic cell diversity is influenced by normal myeloid development even after malignant transformation, and that in leukaemia the

signalling phenotype often revealed a different degree of maturation than predicted by surface phenotype.

Single-cell barcoding and transplantations

As discussed above, single-cell gene and protein expression profiles can describe heterogeneity within in a population. However, these measurements alone cannot demonstrate functional differences between cells. Genetic barcoding of single cells followed by transplantation into lethally irradiated mice can help resolve functional heterogeneities in haematopoietic populations. After transplantation, the in vivo contribution of individual cells to different lineages can be assessed by sequencing to discover barcode identity (Figure 1D). Perié et al. have recently used this technique to investigate the clonal output of different stem and progenitor cells at the single-cell level [18]. In their study, they looked at whether the common myeloid and lymphoid progenitor (CMP and CLP, respectively) divide is the first step of lineage commitment, or whether lineage commitment actually occurs earlier, by genetically barcoding CMPs and tracking in vivo cell fates. CMPs produced highly biased myeloid or erythroid output after 14 days, suggesting that cells are already at an early commitment stage at this point. The authors also transplanted barcoded HSCs and multipotent progenitors (MPPs) into mice to assess their ability to produce myeloid and erythroid cells. They found that production of both cell fates resulted mainly from HSCs and only 20% of MPPs. The remaining MPPs were restricted to a single fate, showing that subdivision in cell fate is already detectable at the MPP stage. Furthermore, the finding that cells are already myeloid or erythroid biased at the CMP stage is supported by published in vivo and in vitro studies [29,30].

In order to interrogate CMP heterogeneity, Paul et al. used MARS-seq to measure gene expression in 2730 myeloid progenitors [29]. The authors adapted the previously described method of Jaitlin et al. to cluster these progenitor cells into groups with distinct expression profiles. By doing so, they identified clusters with broad erythroid or myeloid progenitor characteristics, with no evidence of individual cells expressing sets of genes suggesting priming towards multiple lineages. CMP clusters could be further sub-sorted and both MARS-seq and transplantation assays showed CMPs are not really a cluster of heterogeneous cells with undetermined cell fate, but rather a group of subpopulations primed for one of seven myeloid fates. These findings, obtained by MARS-seq, present a similar conclusion to that of Perié et al., found by genetic barcoding, suggest an explanation for the source of heterogeneity, and question the continued usefulness of the classically defined CMP.

Reconstructing lineage differentiation

During haematopoiesis, cells become increasingly specialised as they commit to one of the several fates corresponding to mature cell types. Isolating and collecting populations at different stages of differentiation, followed by population profiling using techniques such as RNA-seq, goes some way to describing cell differentiation, but is limited by time resolution and must assume that cells are synchronised through the differentiation process. Single-cell profiling demonstrates that large variation exists within an isolated population thought to be homogenous from bulk studies [8,13,20,26]. Single-cell technologies also allow unbiased in vivo profiling of tissues, such as bone marrow or tumour tissues, which contain cells at multiple stages of differentiation. In silico lineage reconstruction takes advantage of the ability to resolve heterogeneities within populations and uses computational methods to infer lineage differentiation based on single-cell data.

Constructing a lineage tree from single-cell data

Even in a system as well characterised as haematopoiesis, the exact structure of the haematopoietic tree remains under debate [27,31]. Single-cell profiling has been used as a tool to address this question based on the assumption that cells at an equivalent stage of differentiation have similar expression profiles. Using single-cell expression profiling, individual differentiating cells can be clustered into groups and the closest groups connected into a structure representing a lineage hierarchy (Figure 2A). The spanning-tree progression analysis of density-normalized events (SPADE) algorithm uses this approach to build lineage hierarchies from flow and mass cytometry data collected from bone marrow cells [32]. This method first calculates a density-dependent sample of the data to ensure that rare populations are not obscured. Cells in this sample are then clustered based on their expression profiles, and the most similar clusters are linked into a tree aiming to represent the lineage hierarchy. Strengths of SPADE are that it includes rare cell populations in the hierarchy and does not require prior information to infer the lineage structure. However, different random density-dependent samples obtained by SPADE lead to different clusters and can therefore produce alternative tree structures, leading to limitations with the stability of this approach.

SPADE was used by Guo et al. to construct a lineage tree resembling the haematopoietic differentiation hierarchy, to investigate the much debated question about the starting point of lineage commitment for HSCs [27]. Recent studies show this point likely occurs before the CMP/CLP split, in contrast to what was initially thought [29,30]. Guo et al. also challenge the view that commitment occurs at the CMP stage, collecting more than 1500 single cells for qPCR and quantifying 280 commonly used surface markers for all cells. By performing unsupervised hierarchical clustering, they showed that gene expression clusters are closely correlated with cell type clusters, and generated individual expression maps for each gene

cluster, which helped to resolve heterogeneity in the loosely defined populations. Progenitors were shown to have high levels of heterogeneity, suggesting a continuum of transcriptional states. The lineage tree constructed with SPADE showed CMPs were found in both megakaryocyte-erythrocyte (MegE) and lymphomyeloid lineages; computational analysis and in vitro validation identified a new surface marker (CD55) to be a valid marker for MegE and lymphomyeloid differentiation at the CMP and MPP stages. MegEs were closely connected to the long-term HSC branch, and through in vitro tracing experiments the authors also showed that megakaryocytic colonies emerge first in HSC cultures, indicating a very early lineage bias and supporting the in silico findings. As such, Guo et al. demonstrate the usefulness of single-cell qPCR and SPADE at a single-cell level to resolve heterogeneity as well as to build on the haematopoietic hierarchy; furthermore, they applied the model to leukaemic stem cells to compare differentiation in healthy and leukaemic cells.

In a more recent study, Spitzer et al. describe their computational method, Scaffold, to arrange immune cells profiled by single-cell mass cytometry into a ‘reference map’ of the murine immune system [33]. This approach involves an initial clustering step followed by construction of a graph using the clustered cells. Scaffold uses a method called force-directed graphs to find a visualisation based on the similarity between cell types. Here similar cell clusters are pulled close together in the force-directed graph, whereas dissimilar clusters lack such a strong attracting force and lie further apart. The resulting graph links cells in a structure that represents the immune system hierarchy. The authors constructed Scaffold maps for cells from different samples, which enabled comparison of immune system organisation in different tissues, genetic backgrounds and species. Circadian rhythm was seen to affect the distributions of the immune cells, with some immune cell populations fluctuating depending on the time of day. The Scaffold method also allows new data to be projected onto

the existing map, thereby providing a reference for future studies and allowing integration of multiple datasets from various tissues, disease states or even different laboratories.

Another recent study questioned the current thinking of the branching point between monocyte-macrophage potential from granulocyte-macrophage potential [34]. By analysing myeloid-restricted pre-granulocyte-macrophage-progenitors (pre-GMs) by scRNA-seq, Drissen et al. suggest that bifurcation is observable at an earlier point than previously recognised. They showed that cells expressing Gata1 display megakaryocyte, erythrocyte, eosinophil and mast cell potential, whereas cells not expressing Gata1 exhibit lymphocyte, neutrophil and monocyte potential. These results demonstrate that the expression of Gata1 could be an early indicator of lineage potential in pre-GMs. Researchers also investigated how lineage bias changes in ageing HSCs and the effect it has on adaptive immunity in older patients. By interrogating HSC transcriptomes using scRNA-seq and Gene Set Enrichment Analysis, Grover et al. confirmed an age-dependent increase in megakaryocyte/platelet programming, showing that an increased molecular and functional platelet bias is a key characteristic of HSC ageing, and found a previously unrecognised subset of aged HSCs with platelet-restricted output [35].

Ordering cells in pseudotime

An exciting extension of inferring differentiation hierarchies is to order single-cell profiles by progress through differentiation. Assuming that gene and protein expression change continuously as cells differentiate, and that a sample contains cells spread at a sufficient density through differentiation, it was hypothesised that single-cell expression profiles could be used to arrange cells in ‘pseudotime’, where the position of a cell in pseudotime corresponds to its progress through differentiation (Figure 2B). Based on these simple

assumptions different algorithms have been designed to solve this computational ordering problem. Trapnell et al. describe the algorithm Monocle, which first performs a dimensionality reduction of the data before constructing a graph on this lower-dimensional representation and finding the minimum spanning tree [21]. Cells are then ordered in pseudotime based on their position in the minimum spanning tree, allowing changes in gene expression pattern throughout pseudotime to be investigated (Figure 2B). Another algorithm, Wanderlust, was applied to single-cell mass cytometry data to capture B-cell development in human bone marrow [36]. Wanderlust constructs a pseudotime ordering by first considering a k-nearest-neighbour graph on the single-cell expression data. The ordering of cells is based on the length of paths through this graph originating from a user-defined starting cell. This algorithm can cope with very large numbers of cells, and uses subsampling methods to obtain stable orderings, avoiding the possibility of 'short circuits' through the data. Bendall et al. used mass cytometry to study 44 parameters in B-cell lymphopoiesis, collecting enough cells to encompass B-cell development with the aim of inferring a developmental trajectory [36]. Using Wanderlust, the authors confirmed that all the landmarks of B-cell lymphopoiesis were correctly ordered. The trajectory also allowed for identification of early populations and ordered them across development. Deoxynucleotidyl transferase, an enzyme involved in IgH locus rearrangement, and CD24 increased prior to B-cell surface marker expression, suggesting their role as novel identifiers of early B-cell populations in the bone marrow. By studying the rearrangement of the IgH locus, Wanderlust was used to identify new early B-cell populations and order them developmentally. Regulatory signalling change was observed in association with coordinated marker expression, and trajectory analysis revealed that expression and signalling changes correspond to developmental checkpoints, involved in IgH locus rearrangement and receptor cross-linking responsiveness. Checkpoints were challenged by pharmacological inhibition, which caused restricted B-cell development. Therefore, by

combining CyTOF with the Wanderlust algorithm, Bendall et al. confirmed the B-cell development hierarchy, identified early B-cell populations, and validated important developmental checkpoints, demonstrating the value of single-cell technologies for reconstructing and validating the haematopoietic differentiation hierarchy.

Regulatory relationships

To better understand how multipotent cells choose between different fates during haematopoietic differentiation, it will be important to define the underlying regulatory programs [37,38]. Transcriptional regulatory networks, for example, are composed of transcription factor proteins and the cis-regulatory modules that they bind to. Identifying these networks can provide information on how regulatory programs control cell fate decisions. However, network reconstruction directly from experimental evidence has so far been limited to the simplest organisms due to the sheer number of possible regulations and complex network structures. Instead, many studies have focused on the more feasible approach of inferring regulatory networks from gene expression data, which requires data to be collected from multiple experimental perturbations or conditions. Network inference from population expression data is therefore constrained by both small sample size and masked heterogeneity within cell types. Single-cell data represent a powerful alternative for identifying new regulatory relationships, as each cell presents an observation with its own expression levels meaning that the number of samples is vastly increased.

Identifying regulatory relationships

Measuring single-cell gene expression provides potentially thousands of observations of gene values in individual cells. Such large sample sizes can be used to identify potential regulatory relationships by considering correlations between genes (Figure 3A). Setting a threshold on

correlation strength can then identify putative networks consisting of links between genes with high correlations. Several studies have calculated correlation between genes from single-cell gene expression data and have identified experimentally validated regulatory relationships between highly correlating genes [13,20].

Modelling regulatory networks

Although key players in cell fate decisions can be identified, decision-making in cells is in fact governed by complex networks of transcription factors with the possibility of combinatorial interactions between elements of a network. A regulatory relationship between two genes cannot necessarily be considered in isolation, but might depend on the presence or absence of additional transcription factors. Logical relations can be abstracted as Boolean functions where expression of a gene is either ‘on’ or ‘off’, forming part of a Boolean network (Figure 3B). With this type of abstraction it becomes possible to model and simulate regulatory networks. Single-cell expression data offers exciting potential in this area, as gene expression levels can be converted to binary data for each cell providing a large number of possible Boolean states (Figure 3C). It has been demonstrated that single-cell gene expression data can be used to computationally infer these Boolean models in systems including embryonic blood development [26] and embryonic stem cells [39]. A drawback of Boolean models is the abstraction of gene expression levels to binary on/off states, which discounts any possible influences of quantitative expression differences. A recent HSPC regulatory network constructed from extensive and quantitative experimental evidence utilised a Bayesian approach and was demonstrated to be a useful tool for modelling transcription factor perturbation in single cells from a myeloid progenitor model cell line [40].

Networks in HSPCs

Interrogating single-cell data has been very useful for identifying previously unrecognised regulatory networks as well as identifying important factors involved in lineage commitment. Pina et al. used single-cell qRT-PCR to investigate self-renewing cells and erythroid- or myeloid-committed progenitors [20]. Gene expression analysis showed lineage commitment to be associated with negative gene regulatory relationships, providing a possible insight into differences in self-renewal and commitment. Ddit3 was identified as a previously unrecognised key player in lineage commitment, positively associated with Gata2 in self-renewal and committed cells, and negatively associated with Cebpa, which is important for neutrophil commitment. Ddit3 knockdown resulted in loss of erythroid function and a switch to myelo-monocytic potential, whereas enforced expression in granulocyte-monocyte progenitors (GMPs) resulted in cells with increased self-renewal properties and reduced myeloid potential. Analysis of both wild-type and Ddit3-overexpressing GMPs by PCA confirmed the experimental results, where Ddit3 positively regulated erythroid fates while negatively regulating myeloid fates. The authors also found the global transcriptional network of GMPs was altered by overexpressing Ddit3, as seen by increased connectivity with Gata2 and stabilisation of primitive MegE precursors preventing myeloid fate. This single-cell study suggested that conflicting lineage-potential programs exist at the point of cell commitment, and identified a key relationship between Gata2 and Ddit3.

Moignard et al. also used qRT-PCR to analyse 18 transcription factors known to play a role in haematopoiesis [13]. The study investigated long-term HSCs as well as lymphoid-myeloid progenitor populations, Pre-MegEs, GMPs and CMPs to identify the relevant networks in HSC to progenitor differentiation. The authors used hierarchical clustering and correlation analysis to look at relationships between transcription factors in all cell populations as well as each population individually, and showed that for individual populations there was a

reduction in negative correlations, suggesting lack of repression may be important for cell fate transitions. The correlation analysis also revealed two new regulatory links: Gata2-Gfi1b and Gata2-Gfi1, highlighting a previously unrecognised regulatory triad between Gata2, Gfi1b and Gfi1, where mutual inhibition between Gfi1b and Gfi1 is regulated by Gata2. The study therefore demonstrates the utility of single-cell network interrogation in finding regulatory networks unidentified in bulk-cell studies, increasing our understanding of cell fate decisions and HSC differentiation.

Linking molecular profiling to cellular function

The information we can conclusively gain about HSCs from both single cell and bulk technologies is limited by the fact that the isolated populations are 40-50% pure in function at best, as seen in single-cell transplantation experiments [41–45]. As such, unless a more complex panel of cell surface markers is used, it is not possible to know whether the cells being investigated are true functional representatives of the target populations, or cells of another identity, expressing the same surface markers but fulfilling a different role. Furthermore, gene expression analyses are retrospective in nature, meaning the cells analysed are no longer available for functional studies; these limitations mean that linking conclusions made about gene expression with functional information is a difficulty in defining HSPC characteristics.

A pipeline to connect gene expression and functional analyses

Transplantation experiments are useful for assessing the ability of a single-cell to repopulate the HSC compartment, whereas single-cell gene expression analysis provides information of cell transcriptional states; the challenge lies in bringing together these two separate sets of information to determine the genetic profiles of cells with specific biological functions. Index

sorting is an important tool for bridging this gap by providing data on the cell surface markers of sorted cells, which can be directly compared between cells used for gene expression and functional assays. The information together can be used, for example, to inform new marker identification and for refining sorting strategies to improve population purity, or to isolate new populations and study subpopulations identified from gene expression analysis.

To link molecular profiling and cellular function, a single-cell processing pipeline can be imagined, from cell sorting to experimental design (Figure 4). By implementing this pipeline, Wilson et al. defined a purification strategy for a HSC subpopulation with a homogenous molecular profile, termed the molecular overlapping population (MolO). To start with, they used four of the most refined HSC purification strategies [43,46–48] to isolate murine HSCs by FACS, collecting index sorting data for all cells. They obtained single-cell gene expression data by qPCR for 48 preselected genes important for HSC biology [8]. Common functional HSCs, or MolOs, were identified bioinformatically as a population consisting of cells from each sorting strategy, weighted based on durable self-renewal and repopulation probabilities as published in the literature [8,43,46–48]. To make a transcriptome-wide investigation, 96 cells were analysed by scRNA-seq and ranked on index data according to the likelihood that their gene expression profiles corresponded to a functional HSC. Index sorting data showed MolO cells were enriched for $CD150^{+}CD48^{-}$ (SLAM) $Sca1^{hi}$ expression; from this finding, a sorting strategy was devised to specifically enrich for MolO cells. Animals with at least 1% donor white blood cells at 16 or 24 weeks were considered to have robust multilineage repopulation with long-term reconstituting HSCs. Multilineage repopulation was seen in mouse transplants and in vitro single-cell culture showed SLAM $Sca1^{hi}$ cells proliferate and differentiate less than SLAM $Sca1^{lo}$ cells. Interrogation of the

complementary single cell transplant and scRNA-Seq data allowed further refinement of sorting gates to isolate HSCs with 67% functional purity. Improving the functional purity of a sorted population will improve our ability to perform cellular function assays, as well as identify new molecular regulators of stem cell function. The pipeline presented in this paper is not only applicable to the haematopoietic system, but can also be used in other cell systems. A similar pipeline was applied by Paul et al. as previously described [29] to capture myeloid progenitors in a broad Lin^- $Sca1^-$ $cKit^+$ gate and use index sorting linked with gene expression analysis to retrospectively identify the populations for further study. Although transplantation studies and other functional assays are useful and indeed necessary for validating molecular profiles of HSPCs, these studies demonstrate how index sorting bridges an important gap in ensuring that the gene expression and functional assays are looking at similar cells, as indicated by forward/side scatter and surface marker expression.

Conclusions and perspectives

Haematopoiesis researchers have been at the forefront of applying single-cell technologies including FACS, CyTOF, qRT-PCR, RNA-seq, MARS-seq and genetic barcoding [6,8,13,18,29,36]. When combined with an array of computational methods, these methods can be used to better understand the function, gene expression and regulatory networks of individual cells, and also to learn about heterogeneity within and between populations, as well as to define how these populations relate to each other.

Whilst HSPCs have been profiled using a range of single-cell techniques, some recent technologies such as Droplet based methods have not yet been published as applied to HSPCs. It is likely that such new technologies will address a major limitation of the older scRNA-seq protocols, which is the significant cost incurred when sequencing high numbers

of cells. DropSeq uses microfluidic droplet generation to first isolate and barcode individual cells, before pooling and sequencing all cells together as a batch to reduce sequencing costs. Although individual methods such as DropSeq [49] and inDrops [50] have technical differences, the principle is the same: cells are encapsulated in nanoliter droplets with DNA-barcoding beads, which attach to genes in each cell and can then be sequenced to obtain gene expression profiles for thousands of cells at a much lower cost.

Based on bulk analysis of different haematopoietic stem and progenitor populations, changes in DNA methylation are known to occur between the different stages of haematopoiesis [51–53] yet so far this has not been investigated at the single-cell level. Recently Guo et al. reported an adaptation of reduced representation bisulfite sequencing (RRBS) allowing profiling of the methylation landscape in single-cells [54]. Their technique, scRRBS, covered an average of 40% of the CpG sites detected by bulk RRBS in mouse embryonic stem cells. However, as the concept of RRBS is to reduce costs by mainly restricting sequencing to genomic regions with high CpG density, even methylation profiling by bulk RRBS will only cover around 10% of all CpG sites [55]. Genome-wide bisulfite sequencing has also been extended to single cells [56] providing methylation scores on up to 48% of CpG sites in individual cells. Additionally, they were also able to distinguish methylation differences between embryonic stem cells grown in standard serum with leukaemia inhibitory factor (LIF) conditions and ground-state pluripotency-inducing conditions (2i plus LIF) [57].

The ability to sequence both genomic DNA and mRNA from the same cell is a new single-cell technology that has not yet been applied to haematopoietic cells. The methods gDNA-mRNA sequencing (DR-seq) [58] and genome and transcriptome sequencing (G&T-seq) [59], as described by Dey et al. and Macaulay et al. respectively, enable the link between

genomic and transcriptomic heterogeneity to be investigated by quantifying both of these features simultaneously. Applied to haematopoiesis, this could be particularly interesting in the context of the development of blood disorders where HSPCs are often seen to have acquired mutations potentially linked to aberrant function.

As described in this review, transcriptomic heterogeneity in HSPCs has been widely reported, posing the question of how this variation is regulated. One factor linked to gene expression is the accessibility of chromatin, a property that can be measured by an assay for transposase-accessible chromatin using sequencing (ATAC-seq) [60]. Mazumdar et al. used ATAC-seq to investigate the role of cohesin mutants in acute myeloid leukaemia cell lines. The authors showed that mutations in cohesin resulted in increased chromatin accessibility at key transcription factor binding sites of ERG, GATA2 and RUNX1 [61]. However, this study was limited to bulk analysis of cells; recently, there have been descriptions of methods able to measure chromatin accessibility at the single-cell level. Buenrostro et al. performed single-cell ATAC-seq (scATAC-seq) by using a microfluidics platform to capture and process individual cells [62]. scATAC-seq could successfully distinguish between different cell lines based on differential chromatin accessibility patterns, indicating its potential for uncovering heterogeneity within populations of haematopoietic cell types. However, the method is limited by coverage: the authors estimate that only 9.4% of promoters are represent in a scATAC-seq library. An alternative approach is described by Cusanovich et al., who use a system of combinatorial barcoding of cells followed by bulk ATAC-seq on the labelled population [63]. The advantage of this approach is that the need to isolate and process single cells is avoided, but at the cost of increased chance that an ATAC-seq profile in fact belongs to more than one cell. The authors estimated that their double barcoding method resulted in around 11% of nuclei being labelled with the same combination of barcodes, although this

rate is dependent on how many nuclei are processed per well. This rate can be reduced at the cost of cellular throughput for an experiment. Altogether, these methods of measuring chromatic accessibility provide another single-cell avenue for studying heterogeneity in haematopoiesis.

Single-cell studies are rapidly providing a greater understanding of cell and population heterogeneity. Sources of heterogeneity that future research will need to take into account include the microenvironment, age and cell cycle status. It is also likely that differences between species will extend to aspects of cellular heterogeneity. Single-cell profiling has begun to give us some insight into how factors such as age [35,64] and species [33] can influence the gene or protein expression profiles of haematopoietic cells. Furthermore, it is self-evident that heterogeneity caused by interactions with, for example, different niches can only be understood if we have information about the specific features that differentiate the various niches. Current high-throughput single-cell profiling protocols rely on the generation of single cell suspensions and therefore destroy all 3-dimensional context of the wider tissue. Future research efforts will need to be devoted to developing technologies that can generate single-cell molecular profiles within the context of an intact tissue.

Another very important issue that still needs to be addressed is that the molecular profiling techniques described above are restricted by their ability to only provide snapshot data, a representation of the cell in a particular gene expression state at a particular point in differentiation. Haematopoietic fate decisions are dynamic processes, where profiles of individual cells change over time. The concept of pseudotime acknowledges this by attempting to link individual profiles in an inferred dynamic trajectory. However, it is also important that the true dynamics of molecular profiles of HSPCs are investigated, as

pseudotime does not represent ‘real’ time but can be influenced by factors such as cell proliferation and numbers. Continuous single-cell imaging allows for the visual and quantitative tracking of a single cell as it progresses through haematopoiesis, and has the ability to link current cell gene expression, protein activity and structure with future function and fate [65]. Linking single-cell molecular profiles with the lower-throughput technique of time-lapse imaging and tracking of individual cells will increase the informative insight gained into the true dynamics of HSPC fate decisions [66,67].

Finally, a vital consideration related to all of the above technologies is that advances in single-cell techniques require the development of new and improved computational methods. For example, it is clear that for scRNA-seq data steps such as normalisation are essential, yet many different methods are used and still have limitations. Developing specialised computational methods for dealing with high-dimensional, and often noisy, single-cell data must remain a priority [68].

Although studying bulk populations is useful for gaining insights into HSPC biology, as the heterogeneous nature of HSPC populations has been recognised, the need for single-cell profiling techniques has also been further established. The value of single-cell techniques is greatly increased by pairing them with computational methods to manipulate and analyse the data, in order to draw meaningful conclusions. Current techniques allow for the study of heterogeneous populations, regulatory relationships, and lineage differentiation. Single-cell transplant assays are a useful tool for the functional validation of gene expression analysis techniques, and index-sorting data acts a bridge to be able to directly compare and integrate the two types of data based on surface marker expression. Through further study at the single-cell level, and by investigating the ability to link computationally-analysed snapshot

data with live cell data, researchers will be able to delve further into the heterogeneity and regulatory networks governing HSPC biology.

Acknowledgements

Research in the authors' laboratory is supported by Cancer Research UK, the Biotechnology and Biological Sciences Research Council, Bloodwise, the Leukemia and Lymphoma Society, a Wellcome Trust Strategic Award (Tracing Early Mammalian Lineage Decisions by Single Cell Genomics) and core support grants by the Wellcome Trust to the Cambridge Institute for Medical Research and Wellcome Trust - MRC Cambridge Stem Cell Institute. FKH and SN gratefully acknowledge the MRC for funding of their studentships.

References

1 Tenen DG (2003) Disruption of differentiation in human cancer: AML shows the way. *Nat. Rev. Cancer* **3**, 89–101.

2 Moignard V & Göttgens B (2014) Transcriptional mechanisms of cell fate decisions revealed by single cell expression profiling. *Bioessays* **36**, 419–26.

3 Klein AM & Simons BD (2011) Universal patterns of stem cell fate in cycling adult tissues. *Development* **138**, 3103–3111.

4 Simons BD & Clevers H (2011) Strategies for homeostatic stem cell self-renewal in adult tissues. *Cell* **145**, 851–862.

5 Lindström S (2012) Flow cytometry and microscopy as means of studying single cells: a short introductional overview. *Methods Mol. Biol.* **853**, 13–5.

6 Osborne GW (2011) Recent advances in flow cytometric cell sorting. *Methods Cell Biol.* **102**, 533–56.

7 Schulte R, Wilson NK, Prick JCM, Cossetti C, Maj MK, Gottgens B & Kent DG (2015)

Index sorting resolves heterogeneous murine hematopoietic stem cell populations. *Exp. Hematol.* **43**, 803–11.

8 Wilson NK, Kent DG, Buettner F, Shehata M, Macaulay IC, Calero-Nieto FJ, Sánchez Castillo M, Oedekoven CA, Diamanti E, Schulte R, Ponting CP, Voet T, Caldas C, Stingl J, Green AR, Theis FJ & Göttgens B (2015) Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell* **16**, 712–724.

9 Hayashi T, Shibata N, Okumura R, Kudome T, Nishimura O, Tarui H & Agata K (2010) Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its "index sorting" function for stem cell research. *Dev. Growth Differ.* **52**, 131–44.

10 Bendall SC, Nolan GP, Roederer M & Chattopadhyay PK (2012) A deep profiler's guide to cytometry. *Trends Immunol.* **33**, 323–32.

11 Behbehani GK, Bendall SC, Clutter MR, Fantl WJ & Nolan GP (2012) Single-cell mass cytometry adapted to measurements of the cell cycle. *Cytometry. A* **81**, 552–66.

12 Chattopadhyay PK & Roederer M (2015) A mine is a terrible thing to waste: high content, single cell technologies for comprehensive immune analysis. *Am. J. Transplant* **15**, 1155–61.

13 Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ, Kinston S, Joshi A, Hannah R, Theis FJ, Jacobsen SE, de Bruijn MF & Göttgens B (2013) Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.* **15**, 363–372.

14 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**, 171–81.

15 Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC & Teichmann SA (2015) The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell* **58**, 610–620.

16 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 (80-.).* **343**, 776 – 779.

17 Naik SH, Schumacher TN & Perié L (2014) Cellular barcoding: a technical appraisal. *Exp. Hematol.* **42**, 598–608.

18 Perie L, Duffy KR, Kok L, Boer RJ De & Schumacher TN (2015) The Branching Point in Erythro-Myeloid Differentiation. *Cell* **163**, 1655–1662.

19 Basu S, Campbell HM, Dittel BN & Ray A (2010) Purification of specific cell population by fluorescence activated cell sorting (FACS). *J. Vis. Exp.*

20 Pina C, Teles J, Fugazza C, May G, Wang D, Guo Y, Soneji S, Brown J, Edén P, Ohlsson M, Peterson C & Enver T (2015) Single-Cell Network Analysis Identifies DDIT3 as a Nodal Lineage Regulator in Hematopoiesis. *Cell Rep.* **11**, 1503–1510.

21 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.

22 van der Maaten L & Hinton G (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605.

23 Haghverdi L, Buettner F & Theis FJ (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 1–10.

24 Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP & Pe'er D (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–52.

25 Woodhouse S, Moignard V, Göttgens B & Fisher J (2015) Processing, visualising and reconstructing network models from single-cell data. *Immunol. Cell Biol.*

26 Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, Nishikawa S-I, 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**, 269–76.

27 Guo G, Luc S, Marco E, Lin TW, Peng C, Kerenyi M a., Beyaz S, Kim W, Xu J, Das PP, Neff T, Zou K, Yuan GC & Orkin SH (2013) Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell* **13**, 492–505.

28 Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, Finck R, Gedman AL, Radtke I, Downing JR, Pe'er D & Nolan GP (2015) Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197.

29 Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, Winter D, Lara-Astiaso D, Gury M, Weiner A, David E, Cohen N, Felicia Kathrine Bratt Lauridsen, Haas S, Schlitzer A, Mildner A, Ginhoux F, Jung S, Trumpp A, Porse BT, Tanay A & Amit I (2015) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677.

30 Notta F, Zandi S, Takayama N, Dobson S, Gan OI, Wilson G, Kaufmann KB, McLeod J, Laurenti E, Dunant CF, McPherson JD, Stein LD, Dror Y & Dick JE (2015) Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116.

31 Adolfsson J, Månsson R, Buza-Vidas N, Hultquist A, Liuba K, Jensen CT, Bryder D, Yang L, Borge O-J, Thoren LAM, Anderson K, Sitnicka E, Sasaki Y, Sigvardsson M & Jacobsen SEW (2005) Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage

commitment. *Cell* **121**, 295–306.

32 Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner R V, Linderman MD, Sachs K, Nolan GP & Plevritis SK (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891.

33 Spitzer MH, Gherardini PF, Fragiadakis GK, Bhattacharya N, Yuan RT, Hotson a. N, Finck R, Carmi Y, Zunder ER, Fantl WJ, Bendall SC, Engleman EG & Nolan GP (2015) An interactive reference framework for modeling a dynamic immune system. *Science (80-.).* **349**, 1259425–1259425.

34 Drissen R, Buza-Vidas N, Woll P, Thongjuea S, Gambardella A, Giustacchini A, Mancini E, Zriwil A, Lutteropp M, Grover A, Mead A, Sitnicka E, Jacobsen SEW & Nerlov C (2016) Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nat. Immunol.* **advance on**.

35 Grover A, Sanjuan-Pla A, Thongjuea S, Carrelha J, Giustacchini A, Gambardella A, Macaulay I, Mancini E, Luis TC, Mead A, Jacobsen SEW & Nerlov C (2016) Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nat. Commun.* **7**, 11075.

36 Bendall SC, Davis KL, Amir ED, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP & Pe D (2014) Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* **157**, 714–725.

37 Göttgens B (2015) Regulatory network control of blood stem cells. *Blood* **125**, 2614–2620.

38 Peter I & Davidson E (2015) *Genomic control process: development and evolution* Academic Press.

39 Xu H, Ang Y-S, Sevilla A, Lemischka IR & Ma'ayan A (2014) Construction and validation of a regulatory network for pluripotency and self-renewal of mouse

embryonic stem cells. *PLoS Comput Biol* **10**, e1003777.

40 Schütte J, Wang H, Antoniou S, Jarratt A, Wilson NK, Riepsaame J, Calero-Nieto FJ, Moignard V, Basilico S, Kinston SJ, Hannah RL, Chan MC, Nürnberg ST, Ouwehand WH, Bonzanni N, de Bruijn MF & Göttgens B (2016) An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. *Elife* **5**.

41 Beerman I, Bhattacharya D, Zandi S, Sigvardsson M, Weissman IL, Bryder D & Rossi DJ (2010) Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5465–70.

42 Challen GA, Boles NC, Chambers SM & Goodell MA (2010) Distinct hematopoietic stem cell subtypes are differentially regulated by TGF-beta1. *Cell Stem Cell* **6**, 265–78.

43 Kent DG, Copley MR, Benz C, Wöhrer S, Dykstra BJ, Ma E, Cheyne J, Zhao Y, Bowie MB, Zhao Y, Gasparetto M, Delaney A, Smith C, Marra M & Eaves CJ (2009) Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential. *Blood* **113**, 6342–50.

44 Kiel MJ, Yilmaz OH, Iwashita T, Yilmaz OH, Terhorst C & Morrison SJ (2005) SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121**, 1109–21.

45 Morita Y, Ema H & Nakauchi H (2010) Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J. Exp. Med.* **207**, 1173–82.

46 Adolfsson J, Borge OJ, Bryder D, Theilgaard-Mönch K, Åstrand-Grundström I, Sitnicka E, Sasaki Y & Jacobsen SEW (2001) Upregulation of Flt3 expression within the bone marrow Lin-Sca1+c-kit+ stem cell compartment is accompanied by loss of self-renewal capacity. *Immunity* **15**, 659–669.

47 Kiel MJ, Radice GL & Morrison SJ (2007) Lack of evidence that hematopoietic stem cells depend on N-cadherin-mediated adhesion to osteoblasts for their maintenance. *Cell Stem Cell* **1**, 204–17.

48 Weksberg DC, Chambers SM, Boles NC & Goodell MA (2008) CD150- side population cells represent a functionally distinct population of long-term hematopoietic stem cells. *Blood* **111**, 2444–51.

49 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.

50 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.

51 Attema JL, Papathanasiou P, Forsberg EC, Xu J, Smale ST & Weissman IL (2007) Epigenetic characterization of hematopoietic stem cell differentiation using miniChIP and bisulfite sequencing analysis. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 12371–6.

52 Rice KL, Hormaeche I & Licht JD (2007) Epigenetic regulation of normal and malignant hematopoiesis. *Oncogene* **26**, 6697–714.

53 Álvarez-Errico D, Vento-Tormo R, Sieweke M & Ballestar E (2014) Epigenetic control of myeloid cell differentiation, identity and function. *Nat. Rev. Immunol.* **15**, 7–17.

54 Guo H, Zhu P, Wu X, Li X, Wen L & Tang F (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–35.

55 Gu H, Bock C, Mikkelsen TS, Jäger N, Smith ZD, Tomazou E, Gnirke A, Lander ES & Meissner A (2010) Genome-scale DNA methylation mapping of clinical samples at

single-nucleotide resolution. *Nat. Methods* **7**, 133–6.

56 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.

57 Ying Q-L, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P & Smith A (2008) The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–23.

58 Dey SS, Kester L, Spanjaard B, Bienko M & van Oudenaarden A (2015) Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* **33**, 285–289.

59 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.

60 Buenrostro JD, Giresi PG, Zaba LC, Chang HY & Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–8.

61 Mazumdar C, Shen Y, Xavy S, Zhao F, Reinisch A, Li R, Corces MR, Flynn RA, Buenrostro JD, Chan SM, Thomas D, Koenig JL, Hong W-J, Chang HY & Majeti R (2015) Leukemia-Associated Cohesin Mutants Dominantly Enforce Stem Cell Programs and Impair Human Hematopoietic Progenitor Differentiation. *Cell Stem Cell* **17**, 675–688.

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

63 Cusanovich DA, Daza R, Adey A, Pliner H, Christiansen L, Gunderson KL, Steemers FJ,

Trapnell C & Shendure J (2015) Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (80-.).* **348**, 910–4.

64 Kowalczyk MS, Tirosh I, Heckl D, Nageswara Rao T, Dixit A, Haas BJ, Schneider R, Wagers AJ, Ebert BL & Regev A (2015) Single cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.*, gr.192237.115.

65 Hoppe PS, Coutu DL & Schroeder T (2014) Single-cell technologies sharpen up mammalian stem cell research. *Nat. Cell Biol.* **16**, 919–27.

66 Eilken HM, Nishikawa S-I & Schroeder T (2009) Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature* **457**, 896–900.

67 Schroeder T (2011) Long-term single-cell imaging of mammalian stem cells. *Nat. Methods* **8**, S30–5.

68 Stegle O, Teichmann SA & Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Publ. Gr.* **16**, 133–145.

Figure legends

Figure 1: Single-cell profiling enables heterogeneities within cell populations to be explored.

A) Heterogeneous populations of cells, represented here using different colours, can be investigated using single-cell analysis. Firstly a population of cells is isolated for single-cell profiling using techniques such as flow cytometry or index sorting. This allows for the listed single-cell profiling methods to be applied. These techniques are chosen depending on the biological question of interest. B) Dimensionality reduction techniques allow heterogeneities within a population of cells to be visualised based on single-cell expression profiles. Plotting

cells in this two dimensional coordinate system allows visualization of heterogeneity within the populations and can confirm that subpopulations separate based on their expression profiles. Techniques such as principal component analysis can successfully separate populations consisting of a mixture of cell types based on single-cell profiles, as displayed in the top panel of (B). When considering samples of cells at different stages of differentiation, depicted ranging from grey to yellow or red along two lineage branches in the bottom panel of (B), the use of techniques such as diffusion maps may be more suitable. Diffusion maps can capture the continuous nature of processes such as differentiation and allow visualisation of branching trajectories in the data. C) Unbiased clustering techniques can be applied to single-cell data to explore similarities between cells. In hierarchical clustering, as shown here, the most similar groups of cells are more closely connected in the dendrogram. This structure then allows us to explore different levels of clustering within the data: for example the cells can be split into three groups that correspond to their cell type. D) Single-cell barcoding techniques allow heterogeneities to be resolved at a functional level. A population of interest can be sorted, for example from mouse bone marrow. Individual cells are then genetically labelled with different barcodes, depicted here in different colours. These barcoded cells can then be transplanted into a lethally irradiated recipient, and after several weeks the host bone marrow can be harvested, sorted using FACS and sequenced to reveal to which haematopoietic cells types each barcoded cell contributes.

Figure 2: Single-cell expression profiles can be ordered to reconstruct lineage differentiation.

Using the assumption that the cells closest in the differentiation process will have the most similar gene or protein expression profiles, methods have been developed with the purpose of reconstructing lineage differentiation from single-cell measurements. A) A population of

cells can contain several subpopulations (represented by different colours) from different stages of lineage differentiation. Individual cells can be clustered into groups based on gene or protein expression profiles. By assigning similarity scores between groups we can construct a graph where each node corresponds to a cell cluster and the edges between nodes are weighted by similarity scores between clusters. This graph then forms the means for finding a reconstruction of the lineage tree, for example by finding the minimum spanning tree as described by Qiu et al. [32]. B) Even in in vitro differentiation experiments not all cells differentiate at the same rate. A population can contain cells at multiple stages of differentiation, here depicted in colours on the spectrum from red to blue. Based on similarities between their expression profiles, these cells can be computationally ordered in pseudotime, a quantity that represents their progress through differentiation. Patterns of gene or protein expression can then be explored along pseudotime allowing the identification of key biological events or factors linked to the differentiation process.

Figure 3: Inferring regulatory relationships from single-cell expression data.

A) Quantities such as correlation between gene pairs can be calculated using single-cell gene expression measurements. As shown in this gene-gene correlation heatmap some pairs will exhibit positive correlation and some pairs negative correlation, suggestive of positive and negative regulatory relationships, respectively. Thresholds can be chosen to select the most strongly correlating gene pairs. Here correlations that do not meet these thresholds are coloured in white. Connections between these highly correlating gene pairs can then be drawn in a network diagram with red or blue lines representing either positive or negative regulation. B) Transcription factors can be part of combinatorial regulatory relationships. If factors A and B both activate gene C this could correspond to two different scenarios, which are represented here using Boolean logic functions. It could be that either A or B alone will

cause activation of C, as shown on the left with the Boolean Or function. The output of an Or function is given in the truth table next to the gate. Alternatively, it could be that binding of both A and B is required for activation of C, as shown by the And gate and truth table. C) Regulatory networks can be modelled using Boolean functions. Gene expression measurements for single cells can be converted into binary (ON/OFF) expression by choosing a threshold. Computational methods applied to this binary data allow inference of regulatory relationships from these data, represented here by Boolean And/Or functions.

Figure 4: Single-cell gene expression analysis can be linked to cellular function using index sorting data.

Molecular profiling and functional assays can be linked together to gain a greater understanding of HSPC biology by taking advantage of index sorting data. A) Heterogeneous populations are sorted using known purification strategies and data on surface marker expression of each individual cell is collected using the index sorting function. B) The cells collected can be interrogated using various techniques to gain a gene expression profile for each cell. These techniques will vary based on the experimenter's interests and include qRT-PCR or scRNA-seq. The gene expression profiles are compared to surface marker data obtained by index sorting. C) A subpopulation of cells can be sorted on a new, refined sorting strategy, which was defined by comparing gene expression and surface marker expression data. D) The subpopulation can be further interrogated by a variety of techniques, such as further gene expression analysis and in vitro and in vivo assays. Index sorting data can again be used to compare the surface marker expression of cells used for the different assays, to link the data together for a complete molecular and functional profile of cells of interest.

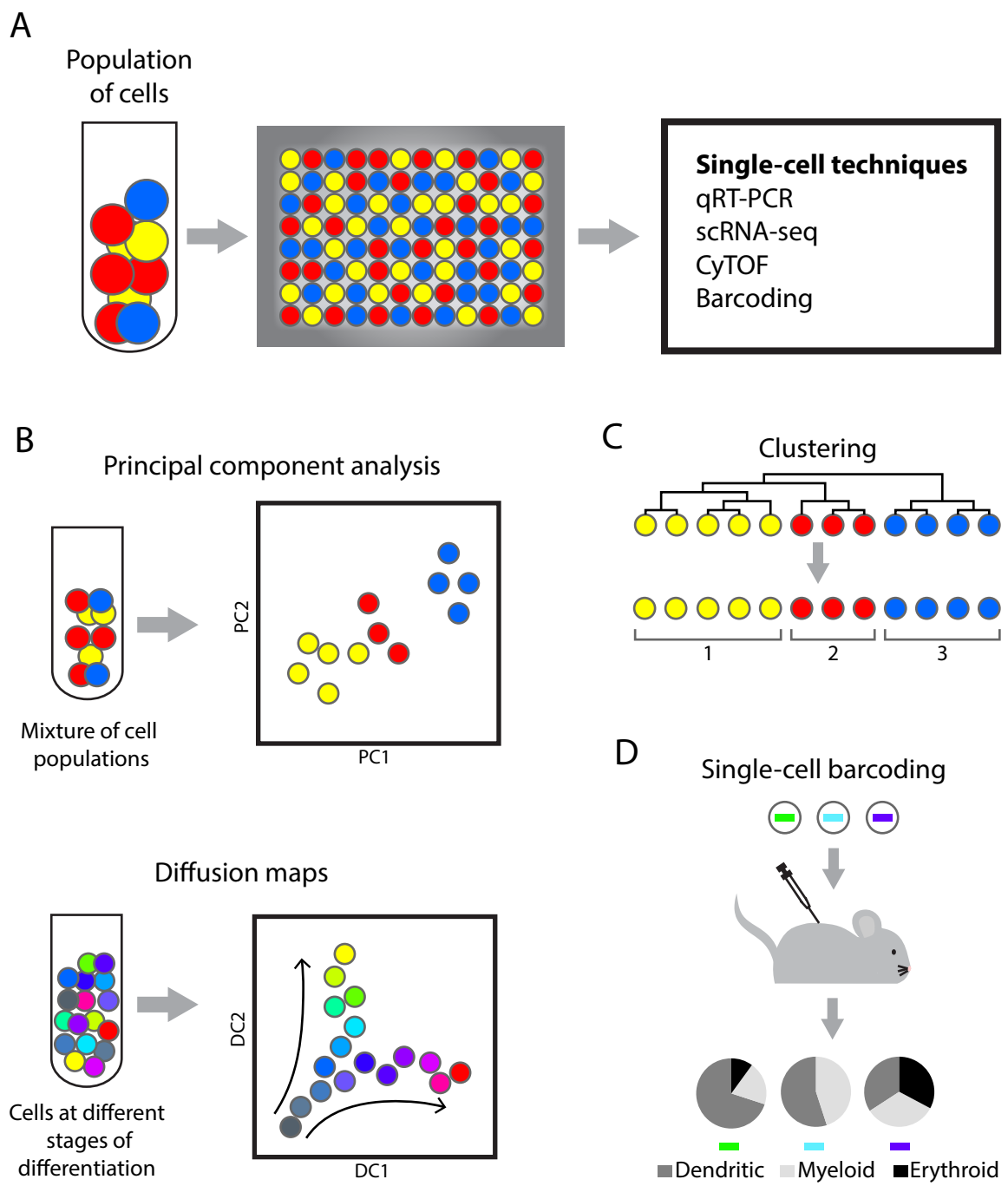
A
Population of cells
Single-cell techniques
qRT-PCR
scRNA-seq
CyTOF
Barcoding
B
Principal component analysis
PC2
PC1
Mixture of cell populations
Diffusion maps
DC2
DC1
Cells at different stages of differentiation
C
Clustering
1
2
3
D
Single-cell barcoding
Dendritic
Myeloid
Erythroid

Figure 1

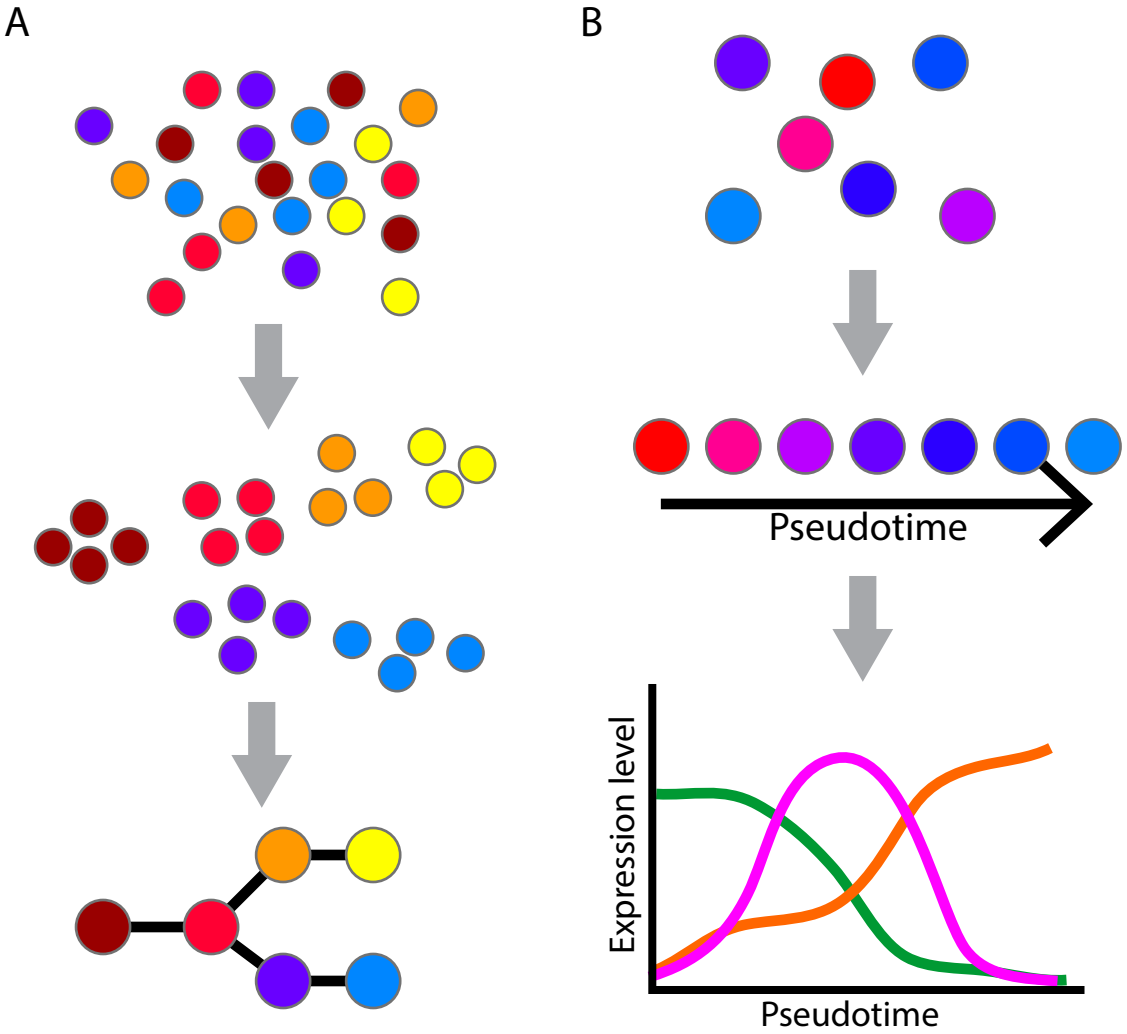

Figure 2

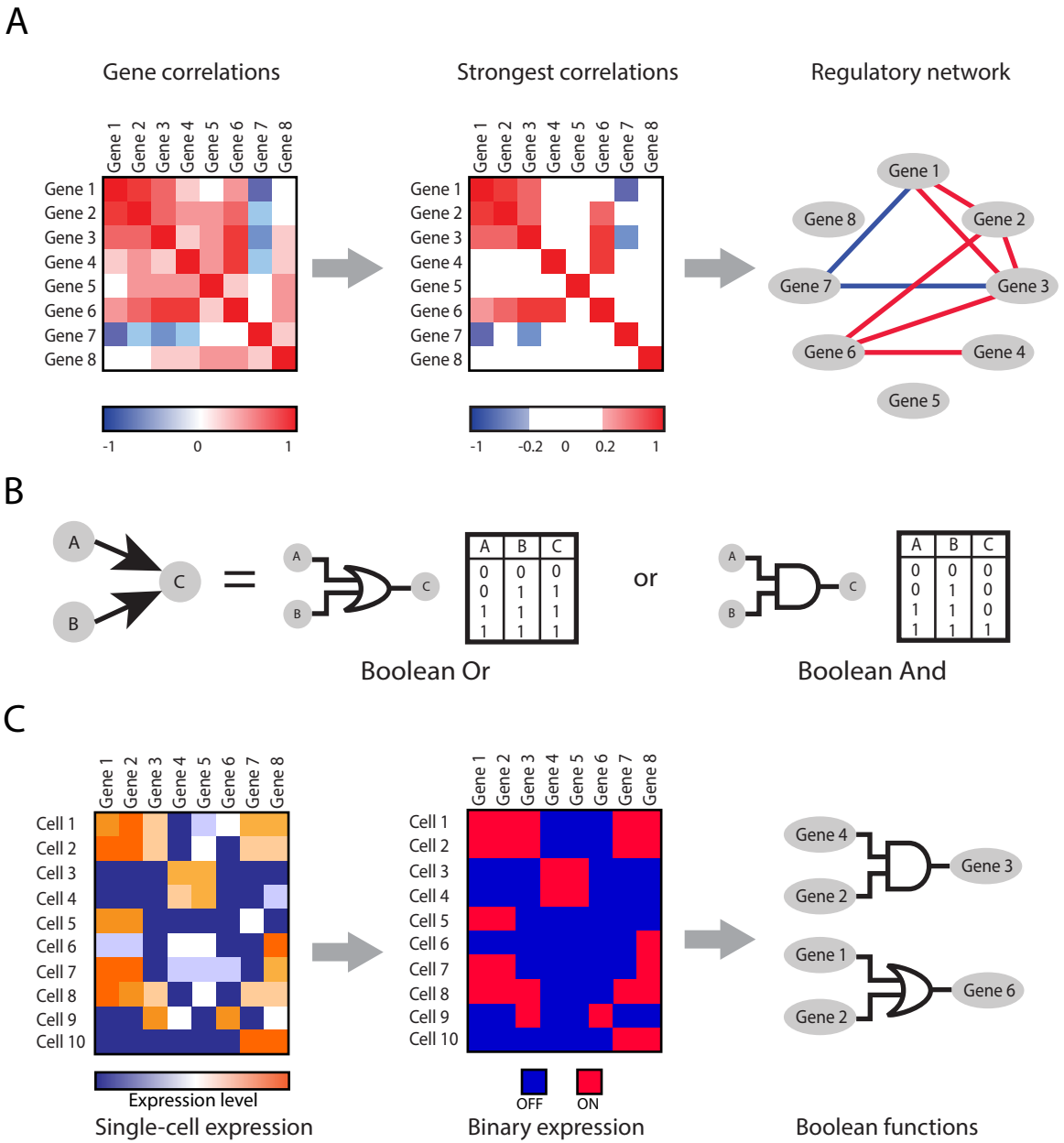
A
Gene correlations
Strongest correlations
Regulatory network
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
-1
0
1
-0.2
0.2
B
A
B
C
=
or
Boolean Or
Boolean And
C
Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
Cell 6
Cell 7
Cell 8
Cell 9
Cell 10
Expression level
Single-cell expression
OFF
ON
Binary expression
Boolean functions

Figure 3

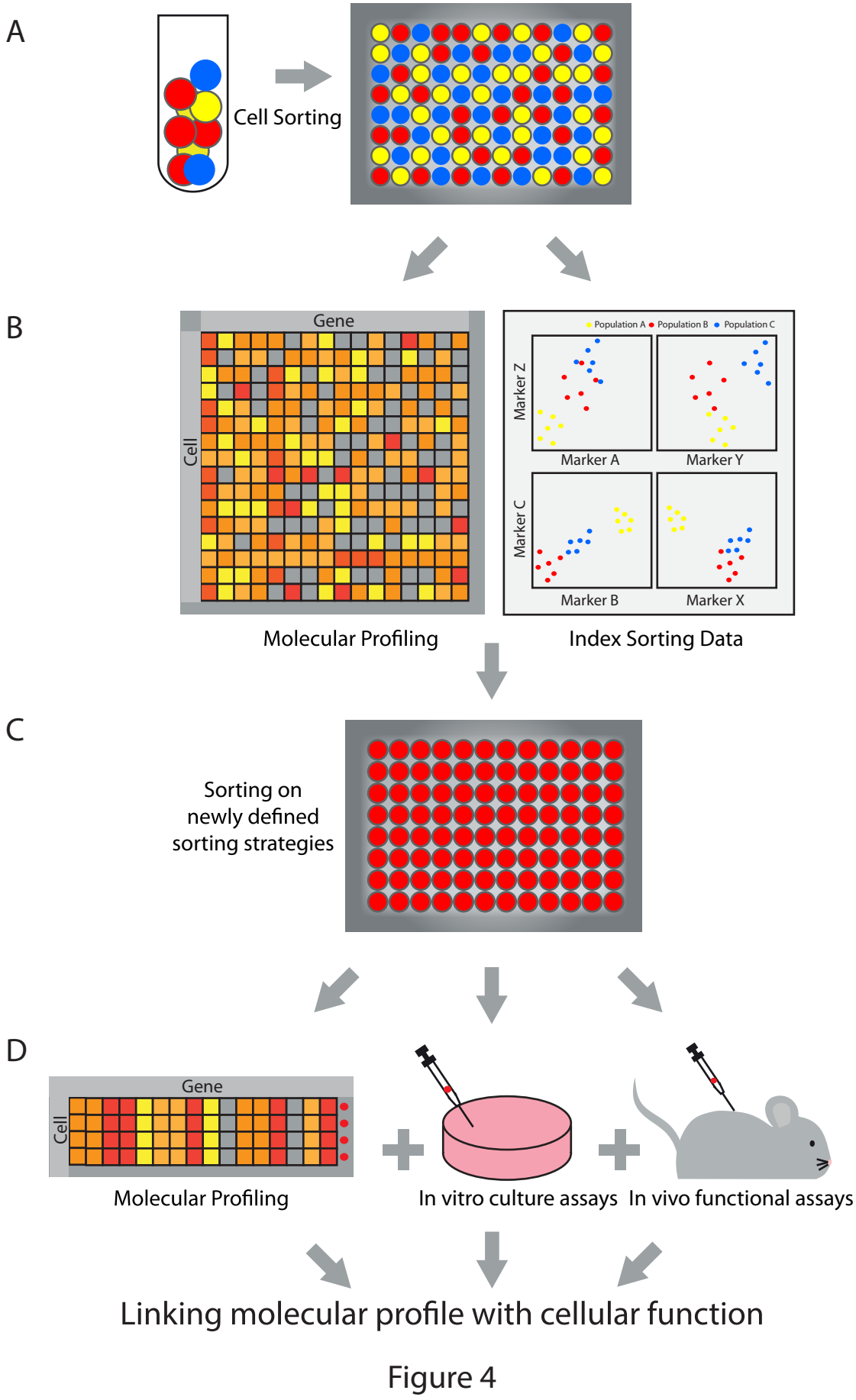
A
Cell Sorting
B
Gene
Cell
Population A
Population B
Population C
Marker Z
Marker A
Marker Y
Marker C
Marker B
Marker X
Molecular Profiling
Index Sorting Data
C
Sorting on
newly defined
sorting strategies
D
Gene
Cell
Molecular Profiling
In vitro culture assays
In vivo functional assays
Linking molecular profile with cellular function

Figure 4