UNDERSTANDING GENE TRANSCRIPTIONAL REGULATION AT SINGLE CELL RESOLUTION

by

Chen Xu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2015

Approved by:

_____
Dr. Zhengchang Su

_____
Dr. Anthony Fodor

_____
Dr. Jun-tao Guo

_____
Dr. Jennifer Weller

_____
Dr. Bao-Hua Song

ABSTRACT

CHEN XU. Understanding gene transcriptional regulation at single cell resolution.
(Under the direction of DR ZHENGCHANG SU)

The recent advance of single-cell technologies has provided an unprecedented opportunity to bring new insights into many complex biological phenomena, such as the regulation of cell differentiation in a multi-cellular organism and cell-to-cell variability in an isogenic population. In this dissertation, we have explored the gene expression regulation using datasets generated by single-cell techniques in three aspects. First, we analyzed a large-scale gene expression dataset measured in individual cells throughout the embryogenesis of *C. elegans* in a nearly continuous time-scale. We revealed many known and novel genes driving lineage divergence at early cell divisions, facilitating a systematic understanding of the fate specification in *C. elegans*. Second, we developed a novel clustering algorithm named SNN-Cliq that utilizes the shared nearest neighbor and graph-theoretic partitioning techniques. Our algorithm has the superiority of handling high-dimensional noisy data in that it allows clustering on a variety of single-cell RNA-sequencing (RNA-Seq) data with high accuracy. Last, using an RNA-Seq technique, we profiled transcriptomes in 51 yeast cells from three treatments. Intriguingly, we found that the transcription variation, or noise, shows distinct features under different treatments for certain functional gene modules and regulatory pathways. Our results also suggest that transcriptional noise is subject to regulation in response to environmental stresses. In summary, this dissertation has contributed to algorithmic development for analyzing various single-cell datasets and deepened our knowledge of transcriptional regulation at the single cell level.

DEDICATION

Science extends my world through time and space.

I dedicate my work to my family, teachers and friends who lead me into science.

ACKNOWLEDGEMENTS

My greatest appreciation goes to my advisor, Dr. Zhengchang Su for his brilliant ideas and inspiration. He encouraged me to gain a wider breadth of knowledge in computer science and statistics. His advice allows me to set the foundation for becoming a bioinformatist with the ability of independent thinking. I am sure his suggestions will also benefit my long-term career goals. He also gave me the opportunity to work with independence and develop my own interests in bioinformatics. In addition, he showed consideration, understanding and support when I met obstacles during my past five-years living aboard.

I want to thank all members in the Su Lab for being such helpful lab-mates and creating such a friendly laboratory environment. Despite the huge work pressure we were facing, we were always ready to give help and suggestions to each other. Dr. Shan Li, Dr. Chuanbin Du, Dr. Xia Dong, Dr. Meng Niu, and Ehsan are all my best friends. They provided enormous pleasure and amusement. I cannot imagine a better PhD experience without them.

I am grateful to the Department of Bioinformatics and Genomics at UNCC, especially my committee members, Dr. Jennifer Weller, Dr. Anthony Fodor, Dr. Jun-tao Guo and Dr. Bao-hua Song. Their inspiring input and encouragement allow me to continuously make progress on research with confidence.

Finally, and most importantly, I dedicate my deepest gratitude to my parents, An Xu and Mingshu Dong, for their endless support and love. It is always not easy for them when I leave home to a distant foreign country for so long. I sincerely hope my achievement can make them proud and happy. I thank my husband, Pengfei Xiao, for his faith in me. He makes me feel confident in my life and work, as well as a good future.

INTRODUCTION

The cell is the functional unit of all kinds of free-living organisms. Multi-cellular organisms contain different types of cells derived through differentiation during embryogenesis, where the genome undergoes regulated transcription and translation, leading to phenotypic heterogeneity in cells carrying out various functions. On the other hand, due to the inherent stochasticity of biochemical processes in a cell, even monoclonal cells that have been cultured under identical conditions can display cell-to-cell variability. However, such heterogeneity is masked in conventional experiments performed at a population level, because the analysis is averaged over thousands or millions of cells. Therefore, despite intensive research, the extent of gene expression heterogeneity and the diversity of cell types in a tissue remain largely unknown (Macaulay and Voet, 2014). Also owning to the lack of reliable single-cell methods, many long-standing biological questions remain unsolved. For example, using population-averaging techniques, it is unclear whether all the observed cellular activities (e.g., signaling pathways) happen in each individual cell, or rather in different subsets of cells. In addition, many important cell types are often in rare quantities and exhibit heterogeneous characters, such as stem cells and tumor cells. In principle, single-cell analysis can dissect a heterogeneous tissue into subpopulations and identify rare cells based on gene expression profiles of individual cells. Consequently, single-cell approaches are essential to gain better understanding of complex biological phenomena.

Due to technical limitations, the early study of single cells is restricted to examining a few number of specific genes in a cell at a time. One of the earliest study conducted by Elowitz et al. used CFP and YFP fluorescent proteins to quantify the intrinsically and

extrinsically originated expression variation of the LacI gene in *E.coli* (Elowitz *et al.*, 2002). Later, Cai *et al*. introduced a microfluidic-based enzymatic assay that traps a living cell in a small volume formed by compression of a flow channel, allowing real-time observation of the fluorescein from the low copy number of β-galactosidase molecules in *E. coli* with single molecule sensitivity (Cai *et al.*, 2006). These single-cell techniques based on reporter genes were also extended to some eukaryotes in which fluorescence can be clearly observed under microscope, such as yeast (Volfson *et al.*, 2006; Rinott *et al.*, 2011) and *C. elegans* (Liu *et al.*, 2009). Beside the determination of variability at the protein level, exploring the mRNAs at single-cell resolution is also of great interest because a considerable part of gene expression variation is contributed by the transcription process. The technique of single mRNA-sensitivity fluorescence in situ hybridization (FISH) allows to count the exact number of mRNAs present in individual cells (Raj *et al.*, 2008; Zenklusen *et al.*, 2008). By monitoring a time-series of gene activity using single-RNA FISH, Zenklusen *et al*. and So *et al*. identified bursts of transcription and characterized the transcription kinetics in yeast and *E. coli* (Zenklusen *et al.*, 2008; So *et al.*, 2011). Another single-cell mRNA profiling technique based on microfluidics qPCR enables the quantification of several hundred transcripts in hundreds of single cells simultaneously (Citri *et al.*, 2012). Using this technique, a study revealed both stable and dynamic transcription factor relationships in a critical regulatory network from five blood stem and progenitor cells (Moignard *et al.*, 2013). The method has also been applied to investigate the gene expression of 48 transcription factors in over 500 cells from 8-cell to 64-cell stage mouse blastocyst (Guo *et al.*, 2010). Single-cell mass cytometry combining mass spectrometry and flow cytometry enables the measure of more than 40 features (binding

antibodies, viability, DNA content, cell size, etc.) simultaneously in thousands to millions of cells in an experiment (Bendall *et al.*, 2011). It was used to trace the lineage trajectory during the development of hematopoietic stem cells through to native B cell (Bendall *et al.*, 2014). Although these single-cell techniques have allowed one to measure a handful of genes from a moderate number of cells, the number of gene measured is still limited by the requirement of distinguishable fluorescent dyes, available primers and antibodies to the targets of interest (Kalisky and Quake, 2011).

In the last decade, there has been a surge in the development of single-cell technologies with radically improved throughput, bringing in single-cell genomics, transcriptomics, proteomics, and metabolomics. For example, single-cell transcriptome profiling has been carried out using RNA-sequencing (RNA-Seq) (Tang *et al.*, 2009). Recently, single-cell RNA-Seq has been widely used in comparing transcriptome profiles of individual cells, and characterizing the dynamic reaction of transcripts to the environment. For instance, by sequencing the individual embryonic cells at different stages of embryos, it is feasible to resolve the cell fate decisions and coordination among individual cells in embryonic development (Tang, Barbacioru, Bao, *et al.*, 2010; Yan *et al.*, 2013; Hashimshony *et al.*, 2012). Single-cell RNA-Seq can successfully differentiate a variety of biologically and clinically important cell types, such as circulating tumor cells (Ramsköld *et al.*, 2012). Moreover, these methods enable the revealing of stochastic and deterministic allele specific expression and alternative splicing of isoforms at single cell resolution in mouse preimplantation embryos (Deng *et al.*, 2014), early blastomeres (Tang *et al.*, 2011), and immune cells (Shalek *et al.*, 2013). Single-cell RNA-Seq also allows characterizing cells'

strategies for coping with environmental changes and cell-to-cell variations under a certain environmental stressor (Chapter 3).

As we have seen, single-cell techniques and analyses have brought new insights into many fundamental biological phenomena. In particular, analytics of single cells have deepened our understanding of the cellular heterogeneity in isogenic populations and underlying mechanisms. The fluctuation of gene expression in identical population of cells, also referred to as 'noise', has been extensively studied in *E. coli* and yeast (Elowitz *et al.*, 2002; Raser and O'Shea, 2004; Colman-Lerner *et al.*, 2005; Bar-Even *et al.*, 2006; Newman *et al.*, 2006; Li *et al.*, 2010; Dong *et al.*, 2011; Dunlop *et al.*, 2008; Hornung *et al.*, 2012; Stewart-Ornstein *et al.*, 2012; Carey *et al.*, 2013). The measures, models, origins, consequences, functions and regulatory roles of noise have recently been reviewed in detail (Raser and O'Shea, 2005; Kalisky and Quake, 2011; Pelkmans, 2012; Munsky *et al.*, 2012; Eldar and Elowitz, 2010; Kaern *et al.*, 2005; Raj and van Oudenaarden, 2008, 2009; Paulsson, 2005; Losick and Desplan, 2008; Maheshri and O'Shea, 2007). In general, gene expression noise is due to the stochastic nature of chemical reactions involving small numbers of molecules in a cell. Specifically, gene expression noise has a wide origin: the random birth and death of mRNAs, the global factors such as cell size and cell cycle, the pathway-specific regulation, etc. can all contribute to the gene expression variation. In addition, it has been found that the functional pathways and transcriptional regulatory mechanisms of a gene can be derived from the noise features (Mettetal *et al.*, 2006; Li *et al.*, 2010; Stewart-Ornstein *et al.*, 2012). In prokaryotes, it has been demonstrated that noise can be beneficial for the adaptation and evolution of cell population, as it expands the range of phenotypes resulting in an increased survival rate for at least some individuals in a

population in sudden environmental changes (Johnston and Desplan, 2010; Eldar and Elowitz, 2010). In eukaryote development, noise leads to heterogeneity in the initial homogenous cell population and allows the selection and propagation of cell-type specific gene expression (Kaern *et al.*, 2005).

Clearly, the rapidly growing single-cell datasets present a tremendous opportunity and challenge to the computational biology community for their analysis to reveal new insights into many biological problems. In this dissertation, I shall utilize the single-cell data and techniques to explore some important biological questions associated with cell fate decisions, cell type identification, and transcriptional noise regulation using novel computational methods. In the first chapter, I shall show-case how to utilize a reporter gene expression dataset recorded by time-lapse microscopy in single embryonic cells to investigate individual cell fate and lineage specification in early embryogenesis of *C. elegans*. In the second chapter, I shall introduce a novel clustering algorithm designed specifically for high-dimensional noisy single-cell gene expression data. I shall demonstrate that this algorithm can accurately dissect the cell population in early embryo development. In the third chapter, I shall present our single-cell RNA-Seq dataset including 51 yeast cells collected under different growth conditions. I shall compare the transcription noise under different stress conditions and explore the role of noise in deriving gene transcription regulation mechanisms as well as regulons. In all, these endeavors have emphasized the power of single-cell techniques in deepening our understanding of transcriptional regulation and its role in embryogenesis and stress responses at single cell level.

TABLE OF CONTENTS

CHAPTER 1: IDENTIFICATION OF GENES DRIVING LINEAGE DIVERGENCE FROM SINGLE-CELL GENE EXPRESSION DATA IN C. ELEGANS

1.1 Abstract

The nematode *Caenorhabditis elegans* (*C. elegans*) is an ideal model organism to study the cell fate specification mechanisms during embryogenesis. It is generally believed that cell fate specification in *C. elegans* is mainly mediated by lineage-based mechanisms, where the specification paths are driven forward by a succession of asymmetric cell divisions. However, little is known about how each binary decision is made by gene regulatory programs. In this study, we endeavor to obtain a global understanding of cell lineage/fate divergence processes during the early embryogenesis of *C. elegans.* We reanalyzed the EPIC data set, which traced the expression level of reporter genes at single-cell resolution on a nearly continuous time scale up to the 350-cell stage in *C. elegans* embryos. We examined the expression patterns for a total of 131 genes from 287 embryos with high quality image recordings, among which 86 genes have replicate embryos. Our results reveal that during early embryogenesis, divergence between sister lineages could be largely explained by a few genes. We predicted genes driving lineage divergence and explored their expression patterns in sister lineages. Moreover, we found that divisions leading to fate divergence are associated with a large number of genes being differentially expressed between sister lineages. Interestingly, we found that the developmental paths of lineages could be differentiated by a small set of genes. Therefore, our results support the notion that the cell fate patterns in *C. elegans* are achieved through stepwise binary

decisions punctuated by cell divisions. Our predicted genes driving lineage divergence provide good starting points for future detailed characterization of their roles in the embryogenesis in this important model organism.

1.2 Introduction

A central goal in developmental biology is to understand the molecular mechanisms of cell fate determination during embryogenesis. The nematode *C. elegans* displays an essentially invariant cell lineage during embryogenesis and gives rise to a constant number (558) of cell nuclei in the newly hatched lava (Sulston *et al.*, 1983). This well-characterized developmental architecture has served as an excellent model system to study the relationship between cell lineage and fate. Cell fate specification generally depends on both lineage-based and spatiotemporal context-based mechanisms (Labouesse and Mango, 1999; Maduro, 2010). In *C. elegans,* the spatiotemporal context for each embryonic cell is uniquely specified by a lineage history defined as a series of divisions leading to the cell. Over the last few decades, numerous efforts have been made in understanding how these two mechanisms are involved in driving the embryo patterning of *C. elegans* (Bowerman *et al.*, 1997; Broitman-Maduro *et al.*, 2006; Edgar *et al.*, 2001; Good *et al.*, 2004; Hunter and Kenyon, 1996; Liu *et al.*, 2009; Maduro and Rothman, 2002; Robertson *et al.*, 2004).

In principle, the functional state of a cell is determined by the expression of a specific combination of genes (Bertrand and Hobert, 2010). During the embryogenesis of *C. elegans*, the dynamic expression of regulatory genes results in a regulatory cascade, which drives the differentiation pathway from the zygote to a terminally differentiated cell through a succession of temporary functional states (Bertrand and Hobert, 2010; Maduro, 2010). Often the transitions of functional states are coupled to developmental decisions

being made about each cell cycle and are closely related to asymmetric divisions directed by the binary cell fate decision systems (Baugh, 2003; Cowing and Kenyon, 1996; Edgar and McGhee, 1988; Kaletta *et al.*, 1997). A general binary decision mechanism adopted by *C. elegans* is the Wnt/β-catenin (SYS-1) pathway (Mizumoto and Sawa, 2007; Phillips and Kimble, 2009). This pathway can cooperate with lineage-specifying factors to transit the functional states forward to the next layer (Bertrand and Hobert, 2010). For example, the GATA factors *med-1/2* are first activated by the maternal factor SKN-1 in EMS to determine the fate of both the E and MS lineages (Bowerman *et al.*, 1992). Because SYS-1 is asymmetrically enriched in the posterior nucleus compared to the anterior nucleus, the POP-1 level is low in E but high in MS. In the E cell, both *end-1* and *end-3* are transcribed under the combined control of SKN-1, the POP-1/SYS-1 complex, and MED-1/2, which ultimately leads to the intestinal differentiation. On the other hand, in the MS cell, a high level of POP-1 allows mesoderm development by repressing the endoderm promoting *end-1/3* genes (Lin *et al.*, 1998; Maduro *et al.*, 2005).

To elucidate developmental processes in *C. elegans*, emerging studies started to characterize the expression profiles of individual progenitors and differentiated cells. For example, the transcriptomes of a few early-stage *C. elegans* blastomeres have been recently reported using single-cell RNA-seq techniques (Hashimshony *et al.*, 2012). However, enormous technological development is still needed before the transcriptomes of every blastomere can be sequenced. Alternatively, Waterston and colleagues have developed a method to trace the cell lineages of embryos expressing fluorescent reporters, enabling the quantification of expression levels of a certain gene in individual cells on a nearly continuous time scale (Bao *et al.*, 2006; Murray *et al.*, 2008). Using this method, these

authors have recently measured the expression levels of more than 100 genes, mostly transcription factors (TFs), at single-cell resolution during embryogenesis up to the 350-cell stage (Mace *et al.*, 2013; Murray *et al.*, 2012). This comprehensive dataset at high temporal and spatial resolution should provide a great opportunity to connect gene expression with lineage specification in a systematic manner. Although significant results have been drawn in the original study, in this paper we explored this dataset by a new computational method from a different perspective. Although the number of genes measured is far from complete to describe the entire embryogenesis, we found that a few of these genes are sufficient to explain the divergence of many sister lineages. Our analysis revealed many known and novel candidate genes driving lineage and fate divergences in early blastomere divisions. Intriguingly, using a small sample of these identified genes, we were able to accurately differentiate the developmental paths leading to later lineages starting from a certain embryo stage. Therefore, our study provides new insights into the regulatory mechanisms governing lineage and fate divergence, and facilitates elucidating the functions of these genes in *C. elegans* embryogenesis. For convenience of discussion in this paper, we refer to a pair of lineages with a common mother cell as 'sister lineages', and relate them to their mother cell by calling them 'daughter lineages'. For example, ABal yields two sister blastomeres after its division: ABala and ABalp, which are the ancestor cells of the ABala and ABalp lineages, respectively. Therefore, the ABala and ABalp lineages are a pair of sister lineages and they are daughter lineages of ABal.

1.3 Material and Methods

1.3.1 Data Processing

The expression data of a total of 141 genes in 345 embryos were downloaded from the Expression Patterns in *C. elegans* (EPIC) database (http://epic2.gs.washington.edu/Epic2/) (Mace, *et al*., 2013; Murray, *et al*., 2012) and the expression levels normalized by the 'local normalization' method (Murray, *et al*., 2012) was used. To pursue a high confidence for the results, we excluded the cells with fewer than half genes (70 out of 141) measured in the dataset (mostly the cells appear in embryos later than the 350-cell stage), which resulted in 735 conceptual embryonic cells. We used the median expression level over all time points measured for a cell to represent the expression level of the gene in the cell.

We defined that a cell expresses a gene if its expression level exceeds the background level (defined as 2,000 units in the original paper). We excluded from the analysis the embryos having 5 or less than 5 cells expressing the tagged gene. For a gene with replicate embryos, we excluded embryos with inconsistent expressions (Supplemental Table S1). Specifically, we used *k*-means clustering method to partition embryos with the same tagged gene in two clusters (k=2) based on the number of cells expressing the gene in an embryo. If the centroid of one cluster is smaller than 0.3 fold of the other, we excluded all embryos in the former cluster.

There is a lag time between the onset of fluorescence and the expression commitment of the corresponding gene, since it takes a few cell cycles for GFP or mCherry to mature before they emit fluorescence (Murray, *et al.*, 2012). This may affect the ability and sensitivity to identify differentially expressed genes. It has been shown that the histone::mCherry reporter protein is highly stable with a lifetime longer than a few cell

cycles (Murray, *et al*., 2012). Thus, its fluorescence persists even if the corresponding native gene is turned off. We took advantage of this fact to overcome the artifacts caused by the delay of fluorescence. Specifically, for each cell and gene, we used the median of the reporter levels in the cell and all its descendants to represent the expression level of the gene in the cell. The modified gene expression levels may better reflect a gene's actual onset time and expression levels. In addition, the method may reduce the effect of inconsistent gene expression caused by technical variability. Another concern is that there are many expression values below the background level (2000) in the data set, which could cause false positives in selecting differentially expressed genes. Therefore, we reassigned them to 2,000 to keep the consistency and continuity of the data (Supplemental Table S2).

1.3.2 Wilcoxon rank sum test and machine-learning classifiers

We used a combination of statistical and classification methods to identify sets of genes that best discriminate a pair of sister lineages. We excluded from this study the divisions leading to one or both sister lineages having fewer than 15 cells with expression data recorded, resulting in a total of 40 divisions up to the sixth round divisions starting from the zygote. The AB/P1 sister lineage pair is also excluded because the reporters of many ubiquitously expressed genes are not expressed in the P1 lineage. Firstly, we performed a non-parametric test, Wilcoxon rank sum test (equivalent to Mann-Whitney U test), to select genes differentially expressed in two sister lineages at a division. To minimize the effect of variations between individual embryos and gene expression noise, we applied a strict selection criterion: a gene is selected for a division only if in all replicate embryos the gene has significantly different ($p\text{-value}<0.05$) expression levels between the two sister lineages. The selected genes are referred to as *informative genes* in this paper.

Secondly, we employed machine-learning methods to further refine the selected informative genes that potentially drive lineage divergence. To this end, we computed the mean expression level for a gene in the same conceptual cell across replicate embryos, to represent the expression level of the gene in the conceptual cell. We employed a decision stump (Iba and Langley, 1992) to calculate the classification error rate for each informative gene in separating a pair of sister lineages. Next, we selected the genes with error rates smaller than 0.15 as *important genes* of a division. For each important gene, we assigned it to one of the two sister lineages in which it has a higher expression level than the boundary point (possible boundary location of two classes detected by the decision stump). To see how sister lineages could be distinguished by a group of genes collectively, we performed random forest (Breiman, 2001) using the TreeBagger command in MATLAB (R2013b). Before creating ensemble of bagged decision trees, we fixed the initial random seed. For each classification, a random forest was grown on 50 trees. The default setting of TreeBagger for classification was used. The minimal leaf size was set to 1 and the square root of total number of genes was selected for each split at random. We calculated the out-of-bag (oob) error to get an unbiased estimate of the classification error.

1.3.3 Information Content Reduction

To quantify the fate constitution of a blastomere, we counted the number of its terminal descendants that fall into each fate category and calculated an information content (IC) for the blastomere defined as $IC = -\sum_{k=1}^{9}\left(P(k)ln\left(P(k)\right)\right)$, where k stands for each fate category and $P(k)$ denotes the fraction of terminal cells within category k. The following nine fate categories were used for the analysis: glia, hypodermal, intestinal, body wall muscle, pharyngeal, arcade, rectal, seam, and neuron (except neurons in pharyngeal). Each

terminal cell can only be assigned to at most one category; hence, the categories are disjoint. The degree of asymmetry of a division was quantified by the averaged amount of IC reduced in two daughter cells compared to the mother cell.

1.3.4 The Relationship between Important Genes and the Developmental History of Lineages

Given a lineage tree derived from a root cell (e.g., AB), we want to compare the developmental paths for lineages that are generated after N (e.g. N=5) rounds of divisions, i.e., the sub-lineages starting from the (N+1)-th (e.g. 6th) level in the tree (e.g. ABalaaa, ABalaap, ABalapa, ABalapp, etc.). To achieve this goal, we first defined the developmental path from the root to a target lineage as a gene expression rule constituted by a series of binary decisions. At a division M along a path, we selected one gene g_i with the minimum error rate in classifying the two sister lineages and identified the classification boundary (value V_i). We assigned an IF function E_i= IF($g_i \leq V_i$) or E_i= IF($g_i \geq V_i$) to each sister lineage according to its expression level of g_i relative to V_i. The IF function returns TRUE when the specified condition is met. For a path L consisting of N divisions leading to a target lineage at the (N+1)-th level, we constructed a Boolean algebra R_L by serially applying the logical operator AND to connect all the E_i functions: $R_L = (E_1 \wedge E_2 \wedge \ldots \wedge E_N)$, which returns TRUE only when all the E_i functions evaluate to TRUE. As a result, each of the 2^N target lineages has a corresponding rule R. To reveal how different the 2^N paths are from each other in generating descendant lineages, we scored each target lineage T by each R, and compared the scores. To score a target lineage T by R_L, each cell in T is judged by R_L and the final score of T on R_L is equal to the number of cells that fail to follow R_L. Hence, the score could be any integer between zero and the number of cells in T. In this way we

converted the task of comparing paths to a well-known assignment problem. We solved it by adopting the Hungarian algorithm (MATLAB implementation), which finds the minimum weight matching of a bipartite graph and meanwhile optimizes the assignments by minimizing the sum of scores. Then from the assignment results, we could conclude how similar or diverse the paths are in defining target lineages.

1.4 Results

In principle, a lineage specification process is driven by a regulatory cascade that diversifies daughter cells at each division along the lineage path and ultimately leads to a complete cell differentiation and fate determination. In this work, we attempt to understand how the cell fates are conferred by a series of blastomere divisions by identifying candidate genes that potentially drive lineage divergence. Toward this goal, we identified putative determinant genes at every cell division along a developmental time-line. Specifically, we looked for genes that could best explain the divergence between two sister lineages yielded at each of the selected 40 divisions.

1.4.1 An Effective Quality Control Method Facilitates the Analysis of Single-Cell Gene Expression Data

All the analysis in this work are based on the EPIC data set (Mace, *et al*., 2013; Murray, *et al*., 2012), which were collected by tracing gene expression levels from 345 *C. elegans* embryos, each was constructed to express a reporter for a specific gene (Supplemental Table S1). Among the 141 reporter genes, 42 were only recorded in a single embryo without replicates; the rest 99 genes were measured in more than one (2-16) embryo. Expression levels of replicates are highly concordant, but for some genes there still exists variability. The original paper examined the consistency of gene expression levels from

reporters that were analyzed in multiple embryos, in regard to the number of strongly expressing cells (Murray, *et al*., 2012). They found that most reporters (~80%) were highly concordant. For the remaining reporters, the variability was largely due to the replicate embryos constructed from different strains, where one strain was overall less bright than the other. Alternatively, the onset of the reporters was near the end of the embryo stage, thus the detection was less reliable. In addition, some genes in EPIC were measured in two reporter constructs (promoter fusion and protein fusion). For most genes, the expression patterns of the two different reporters are similar but not identical (Murray, *et al*., 2012). Since expression levels from different embryos are inevitably variable owing to the stochastic nature of gene expression and technical variability, it is essential to control the data quality and minimize the technical variation before a meaningful conclusion can be drawn on the gene expressions.

Firstly, we intended to identify and exclude embryos which are from strains or constructs that behave differently and are much less bright than other replicates. The measurements of reporter intensity from these embryos would not be reliable. In addition, since we used a rigid background level (2000 intensity units) for all embryos, replicates with different overall reporter intensity should not be considered together. Therefore, using the quality control procedure detailed in MATERIALS AND METHODS, we filtered out the embryos with very few cells expressing the gene (>2000 intensity units) or significantly less cells expressing the gene compared to replicate embryos (Supplemental Table S1). As a result, 58 embryos from 31 genes were filtered out; 287 embryos from 131 genes were included for further analysis. Among the 131 genes, 86 genes have two to nine replicate embryos; the rest 45 genes have only a single embryo included, mostly (35/45) because of

a lack of replicate in the original data set. Although we did not exclude these 45 genes with a single recording from further analysis for completeness, we clearly labeled them throughout the following results to bring them to notice. Among the 58 embryos filtered out, 24% have no cell expressing; 50% have less than ten cells with expression over the background level; the rest embryos have moderate number of cells expressing the gene but significantly less than other replicate embryos, mainly because of different constructs or strains used. It is worth noting that, using unsupervised clustering method could differentiate strains or constructs that act differently. For example, *lin-26* was tagged in four embryos, each with 8, 15, 268 and 297 cells expressing the gene (level>2,000). The k-means clustering partitioned the four embryos into two clusters, [8, 15] and [268, 297]. Since the centroid of the former cluster (11) is significantly smaller (<0.3 fold) than that of the latter cluster (282), both embryos in the former cluster were filtered out. In fact, embryos in the former cluster were constructed from a different strain than the embryos in the latter cluster.

We next excluded genes with inconsistent expression patterns among replicate embryos and meanwhile selected informative genes that likely account for lineage divergence at a division. For each of the 40 divisions, we identified genes that are significantly differentially expressed by two sister lineages using Wilcoxon rank sum test in each embryo; we only considered a genes as an informative gene if all replicate embryos reach the significance level (p-value<0.05) at the division. As summarized in Supplemental Table S3, for most divisions, majority of the 131 genes do not satisfy this selection criterion and thus are excluded; the number of informative genes ranges from 15 (ABarppa/ABarppp) to 79 (EMS/P2) for a division. The sample size (number of cells in a lineage) positively

correlates with the number of informative genes (ρ=0.60), which is expected since a larger sample size would make it easier to reach the significance level. This additional filtering procedure ensures that the informative genes selected for a division are the genes with high reproducibility and consistent differential expression patterns (p-value<0.05) among all its replicates at the division.

1.4.2 Divergence of Sister Lineages is Coupled with Cell Divisions

During *C. elegans* embryogenesis, developmental pathways are largely directed by asymmetric cell divisions, which give rise to two sister lineages with diverse fates (Baugh, 2003; Cowing and Kenyon, 1996; Edgar and McGhee, 1988; Kaletta *et al.*, 1997). The divergence is executed and maintained by regulatory programs that are manifested as differential expression patterns between two sister lineages. To see whether the divergence of sister lineages could be explained by the selected informative genes, we adopted a supervised classification method, random forest, at each division. Here the classification objects are binary labeled cells according to which of the two sister lineages they belong to, and the features are the gene expression levels. As shown in Figure 1.1A, the error rates of classifications for the 40 pairs all drop rapidly as the number of decision trees increases and are eventually stabilized near zero. By contrast, when the class labels on the cells are randomly permuted, the random forest with the same settings fails to classify the two lineages, as its performance is no better than random guessing. In the case of the ABarppa/ABarppp sister lineages, the classification on permuted data yields an error rate around 0.5 and could not be brought down when more trees are grown (Figure 1.1B). To verify our criterion for selecting informative genes, we compared the classification results using all of the genes (131) to those using only the selected informative genes. As shown

in Supplemental Table S4, although the informative genes selected for each division are fewer, they generally have at least the same classification power as all the 131 genes. For example, only 15 out of the 131 genes are selected as informative genes for the ABarppa/ABarppp sister lineages. However, the error rate drops even faster using the 15 informative genes than using all the genes (Figure 1.1B). These results support the model that the transition of gene expression profiles is made about each cell cycle through asymmetric divisions in *C. elegans*. Moreover, our selection criteria can largely identify genes that likely contribute to the divergence of sister lineages.

1.4.3 Divergence of Sister Lineages Can be Explained by a Few Important Genes

To further narrow down the genes explaining the divergence of a pair of sister lineages, we evaluated each informative gene for its capability to separate a pair of sister lineages. We considered an informative gene as an important gene if it could discriminate a pair of sister lineages with an error rate less than 0.15. We assigned each important gene to one of the two sister lineages in which it has a higher expression level than the boundary point detected by the decision stump (Supplemental Table S5). About 40% of the informative genes are selected as important genes; the number of important genes identified for a division ranges from 0 (ABal/ABar and ABpl/ABpr) to 53 (MS/E). The inability to identify important genes for ABal/ABar and ABpl/ABpr is understandable as ABa and ABp are precursors of analogs, each producing a group of approximately equivalent cells that result in bilateral symmetry in the nervous system (Sulston, 1983). Remarkably, 108 (82.4%) of the 131 genes in our analysis are identified as important genes for at least one sister lineage pair. On average, a gene is identified as an important gene for about five divisions; about 10% of the genes are identified as important genes in more than 10 divisions. For example,

tbx-11 is identified as an important gene for 17 pairs of sister lineages (Supplemental Table S5).

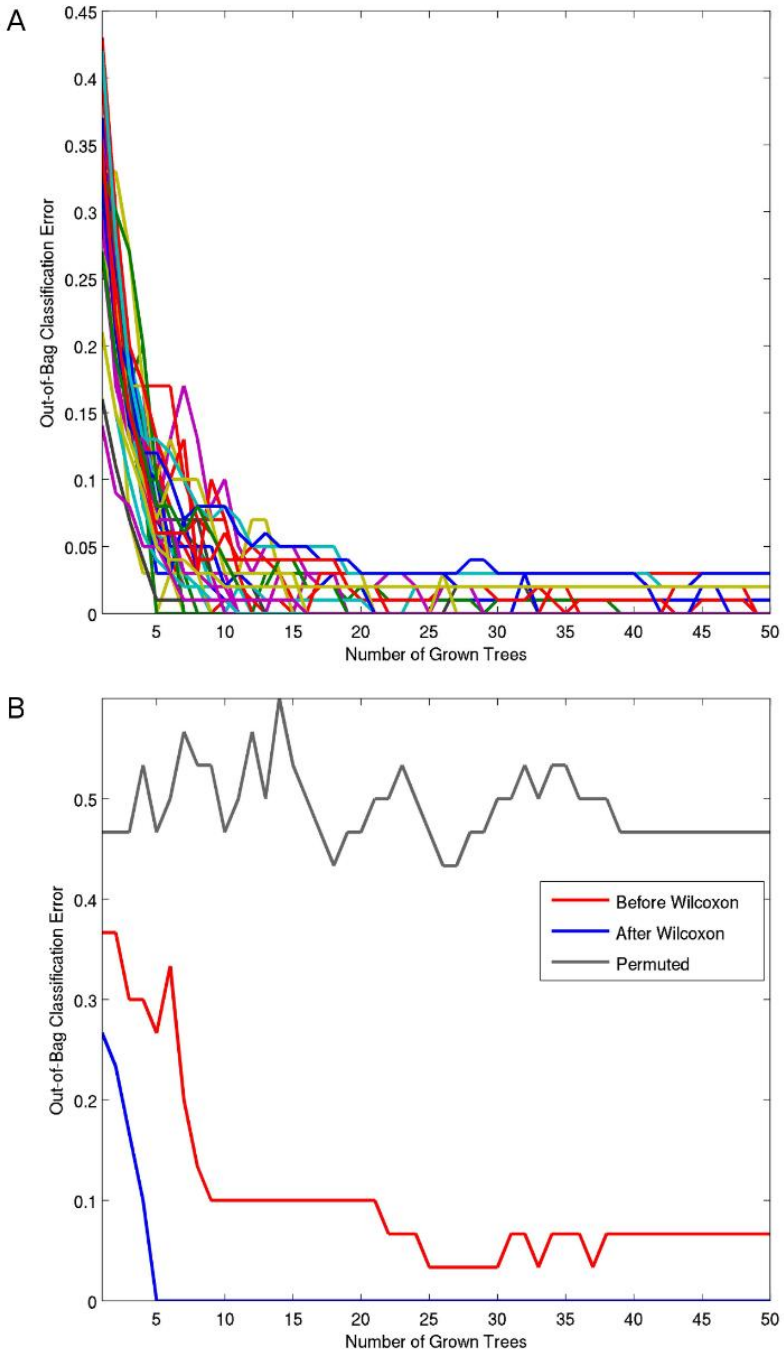

Figure 1.1. The classification error rate changes as the number of decision trees increase in the random forests. (A) The Out-of-Bag (OOB) error rates as a function of the number of decision tree grown in the random forests for classifying the 40 pairs of sister lineages.

Figure 1.1. (Continued) (B) The random forest can successfully differentiate ABarppa/ABarppp sister lineages using either all the 131 genes or only the 15 informative genes selected by Wilcoxon rank sum test. However, the classification fails when class labels on cells are permutated, as the OOB error rate fluctuates around 0.5 even when 50 trees are grown.

To illustrate how the important genes contribute to discriminating a sister lineage pair, we conducted principal component analysis (PCA) on the expression levels of the important genes (Supplementary Figure S1). Shown in Figure 1.2A is the case for the ABala/ABalp sister lineages, in which 11 of the 43 informative genes are identified as important genes. Intriguingly, the cells from the ABala/ABalp pair are well separated into two clusters by the first and second principle components (PCs) of the 11 important genes (Figure 1.2A). In contrast, when the rest 32 genes with error rates over 0.15 are used for PCA, the resulting boundary between the ABala and ABalp lineages is much less clear (Figure 1.2B). In addition, as indicated earlier, although the sample size correlates positively with the number of informative genes selected by Wilcoxon rank sum test (ρ=0.60), it does not show such correlation with the number of important genes selected by decision stump (ρ=-0.21).These results suggest that decision stump is able to effectively identify genes making the biggest contribution to separating sister lineages, thereby accurately refining the selection of genes that potentially drive the lineage divergence.

Importantly, as shown in Supplemental Table S5, many of the important genes identified for discriminating sister lineages are in excellent agreement with their known functions. We illustrate this using a few well-studied examples. First, we have identified *med-2 and pal-1* to be important genes in discriminating the EMS/P2 sister lineage pair. Based on the expression level relative to the boundary point, we assigned *med-2* to the EMS lineage and *pal-1* to the P2 lineage. Consistent with these predictions, it has been shown that in the P2

lineage, the C and D fate specifications require the CAUDAL-like TF PAL-1 for body wall muscle and hypodermal development (Edgar *et al.*, 2001; Hunter and Kenyon, 1996). On the other hand, the body wall muscle cells derived from the MS cell in the EMS lineage do not depend on the activation of *pal-1* but *med-1/2*, which are necessary and sufficient to program mesendoderm development (Maduro *et al.*, 2001). Moreover, in *med* depleted embryos, EMS descendants adopt the C fate. Hence, our results are consistent with these early findings that the differentiation program of P2 is driven by *pal-1* whereas that of EMS is driven by *med-2*. However, since *med-1* is not measured in the original study, it is not included in our analysis here. Second, we identified the *med-1/2* target genes *end-1/3* and *tbx-35* as important genes in discriminating the MS/E sister lineage pair, and assigned *tbx-35* to the MS lineage and *end-1/3* to the E lineage. In consistent with these predictions, it has been shown that *end-1/3* are the earliest expressed genes in the E lineage and the END-1/3 regulation defines the separation of the E lineage from its sister lineage MS by contributing to the intestinal fate commitment (Zhu *et al.*, 1997). On the contrary, *tbx-35*, a T-box TF, is repressed in the E lineage but activated in the MS lineage to specify the MS-derived pharynx and body wall muscle fates (Broitman-Maduro *et al.*, 2006). Third, it has been reported that TBX-35 acts through regulators PHA-4 and HLH-1 to specify pharynx and muscle fates, respectively (Gaudet and Mango, 2002; Krause *et al.*, 1990; Lei *et al.*, 2009). PHA-4 first appears close to the time point at which a pharyngeal clonal forms and then present in all pharyngeal cells derived from the ABalp, ABara, MSaa and MSpa lineages (Horner *et al.*, 1998; Kalb *et al.*, 1998). HLH-1 is a potent myogenic factor whose expression is detected in the C, D and MS blastomeres and descendants (Fukushige *et al.*, 2006; Krause *et al.*, 1990). Our results are in good agreement with these early findings, as

we assigned *pha-4* to the ABalpa, ABara, MSaa and MSpa lineages which are all precursors of the pharynx tissue. Besides, *hlh-1* was assigned to MSap and Cpp lineages which are both muscle clones. Fourth, in addition to *hlh-1*, we also identified another previously described body wall myogenic factor gene *hnd-1* (Fukushige and Krause, 2005) as an important gene in the Cpp, MSap and MSpp lineages, which are developmental clones for body wall muscles. Fifth, previous studies have characterized *nhr-25* as a hypodermal tissue marker (Asahina *et al.*, 2000; Moore *et al.*, 2013) and we assigned it to the Cpa lineage, a clone fated to be hypoderm. Finally, we found *tbx*-37/38 as important genes in segregating the ABa lineage from the ABp lineage, which is in agreement with their known roles in mesodermal induction for specifying the pharynx fate in the ABa lineage (Good *et al.*, 2004). In conclusion, our results indicate that the important genes picked by our algorithm are likely to play a role in the lineage specification process; their divergent expression levels in two sister lineages might activate distinct transcription programs that drive further fate specification processes.

1.4.4 Effective Lineage Classifications are Associated with Fate Divergences

In essence, gene expression has a high correlation with cell fate, as terminally differentiated cells must express determinant/signature genes to fulfill certain functions. It would be interesting to evaluate the relationships between gene expression and fate divergences. We began by determining the degree of asymmetry at each division, since an asymmetric division that produces two daughter lineages with different developmental potentials is a general mechanism for fate specification. We computed an Information Content (IC) for each embryonic cell based on the tissue type constitution of its terminal descendants (Supplemental Table S6). Given that the IC could capture the differentiation

potential of a cell, the level of IC reduction in the daughter cells relative to their mother cell is a measurement of the degree of asymmetry of the division. For example, Cp generates eight hypodermal cells and 16 body wall muscle cells, having an IC of 0.64 nats. Its daughter cells, Cpa which generates only hypodermal cells (eight), and Cpp which generates only body wall muscle cells (16), both have an IC of 0 nats.

This significant reduction of IC from 0.64 to 0 is consistent with the fact that the division is highly asymmetric. In addition, the 0 IC of the daughter cells reflects the fact that both of them become clonal producing precursor cells of single fate. In contrast, the E cell (produces 20 intestinal cells) undergoes a symmetric division, as its daughter cells Ea and Ep carry no difference in their progeny constitution (each produces 10 intestinal cells). In agreement with the symmetric division, there is no IC reduction in both of the daughter cells (E, Ea and Ep all have an IC 0). Therefore, the IC reduction can quantify the degree of asymmetry for a division.

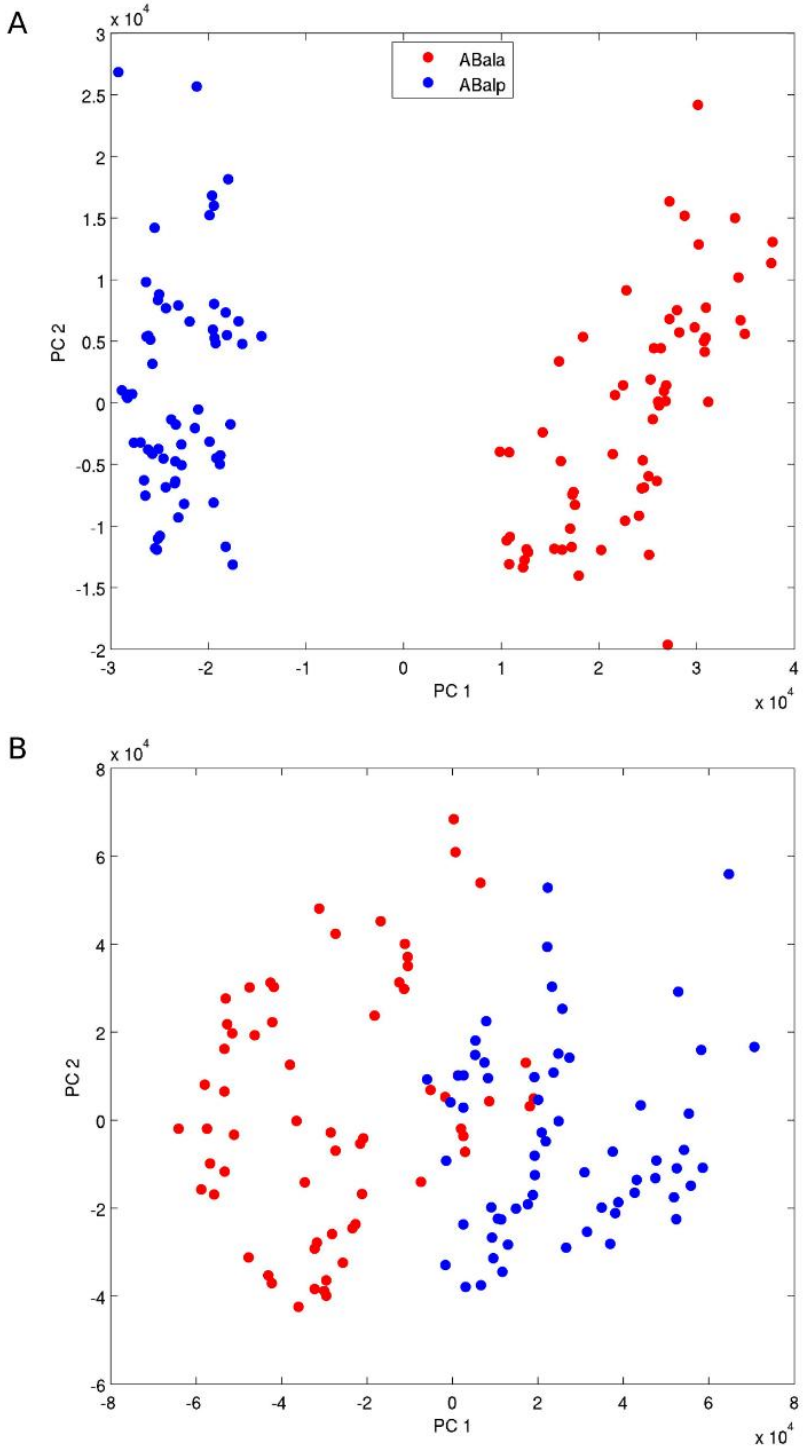

Figure 1.2. Few important genes can explain the divergence between sister lineages. (A) Projections of cells in the ABala/ABalp lineages on the first (x-axis) and second (y-axis) principle components of the 11 important genes with error rates below 0.15. Cells are colored according to their lineage origin (ABala: red, ABalp: blue). (B) PCA projections using the rest 32 informative genes with error rates over 0.15.

We further found that the number of important genes of a sister lineage pair is positively correlated (ρ=0.49) with the mean IC reduction in its daughter cells (Figure 1.3). The fact that many genes are identified to be important genes at some divisions is associated with functional asymmetry of the sister lineages. For example, among all of the 40 classifications, MS/E has the largest number of important genes (53). In fact, MS and E sister lineages are derived from a highly asymmetric division and the mean IC reduction (0.71) is the highest among all the 40 divisions. By contrast, the fact that few genes are identified to be important genes at some divisions are associated with functional symmetry of sister lineages. The divisions with the lowest number of important genes identified (0~3) are the ones that yield functionally symmetric lineages, such as MSa/MSp, Ca/Cp, ABpl/ABpr, and ABal/ABar. In the 13 cases where the numbers of important genes are below 10, the divisions have relatively low mean IC reductions (0.12 in average), while in the remaining 27 cases with equal to or more than 10 important genes, the divisions have relatively high mean IC reductions (0.20 in average). These results indicate that the important genes identified at asymmetric divisions might play roles in fate specifications. However, as shown in Figure 1.3, there are also some relatively symmetric divisions having many important genes. In these cases, the important genes might act as upstream regulators functioning before the asymmetric divisions take place and the fate divergence can be observed.

1.4.5 Lineage Identities Can be Uniquely Defined by a Few Important Genes

It is generally believed that lineage specifications in *C. elegans* are largely achieved by a series of stereotyped cell divisions from the zygote through to terminally differentiated cells. We hypothesize that the specification/differentiation pathways are uniquely defined

for each lineage, in which a cascade of binary decisions made at ancestral nodes/cells collectively define and restrict the characters of descendant cells. If this is the case, we wonder what genes are required to uniquely distinguish the developmental paths leading to descendant lineages. We addressed this question using the AB lineages, as we have learned the important genes at all the divisions up to the fifth round starting from AB (e.g., the divisions of AB, ABa, ABal, ABala and ABalaa). We compared the developmental paths of the 32 lineages derived from the fifth round of AB divisions (i.e., the lineages of ABalaaa, ABalaap, ABalapa, etc., and we refer to them as the AB32 lineages hereafter). Specifically, for each developmental path from AB to an AB32 lineage, e.g., the ABalaaa lineage, we defined a corresponding gene expression rule consisting of binary decisions of one important gene (with the lowest classification error rate) at each of the five divisions along the path. For example, the path from AB to ABalaaa is collectively defined by the expression patterns of the five important genes at their respective divisions, i.e., divisions of AB, ABa, ABal, ABala and ABalaa. Because some genes were repeatedly identified as important genes at different divisions, we only used a total of 14 important genes, all measured and analyzed in multiple embryos, to define all the 32 rules (Supplemental Table S7). We scored each AB32 lineage by each rule to see how well the 32 lineages could be differentiated from one another by the scores using the Hungarian assignment algorithm (See MATERIALS AND METHODS). Remarkably, the lineages are all correctly assigned to their corresponding paths. This finding demonstrates that the combined expression patterns of a small set of genes are sufficient to distinguish developmental paths starting from a common ancestor cell.

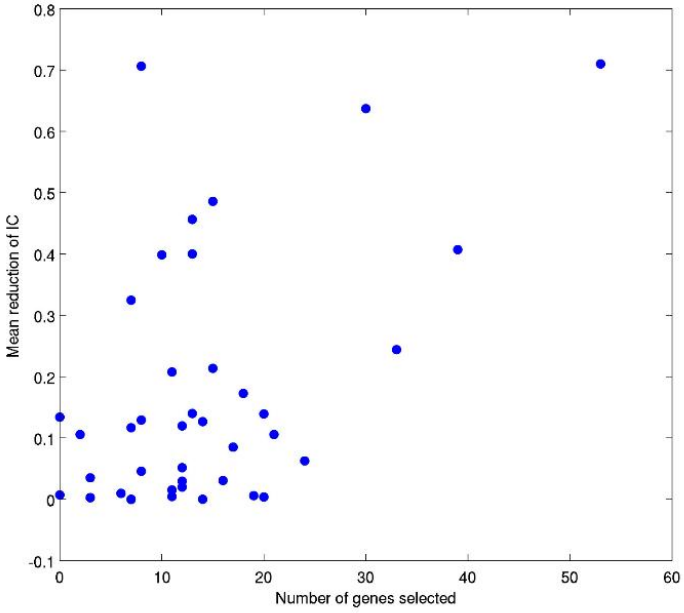

Figure 1.3. Number of important genes is positively correlated with the degree of asymmetry of a division. The x-axis is the number of important genes selected at a division. The y-axis is the mean IC reduction of the division.

1.5 Discussion

In this study, we reanalyzed the EPIC data set containing expression levels for 141 genes measured in individual cells in *C. elegans* embryos up to the ~350-cell stage. Although the relatively small number of genes and short period of monitoring time are clearly not sufficient to explain all developmental details occurring during embryogenesis, this large-scale dataset provides a substantial resource for analyzing the relationships between gene expression and lineage specification in early embryogenesis. However, careful data processing must be taken in order to draw meaningful conclusions due to the unavoidable technical artifacts in the data set. Specifically, since there is a time lag of about 30mins for a detectable fluorescence after the expression commitment (Murray, *et al.*, 2012), a modification of the expression levels is necessary to overcome the delay. Considering the

fact that the fluorescent proteins are highly stable after being produced, and that newly produced cells do not grow in size during embryogenesis, the concentration of a reporter in a cell is largely the same as in its descendant cells if divisions are symmetric. On the other hand, if divisions are not necessarily symmetric, the concentration of the reporter in the cell is roughly the median of the reporter levels in the descendant cells. Thus we used the median level of the reporters in the cell and all its descendants to infer the concentration of the corresponding gene in the cell. In principal, using the median could also reduce the effect of technical variability in the measurements of fluorescence levels. Indeed, we found that this modification largely enhanced the sensitivity to identify differentially expressed genes by the Wilcoxon rank sum test compared to using the direct measurements (data not shown), suggesting that the modification could largely correct the delay of fluorescence without introducing considerable new artifacts.

Moreover, studies of differential gene expression typically select causal genes according to fold change or test statistic rank. However, they both have limitations for this study. First, in the fold-change method, a gene is considered to be significantly differentially expressed if the ratio of its mean expression level in one sample to that in the other sample exceeds an arbitrary cutoff value (Cui and Churchill, 2003). One drawback using fold-change is the lack of associated statistic indicating the confidence level of the results. Moreover, when genes are expressed at low levels, fold-change suffers from high noise ratio which can result in high false discovery rates (FDRs). It may also fail for highly expressed genes, where small changes may be real but are rejected (Cui and Churchill, 2003; Tusher *et al.*, 2001). Second, since there are only few replicate embryos for most genes in the dataset, to fully utilize the invaluable information in each embryo, we chose

not to simply take averages of replicates followed by a ranking of genes by p-value. Instead, we applied statistical hypothesis testing in each embryo, and selected a gene as an informative gene only if all replicates reach the significance level. This reduces the effect of variations, thereby ensuring genes selected have a high consistency of expression patterns among replicate embryos. However, we wanted to only consider genes that could have sister lineages well separated with a clear boundary. Therefore, we further narrowed down the genes that are likely to drive the divergence of sister lineages according to their classification error rates. We resorted to the feature selection method, decision stump, to identify the important genes for each sister lineage. The task of feature selection is a widely addressed problem, where one has class-labeled data and wants to figure out which features best discriminate among the classes. Here the classes are lineages and the features are gene expression levels. The task is to select sets of genes that can best discriminate sister lineages. After applying an error rate cutoff in gene selection, the number of important genes is no longer correlated with the sample size (the number of cells in lineages). Besides, PCA projections on important genes result in a larger distance between two sister lineages compared to PCA on other informative genes, suggesting that important genes contribute significantly to the divergence of sister lineages. In summary, our results demonstrate that the combination of statistical hypothesis testing and decision stump could lead to better results than what would be resulted by using either one alone.

To evaluate the selected genes for discriminating a pair of sister lineages, we performed classification analyses on the sister lineage pairs using random forest, which is a strong classifier working by growing many decision trees and choosing the classification as the mode of outputs of individual trees (Breiman, 2001). Our successful classification of the

sister lineages at each division strongly support the model that cell lineage specification in *C. elegans* occurs sequentially through a cascade of binary decisions with each division diverging the daughter cells further, eventually leading to complete differentiations of terminal cells. Moreover, we found that some genes were repeatedly identified as important genes in various divisions. It raises an interesting question of whether their asymmetric expressions could be downstream effects of the transcriptional regulation by the POP-1/TCF, which is part of a general anterior/posterior coordinating system that acts in an iterative manner to differentiate sister cells (Lin *et al.*, 1998). Based on the classification boundary point set by the decision stump, we assigned each important gene to the sister lineage where it shows a higher expression. This allows a further investigation of the correlation between gene expression and lineage fates. It is worth noting that some important genes are clonally expressed in one sister lineage (bold typed in Supplemental Table S5). It would be interesting to reveal their roles in lineage/fate specification by either gain-of-function or loss-of-function experiments.

In *C. elegans*, multiple lines of evidence support the strong effect of lineages on fate specification (reviewed in Maduro, 2010). Thus, in our study, we sought to associate differentially expressed genes with fate divergent processes. It has been observed that organ/tissue identity genes are active at as early as 50- or 80-cell stage well before any overt cell differentiation and from then on their activities are maintained through adulthood (Labouesse and Mango, 1999; Maduro, 2010). In this regard, although the EPIC is limited to the 350-cell stage and does not include the final round of divisions when complex cell migrations and tissue/organ formations take place, the dataset can be a useful resource to reveal possible genes related to fate specifications. To identify novel genes related to fate

divergence, we determined whether or not a division is likely to be asymmetric based on IC reductions in the daughter cells. Indeed, we found a positive correlation between the number of important genes identified and the degree of asymmetry of a division. The result is expectable based on the knowledge of the relationship between gene expression patterns and cell fates that, if majority of the progenies in sister lineages are fated to the same tissue, the gene expression in the two lineages should be similar. The result also endorses our method for identifying important genes, and indicates that the EPIC data set contains many genes related to fate specification processes. Furthermore, we found that many of the important genes selected at divisions with relatively high IC reductions (>0.3) are known for their roles in initiating the fate specification processes in the corresponding lineages. We suppose those of unknown functions at such divisions to be novel lineage-specifying regulators and thus warrant further experimental investigations. However, the method is not valid for tissues without a clear lineage relation. One example is the nervous system where the majority of neurons are derived non-clonally from multiple lineages (Hobert, 2005). It is also worth noting that the IC is calculated based on an arbitrary and incomplete categorization of cell types, and the major tissue categories could be further subdivided. For example, pharynx is constituted by multiple types of cells such as neurons, epithelial and muscle cells. Clearly, different categorizations of the cells can lead insights into different aspects of the lineage and fate specification programs.

We modeled the developmental paths as a binary decision tree, with each level of divisions separating the lineages further. In this way, the entire tree can be built into a complete scoring system, where all the internal decisions contribute to the scoring rule for the lineages at the target level. For example, the expression pattern of the ABalaaa lineage

(6th level) must follow all the binary decisions made at the division of AB (1st level), ABa (2nd level), ABal (3rd level), ABala (4th level) and ABalaa (5th level). We found that a small number of genes (14) are sufficient to uniquely define all the developmental paths derived from a common ancestral cell. This result is in excellent agreement with the notion that lineage histories play crucial roles in cell fate specifications. In other words, the cell fates are specified by stepwise instructions directed by the binary decisions occurring at each division leading to the cell. We expect that with more single-cell gene expression data available, more biological insights into cell fate specifications can be revealed by similar analyses.

1.6 Conclusions

In the classic lineal control model of *C. elegans* embryo development, most blastomere and terminal identities stem from consecutive binary diversifications (Kaletta *et al.*, 1997). It would be highly valuable to dissect the architecture of regulatory cascades and reveal genes that play essential roles in driving the divergence of two lineages generated at each cell division. The major challenge to the goal is how to develop an effective approach to analyze highly noisy single-cell gene expression data with no or few replicates. Using a combination of careful data processing, non-parametric statistical test and classification methods, we were able to identify potential genes that distinguish sister lineages generated in the early embryogenesis in *C. elegans*. Intriguingly, we found that only a small set of genes is sufficient to discriminate a pair of sister lineages. With the availability of single-cell expression data for more genes and cells in later embryogenesis, more biological insights into cell lineage/fate specification can be revealed by similar analysis. Such

decoding of regulatory architecture during embryogenesis can eventually lead to a comprehensive understanding of the lineage/fate specification processes in embryogenesis.

CHAPTER 2: IDENTIFICATION OF CELL TYPES FROM SINGLE-CELL TRANSCRIPTOMES USING A NOVEL CLUSTERING METHOD

2.1 Abstract

The recent advance of single-cell technologies has brought new insights into complex biological phenomena. In particular, genome-wide single-cell measurements such as transcriptome sequencing enable the characterization of cellular composition as well as functional variation in homogenic cell populations. An important step in the single-cell transcriptome analysis is to group cells that belong to the same cell types based on gene expression patterns. The corresponding computational problem is to cluster a noisy high dimensional dataset with substantially fewer objects (cells) than the number of variables (genes). In this paper we describe a novel algorithm named SNN-Cliq that clusters single-cell transcriptomes. SNN-Cliq utilizes the concept of shared nearest neighbor that shows advantages in handling high dimensional data. When evaluated on a variety of synthetic and real experimental datasets, SNN-Cliq outperformed the state-of-the-art methods tested. More importantly, the clustering results of SNN-Cliq reflect the cell types or origins with high accuracy. The algorithm is implemented in MATLAB and Python. The source code can be downloaded at http://bioinfo.uncc.edu/SNNCliq.

2.2 Introduction

The recent advance of single-cell measurements has deepened our understanding of the cellular heterogeneity in homogenic populations and the underlying mechanisms (Kalisky and Quake, 2011; Pelkmans, 2012; Raser and O'Shea, 2004). With the rapid adaption of

single-cell RNA-Seq techniques (Saliba *et al.*, 2014), enormous transcriptome datasets have been generated at single-cell resolution. These datasets present a tremendous opportunity and challenge to the computational biology community for their analysis to reveal new insights into many biological problems, for example, to elucidate cell types in complex tissues. A straightforward approach to this problem would be to partition the cells into well separated groups via clustering techniques, so that cells (data points) in the same group exhibit similar gene expression levels (attributes). However, the high variability in gene expression levels even between cells of the same type (Buganim *et al.*, 2012; Guo *et al.*, 2010; Hashimshony *et al.*, 2012; Shalek *et al.*, 2013) can confound this seemly straightforward clustering approach. In addition, single-cell RNA-Seq data is generally in tens of thousands dimensions, which can substantially further complicate the clustering problem. In particular, usually only a few out of thousands genes are significantly differentially expressed in distinct cell types. Consequently, when clustering on the whole transcriptome, many genes would be regarded as irrelevant attributes and may even impede the identification of cell types.

It has been claimed that for a broad range of data distributions, the conventional similarities (such as Euclidean norm or Cosine measure) become less reliable as the dimensionality increases (Beyer *et al.*, 1999). The reason is that all data become sparse in high dimensional space and therefore the similarities measured by these metrics are generally low between objects (Beyer *et al.*, 1999). Accordingly, many clustering methods based on these measures are not effective enough for high dimensional data with few objects. An alternative similarity measure utilizes the ranking induced by a specified primary similarity. One commonly used secondary similarity is based on the notion of

shared nearest neighbor (SNN), which takes into account the effect of surrounding neighbor data points. More specifically, the similarity between a pair of data points is a function of their intersection of the fixed-sized neighborhoods determined by the primary measure (e.g. Euclidean norm). It has been demonstrated that in high dimensionality, SNN measures are more robust and result in more stable performances than the associated primary measures (Houle *et al.*, 2010). SNN techniques have been successfully applied to some clustering problems (Ertöz *et al.*, 2003; Guha *et al.*, 2000; Jarvis and Patrick, 1973). Inspired by these earlier applications, we define a new similarity between two data points based on the ranking of their shared neighborhood.

By representing data as a similarity graph in which nodes correspond to data points and weighted edges represent the similarities between data points, the clustering task can be achieved through partitioning the graph into homogeneous and well-separated subgraphs. That is, the nodes in the same subgraph have high interconnectivity, while nodes from different subgraphs have few connections in between. Several graph theory-based algorithms have been applied to clustering problems in earlier studies. One of the best-known graph-theoretic divisive clustering methods first finds the minimal spanning tree (MST), and then splits the tree by removing inconsistent edges with weights larger than the average in neighborhood (Zahn, 1971). Another algorithm called Chameleon first divides a graph into several subsets via a multi-level procedure, and then repeatedly combines these subsets to the ultimate clustering solution (Karypis *et al.*, 1999). However, the partitioning schemes used in these methods all require a prior knowledge of the number of subsets to be produced or the sizes of the partitions. Some other approaches avoid this problem by making assumptions about when to stop the recursive partition. For example,

the HCS clustering method (Hartuv and Shamir, 2000) defines a cluster as a highly connected subgraph (HCS) with a connectivity (the minimum number of edges to be removed to disconnect a graph) above half the number of nodes. The method iteratively cuts an unweighted graph using the minimum-cut algorithm until such subgraphs are produced. However, the algorithm produces many singletons for a sparse graph, although it includes a singleton adoption step. Besides, it does not separate clusters completely for certain data structures in our hand (see below).

To overcome the limitations of these existing algorithms, we developed a quasi-clique-based clustering algorithm inspired by our earlier work (Zhang *et al.*, 2009) to identify tight groups of highly similar nodes that are likely to belong to the same genuine clusters. Combining this algorithm with the SNN-based similarity measure, our method called SNN-Cliq is able to automatically determine the number of clusters in the data. Moreover, it can identify clusters of different densities and shapes, which is considered to be one of the hardest issues in clustering problems. Additionally, it requires few input parameters and finding a valid parameter setting is generally not hard. Most importantly, SNN-Cliq shows great advantages over traditional methods especially in clustering high-dimensional single-cell gene expression datasets.

2.3 Methods

By incorporating the concept of SNN in similarity measures, we model data as an SNN graph, with nodes corresponding to data points (e.g. vectors of gene expression levels of individual cells) and weighted edges reflecting the similarities between data points. We then find the ultimate clustering solution by using graph-theoretic techniques to cluster the

sparse SNN graph. The SNN-Cliq is carried out in the following steps and is schematically shown in Figure 2.1.

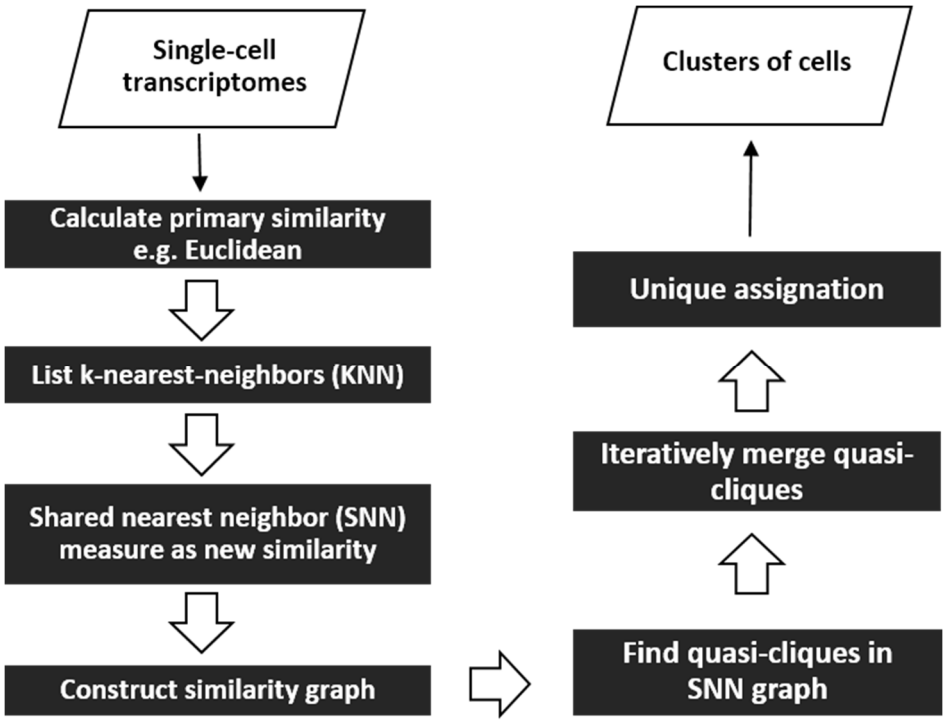

Figure 2.1. An overview of the SNN-Cliq algorithm.

Step 1: Construct an SNN graph

We first compute a similarity matrix using Euclidean distance (other suitable measures can also be used instead) between pairs of data points (e.g. a point is a cell and the distance between points is calculated using the vectors of gene expression levels in the cells). Next, for each data point x_i, we list the k-nearest-neighbors (KNN) using the similarity matrix, with x_i itself as the first entry in the list. To construct an SNN graph, for a pair of points x_i and x_j, we assign an edge $e(x_i, x_j)$ only if x_i and x_j have at least one shared KNN. The weight

of the edge e(x_i, x_j) is defined as the difference between k and the highest averaged ranking of the common KNN:

$$w(x_i, x_j) = \max\{ k - \frac{1}{2}(rank(v, x_i) + rank(v, x_j)) \mid v \in NN(x_i) \cap NN(x_j)\} \quad (1)$$

where k is the size of the nearest neighbor list, and $rank(v, x_i)$ stands for the position of node v in x_i's nearest neighbor list NN(x_i). Note that a closer neighbor v is higher ranked but the value of $rank(v, x_i)$ is lower. For example, $rank(x_i, x_i)$=1 because x_i is ordered first in x_i's nearest neighbor list.

Therefore, this SNN graph captures the similarity between two nodes in terms of their connectivity in the neighborhood. In other words, unlike the primary similarity, in our measure, the similarity between two nodes needs to be confirmed by their closeness to other nodes (common nearest neighbors). The rationale behind SNN is that the ranking of nodes is usually still meaningful in high dimensional space though the primary similarity might not (Houle *et al.*, 2010). The ranking of shared neighbors of two nodes in a genuine cluster is expected to be high, thus leading to a highly weighed edge. In contrast, the ranking of shared neighbors of two nodes from different clusters is expected to be low, resulting in a lowly weighted edge. Moreover, SNN graphs are usually sparse, thus allowing for scaling to large datasets.

Step 2: Identify clusters in the SNN graph

In a recent application, we proposed an algorithm for graph partition by finding maximal cliques (Zhang *et al.*, 2009). A maximal clique is a complete (fully connected) subgraph that is not contained in a larger clique. Although enumerating all the maximal cliques in a graph is an NP-hard problem, maximal cliques associated with each node can be efficiently found by a heuristic approach (Zhang *et al.*, 2009). However, cliques are rare in SNN

graphs due to the general sparsity. We instead search for quasi-cliques, which are dense enough but not necessarily complete. Our graph clustering method consists of two steps. Firstly, we extract local maximal quasi-cliques associated with each node in the subgraph induced by the node. We then construct clusters through merging these quasi-cliques and assigning nodes to unique clusters.

2.3.1 Find Quasi-cliques in the SNN Graph

Given an SNN graph, we use a greedy algorithm to find a maximal quasi-clique associated with each node (Figure 2.2). Firstly, for a subgraph S induced by a node v (S consists of v, all its neighbor nodes and associated edges), we find a dense quasi-clique in S. To this end, for each node s in S, we compute a local degree d as the number of edges incident to s from the other nodes in S. We select the s_i with the minimum degree d_i among all the nodes in S and remove s_i from S if $d_i / |S| < r$, where $|S|$ is the size of the current subgraph S and r is a predefined threshold ($r \in (0, 1]$). We then update d for the remaining nodes and repeat the process until no more nodes can be removed. If the final subgraph S contains more than three nodes, i.e. $|S| \geq 3$, we call it the quasi-clique for v.

After all possible quasi-cliques are found, we eliminate redundancy by deleting quasi-cliques that are completely included in other quasi-cliques. The parameter r defines the connectivity in the resulting quasi-cliques. A higher value of r would lead to a more compact subgraph, while a lower value of r would result in a less dense subgraph. One can try different values of r to explore the cluster structures or optimize the results, but we found that when r=0.7 the method performed well in all of the problems tested (see below). In fact, because of the following merging step, adjusting r in a certain range would not lead to substantial differences in the results.

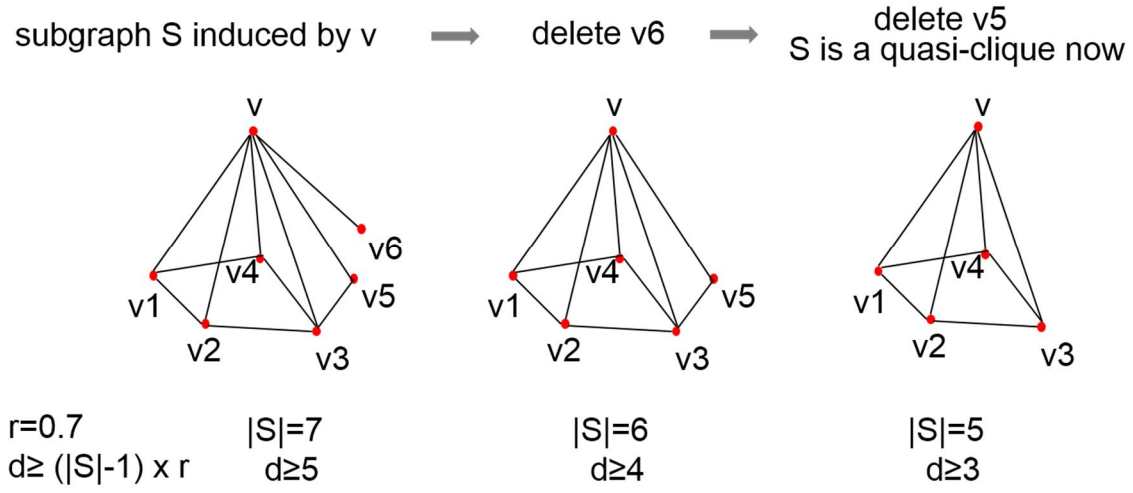

Figure 2.2. Finding quasi-clique associated with a node. A schematic to illustrate how to greedily find a quasi-clique associated with a node v in the subgraph induced by v and its neighbors. Initially, S includes seven nodes and v6 has the minimum degree (d_{v6}=1). Since a quasi-clique of r=0.7 requires a node to have at least 5 neighbors (d≥5), v6 is removed from S. In the new S with six nodes, the threshold of degree becomes d≥4. After v5 is deleted, all nodes (v, v1, v2, v3 and v4) connect to enough neighbors (d≥3) and S becomes a quasi-clique.

2.3.2 Identify Clusters by Merging Quasi-cliques

We identify clusters in the SNN graph by iteratively combining significantly overlapping subgraphs starting with the quasi-cliques. For subgraphs S_i and S_j, the overlapping rate $O_{i,j}$ is defined as the size of their intersection divided by the minimum size of S_i and S_j:

$$O_{i,j} = \frac{|S_i \cap S_j|}{\min(|S_i|, |S_j|)} \quad (2)$$

We initialize the set of subgraphs to be all the quasi-cliques and merge S_i and S_j if $O_{i,j}$ exceeds a predefined threshold m ($m \in (0, 1]$). In all the applications in this paper, we set m to 0.5. After each merging, we update the current set of subgraphs and recalculate pair-

wise overlapping rates if necessary. This process is repeated until no more merging can be made, and the final set of subgraphs are our identified clusters. Since a subgraph may overlap with multiple other subgraphs and merging in different orders may lead to distinct results, we give high priority to the pair with the largest total size $|S_i|+|S_j|$. In this way, a larger cluster is promised and would not likely be split into small ones.

2.3.3 Assign Nodes to Unique Clusters

The iterative merging stops when no pairs of clusters have an overlapping rate greater than m. However, the clusters may still have small overlaps, resulting in some nodes appearing in multiple clusters. However, for many problems such as clustering single-cell transcriptomes that we intend to address in this paper, one would prefer a 'hard clustering' (each data point belongs to exactly one cluster) over a 'fuzzy clustering' (each data point can belong to more than one clusters). To this end, for each candidate cluster C that the target node v is in, we calculate a score measuring the proximity between C and v, defined as the averaged weights on the edges incident to v from nodes in C:

$$Score(C,v)=\frac{1}{|C|}\sum_{i=1}^{|C|}w(c_i,v) \tag{3}$$

where c_i is a node in C. Then we assign v to the cluster with the maximum score and eliminate v from all the other candidate clusters. The assignation will change the cluster composition and may produce clusters with less than three nodes. In this circumstance, these data points are considered to be singletons. However, we did not observe such cases in our applications.

2.3.4 Time Complexity of the Algorithm

The most time-consuming step of SNN-Cliq is to construct the SNN graph, which requires $O(n^2)$ time, where n is the number of data points. Despite this, this step can be still

fast for single-cell transcriptome dataset, since n is usually quite small compared to the number of variables (genes/transcripts). The time complexity for finding a quasi-clique induced by a node is $O(d_v^2)$, where d_v is the degree of the node. Since d_v is usually much smaller than n in a sparse SNN graph, the entire cost of finding quasi-cliques for n nodes is bounded by $O(n)$. Moreover, this step can be easily accelerated by parallelization, since there is no data dependency in the process of finding quasi-cliques associated with each node. The merging step does not scale with n and is rather faster, since the overlaps of quasi-cliques only account for a small portion and are related to the cluster structures rather than n.

2.3.5 Validation Measures

We use three external validation measures, Purity, Adjusted Rand Index (Hubert and Arabie, 1985) and F_1 score (van Rijsbergen, 1974), to evaluate the performance of the clustering methods. Let U be the set of genuine classes (cell types) and V be the set of our computed clusters. Purity first assigns each cluster v_i to the class u_j that is the most frequent in the cluster. Then the total number of correctly assigned objects (cells) is divided by the total number of objects in the dataset (N):

$$Purity = \frac{1}{N}\sum_i \left(v_i \cap u_j\right). \tag{4}$$

ARI is one of the most successful measure of the agreement between two partitions with different number of classes/clusters. It is computed by:

$$ARI = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{N}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}, \tag{5}$$

where 'a' is the number of pairs of objects in the same class in U and the same cluster in V; 'b' is the number of pairs in the same class in U but not the same cluster in V; 'c' is the number of pairs that are not in the same class in U but in the same cluster in V; 'd' is the number of pairs that are neither in the same class in U nor in the same cluster in V. The F_1 score is the harmonic mean of precision and recall:

$$F_1 score = \frac{2 \cdot a}{2 \cdot a + b + c}. \quad (6)$$

Data points that are treated as noise or singletons are excluded from the calculation of Purity. In calculating ARI and F1 score, noise or singletons are treated as individual clusters.

2.3.6 Novelty of SNN-Cliq

The similarity measure in SNN-Cliq is based on the technique of shared nearest neighbor (SNN), which has been applied to several recent clustering applications (Ertöz *et al.*, 2003; Guha *et al.*, 2000; Jarvis and Patrick, 1973). Depending on the problem, different SNN similarity functions were proposed. For example, the similarity between objects x_i and x_j can be simply defined to be the intersection size of their k-nearest-neighbor list (Houle, *et al.*, 2010). Other functions take the ordering of the nearest neighbors into account. In a density based clustering approach, Ertöz, *et al*. (2003) took the sum of the similarities of a point's nearest neighbors as a local density measure: $strength(x_i, x_j)=\sum(k+1-rank(v, x_i))(k+1-rank(v, x_j))$, where v is a shared neighbor and $rank(v, x_i)$ is the position of v in x_i's list. In our paper, we define a new SNN function that only considers the ordering of the common neighbor that is on average the closest to x_i and x_j (the function is present in Methods 2.1). It emphasizes the closeness between points instead of the local density, thereby not discarding points in very low density regions. In addition, we believe

that this SNN function is more tolerant to changes in the parameter *k*. Finally, our function also extends the concept of SNN to construct a weighted similarity graph.

Furthermore, although the graph clustering step in the SNN-Cliq method is inspired by Zhang et al (Zhang *et al.*, 2009), they differ in many ways due to the differences of target graphs and ultimate goals. Zhang's method aims to cut down a large and dense graph to small parts for the purpose of computational efficiency in further steps; thus, it allows overlaps between resulting subgraphs. However, SNN-Cliq aims to partition a sparse graph into distinct clusters with no overlap in between. We delineate the differences between the two algorithms in the following three points.

First, Zhang's method starts by identifying cliques in a graph, because the graph it deals with is dense and large. By contrast, SNN-Cliq starts by searching for quasi-cliques that allow missing edges between nodes in a subgraph, because the graphs we deal with are usually sparse due to the similarity is calculated by shared nearest neighbor. Second, Zhang's method iteratively combine cliques/subgraphs by checking with two criteria: $|S1 \cap S2| / \min(|S1|, |S2|) > 0.9$ and $|S1 \cap S2| / \max(|S1|, |S2|) > 0.7$. As a result, S1 and S2 are only merged when the intersection size is large enough in both subgraphs. The high threshold (0.7 and 0.9) used will fail to merge many overlapping subgraphs, but this does not affect their results since their goal is to cut a dense graph instead of a hard clustering. In fact, their resulting subgraphs are still very dense and are similar to the quasi-cliques we find in the first step. In SNN-Cliq, we only require one criterion: $|S1 \cap S2| / \min(|S1|, |S2|) > 0.5$, to merge subgraphs. The purpose of this design is to allow the quasi-cliques to grow into non-spherical clusters. Finally, in the case of a node appearing in multiple clusters,

Zhang's method does not assign a node into a particular cluster. By contrast, SNN-Cliq always allocates a node to the nearest cluster to achieve hard clustering.

2.4 Results

2.4.1 Performance on Synthetic Datasets

Firstly, we illustrated the effect of the parameters on SNN graphs and clustering results using a synthetic two dimensional (2-D) dataset consisting of six perceptually distinct groups (two high-dense, two mid-dense and two low-dense clusters) (Figure 2.3A–C). The dataset was generated manually by randomly placing points on a 2-D space, and then the coordinates were retrieved. The class labels were given according to an intuitively good clustering way. Figure 2.3A–C show the resulting SNN graphs for k=5, 8 and 10, respectively. With the increase in k from 5 (Figure 2.3A) to 8 (Figure 2.3B), more edges were present in the SNN graph, connecting nodes in the same or from different clusters. However, in spite of the differences in the SNN graphs, clustering outputs stayed the same (six clusters). When k became even greater than the average size of the clusters (k=10 in Figure 2.3C), the method started to combine similar clusters in the low- to mid-dense regions. We further systematically evaluated k on a wide range (k=3–25) (Figure 2.4A). The minimum value of a valid k is three, because a node needs at least two other neighbors to form a quasi-clique. When k was too large ($k \geq 9$), clusters might not be thoroughly separated; on the other hand, when k was too small (k=3 and 4), a genuine cluster might be split into parts (Figure 2.4A). These results demonstrate that SNN-Cliq is relatively robust with respect to the changes in k to a certain extent. A valid choice of k depends on both the size and density of data. In general, a large and high-density dataset usually requires a relatively high k value compared to a sparse and low-density dataset. The parameters r and

m both control the compactness of subgraphs, thus can be used to adjust the granularity of resulting clusters (Figure 2.4B–E). Altering *r* or *m* usually has the same effect. As shown in Figure 2.4B–E, the correct clustering could be achieved by many different combinations of *k*, *r* and *m* settings; however, when *r*=0.7 and *m*=0.5 the method had a higher tolerance to changes in *k*. Therefore, in the following applications we set *r*=0.7 and *m*=0.5.

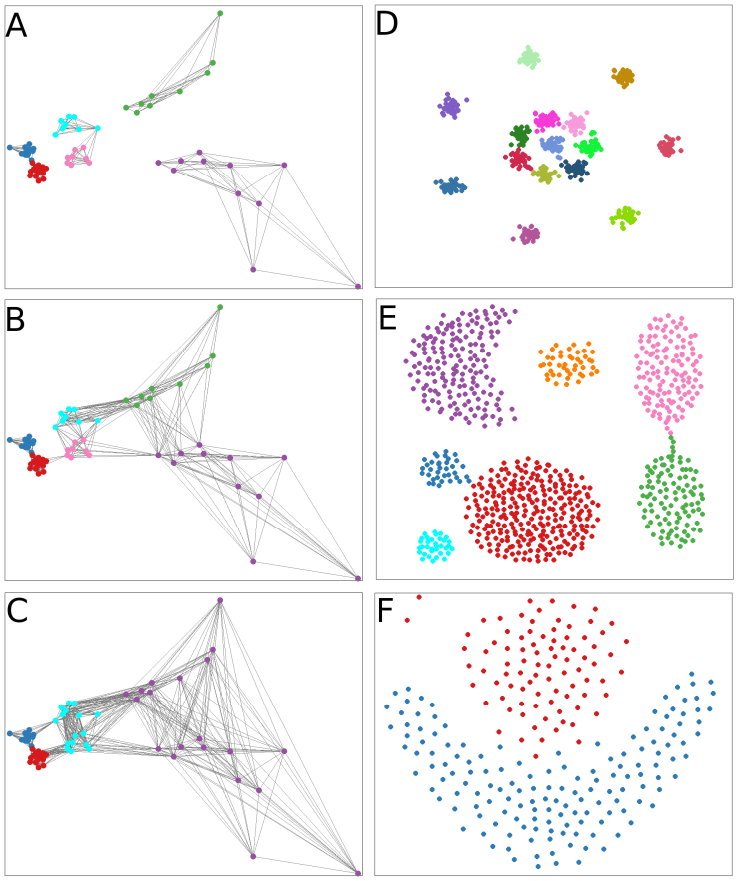

Figure 2.3. (A-C) Shared Nearest Neighbor (SNN) graphs constructed with k=5 (A), 8 (B) and 10 (C) for a synthetic 2-D dataset containing six perceptual clusters with high-, mid- and low- densities. Edge weights are not shown for clarity. (D-F) Performance of SNN-Cliq on three synthetic 2-D datasets with distinct structures. Dataset are from (Veenman *et al.*, 2002) (D), (Gionis *et al.*, 2007) (E), and (Fu and Medico, 2007) (F). Data points grouped in the same cluster by the algorithm are shown in the same color.

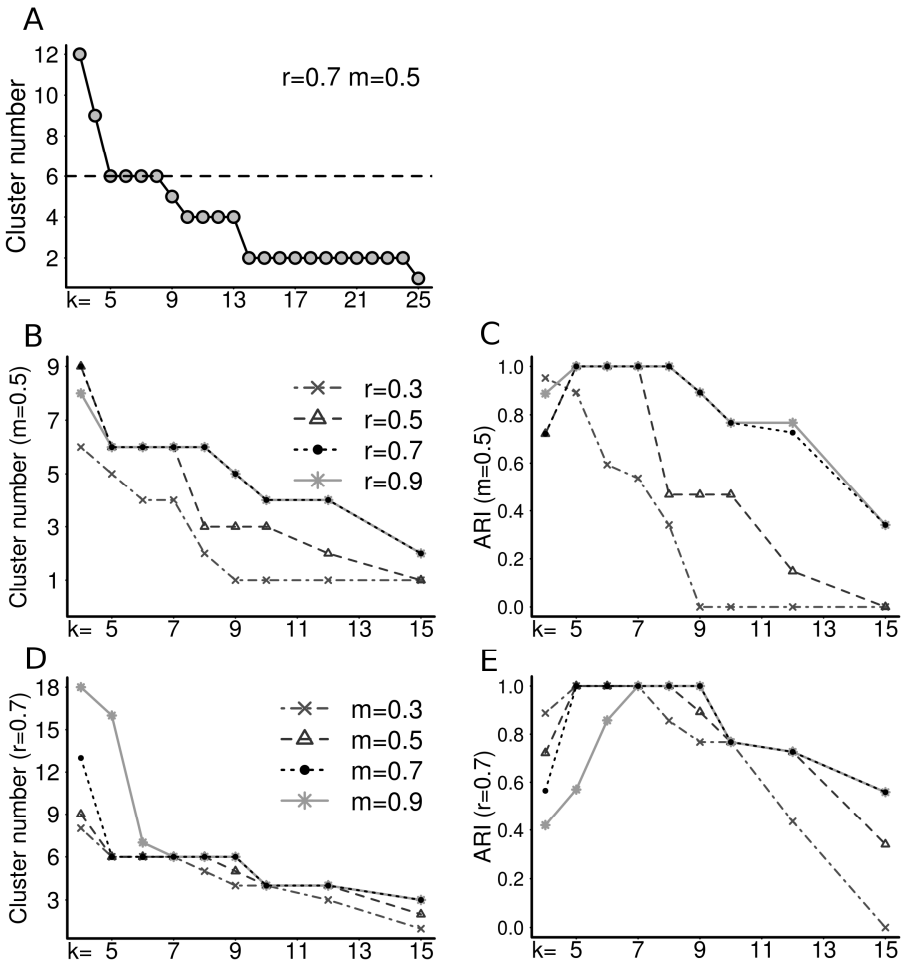

Figure 2.4. The effects of parameters on the clustering results of the synthetic dataset shown in Figure 2.1A. (A) The number of clusters detected as a function of k. (B–E) The number of clusters and Adjusted Rand Index (ARI) (see Supplementary Text for how it is calculated) at different parameter settings.

To demonstrate the applicability of SNN-Cliq, we tested it on several datasets with distinct structures presented in Figure 2.3D–F. The dataset shown in Figure 2.3D is composed of 15 similar 2-D Gaussian clusters that are positioned in rings (Veenman, *et al*., 2002). With *k*=15–35, we obtained the same correct clustering result as the original paper did (Veenman, *et al*., 2002). The dataset shown in Figure 2.3E contains clusters of arbitrary shapes and clusters connected by narrow bridges (Gionis, *et al*., 2007). SNN-Cliq

successfully determined the seven clusters as long as *k*=20–30. In contrast, applying HCS (from the RBGL package in R) (Carey *et al.*, 2011) to the SNN graphs failed to break the bridges, although a wide range of *k* was tested (Supplementary Figure S3A). The dataset shown in Figure 2.3F consists of two clusters with hardly defined border and shape, which represents a difficult case of clustering (Fu and Medico, 2007). Nonetheless, SNN-Cliq successfully separated the two distinct groups by breaking the bordering area with k=25, which agrees with an intuitively good clustering for this dataset. By contrast, using HCS on the SNN graph failed to give a result compliant with visual intuition (Supplementary Figure S3B).

2.4.2 Performance on Single-cell Transcriptome Datasets

It is generally believed that different cell types in multicellular organisms express distinct sets of genes, as is often manifested by traditional cell-population based assays. However, it has been shown that individual cells of the same type display inevitable cell-to-cell variations due to the stochastic nature of biochemical processes (Kalisky and Quake, 2011; Pelkmans, 2012). Such variability, also referred to as 'noise', makes the identification of the type of a cell on the basis of its transcriptome nontrivial. Moreover, as the small copy number of RNA molecules in a cell may lead to random loss of transcripts during library preparations, there is a notable technical noise in single-cell transcriptomes (Brennecke *et al.*, 2013). Therefore, we want to know whether or not individual cells could be grouped according to their cell types using the measured transcriptomes. We tested SNN-Cliq for such capability using three single-cell RNA-Seq datasets generated by different techniques in a variety of cell types in human and mouse (Deng *et al.*, 2014; Ramsköld *et al.*, 2012; Yan *et al.*, 2013). In the original papers, the authors have clustered the cells by hierarchical

clustering or projected the cells onto the first two principal components derived from a principal component analysis (PCA). Although these analyses revealed general relationships between cells, they lacked a clear grouping description of cells. To extend these studies and explore the valuable data further, we shall present the cell clustering results obtained by SNN-Cliq and compare them with those of two widely used clustering algorithms. One is K-means (MacQueen, 1967), a partition-based clustering technique that is suitable for spherical shaped clusters of similar sizes and densities. Another is DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester *et al.*, 1996), which clusters density-connected points and discards as noise the points having less than a user defined number (MinPts) of neighbors in a given radius (Eps). In addition, we shall compare our quasi-clique-based method with HCS in partitioning SNN graphs.

2.4.2.1 Human Cancer Cells

The first dataset was generated by Ramsköld *et al*. (2012) using a single-cell RNA-Seq protocol called Smart-Seq, which significantly improved read coverage across transcripts. The dataset includes transcriptomes of human embryonic stem cells hESC (n=8), putative melanoma CTCs (n=6) isolated from peripheral blood, melanoma cell lines SKMEL5 (n=4) and UACC257 (n=3), prostate cancer cell lines LNCap (n=4) and PC3 (n=4), and bladder cancer cell line T24 (n=4). We downloaded the normalized gene expression levels in RPKM (reads per kilobase of transcript per million mapped reads) from the Gene Expression Omnibus (GEO) database. Since technical variability in the measurements of gene expression levels becomes pronounced for lowly expressed genes due to random loss of transcripts (Ramsköld, *et al*., 2012), excluding such genes before analysis could enhance the reliability of results. As suggested by the original paper, we used genes with an

averaged RPKM≥20 for the analysis, involving 3,582 genes. To reduce the effects of highly expressed genes, we log-transformed the RPKMs, i.e. $\log_2(x+1)$. The gene expression variability is illustrated in Supplementary Figure S4. Because of the small number of cells in the dataset, we set k=3; r and m are at default values (r=0.7, m=0.5). As shown in Figure 2.5A, SNN-Cliq yielded six clusters, with five clusters each corresponding to a unique cell type and one cluster including cells of SKMEL5 and UACC257. However, both SKMEL5 and UACC257 are melanoma cell lines and the difference between them should be relatively small.

To compare our quasi-clique-based method with HCS in partitioning SNN graphs, we applied HCS on the same SNN graph. As shown in Figure 2.5A, HCS discarded four (shown in black) of the six CTC cells as singletons. To compare our entire algorithm with other methods in capturing the cell types, we applied K-means from MATLAB and DBSCAN from Python module scikit-learn-0.15.0 (Pedregosa *et al.*, 2011) to the log-transformed RPKMs, also with Euclidean norm as the similarity measure. Although K-means was performed with the correct parameter (K=7), the clusters found were either formed by cells of multiple types or a portion of cells of a certain type (Figure 2.5A). For example, CTC and SKMEL5 cells were all in one cluster, while hESC cells were partitioned into two different clusters. To give DBSCAN some advantages, we tried different sets of parameters (MinPts, Eps) and reported the one giving the best result (MinPts=3, Eps=150). However, DBSCAN only found two different clusters; one cluster agreed with the type hESC and the other cluster was a mixture of six cell types (Figure 2.5A). We further compared these methods using three external evaluation measures, Purity, Adjusted Rand Index (ARI) and F_1 score (see Supplementary Text for how they are

calculated). As shown in Figure 2.6A, the performance of SNN-Cliq was better than the other methods in all the three measures.

2.4.2.2 Human Embryonic Cells

The second dataset was produced by Yan and colleagues using a single-cell RNA-Seq approach that showed high sensitivity and reproducibility (Yan, *et al*., 2013). The dataset includes transcriptomes of human oocytes and cells in early embryos at seven crucial developmental stages: metaphase II oocyte (n=3), zygote (n=3), 2-cell-stage (n=6), 4-cell-stage (n=12), 8-cell-stage (n=20), morula (n=16) and late blastocyst at hatching stage (n=30). For each stage, two to three embryos were used. We applied SNN-Cliq with the same parameterization as before (k=3, r=0.7 and m=0.5) to the log-transformed RPKMs of 19,591 known RefSeq genes with RPKM>0.1 in at least one cell. As shown in Figure 2.5B, SNN-Cliq successfully clustered the cells from the same developmental stages, except for a few cells being mixed into neighboring stages, i.e., two morula cells were placed in the 8-cell-stage cluster and four 4-cell-stage cells were placed in the 2-cell-stage cluster. SNN-Cliq partitioned the 8-cell-stage cells into three different clusters. Intriguingly, the splitting reflects their distinct embryo origins (embryo 1, 2 and 3), as cells from the same embryo form their own cluster. It indicates the notable differences between individual embryos at this developmental stage. Similarly, the morula cells were split into different clusters for the two embryos. Interestingly, morula cells from embryo#2 were further partitioned into two clusters, indicating that heterogeneous expression patterns and possible cell differentiations might have occurred at this stage.

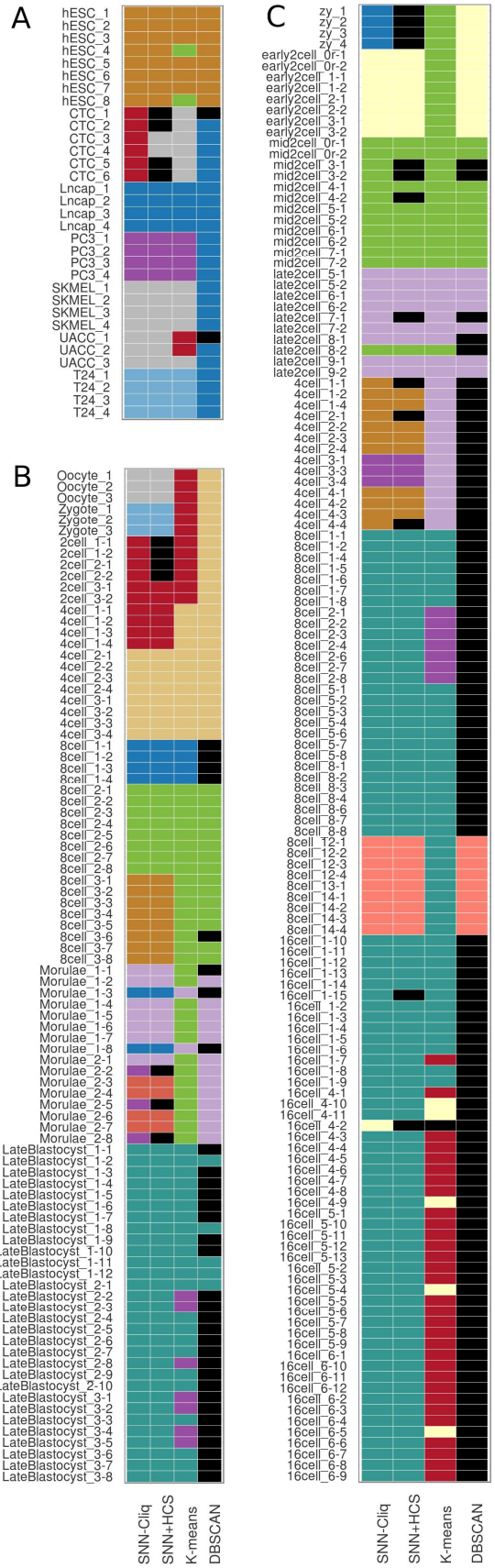
A
hESC_1
hESC_2
hESC_3
hESC_4
hESC_5
hESC_6
hESC_7
hESC_8
CTC_1
CTC_2
CTC_3
CTC_4
CTC_5
CTC_6
Lncap_1
Lncap_2
Lncap_3
Lncap_4
PC3_1
PC3_2
PC3_3
PC3_4
SKMEL_1
SKMEL_2
SKMEL_3
SKMEL_4
UACC_1
UACC_2
UACC_3
T24_1
T24_2
T24_3
T24_4
B
Oocyte_1
Oocyte_2
Oocyte_3
Zygote_1
Zygote_2
Zygote_3
2cell_1-1
2cell_1-2
2cell_2-1
2cell_2-2
2cell_3-1
2cell_3-2
4cell_1-1
4cell_1-2
4cell_1-3
4cell_1-4
4cell_2-1
4cell_2-2
4cell_2-3
4cell_2-4
4cell_3-1
4cell_3-2
4cell_3-3
4cell_3-4
8cell_1-1
8cell_1-2
8cell_1-3
8cell_1-4
8cell_2-1
8cell_2-2
8cell_2-3
8cell_2-4
8cell_2-5
8cell_2-6
8cell_2-7
8cell_2-8
8cell_3-1
8cell_3-2
8cell_3-3
8cell_3-4
8cell_3-5
8cell_3-6
8cell_3-7
8cell_3-8
Morulae_1-1
Morulae_1-2
Morulae_1-3
Morulae_1-4
Morulae_1-5
Morulae_1-6
Morulae_1-7
Morulae_1-8
Morulae_2-1
Morulae_2-2
Morulae_2-3
Morulae_2-4
Morulae_2-5
Morulae_2-6
Morulae_2-7
Morulae_2-8
LateBlastocyst_1-1
LateBlastocyst_1-2
LateBlastocyst_1-3
LateBlastocyst_1-4
LateBlastocyst_1-5
LateBlastocyst_1-6
LateBlastocyst_1-7
LateBlastocyst_1-8
LateBlastocyst_1-9
LateBlastocyst_1-10
LateBlastocyst_1-11
LateBlastocyst_1-12
LateBlastocyst_2-1
LateBlastocyst_2-2
LateBlastocyst_2-3
LateBlastocyst_2-4
LateBlastocyst_2-5
LateBlastocyst_2-6
LateBlastocyst_2-7
LateBlastocyst_2-8
LateBlastocyst_2-9
LateBlastocyst_2-10
LateBlastocyst_3-1
LateBlastocyst_3-2
LateBlastocyst_3-3
LateBlastocyst_3-4
LateBlastocyst_3-5
LateBlastocyst_3-6
LateBlastocyst_3-7
LateBlastocyst_3-8
SNN-Cliq
SNN+HCS
K-means
DBSCAN
C
zy_1
zy_2
zy_3
zy_4
early2cell_0r-1
early2cell_0r-2
early2cell_1-1
early2cell_1-2
early2cell_2-1
early2cell_2-2
early2cell_3-1
early2cell_3-2
mid2cell_0r-1
mid2cell_0r-2
mid2cell_3-1
mid2cell_3-2
mid2cell_4-1
mid2cell_4-2
mid2cell_5-1
mid2cell_5-2
mid2cell_6-1
mid2cell_6-2
mid2cell_7-1
mid2cell_7-2
late2cell_5-1
late2cell_5-2
late2cell_6-1
late2cell_6-2
late2cell_7-1
late2cell_7-2
late2cell_8-1
late2cell_8-2
late2cell_9-1
late2cell_9-2
4cell_1-1
4cell_1-2
4cell_1-4
4cell_2-1
4cell_2-2
4cell_2-3
4cell_2-4
4cell_3-1
4cell_3-3
4cell_3-4
4cell_4-1
4cell_4-2
4cell_4-3
4cell_4-4
8cell_1-1
8cell_1-2
8cell_1-4
8cell_1-5
8cell_1-6
8cell_1-7
8cell_1-8
8cell_2-1
8cell_2-2
8cell_2-3
8cell_2-4
8cell_2-6
8cell_2-7
8cell_2-8
8cell_5-1
8cell_5-2
8cell_5-3
8cell_5-4
8cell_5-6
8cell_5-7
8cell_5-8
8cell_8-1
8cell_8-2
8cell_8-3
8cell_8-4
8cell_8-6
8cell_8-7
8cell_8-8
8cell_12-1
8cell_12-2
8cell_12-3
8cell_12-4
8cell_13-1
8cell_14-1
8cell_14-2
8cell_14-3
8cell_14-4
16cell_1-10
16cell_1-11
16cell_1-12
16cell_1-13
16cell_1-14
16cell_1-15
16cell_1-2
16cell_1-3
16cell_1-4
16cell_1-5
16cell_1-6
16cell_1-7
16cell_1-8
16cell_1-9
16cell_4-1
16cell_4-10
16cell_4-11
16cell_4-2
16cell_4-3
16cell_4-4
16cell_4-5
16cell_4-6
16cell_4-7
16cell_4-8
16cell_4-9
16cell_5-1
16cell_5-10
16cell_5-11
16cell_5-12
16cell_5-13
16cell_5-2
16cell_5-3
16cell_5-4
16cell_5-5
16cell_5-6
16cell_5-7
16cell_5-8
16cell_5-9
16cell_6-1
16cell_6-10
16cell_6-11
16cell_6-12
16cell_6-2
16cell_6-3
16cell_6-4
16cell_6-5
16cell_6-6
16cell_6-7
16cell_6-8
16cell_6-9
SNN-Cliq
SNN+HCS
K-means
DBSCAN

Figure 2.5. Comparison of the clustering results from different algorithms on the human cancer cell dataset (Ramsköld *et al*., 2012) (A), human embryonic cell dataset (Yan *et al*., 2013) (B) and mouse embryonic cell dataset (Deng *et al*., 2014). In the heatmap, each row stands for an individual cell; each column corresponds to the clustering result produced by one of the four methods. Cells that are grouped in the same cluster by a method are displayed in the same color in the column. Cells that are treated as noise or singletons by the method are shown in black in the column. The embryo origins of cells from the same stage are distinguished by the first number in the cell names.

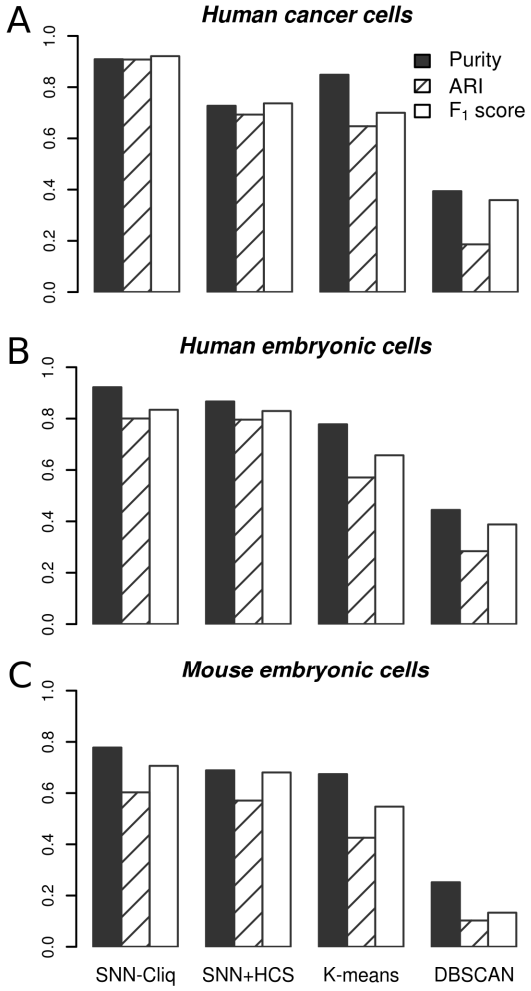

Figure 2.6. Evaluation of clustering algorithms by external validation measures, Purity, ARI and F_1 score. The gold standard of classes is determined by cell types or developmental stages. For mouse embryonic cell dataset, gold standard also considers the library preparation technique (Smart-Seq or Smart-Seq2).

Applying HCS to the SNN graph yielded very similar results to our graph clustering method (Figure 2.5B). However, it failed to recover the 2-cell-stage because most cells at this stage were discarded as singletons (shown in black in Figure 2.5B). Although K-means was conducted with the correct parameter (K=7), it lumped all the cells from oocyte, zygote and 2-cell-stage into a single cluster, and failed to differentiate morula and 8-cell-stage (Figure 2.5B). The results given by DBSCAN (MinPts=5, Eps=150) were not compliant with the cell identities in most of the cases; furthermore, a large number of cells, in particular the late blastocyst cells, were assigned to noise (Figure 2.5B). Evaluations using objective measures also show that SNN-Cliq outperformed the other methods (Figure 2.6B).

2.4.2.3 Mouse Embryonic Cells

The last dataset was generated by Deng and colleagues (Deng, *et al*., 2014) using Smart-Seq (Ramsköld, *et al*., 2012) or its updated form Smart-Seq2 (Picelli *et al.*, 2013). The dataset consists of transcriptomes for individual cells isolated from mouse (CAST/EiJ x C57BL/6J) embryos at different preimplantation stages. We obtained RPKMs for a total of 135 cells from GEO, including zygote (n=4), early 2-cell-stage (n=8), mid 2-cell-stage (n=12), late 2-cell-stage (n=10), 4-cell-stage (n=14), 8-cell-stage (n=37) and 16-cell-stage (n=50). A total of 19,703 RefSeq genes with RPKM>0.1 in at least one cell were included for the analysis. We conducted SNN-Cliq with the same parameter setting as before (k=3, r=0.7 and m=0.5). SNN-Cliq successfully recovered zygote, early 2-cell, mid 2-cell, late 2-cell and 4-cell stages with only few misclassification, i.e., a late 2-cell-stage cell and a 16-cell-stage cell were placed in wrong clusters (Figure 2.5C). However, the 8-cell and 16-cell stages could not be differentiated. It is interesting to note that nine cells at 8-cell stage

were separated into another cluster instead of being lumped in the 8–16-cell cluster. Surprisingly, a closer look into their RNA-seq protocols reveals that the libraries of these nine cells were exclusively prepared by Smart-Seq2, while all the other libraries were prepared by Smart-Seq (recorded in GSE45719). Thus the separation might be at least partially caused by the technical variations of different library preparation protocols. Applying HCS to the same SNN graph yielded similar results to ours in many aspects (Figure 2.5C). However, the entire zygote stage was missing because of the singleton problem. Both K-means (K=7) and DBSCAN (MinPts=3, Eps=130) could not separate cell stages effectively; multiple stages were often jointed together. In addition, DBSCAN produced too many noise cells. Again, SNN-Cliq outperformed the other methods in all the three evaluation criteria (Figure 2.6C).

2.5 Discussion

In single-cell transcriptome analysis, it is often desired to group individual cells based on their gene expression levels, so that each group corresponds to a cell type with specific functions. Such analysis could help to characterize cell compositions in tissues and distinguish developmental stages, thereby leading to a better understanding of the physiology and pathology of the tissues and the developmental process. An ideal clustering method for genome-wide single-cell data should be able to distinguish cell types from highly noisy gene expression levels due to the unavoidable biological and technical variations. Aimed at this goal, we have presented a clustering algorithm SNN-Cliq based on a new SNN graph and quasi-clique finding techniques (the novelty of SNN-Cliq is described in Supplementary text).

SNN-Cliq possesses some notable features worthy of noting. First, it has low polynomial complexity ($O(n^2)$) and is efficient in practice. Therefore, it is fast enough to handle large datasets, including the ever-increasing number of single-cell transcriptome datasets in a foreseeable future. Second, SNN-Cliq does not require users to specify the number of clusters to be produced; instead, it automatically determines the cluster number in a dataset. Third, it is easy to use in terms of parameter settings. We have demonstrated that finding a valid value of *k* is usually not hard and altering *k* in a certain range will not largely affect the results for many clustering problems. To allow more flexibility, SNN-Cliq provides two granularity parameters *r* for finding quasi-cliques and *m* for merging clusters, which can fine-tune the clustering outputs.

SNN-Cliq has outstanding performance on both the synthetic and real experimental datasets evaluated. Since the algorithm does not make any assumptions on the structure of clusters, it can handle data with various shapes and densities as demonstrated on the three synthetic datasets. Furthermore, the evaluation on single-cell RNA-seq datasets clearly demonstrates that SNN-Cliq could generate desirable solutions with high accuracy and sensitivity, outperforming the other algorithms tested (Figure 6A–C). For instance, for the human cancer cell dataset, SNN-Cliq can detect more cell types than the other methods. For the human and mouse embryo datasets, the clustering of embryonic cells according to their developmental stages can be explained by the extensive changes in gene expression over time during early embryonic development. In both human and mouse, the switch from maternal to embryonic genome control is marked by rapid clearance of maternally inherited transcripts and activation of embryonic genome-derived transcription (Telford *et al.*, 1990). In human, the maternal-zygotic transition occurs during the 4-cell to 8-cell stage

(Yan *et al*. 2013). Compared to the vast changes of gene expression over time, the expression patterns are generally homogeneous between cells from the same developmental stage (Supplementary Figure S4). In mouse preimplantation development, two major waves of de novo transcription occur before the 8-cell stage. One corresponds to the maternal-zygotic transition at the 2-cell stage; another mid-preimplantation activation occurs during the 4-cell to 8-cell stage, preparing for the overt morphological changes in subsequent stages (Hamatani *et al.*, 2004). During the 8-cell to 16-cell stage, embryos embark on compaction and establishment of cellular contact, followed by lineage differentiation at blastocyst stage (Wang *et al.*, 2004). The cell-to-cell variability at this phase revealed by the correlation heatmap (Supplementary Figure S4) is consistent with the embryo's need to develop increasingly diverse cells. However, a relatively small number of genes undergo expression changes between the 8-cell and 16-cell stages (Wang *et al*. 2004; Hamatani *et al*. 2004), which may explain the lump of the two stages into one cluster. In addition to detecting the cell stages, SNN-Cliq can recognize cells that were isolated from different embryos and cells that were generated by different library preparation protocols. In particular, SNN-Cliq does not discard data points in regions of low density, as other methods often do by treating them as noise or singletons.

CHAPTER 3. UNDERSTANDING THE TRANSCRIPTIONAL NOISE IN YEAST USING SINGLE-CELL TRANSCRIPTOMES

3.1 Abstract

The stochastic gene expression, or gene expression "noise" has been studied extensively during the last decade. It is now widely recognized that the gene expression noise is a major source of the phenotypic variation of isogenic cells grown in the same environment. This brings into question the relationship between gene expression noise and transcriptome states. Due to the noise can be propagated in gene regulatory networks, it is suggested that functionally related genes such as regulons and pathways can be derived from noise profiles. However, this method has not been extensively examined at a genome scale on the mRNA level. In this project, we will evaluate the gene transcriptional noise under multiple culture conditions in yeast *S. cerevisiae.* Using a single-cell RNA-Seq method, we sequenced 51 yeast cells from three treatments (hypertonic condition, isotonic condition and amino acid starvation) along with five samples of bulk RNA at different dilution. The transcriptomes were sequenced to a sufficient read-depth and cells with low quality data were filtered by computational methods. Our results show that the single cells from the same treatment can be clustered together based on their transcriptomes and that different treatments show distinct transcriptome variability. In addition, we find that treatments can induce distinct noise profiles for some functional modules and regulatory pathways. In conclusion, our results indicate that transcriptional noise may be subject to regulation in response to environmental stresses. We believe that the analysis of noise under different

conditions can lead to a better understanding of gene transcription and regulation in isogenic cells.

3.2 Introduction

In multi-cellular organisms, cells undergo regulated differentiation processes, leading to cell heterogeneity in tissues and organs carrying out various functions. On the other hand, it has been well known that isogenic cells grown in the same condition also exhibit considerable variations in phenotypes. Such variation is the consequence of the inevitable stochastic nature of biochemical processes that depend on low copy molecules in individual cells (Kalisky and Quake, 2011). It is necessary to investigate behaviors and functions of individual cells. Indeed, earlier researches conducted at the single-cell level have brought new insights into many biological phenomena, fundamentally changing our ways of thinking and practice. In both single-cellular organisms such as bacteria and yeast, as well as multi-cellular organisms such as vertebrates, stochasticity can be advantageous and is incorporated into developmental process to generate cell diversity, resulting in a variety of functional consequences that may be difficult to achieve by deterministic mechanisms (Johnston and Desplan, 2010). For example, bacteria utilize stochasticity to enhance the survival chance in case of environmental changes by allowing a variety of cellular states in the population (St-Pierre and Endy, 2008; Losick and Desplan, 2008). Therefore, understanding the origins and consequences of noise is of great importance for elucidating many fundamental biological processes, e.g. cell differentiation, development, evolution, and bacteria and cancer cell drug resistance.

The expression noise (η) of a gene in a cell population is usually quantified by the standard deviation (σ) of the gene's expression level divided by the mean (μ). The

expression noise originates from both intrinsic and extrinsic sources (Elowitz *et al.*, 2002; Swain *et al.*, 2002; Paulsson, 2004). The intrinsic noise arises from the small copy number of molecules carrying out the gene expression processes in a cell. It has been shown that mRNA is produced in a burst manner; the burst size, frequency and rates all contribute to the intrinsic noise (Raj *et al.*, 2006; Maheshri and O'Shea, 2007). In contrast, the extrinsic noise is mainly due to fluctuations in other cellular components, including the global factors such as the cell size and shape, and the pathway-specific factors such as the upstream regulators in a specific signal transduction pathway (Raser and O'Shea, 2004).

A study based on single-cell proteomics in yeast found that for proteins with low to medium abundances, their expression noise is inversely proportional to the abundance and is dominated by the intrinsic noise, e.g. stochastic production/degradation of mRNAs. In contrast, the extrinsic noise makes a significant contribution to the total noise for proteins with high abundances and is uncorrelated with the protein abundance (Newman *et al.*, 2006). Instead, extrinsic noise is strongly correlated with the modes of transcriptional regulation. For example, genes regulated by transcription factors that act on chromatin structure to activate genes present high protein fluctuations. In addition, proteins in different functional modules exhibit distinct noise levels: the stress-response genes show high expression noise, while the ribosomal protein genes have low variation (Newman *et al.*, 2006). These conclusions are further supported by another study on 43 GFP-tagged proteins in yeast cultured in 11 growth conditions (Bar-Even *et al.*, 2006). These authors showed that the general relationship between noise level and abundance can be explained by the random birth and death of individual mRNAs. However, the 'noise residual', which is the deviation from this relation, depends on gene functions (Bar-Even *et al.*, 2006).

Another study further explored this phenomenon and identified 'noise regulons' through analyzing the noise correlations among 182 GFP-tagged proteins under a culture condition (Stewart-Ornstein *et al.*, 2012). Their results strongly suggest that noise can be a powerful tool to infer regulons and functional modules.

However, these studies measuring expression noise at the protein level may have limitations in tracking the actual transcriptional state of genes because the fluctuations of the mRNA levels could be masked by the long half-life of proteins. Additionally, due to technical hurdles, only a limited number of genes can be studied using proteomics-based approaches. To fill these gaps, in this study, we instead measured the transcriptomes in individual yeast cells using an RNA-Seq technique. More specifically, we use the budding yeast as the model organism to delineate the relationship between transcriptional noise and transcription level in a genome scale under different treatments. Our study also showcases that transcriptional noise can be used to elucidate the function modules and regulatory pathways. Thus, our results furthers the understanding of the transcriptional regulation mechanisms as well as of the determinants and biological significance of transcriptional noise.

3.3 Results

3.3.1 Single-cell Transcription Profiling

Using a single-cell RNA-Seq technique, we sequenced the transcriptomes of 51 yeast cells from three treatments: amino acids starvation (AA starvation) (n= 20), isotonic (n=12) and hypertonic conditions (n=19). For the purpose of quality evaluation, we also sequenced the RNA libraries prepared from 5 pg, 10 pg, 20 pg, 1,000 pg and 10,000 pg bulk mRNA extracted from a population of cells under AA starvation. The libraries of 5 pg and 10 pg

mRNA are to mimic those from single cells, as a cell is estimated to contain ~5–10 pg mRNAs. The libraries of 1,000 pg and 10,000 pg mRNA are to mimic those prepared using conventional cell population based RNA-Seq methods. The mapping results of the reads to the genome are summarized in Table 3.1. An average of 50% (ranging from 12% to 69%) of the processed reads were uniquely mapped to the genome, resulting in an average of 10.75 million uniquely mapped reads in a library. Our single-cell RNA-Seq method resulted in a comparable library size to those from cell population based data.

Table 3.1. Summary of reads mapping results.

Cell	***Total***	***Unique***	***%***	***Treatment***
D19	10772086	4898896	45.5	AA starvation
D1	14810328	7933130	53.6	AA starvation
D20	9615756	5468320	56.9	AA starvation
D22	13820378	9354376	67.7	AA starvation
D23	13820962	8982239	65.0	AA starvation
D24	14926756	8368256	56.1	AA starvation
D27	16666368	10066387	60.4	AA starvation
D28	12490922	7324409	58.6	AA starvation
D29	13438936	8690773	64.7	AA starvation
D30	23799706	13562585	57.0	AA starvation
D3	11266478	6985357	62.0	AA starvation
D7	16052474	9181288	57.2	AA starvation
D9	20192678	12546071	62.1	AA starvation
D26X	14749854	6615432	44.9	AA starvation
D91X	19775200	11295484	57.1	AA starvation
D93X	18026718	11061295	61.4	AA starvation
D94X	18543286	11771929	63.5	AA starvation
D95X	19815440	12623451	63.7	AA starvation
D96X	18941278	11622814	61.4	AA starvation
D97X	16512378	8786472	53.2	AA starvation
C80X	17475932	6639685	38.0	Isotonic
C81X	21190142	9051841	42.7	Isotonic
F20X	18874810	11293078	59.8	Isotonic
F21X	15188760	9887713	65.1	Isotonic
F22X	15377178	9890889	64.3	Isotonic
F23X	17434512	8998306	51.6	Isotonic

F24X	20275408	12671109	62.5	Isotonic
F30X	22585884	8369197	37.1	Isotonic
F3X	18533398	9201565	49.6	Isotonic
F40X	21285726	10159072	47.7	Isotonic
F55X	22567110	7810604	34.6	Isotonic
F9X	12318022	6615156	53.7	Isotonic
E10X	14816920	8040547	54.3	Hypertonic
E11X	16187454	4958802	30.6	Hypertonic
G36	42564972	19357757	45.5	Hypertonic
G50	16744672	9101517	54.4	Hypertonic
G53	20617022	6969212	33.8	Hypertonic
G55	17152984	8085432	47.1	Hypertonic
G56	20096004	8087153	40.2	Hypertonic
G57	19752432	6897991	34.9	Hypertonic
A23	26949453	14248479	52.87	Hypertonic
A25	21321783	8693804	40.77	Hypertonic
A26	23186521	9804524	42.29	Hypertonic
A27	41788194	14318736	34.27	Hypertonic
A28	21842906	4951337	22.67	Hypertonic
A29	13823123	3151752	22.80	Hypertonic
A34	30782075	8196010	26.63	Hypertonic
A35	45827457	22900405	49.97	Hypertonic
A44	37507659	4585466	12.23	Hypertonic
A47	18807922	3803587	20.22	Hypertonic
A7	101048198	49690769	49.18	Hypertonic
H6P_11(10000 pg)	15965554	11047853	69.2	AA starvation
H6P_12(1000 pg)	29815776	17175386	57.6	AA starvation
H6P_14(20 pg)	23819532	16063645	67.4	AA starvation
H6P_15(10 pg)	40244016	21455019	53.3	AA starvation
H6P_16(5 pg)	35534074	22821748	64.2	AA starvation

The 1st column shows the total number of reads in the library. The 2nd column shows the number of uniquely mapped reads to the genome.

To see whether the sequencing depth is sufficient to detect transcribed genes, we randomly sampled different numbers of mapped reads from the five bulk RNA libraries and computed the percentage of genes whose mRNA was detected for each dataset. As shown in Figure 3.1A, the number of genes detected approached saturation when around 4–5 million reads were sampled for each dataset. This suggests that for most of our single-cell libraries, the sequencing depth should be more than sufficient to detect transcribed

genes. The detection rates of the five bulk RNA libraries show a clear dose-dependent relationship; datasets sampled from 5 pg and 10 pg libraries have markedly lower gene detection rates (Figure 3.1A). This is as expected since when the starting amount of RNAs is as low as the level in a cell, the effect of random loss in the library preparation process could become pronounced. However, there are fewer genes detected in the 10 pg library than in the 5 pg library, indicating the considerable technical variation in the protocol.

Next, to see if combining single-cell reads can recapitulate the sensitivity of the bulk mRNA profile, we created synthetic ensemble datasets by computationally pooling raw reads from a set of single cells (all in AA starvation treatment) to mimic bulk RNA-Seq experiments. As shown in Figure 3.1B, when reads from as few as 5 cells were combined, the detection rate already reaches saturation. Moreover, the detection rate at the saturation point is comparable to that of the bulk 1,000 pg and 10,000 pg libraries. This clearly demonstrates that our single-cell RNA-Seq method is highly efficient to detect low-copy number mRNA, presumably because of the overwhelmingly large amount of reagents used in library preparation which may alleviate the effects of low copy number molecules in a cell.

Our single-cell RNA-Seq data may suffer from bias toward increased coverage at the 3’ end, because of the oligo(DT) primers used in the first-strand cDNA synthesis (Tang et al. 2009; Tang et al. 2010) Therefore, we evaluated the possible bias of read coverage along a gene body for all libraries. As expected, the read coverage declines toward the 5'-end for all libraries (Figure 3.2). Interestingly, the extent of bias is related to the starting amount of mRNA, as illustrated by the coverage curves of the bulk libraries: the lower the input amount mRNA, the more bias of reads to the 3’-end. Although the single-cell libraries may

vary in the extent of bias, many of them have an extent comparable to the bulk libraries of 20 pg, indicating that a single cell might contain about 20 pg mRNA.

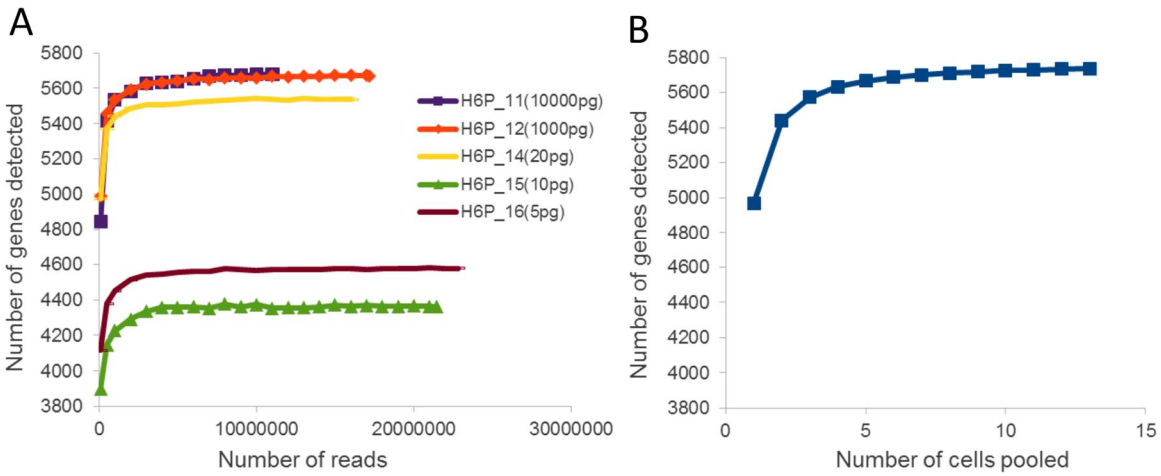

Figure 3.1. (A) Saturation of the five bulk RNA libraries for detecting transcribed genes. The libraries were prepared using different amount (5 pg, 10 pg, 20 pg, 1,000 pg and 10,000 pg) of input mRNA extracted from a population of cells under AA starvation. (B) Saturation curve from single-cell ensembles.

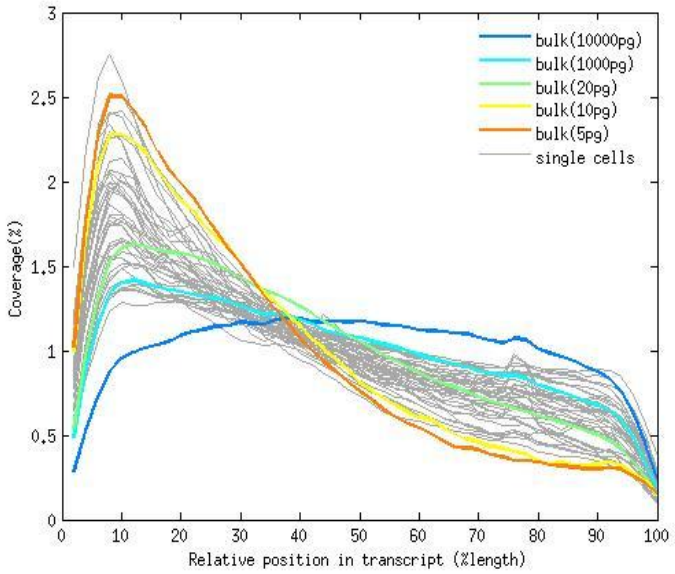

Figure 3.2. 3'-to-5' sequence coverage. For each library, the averaged relative coverage is shown at each relative position along the length of transcripts from the 3'-end to the 5'-end. Only mRNAs longer than 1kb from genes with a single non-overlapping exon were included for this analysis.

3.3.2 Library Quality Assessment

Since single-cell RNA-Seq results are sensitive to multiple factors during library preparation, we evaluated each library for its quality by several assessment criteria: complexity, evenness of coverage, continuity of coverage, sensitivity and correlation (Figure 3.3A). The library complexity is defined as the number of distinct (unique) read's start positions mapped to the genome (Levin *et al.*, 2010). To directly compare the library complexity, we randomly sampled the same number of reads (one million) from each library. The complexity of our libraries ranges from 0.20 to 0.63 with an average of 0.50. Some libraries show obviously lower complexity than others, indicating a strong bias in fragment amplification and an insufficient sampling of the mRNA present in the cell. The evenness is defined as the averaged coefficient of variation (CV) of the read distribution along a gene body (Levin *et al.*, 2010). Since transcripts of low copy number are subject to uneven coverages, we only used the top 50% highly expressed genes for this measure. A few libraries show highly uneven read distributions with a CV being 2-fold higher than other libraries. We note that these libraries also display lower complexities.

The sampling and distribution of reads have a significant influence on the detection rate of a library. As shown in Figure 3.3A, for most of the libraries (36 out of 56), over 4000 genes were detected. The bulk libraries of 1,000 pg and 10,000 pg mRNA input have the highest detection rate (~5700). Over 5000 genes were also detected in many of the single cell libraries, despite a low starting amount. However, there are six single-cell libraries with significantly lower detections (<2000 genes). Since these libraries also show deviated results in complexity and evenness measurements, we speculate that the number of transcribed genes may be under-estimated in these cells because of the affected library

quality. To determine the reproducibility of gene transcription level, for each cell, we calculated the averaged correlation coefficient between the cell and the other cells in the same treatment. The highest averaged correlation coefficients are obtained by the bulk samples: 1,000 pg (ρ=0.80), 20 pg (ρ=0.80), 10,000 pg (ρ=0.76) and 5 pg (ρ=0.67). Except for a few cells, most cells have a moderate correlation (mean=0.52). Last, the continuity of coverage is determined by the number of gaps along the exons of a gene, where a gap is defined as a continuous length of $\geq$5 bases without any reads mapped. We then took a weighted average of this measure across all the genes according to their normalized read coverage in RPKM values. Except for one library, all libraries have continuous read coverage with only one gap in average. Notably, libraries with higher sequencing depth are likely to have fewer number of gaps (e.g. A7, A23, A27 and A35).

In general, these measurements of library quality from different aspects are highly correlated. The quality of single-cell libraries are comparable to that of bulk libraries except for a few (14) cells denoted by darker bars in Figure 3.3A. To better visualize the library qualities, principal component analysis (PCA) was performed on the five measurements. The projection of the cells on the first and second PCs groups the 42 libraries in high qualities into a large dense cluster (Figure 3.3B). The other 14 cells with consistently bad measures are located away from this cluster and thus are excluded from the subsequent analyses.

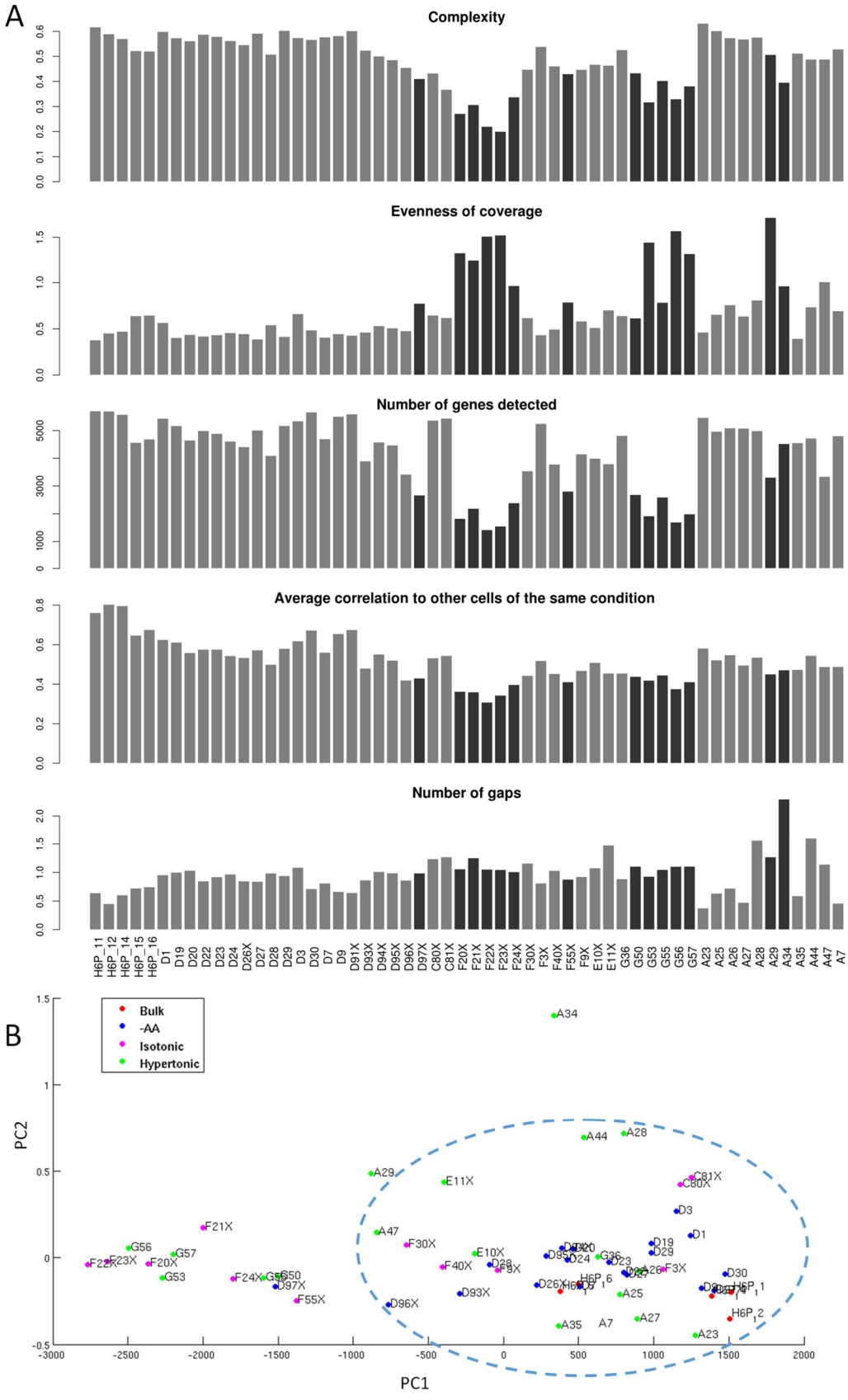
A
Complexity
Evenness of coverage
Number of genes detected
Average correlation to other cells of the same condition
Number of gaps
H6P_11 H6P_12 H6P_14 H6P_15 H6P_16 D1 D19 D20 D22 D23 D24 D26X D27 D28 D29 D3 D30 D7 D9 D91X D93X D94X D95X D96X D97X C80X C81X F20X F21X F22X F23X F24X F30X F3X F40X F55X F9X E10X E11X G36 G50 G53 G55 G56 G57 A23 A25 A26 A27 A28 A29 A34 A35 A44 A47 A7
B
Bulk
-AA
Isotonic
Hypertonic
PC2
PC1

Figure 3.3. (A) Quality measurements for each library. The dark bars denote the libraries that are excluded from further analysis due to low quality. (B) PCA on the five quality measurements displayed using the first two components. The libraries outside the defined area are considered to have low qualities.

3.3.3 Cells are Separated into Clusters Based on Their Transcriptomes

As shown in Figure 3.4A, the abundance of 5834 transcripts show relatively high correlation between cells from the same treatment; an average Pearson's correlation coefficient (PCC) of 0.91±0.06, 0.75±0.09, 0.79±0.12 and 0.73±0.14 is observed for bulk mRNA, AA starvation, isotonic and hypertonic conditions, respectively. To further characterize the transcriptome features under different treatments, we performed PCA analysis (Figure 3.4B) and hierarchical clustering on the 42 libraries (Figure 3.5). The results suggest that individual cells as well as bulk libraries under AA starvation form a cluster and are clearly separated from the cells from the other two conditions. However, the cells from isotonic and hypertonic conditions could not be differentiated from each other, indicating they have similar transcriptomes.

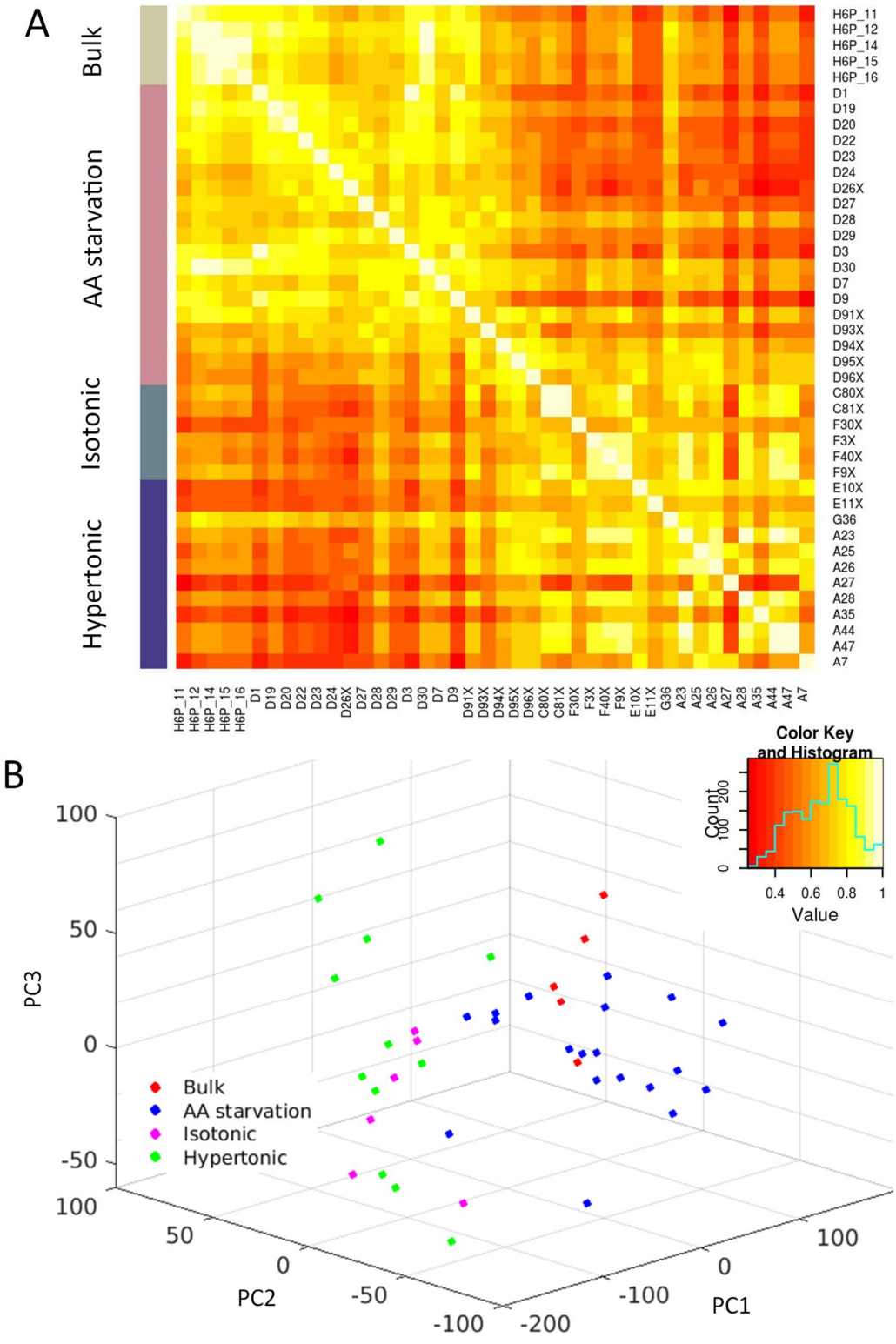

Figure 3.4. (A) Correlation of the transcription levels of cells/samples measured by Pearson's correlation coefficient based on log transformed RPKM. Yellow indicates a high pairwise correlation and red a low correlation. (B) PCA analysis on log transformed

RPKMs. Cells/samples are projected onto the top three principal components. Red dots denote bulk mRNA samples, blue dots AA starvation, magenta dots isotonic and green dots hypertonic condition-treated cells.

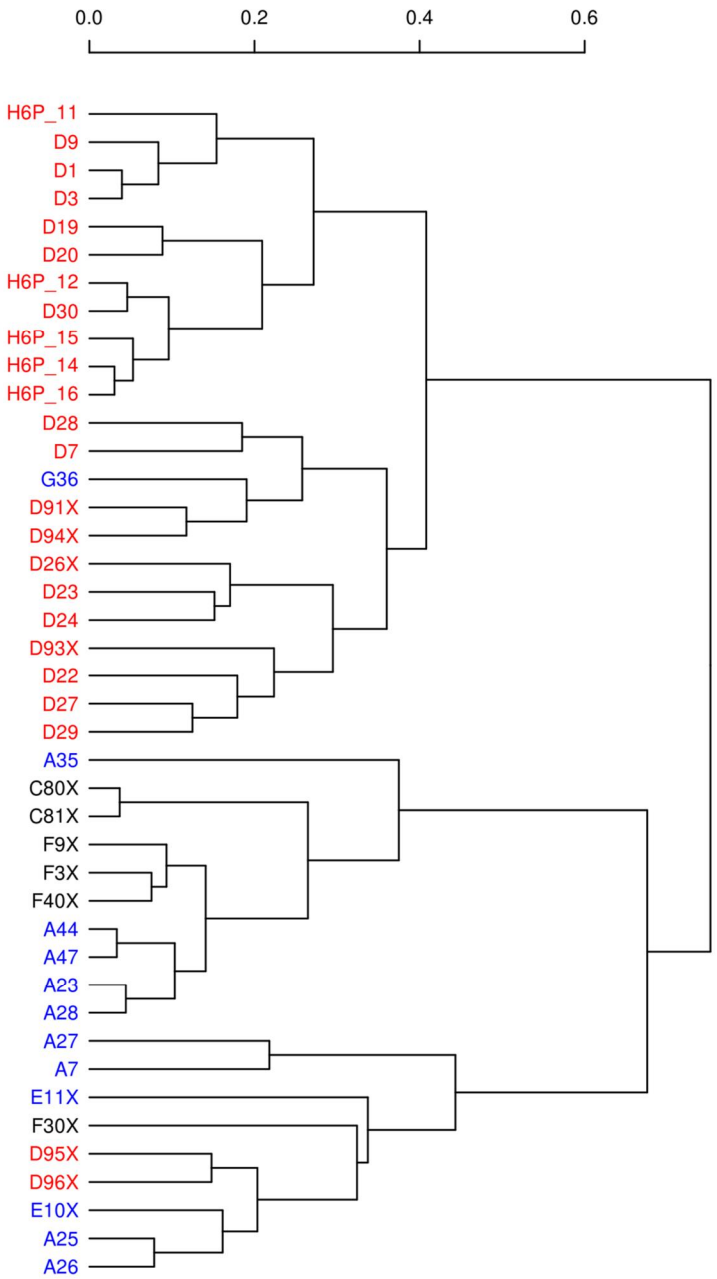

Figure 3.5. Hierarchical clustering of the samples and cells based on Pearson's correlation coefficient values between transcriptomes. AA starvation, isotonic and hypertonic libraries are shown in red, black and blue, respectively.

3.3.4 Differential Gene Transcription between Treatments

We used edgeR (Robinson *et al.*, 2010) to identify differentially transcribed genes between cells from different treatments (p-value<0.05, FDR<5%). Compared to the isotonic condition, 93 genes were down regulated and 50 genes were up regulated under AA starvation. Contrastingly, only 12 genes were down regulated and 6 genes were up regulated in hypertonic condition. Significantly enriched Gene Ontology (GO) biological process terms and KEGG pathways were identified in each set of differentially regulated genes using DAVID 6.7 (Database for Annotation, Visualization and Integrated Discovery) (Huang *et al.*, 2009). False discovery rate was controlled in the multiple comparisons (FDR<5%). As show in Table 3.2, in AA starvation, substantial genes related to protein metabolic process were differentially regulated. However, there is no significant functional annotations enriched in the hypertonic condition compared to the isotonic condition. This is consistent with the results from the above PCA (Figure 3.4) and clustering (Figure 3.5) analyses, where cells under isotonic and hypertonic conditions cannot be separated.

Table 3.2. GO enrichment for differentially expressed genes.

AA STARVATION UP REGULATED	**P-VALUE**	**FDR (%)**
Sulfur Metabolic Process	7.05E-09	7.89E-06
Heterocycle Metabolic Process	7.70E-04	8.57E-01
Cellular Amino Acid And Derivative Metabolic Process	1.10E-03	1.23E+00
Cellular Ketone Metabolic Process	4.69E-03	5.12E+00
Organic Acid Metabolic Process	5.19E-03	5.65E+00
KEGG: Sulfur Metabolism	6.88E-06	4.98E-03
AA STARVATION DOWN REGULATED		
Organic Acid Metabolic Process	5.24E-16	6.77E-13
Cellular Ketone Metabolic Process	1.76E-14	2.14E-11
Cellular Amino Acid And Derivative Metabolic Process	1.00E-11	1.22E-08
Amine Metabolic Process	2.81E-11	3.41E-08
Alcohol Biosynthetic Process	1.01E-09	1.22E-06

Cellular Aromatic Compound Metabolic Process	4.22E-09	5.13E-06
Alcohol Catabolic Process	1.24E-08	1.50E-05
Carbohydrate Catabolic Process	1.95E-07	2.36E-04
Cellular Biosynthetic Process	5.03E-07	6.10E-04
Monosaccharide Metabolic Process	1.30E-04	1.58E-01
Generation Of Precursor Metabolites And Energy	1.99E-04	2.41E-01
Cellular Carbohydrate Metabolic Process	5.59E-04	6.76E-01
Carbohydrate Metabolic Process	7.20E-04	8.71E-01
Carbohydrate Biosynthetic Process	1.31E-03	1.57E+00
KEGG: Glycolysis / Gluconeogenesis	1.25E-07	1.15E-04
KEGG: Lysine Biosynthesis	5.06E-06	4.65E-03
KEGG: Tyrosine Metabolism	6.23E-03	5.59E+00

3.3.5 Transcriptional Noise

We estimated the transcriptional noise (η) of a gene by the coefficient of variation (CV) of its mRNA levels in RPKM, i.e. the standard deviation of transcription levels among the total number of cells (σ) divided by the mean abundance (<mRNA>). As suggested by the above results that transcription levels have no significant differences between hypertonic and isotonic cells, we combined the cells in these two conditions in a single group in the following analysis. Consistent with the proposed stochastic model of gene transcription (see Methods), the transcriptional noise and mean abundance display a scaling relationship (Figure 3.6A-B). However, in addition to the stochastic fluctuation scaling with 1/<mRNA>, transcriptional noise may also originate from other sources, as shown by the deviation from the expected noise level inferred by regression (Figure 3.6A-B). We speculate that the transcription fluctuations at low abundances is largely due to the technical variability. Thus, the genes with RPKM<10 were excluded from the analysis. As shown in Figures 3.6A and 3.6B, in general, noise is inversely proportional to the mean abundance at moderate abundances. At high abundances, extrinsic noise dominates the total noise, leading to uncorrelated noise level to the mean at the right tail (Figure 3.6A-B).

We further explored the noise patterns for a few metabolic pathways under different growth conditions. In some pathways, the noise levels were substantially reduced in AA starvation (Figure 3.7). For example, genes in PWY3O-188, the pathway of aerobic respiration–electron transport chain, exhibit reduced noise levels while the mean transcript abundance remains the same (Figure 3.7). An early study demonstrated that the starvation of essential amino acids in yeast cells could result in a high burden of reactive oxygen species (ROS), whose formation is linked to the mitochondrial respiratory chain (Eisler et al. 2004). A later study suggested that the induction of the transcriptional programs associated with respiration could exert a protective effect against oxidative damages and related stresses during nutritional shortages in yeast (Petti *et al.*, 2011). These pieces of evidence may explain the reduced noise in PWY3O-188 in AA starvation. We note that for these pathways the mean abundance remains the same between treatments.

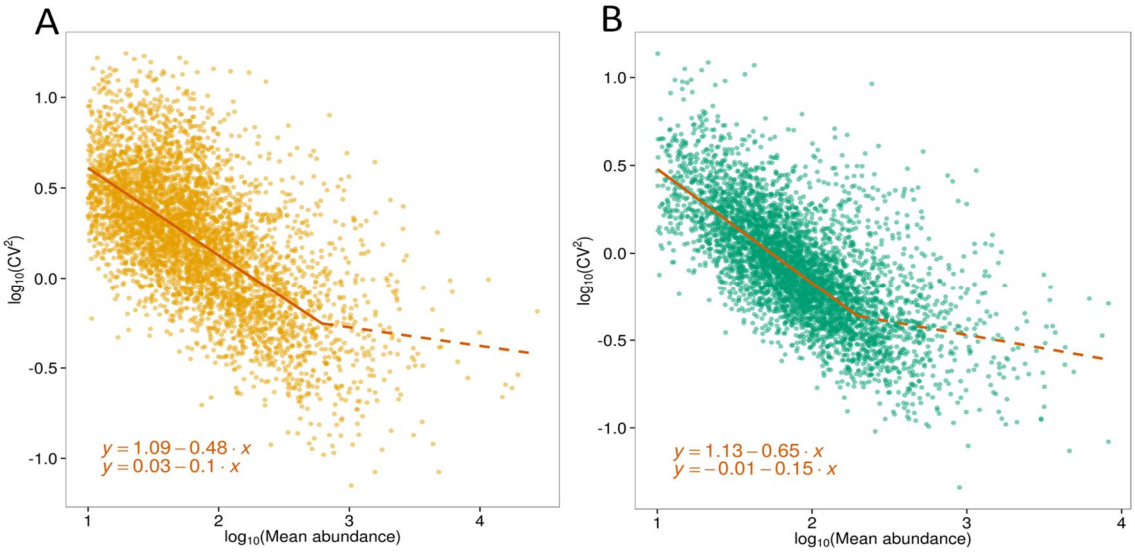

Figure 3.6. Modeling of transcriptional noise. (A-B) CV^2 as a function of the mean transcript abundance in cells under AA starvation (A) and isotonic/hypertonic (B). Only genes with RPKMs over 10 are shown because of the noticeable technical noise at low abundances. Solid lines correspond to the robust linear regression fitting. The two fitting sections are selected to allow the two fitted lines cross on the boundary.

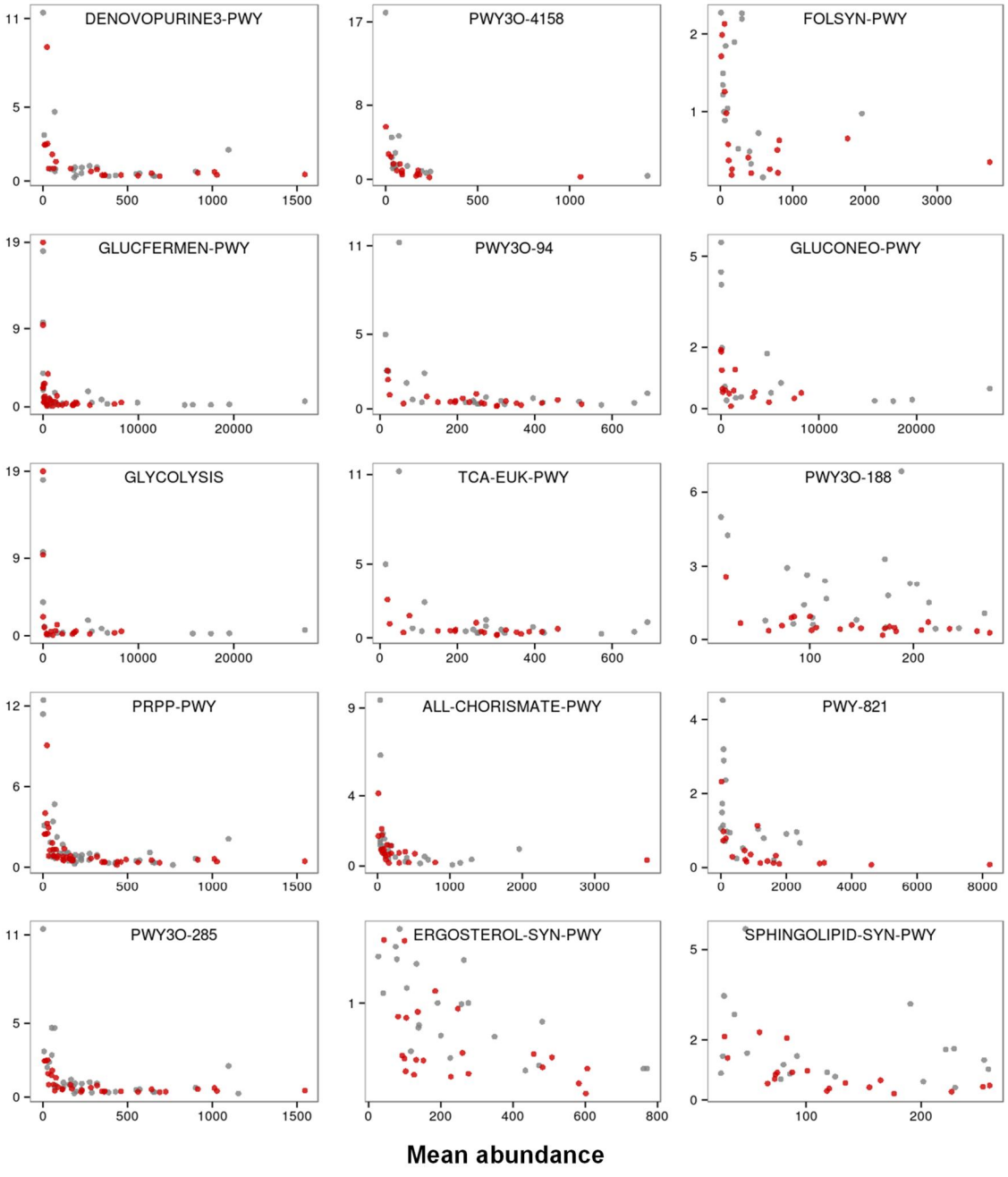

Figure 3.7. Pathway specific relationship between the η^2 (y-axis) and mean abundance (x-axis). Red denotes the AA starvation; gray denotes the isotonic/hypertonic.

The gene-specific noise is quantified by the 'noise residuals' (Bar-Even *et al.*, 2006), defined as the vertical distance from the point to the fitted line (Figure 3.6A-B). Unlike the noise level, the noise residual (NR) is independent of the mRNA abundance and thus can be compared across genes with different transcription level. It has been shown that such adjusted noise level of gene expression can display module-specific patterns to different transcription activation mode or functional pathways (Bar-Even *et al.*, 2006; Newman *et al.*, 2006). However, these earlier analyses have limitations because only few genes were measured or only one growth condition was used. Therefore, we want to extend these studies in a genome scale at the mRNA level under different conditions. The NR values follow a normal distribution centered at zero because of the assumption imposed by linear regression. A gene is then defined to be highly noisy if the corresponding NR>0 and lowly noisy if NR<0.

We investigated the relationship between gene modules and NR levels by hypergeometric distribution ($p<0.05$) under each growth condition. To this end, we selected several regulons or functional groups that are likely to be associated with high noise or low noise. For example, SAGA-dominated genes are tightly connected to environmental stress responses (Huisinga and Pugh, 2004) and are known for their high expression variation (Newman *et al.*, 2006). In agreement with this, our results also show that SAGA-dominated module is significantly enriched in the high noise genes in AA starvation (Table 3.3). We suspect that the high noise presented by ribosomal protein genes mirror the variability of cell states in response to environmental stimuli, considering the fact that the cell size and growth rate can affect the number of mRNAs of ribosomal protein genes in a cell (Warner, 1989).

Table 3.3. Summary of highly noisy and lowly noisy genes in functional modules.

Module	# module genes	AA starvation		Isotonic/ hypertonic	
		# genes	p-value	# genes	p-value
Enrichment in high noise genes					
SAGA-dominated	452	270	3.40E-7	232	0.0812
Ribosomal large subunit biogenesis	35	25	5.02E-3	17	0.544
Ribosomal small subunit biogenesis	25	18	1.49E-2	12	0.582
Ribosomal biogenesis	57	42	9.15E-5	27	0.596
Ribosome	242	166	7.77E-11	141	7.17E-4
Translation	214	161	3.32E-16	103	0.542
Enrichment in low noise genes					
Chromatin modification	95	65	5.23E-4	51	0.402
Chromatin remodeling	54	38	3.51E-3	31	0.249
Transcription DNA-templated	411	248	9.18E-5	205	0.819

The 2nd column shows the total number of genes in the corresponding module. The 3rd and 4th columns show the number of module genes that are categorized as high noise or low noise and the corresponding p-value (gray marks significance) under the AA starvation treatment. The 5th and 6th columns are the same as the 3rd and 4th columns except the cells are under isotonic/hypertonic condition.

These results suggest that comparing the NR levels at different growth conditions can be a powerful method to understand the functions and benefits of transcriptional noise. As shown in Figure 3.8, the NR levels are generally independent of the transcript abundances, as suggested by the almost uniformly distributed gray dots in the NR vs. mean plot. However, the modules display distinct NR features under different growth conditions. For example, the purple and black dots denoting the genes of ribosomal large and small subunit biogenesis move left (indicating lower abundance) and upward (indicating greater NR) in AA starvation compared to in isotonic/hypertonic conditions (Figure 3.8). To characterize the extent of changes, we applied Wilcoxon rank sum test to the NR values for each module (p-value<0.05). At the same time, we show how many genes in a module are significantly

down or up regulated in AA starvation compared to those under isotonic/hypertonic conditions (p-value<0.05, FDR<5%; in total 595 genes were down regulated and 154 genes were up regulated comparing AA starvation to isotonic/hypertonic conditions) (Table 3.4). In consistent with an early finding that translatable mRNAs for many ribosomal proteins are decreased during amino acid starvation (Warner 1989), we also found that the transcription of ribosomal protein genes were largely down regulated in AA starvation (135 out of 264 genes in the module). However, genes in the ribosome module had higher NR levels in AA starvation than in isotonic/hypertonic conditions (p-value<0.05). The same is true for the translation module and ribosome biogenesis module, suggesting a greater variability of cell states under adverse environment. On the contrary, the transcription module has a balanced number of up and down regulated genes, while the NR level is decreased (Table 3.4). This suggests that the increased variation of the above modules may not be originated from the overall control of transcription; instead, they may be subject to module-specific regulation. These results echo the above module enrichment results, indicating that other than transcriptional abundance, transcriptional noise is also subject to regulation and may play a role in response to environmental stress.

Table 3.4. Results for comparison of functional modules between treatments.

Module	**# module genes**	**Abundance**		**Noise residual**	
		down (595)	up (154)	p-value	
Ribosomal large subunit biogenesis	75	13	1	↑	2.87E-2
Ribosomal small subunit biogenesis	46	12	0	↑	2.03E-2
Ribosomal biogenesis	121	25	1	↑	5.16E-4
Ribosome	264	135	3	↑	3.71E-3
Translation	240	141	0	↑	7.92E-8
Transcription DNA-templated	495	14	11	↓	2.15E-3

Carbohydrate metabolic process	116	14	2		1.55E-1
Cellular amino acid biosynthetic process	104	36	13		7.39E-2
Amino acid transport	74	8	8		7.66E-1
Cellular response to oxidative stress	70	9	1		6.47E-1

The 2nd column shows the total number of genes in the corresponding module. The 3rd and 4th columns show the number of genes with significantly down or up regulated transcription in AA starvation compared to isotonic/hypertonic (results from transcriptome comparison by edgeR).The symbol in the 5th column indicates increased (↑) or decreased (↓) NR level in AA starvation compared to Isotonic/Hypertonic. The p-value in the 6th column is calculated by Wilcoxon rank sum test (gray marks significance).

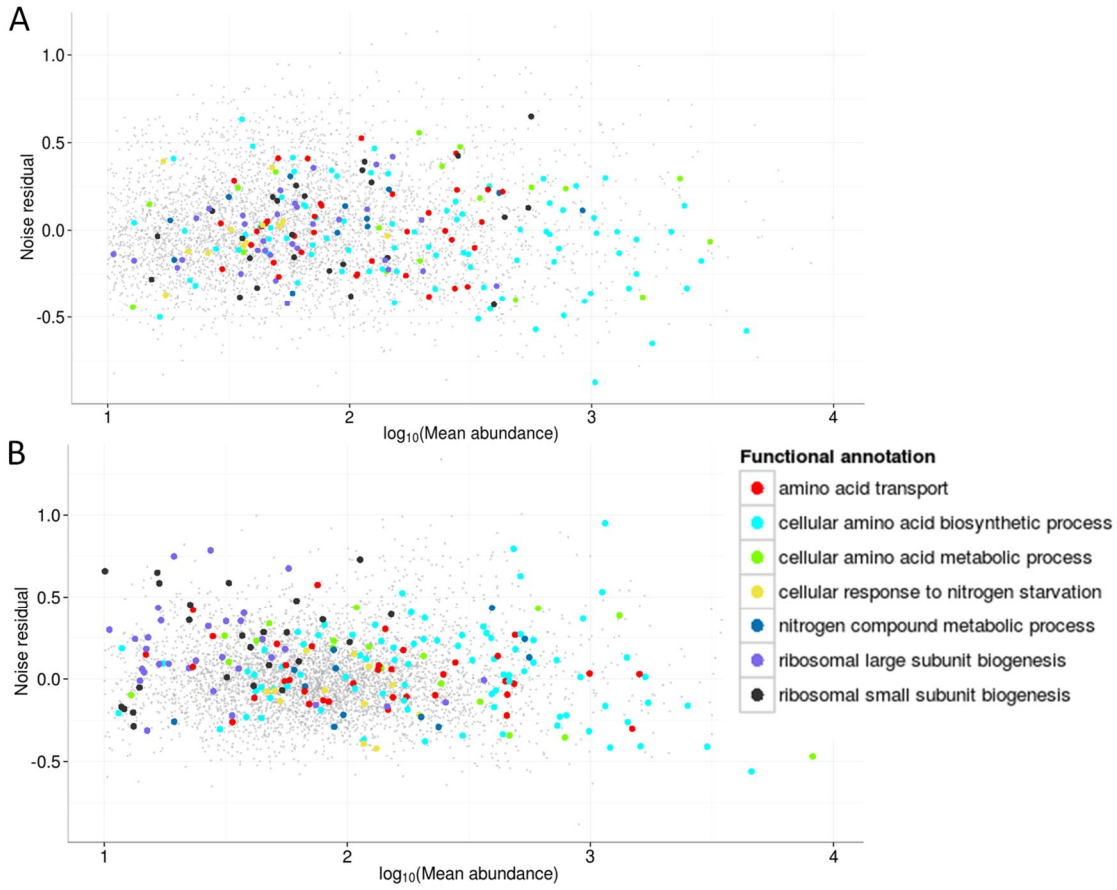

Figure 3.8. Noise residuals (NR) for specific gene modules at isotonic/hypertonic (A) and AA starvation (B). The indicated functional modules are shown in large dots in different colors, and all the other detected genes are shown in small dots in grey.

3.4 Discussions

Single-cell RNA-Seq has become a powerful tool to address important biological problems including cell type identification, understanding mechanisms of gene transcriptional regulation and characterization of functionally related genes. However, at the level of current technology, observed cell-to-cell variation may be confounded by technical variability. Besides, the cell size, state and other factors may also contribute to the transcriptome variability. Therefore, a careful quality control is critical before any formal analysis to minimize the effects of technical variability. In this paper, we utilized five measures (complexity, evenness of coverage, detection rate, correlation and gaps on coverage) to evaluate the quality of reads from different aspects. Based on PCA analysis of cells/samples using these measures, we identified and filtered out several cells with low-quality reads. Indeed, this filter not only enhances the reliability of our analysis but also yields more significant results (comparative results not shown).

Another critical aspect of inference from single-cell transcriptome data is the depth of read coverage, which has an effect on the number of genes that can be detected as has been noted in cell population based libraries (Tarazona *et al.*, 2011). To evaluate the read-depth, we additionally prepared a series of diluted bulk RNA libraries, starting from 10 pg which is considered to be near the amount of mRNA in single cells. From the saturation analysis using the reads from the 10 pg bulk library (Figure 3.1A), we deduced that a minimal sequencing depth of 4×10^6 reads is required to detect almost all genes that can be quantitatively characterized at this starting quantity of mRNA. In fact, our single-cell libraries performed better and many showed higher detection rates comparable to the 1,000 pg or even 10,000 pg bulk libraries.

The treatment-specific characteristics of the single-cell transcriptomes recapitulate the population-level data, as illustrated by the correlation and differential transcription analysis. Cells grown under a certain treatment exhibit higher correlation of transcription levels, while different treatments lead to distinct transcription of relevant genes in response to the treatments. This indicates that transcriptome analysis at the single-cell level is biologically meaningful in respect of growth conditions. Our observation that the isotonic and hypertonic cells cannot be clearly differentiated using their transcriptomes is consistent with a previous finding using microarray to characterize gene expression on diverse environmental transitions (Gasch *et al.*, 2000). This study found that when cells were transferred from standard isotonic to hypertonic (1 M sorbitol) solution, the change in the expression of the genes participating in environmental stress response is only transient (Gasch *et al.*, 2000). Similarly, there was only subtle transient change for these genes when cells having adapted to the hypertonic solution (1 M sorbitol) were transferred back to standard isotonic solution (Gasch *et al.*, 2000). In our experiment, cells under all treatment (AA starvation, hypertonic and isotonic) were exposed to hypertonic (1 M sorbitol) solution for about one hour after being harvested from log-phase growth in YPD in a procedure to remove the cell wall. Subsequently, cells in the isotonic treatment were exposed to a sorbitol lacking solution for at least an hour before being collected, at which point the cells have adapted to the isotonic treatment and gene expression were back to the level as before. Therefore, we combined the cells under isotonic and hypertonic treatments as a single group for analyses.

The PCA analysis on all cells reveals complex transcriptome variability. Cells from AA starvation and isotonic/hypertonic were largely separated. However, cells of the same

treatment did not form a compact cluster; instead, they spread out in a line, indicating considerable variation in between. Beside the biological noise, another possible factor may be that the exposure time in a treatment varies for cells, ranging from one hour to five hours after the transfer to the treatment. The exposure time was not recorded so that the dynamic changes of gene expression over time are not investigated in the current study. However, in the future study, the exposure time should be controlled or recorded for a time-series gene expression study.

We explored the general relationship between the transcriptional noise and mean abundance. Consistent with earlier studies, we found that the major factor governing transcriptional noise is the abundance. In theory, the transcriptional noise (η) due to the stochastic mRNA birth and death with constant probabilities per second is related to the mean number of mRNA by $\eta^2=1/\mu$, or $\log\eta^2=-\log\mu$. Other studies have shown that this relationship is also held for proteins expressed at low and moderate levels. At high abundance, however, the noise is almost unrelated to the mean (Newman *et al.*, 2006). In our study, the genes with low mRNA abundance (RPKM<10) were excluded because the fluctuations of the low-copy mRNAs were likely dominated by technical noise. In the region of moderate abundance, the slope estimated by linear regression is around -0.5, which is larger than the ideal case (-1). We assume the deviation may be due to high extrinsic noise or technical variation. At high abundance, the noise level is barely related to the mean, where the major contribution of noise comes from extrinsic sources, including stochastic activities of global and pathway factors.

The analysis of transcriptional variation at different growth conditions permitted us to explore the effects of environmental factors on the noise profiles of genes in relevant

pathways. In general, the genes in the same pathway show similar noise features under different conditions, indicating propagation of noise in the pathway (Pedraza and van Oudenaarden, 2005). Some pathways present lower noise levels in AA starvation than under the isotonic/hypertonic treatment, such as the pathway of aerobic respiration–electron transport chain. The lower noise of this pathway might arise from additional regulation or control induced by the stress factor, given the fact that activation of this pathway is involved in alleviating the oxidative stress in cells lacking amino acids (Petti *et al.*, 2011). We also found that many gene modules examined are noisier in AA starvation than in the isotonic/hypertonic treatment, e.g. genes involved in ribosome biogenesis and translation. Note that the higher noise level should not be attributed to the lower transcription level in AA starvation, since the noise residual (NR) is shown to be independent of the mRNA abundance. The higher noise level may explain the enhanced phenotypic diversity often observed when cells are stressed. The variability of cell responses to stress conditions may permit a population to maximize the chances of at least some cells' survival in an adverse environment (Raj and van Oudenaarden, 2008). Nevertheless, we assume that the different exposure time in treatment may result in different growth rates and cell sizes, which are two important factors affecting the level of mRNAs of ribosomal protein genes (Warner, 1989). In contrast to our result that the mRNAs of ribosomal protein genes are correlated with high level noise, a previous study on protein level found that the ribosomal proteins exhibit low variations (Newman *et al.*, 2006). The discrepancy might be due to the different experimental protocols. First, their cells were grown under a nutrients-rich normal condition, while our cells were under environmental stresses. Second, the post-transcriptional control mechanisms could

compensate for the mRNA variation and provide additional control to the final protein levels (Warner, 1989). Therefore, the variation on protein level may not faithfully reflect the variation of the mRNA levels.

In summary, our results indicate that transcriptional noise profiles of genes reflect their functional states and are subject to regulation. In addition, transcriptional noise can be used to understand transcriptional regulation and gene functions. The generalizability of these conclusions remains to be seen using a larger dataset collected under more conditions.

3.5 Future Works

We will further investigate the noise patterns using more cells under more growth conditions. Since the noise pattern of a gene is distinguished by the functional module it belongs to, we want to ask whether it is also true conversely, i.e., whether functional modules can be deduced from the transcriptional noise profiles. We hypothesize that with more growth conditions explored, more functional modules may be inferred from the noise analysis. Since regulons are dynamically regulated according to the external environments the cells are exposed to, it may also be of interest to know how many different growth conditions is needed to explore all possible regulons in yeast.

Our results show that the library qualities may vary due to the inevitable technical noise. Therefore, we will use RNA spike-in as an external control to estimate the effect of technical noise on genes at different expression levels. Based on the spike-in information, we expect that we can better control the quality by excluding genes with expression levels below a defined threshold, thereby assuring the accuracy of subsequent analysis. In addition, the absolute counts of a transcript can be estimated using the spike-in as reference. Normalization methods such as RPKM are based on the assumption that all cells have

similar amounts of total RNA. Unfortunately, this assumption does not hold especially when cells are cultured under different conditions and at different stages of a cell cycle. Therefore, using normalized transcription abundance without standardized controls may lead to erroneous interpretations in the subsequent gene expression analysis (Lovén *et al.*, 2012). Therefore, estimating the absolute counts instead of using a normalized abundance can be more reliable, particularly for single-cell analysis where the results can be quite sensitive to the starting amount of mRNA. Furthermore, we will extend the study by including replicate amplifications of bulk RNA diluted to near single-cell quantities. This will help to estimate the range of technical variation arising during amplification and sequencing preparation, which can be used as an additional way for quality control.

3.6 Methods

3.6.1 Experimental Methods

3.6.1.1 Cell Culture and Spheroplasts Preparation

A monoclone of the yeast strain S288C (ATCC) was selected using an YPD based agar (10% yeast extract, 20% peptone, 2% glucose and 20% agar) petri plate and stocked at -80 °C until use. To wake up cells, 30 µl thawed yeast stock inoculated in 3 ml YPD medium (1% yeast extract, 2% peptone and 2% glucose) was incubated overnight at 30 °C and 250 rpm. Cells were then expanded at 30 °C and 250 rpm after a 1:50 dilution in the YPD medium until mid-logarithmic phase (OD_{600} between 0.5 and 0.8). Five OD unit (ODU) cells were collected by centrifugation (500 g, 5 min) at room temperature. The cells were resuspended in autoclaved water and collected by centrifugation (500 g, 5 min) at room temperature. The cells were then resuspended in in the softening medium (100 mM Hepes-KOH, pH 9.4, 10 mM Dithiothreitol) and incubated in room temperature for 15 min. The

cells collected by centrifugation (500 g, 5 min) at room temperature were then resuspended in Spheroplasts (S) medium (1× YNB, 2% glucose, 1x amino acids, 50 mM Hepes-KOH, pH 7.2, and 1 M sorbitol) (Dunn and Wobbe, 2001) to a concentration of 5 ODU/ml. Zymolyase 100T was added to the spheroplasts suspension to a final concentration of 2 µl/ODU, followed by 60 min incubation at 30°C to remove the cell wall. After two washes in S medium by centrifugation (500 g, 5 min) at room temperature, spheroplasts were re-suspended to 5 ODU/ml in the desired treatment solution: AA starvation: S medium (with 1.0 M Sorbitol) without amino acid; carbon starvation: S medium (with 1.0 M Sorbitol) without glucose; hypertonic: S medium with 1.0 M Sorbitol; isotonic condition: S medium without sorbitol. Cells were exposed to the treatment for at least 30 min (up to 5 hours) before harvest.

3.6.1.2 Single Cell Harvest

Half mL of the spheroplasts were placed on a polylysine coated circular cover slip (2 mm diameter) in a petri dish for 5 min at room temperature (23 °C). The cover slip was broken in the center by a forceps, and a small piece of cover slip was transferred to a 30 µl perfusion chamber which was constantly perfused by a desired solution by gravity feeding. The solution change time in the chamber was about 20 sec. Single cells were harvested using a path clamp electrode pipette using a micromanipulator (ROE-200, Sutter) under an inverted microscope (Olympus 1X71) placed on a vibration isolation table (TMC). A cell was harvested in less than 10 nl perfusion solution.

3.6.1.3 Single Cell RNA-Seq Library Preparation

Our method is based on Tang et al. (Tang, Barbacioru, Bao, *et al.*, 2010; Tang, Barbacioru, Nordman, *et al.*, 2010) with modifications to prepare multiplex sequencing

libraries using Illumina Nextera XT Kit. Briefly, a harvested cell was quickly transferred using a home-made microinjection system to 200 µl Eppendorf tube containing 4 µl cell lysis buffer (0.9× PCR Buffer II, 3 mM $MgCl_2$, 0.45% NP40, 4.5 mM DTT, 0.18 U/µl SUPERase-In, 0.36 U/µl Rnase Inhibitor, 12.5 nM AUP1 primer, 2 mM dNTP). The cell was lysed at 70 °C for 90 sec, then placed on ice and stored at -80 °C until use. A cell lysate was thawed on ice, and 1 µl reverse transcription mix was added (13.2 U/ µl SuperScript III Reverse transcriptase, 0.4 U/µl Rnase Inhibitor, and 0.07 µg/µl T4 gene 32 protein). The first strand cDNA was synthesized by incubating the tube at 50 °C for 30 min, followed by inactivation of the reverse transcriptase at 70 °C for 10 min, and then the tube was cooled on ice. Free AUP1 primers were removed by adding 1 µl ExoSAP (Affymetrix) to the tube and incubating at 37 °C for 15 min, followed by inactivation of the ExoSAP at 80 °C for 15 min. This step would leave the AUP1 sequences at the 5'-end cDNA intact. A polyA tail was then added to the 3'-end of the first strand cDNA by adding 6 µl TdT mixture (1× PCR Buffer II, 1.5 mM $MgCl_2$, 3 mM dATP, 0.75 U/µl Terminal Transferase and 0.1 U/µl Rnase H) and incubating at 37 °C for 15 min, followed by inactivation of the enzyme at 70 °C for 10 min. The resulting products (12 µl) were then divided into two equal portions (each 6 µl), and each was mixed with 19 µl second strand buffer (1× High Fidelity PCR Buffer, 2 mM $MgSO_4$, 0.2 mM each dNTP, 0.3 µM AUP2 primer, and 0.1 U/µl high fidelity Platinum Taq DNA polymerase). The two tubes were subject to one PCR cycle (30 sec at 95 °C, 2 min at 50 °C and 6 min at 72 °C) to synthesize the second-strand cDNA in the form of 5'-AUP2-T24-cDNA-A24-AUP1-3'. Nineteen µl PCR mixture (1× High Fidelity PCR Buffer, 2 mM $MgSO_4$, 0.25 mM each dNTP, 2 µM AUP1 Primer, 2 µM AUP2 Primer, 0.1 U/µl Platium Taq DNA Plymerase High Fidelity) was added to each tube, which brings

the volume of each reaction to 44 μl, and cDNA was amplified by 18 PCR cycles (98 °C for 5 sec, 67 °C for 1 min and 72 °C for 6 min). The resulting cDNA from two reactions were combined (total 88 μl) and were further subject to 12 cycles of PCR with two duplicates, each with 2.4 μl sample and 87.6 μl PCR mixture (1× High Fidelity PCR Buffer, 2 mM $MgSO_4$, 0.375 mM each dNTP, 1 μM AUP1 Primer, 1 μM AUP2 Primer, 0.1 U/μl Platium Taq DNA Plymerase High Fidelity). The products were then combined and cDNA was revolved on a 1% ager gel (25 μl sample per lane). The band between 300 bases to the loading well was cut and cDNA was purified using a QIAquick gel purification Kit, followed by magnetic beads (GE Health) purification (10:7 sample to beads ratio). After quantification using a Bioanalyzer (Agilent High Sensitivity DNA Kit), the libraries were then prepared using an Illumina Nextera XT or TruSeq (libraries names start with 'A') DNA Sample Preparation Kit according to the vender's guide. The libraries were sequenced on an Illumina HiSeq2000 or HiSeq2500 machine (100 base-paired reads). Bulk RNA was extracted from population spheroplasts under the same treatments as single cells using a yeast RiboPure^{TM} RNA Purification Kit (Amion). Different amount of purified bulk mRNA (5 pg, 10 pg, 20 pg, 1,000 pg and 10,000 pg) were used to construct sequencing libraries in the same way as for single-cell libraries.

3.6.2 Characterization of Single-cell Transcriptomes

The raw reads were preprocessed by Cutadapt (version 1.2.1) (Martin, 2011) to remove adapter sequences if present. Reads were then mapped to the S. cerevisiae reference genome (Ensembl release 16) using TopHat (version 2.0.9) (Trapnell *et al.*, 2009). Raw counts of gene transcription level were quantified from the uniquely mapped reads by a custom program developed in-house, followed by normalization in RPKM (read per

kilobase coding region per million mapped reads) (Mortazavi *et al.*, 2008). A gene is considered expressing in a cell if the corresponding RPKM is over 0.1.

3.6.3 Model of Stochastic Gene Transcription

The fluctuations of mRNA levels can be modeled by stochastic formulation assuming that genes transit stochastically between active state (on) and inactive state (off) for transcription. For a gene with *g* copies (*g*=1 if the gene has no duplicate in a diploid) and each independently switches on and off with constant rates *a* and *b*, the mRNA fluctuations can be modeled as

$$\eta^2 = \frac{1}{< m >} + \frac{1 - P_{on}}{< g >} \frac{\tau_g}{\tau_m + \tau_g}$$

where <m> is the average level of mRNA; $P_{on} = \frac{a}{a+b}$ is the stationary probability that a gene is on ; $\tau_g = \frac{1}{a+b}$ is an effective time-constant for changes in gene activity; τ_m is the transcript lifetime (Paulsson, 2005). The first term reflects the random birth and death processes that occur in a manner following Poisson statistics. The second term reflects the random transitions in gene activity. If the cellular factors in the second term are constant, the gene transcription noise scales to 1/<m>. For an mRNA species at low abundance, its copy number variation is mainly due to spontaneous Poisson fluctuations. However, for an mRNA species at high abundance, its copy number fluctuation from the 1/<m> term can be negligible; on the other hand, fluctuations due to the global or pathways factors will dominate, resulting in a total noise level deviated from the 1/<m> trend.

3.6.4 Noise Analysis

Documented yeast pathways were downloaded from the SGD Yeast Pathways Database (pathway.yeastgenome.org). Only the pathways with 15 or more expressed genes were

used. We applied two separate robust linear regressions (rlm function in R), which iteratively reweights least squares with a bisquare weighting function, to the regions of moderate level noise and high level noise. The boundary of the two regions was selected to result in an intersection of the two fitted lines on the boundary. Genes associated with a specific functional annotation were downloaded by searching ‘gene and gene products’ from AmiGO2 with a ‘direct annotation filter’ (amigo.geneontology.org/amigo/search/bioentity).

CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

In this dissertation, we focused on developing novel computational methods to allow extensive investigations on the transcriptional regulation using single-cell gene expression dataset. Basically, computational methods are in great demand in the rapidly rising single-cell field and we endeavor to fill such needs. On the other hand, our works presented in Chapters 1-3 clearly have demonstrated the advantages of single-cell techniques in solving complex biological questions.

In Chapter 1, we designed computational methods to dissect the architecture of regulatory cascades and to reveal genes that play essential roles in driving the divergence of two lineages generated at each cell division in the embryonic development of *C. elegans*. To this end, we analyzed the EPIC dataset, which traced the expression level of reporter genes at single-cell resolution on a nearly continuous time scale up to the 350-cell stage in *C. elegans* embryos. We emphasized the importance of quality filter and data processing to compensate the delay of fluorescence before any type of analysis on the data. After carefully excluding dubious measurements and recalculating the expression value in each conceptual cell, we used a combination of statistical and classification methods to identify genes that best discriminate a pair of sister lineages yielded from a cell division. This chapter demonstrates how to use single-cell reporter gene data to decode regulatory architecture during embryogenesis, which would eventually lead to a comprehensive

understanding of the lineage/fate specification processes in embryogenesis. The work has been published in Developmental Biology (Xu and Su, 2014).

In Chapter 2, we designed a novel clustering algorithm, named SNN-Cliq, utilizing the concept of shared nearest neighbor and the technique of graph partition. SNN-Cliq is robust to cluster single-cell RNA-Seq data from various genomes. Using SNN-Cliq, we managed to identify cell types and developmental stages from transcriptomes. Beyond the high performance of our method, we brought into attention the pitfalls and obstacles that could prevent an effective clustering on high-dimensional single-cell transcriptomes. Our work is also an inspiring initiation of designing clustering algorithms for single-cell omic data. The work has been published in Bioinformatics (Xu and Su, 2015) and the program has been downloaded hundreds of times.

In Chapter 3, we aimed to delineate the gene transcriptional noise by analyzing single-cell transcriptomes in yeast under different environmental stresses. We observed different treatments show distinct transcriptome and transcription variations. Most importantly, this work is an effort to fill gaps in our understanding of the consequences of stochastic transcriptional regulation to individual cells and for populations. As a result, we quantified and compared the transcriptional noise in metabolic pathways and functional modules under each culture condition. Our results indicate that transcriptional noise profiles reflect the gene functions and are subject to regulation in response to environmental stresses. Still, many open questions remain in this project. One critical question is can we predict the gene functional modules solely from the noise profiles in a *de novo* way at the genome scale. Another is how to effectively separate biological noise from technical noise. To answer this question, we plan to add RNA spike-ins as an external control to estimate the effect of

technical noise on genes at different expression levels. We would also like to expand our study by involving more growth conditions, such as glucose starvation. We hypothesize that more functional modules may be inferred from the noise analysis by exploring more growth conditions.

To summarize, this dissertation is an extensive investigation of gene transcriptional regulation mechanisms in model organisms such as *C. elegans* and yeast at single-cell resolution. Note that the methods developed in these works can also be applied to other genomes. Most importantly, the single-cell technique is pivotal in solving some long-lasting biological questions such as drug resistance and cancer progression, and our works could become a useful guide and tools in conquering these problems.

REFERENCES

Asahina,M. *et al.* (2000) The conserved nuclear receptor Ftz-F1 is required for embryogenesis, moulting and reproduction in Caenorhabditis elegans. *Genes to Cells*, **5**, 711–723.

Bao,Z. *et al.* (2006) Automated cell lineage tracing in Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 2707–12.

Bar-Even,A. *et al.* (2006) Noise in protein expression scales with natural protein abundance. *Nat. Genet.*, **38**, 636–43.

Baugh,L.R. (2003) Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development*, **130**, 889–900.

Bendall,S.C. *et al.* (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, **332**, 687–96.

Bendall,S.C. *et al.* (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, **157**, 714–25.

Bertrand,V. and Hobert,O. (2010) Lineage programming: navigating through transient regulatory states via binary decisions. *Curr. Opin. Genet. Dev.*, **20**, 362–8.

Beyer,K. *et al.* (1999) When Is 'Nearest Neighbor' Meaningful? In, Beeri,C. and Buneman,P. (eds), *ICDT '99 Proceedings of the 7th International Conference on Database Theory*. Springer-Verlag London, UK, pp. 217–235.

Bowerman,B. *et al.* (1992) skn-1, a maternally expressed gene required to specify the fate of ventral blastomeres in the early C. elegans embryo. *Cell*, **68**, 1061–75.

Bowerman,B. *et al.* (1997) The maternal par genes and the segregation of cell fate specification activities in early Caenorhabditis elegans embryos. *Development*, **124**, 3815–26.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

Brennecke,P. *et al.* (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, **10**, 1093–5.

Broitman-Maduro,G. *et al.* (2006) Specification of the C. elegans MS blastomere by the T-box factor TBX-35. *Development*, **133**, 3097–106.

Buganim,Y. *et al.* (2012) Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, **150**, 1209–22.

Cai,L. *et al.* (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature*, **440**, 358–62.

Carey,L.B. *et al.* (2013) Promoter sequence determines the relationship between expression level and noise. *PLoS Biol.*, **11**, e1001528.

Carey,V. *et al.* (2011) RBGL: An interface to the BOOST graph library. *R Packag. version 1.40.1, cran.r-project.org*.

Citri,A. *et al.* (2012) Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nat. Protoc.*, **7**, 118–27.

Colman-Lerner,A. *et al.* (2005) Regulated cell-to-cell variation in a cell-fate decision system. *Nature*, **437**, 699–706.

Cowing,D. and Kenyon,C. (1996) Correct Hox gene expression established independently of position in Caenorhabditis elegans. *Nature*, **382**, 353–6.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.

Deng,Q. *et al.* (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–6.

Dong,D. *et al.* (2011) Gene expression variations are predictive for stochastic noise. *Nucleic Acids Res.*, **39**, 403–13.

Dunlop,M.J. *et al.* (2008) Regulatory activity revealed by dynamic correlations in gene expression noise. *Nat. Genet.*, **40**, 1493–8.

Dunn,B. and Wobbe,C.R. (2001) Preparation of protein extracts from yeast. *Curr. Protoc. Mol. Biol.*, **Chapter 13**, Unit13.13.

Edgar,L.G. *et al.* (2001) Zygotic expression of the caudal homolog pal-1 is required for posterior patterning in Caenorhabditis elegans embryogenesis. *Dev. Biol.*, **229**, 71–88.

Edgar,L.G. and McGhee,J.D. (1988) DNA synthesis and the control of embryonic gene expression in C. elegans. *Cell*, **53**, 589–99.

Eldar,A. and Elowitz,M.B. (2010) Functional roles for noise in genetic circuits. *Nature*, **467**, 167–73.

Elowitz,M.B. *et al.* (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–6.

Ertöz,L. *et al.* (2003) Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. *Proc. Second SIAM Int. Conf. Data Min.*

Ester,M. *et al.* (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*., pp. 226–231.

Fu,L. and Medico,E. (2007) FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*, **8**, 3.

Fukushige,T. *et al.* (2006) Defining the transcriptional redundancy of early bodywall muscle development in C. elegans: evidence for a unified theory of animal muscle development. *Genes Dev.*, **20**, 3395–406.

Fukushige,T. and Krause,M. (2005) The myogenic potency of HLH-1 reveals wide-spread developmental plasticity in early C. elegans embryos. *Development*, **132**, 1795–805.

Gasch,A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–57.

Gaudet,J. and Mango,S.E. (2002) Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. *Science*, **295**, 821–5.

Gionis,A. *et al.* (2007) Clustering aggregation. *ACM Trans. Knowl. Discov. Data*, **1**, 4–es.

Good,K. *et al.* (2004) The T-box transcription factors TBX-37 and TBX-38 link GLP-1/Notch signaling to mesoderm induction in C. elegans embryos. *Development*, **131**, 1967–78.

Guha,S. *et al.* (2000) Rock: A robust clustering algorithm for categorical attributes. *Inf. Syst.*, **25**, 345–366.

Guo,G. *et al.* (2010) Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell*, **18**, 675–85.

Hamatani,T. *et al.* (2004) Dynamics of Global Gene Expression Changes during Mouse Preimplantation Development. *Dev. Cell*, **6**, 117–131.

Hartuv,E. and Shamir,R. (2000) A clustering algorithm based on graph connectivity. *Inf. Process. Lett.*, **76**, 175–181.

Hashimshony,T. *et al.* (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, **2**, 666–73.

Horner,M.A. *et al.* (1998) pha-4, an HNF-3 homolog, specifies pharyngeal organ identity in Caenorhabditis elegans. *Genes Dev.*, **12**, 1947–1952.

Hornung,G. *et al.* (2012) Noise-mean relationship in mutated promoters. *Genome Res.*, **22**, 2409–17.

Houle,M.E. *et al.* (2010) Can shared-neighbor distances defeat the curse of dimensionality? In, Gertz,M. and Ludäscher,B. (eds), *Scientific and Statistical Database Management: 22nd International Conference, SSDBM 2010, Heidelberg, Germany, June 30–July 2, 2010. Proceedings*. Springer Berlin Heidelberg, pp. 482–500.

Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Huisinga,K.L. and Pugh,B.F. (2004) A Genome-Wide Housekeeping Role for TFIID and a Highly Regulated Stress-Related Role for SAGA in Saccharomyces cerevisiae. *Mol. Cell*, **13**, 573–585.

Hunter,C.P. and Kenyon,C. (1996) Spatial and Temporal Controls Target pal-1 Blastomere-Specification Activity to a Single Blastomere Lineage in C. elegans Embryos. *Cell*, **87**, 217–226.

Iba,W. and Langley,P. (1992) Induction of One-Level Decision Trees. 233–240.

Jarvis,R.A. and Patrick,E.A. (1973) Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans. Comput.*, **C-22**, 1025–1034.

Johnston,R.J. and Desplan,C. (2010) Stochastic mechanisms of cell fate specification that yield random or robust outcomes. *Annu. Rev. Cell Dev. Biol.*, **26**, 689–719.

Kaern,M. *et al.* (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.*, **6**, 451–64.

Kalb,J. *et al.* (1998) pha-4 is Ce-fkh-1, a fork head/HNF-3alpha,beta,gamma homolog that functions in organogenesis of the C. elegans pharynx. *Development*, **125**, 2171–2180.

Kaletta,T. *et al.* (1997) Binary specification of the embryonic lineage in Caenorhabditis elegans. *Nature*, **390**, 294–8.

Kalisky,T. and Quake,S.R. (2011) Single-cell genomics. *Nat. Methods*, **8**, 311–4.

Karypis,G. *et al.* (1999) Chameleon: hierarchical clustering using dynamic modeling. *Computer (Long. Beach. Calif).*, **32**, 68–75.

Krause,M. *et al.* (1990) CeMyoD accumulation defines the body wall muscle cell fate during C. elegans embryogenesis. *Cell*, **63**, 907–19.

Labouesse,M. and Mango,S.E. (1999) Patterning the C. elegans embryo: moving beyond the cell lineage. *Trends Genet.*, **15**, 307–313.

Lei,H. *et al.* (2009) Caudal-like PAL-1 directly activates the bodywall muscle module regulator hlh-1 in C. elegans to initiate the embryonic muscle gene regulatory network. *Development*, **136**, 1241–9.

Levin,J.Z. *et al.* (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods*, **7**, 709–715.

Li,J. *et al.* (2010) Exploiting the determinants of stochastic gene expression in Saccharomyces cerevisiae for genome-wide prediction of expression noise. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 10472–7.

Lin,R. *et al.* (1998) POP-1 and anterior-posterior fate decisions in C. elegans embryos. *Cell*, **92**, 229–39.

Liu,X. *et al.* (2009) Analysis of cell fate from single-cell gene expression profiles in C. elegans. *Cell*, **139**, 623–33.

Losick,R. and Desplan,C. (2008) Stochasticity and cell fate. *Science*, **320**, 65–8.

Lovén,J. *et al.* (2012) Revisiting global gene expression analysis. *Cell*, **151**, 476–82.

Macaulay,I.C. and Voet,T. (2014) Single cell genomics: advances and future perspectives. *PLoS Genet.*, **10**, e1004126.

Mace,D.L. *et al.* (2013) A high-fidelity cell lineage tracing method for obtaining systematic spatiotemporal gene expression patterns in Caenorhabditis elegans. *G3 (Bethesda).*, **3**, 851–63.

MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. In, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, Berkeley, Calif., pp. 281–297.

Maduro,M.F. (2010) Cell fate specification in the C. elegans embryo. *Dev. Dyn.*, **239**, 1315–29.

Maduro,M.F. *et al.* (2001) Restriction of Mesendoderm to a Single Blastomere by the Combined Action of SKN-1 and a GSK-3β Homolog Is Mediated by MED-1 and -2 in C. elegans. *Mol. Cell*, **7**, 475–485.

Maduro,M.F. *et al.* (2005) The Wnt effector POP-1 and the PAL-1/Caudal homeoprotein collaborate with SKN-1 to activate C. elegans endoderm development. *Dev. Biol.*, **285**, 510–523.

Maduro,M.F. and Rothman,J.H. (2002) Making worm guts: the gene regulatory network of the Caenorhabditis elegans endoderm. *Dev. Biol.*, **246**, 68–85.

Maheshri,N. and O'Shea,E.K. (2007) Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu. Rev. Biophys. Biomol. Struct.*, **36**, 413–34.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, pp. 10–12.

Mettetal,J.T. *et al.* (2006) Predicting stochastic gene expression dynamics in single cells. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 7304–9.

Mizumoto,K. and Sawa,H. (2007) Two betas or not two betas: regulation of asymmetric division by beta-catenin. *Trends Cell Biol.*, **17**, 465–73.

Moignard,V. *et al.* (2013) Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.*, **15**, 363–72.

Moore,J.L. *et al.* (2013) Systematic quantification of developmental phenotypes at single-cell resolution during embryogenesis. *Development*, **140**, 3266–74.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–8.

Munsky,B. *et al.* (2012) Using gene expression noise to understand gene regulation. *Science*, **336**, 183–7.

Murray,J.I. *et al.* (2008) Automated analysis of embryonic gene expression with cellular resolution in C. elegans. *Nat. Methods*, **5**, 703–9.

Murray,J.I. *et al.* (2012) Multidimensional regulation of gene expression in the C. elegans embryo. *Genome Res.*, **22**, 1282–94.

Newman,J.R.S. *et al.* (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, **441**, 840–6.

Paulsson,J. (2005) Models of stochastic gene expression. *Phys. Life Rev.*, **2**, 157–175.

Paulsson,J. (2004) Summing up the noise in gene networks. *Nature*, **427**, 415–8.

Pedraza,J.M. and van Oudenaarden,A. (2005) Noise propagation in gene networks. *Science*, **307**, 1965–9.

Pedregosa,F. *et al.* (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Pelkmans,L. (2012) Cell Biology. Using cell-to-cell variability--a new era in molecular biology. *Science*, **336**, 425–6.

Petti,A.A. *et al.* (2011) Survival of starving yeast is correlated with oxidative stress response and nonrespiratory mitochondrial function. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, E1089–98.

Phillips,B.T. and Kimble,J. (2009) A new look at TCF and beta-catenin through the lens of a divergent C. elegans Wnt pathway. *Dev. Cell*, **17**, 27–34.

Picelli,S. *et al.* (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, **10**, 1096–8.

Raj,A. *et al.* (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, **5**, 877–9.

Raj,A. *et al.* (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.*, **4**, e309.

Raj,A. and van Oudenaarden,A. (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, **135**, 216–26.

Raj,A. and van Oudenaarden,A. (2009) Single-molecule approaches to stochastic gene expression. *Annu. Rev. Biophys.*, **38**, 255–70.

Ramsköld,D. *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–82.

Raser,J.M. and O'Shea,E.K. (2004) Control of stochasticity in eukaryotic gene expression. *Science*, **304**, 1811–4.

Raser,J.M. and O'Shea,E.K. (2005) Noise in gene expression: origins, consequences, and control. *Science*, **309**, 2010–3.

Van Rijsbergen,C.J. (1974) Foundation of evaluation. *J. Doc.*, **30**, 365–373.

Rinott,R. *et al.* (2011) Exploring transcription regulation through cell-to-cell variability. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 6329–34.

Robertson,S.M. *et al.* (2004) Identification of lineage-specific zygotic transcripts in early Caenorhabditis elegans embryos. *Dev. Biol.*, **276**, 493–507.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–40.

Saliba,A.-E. *et al.* (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.*, **42**, 8845–8860.

Shalek,A.K. *et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**, 236–40.

So,L.-H. *et al.* (2011) General properties of transcriptional time series in Escherichia coli. *Nat. Genet.*, **43**, 554–60.

Stewart-Ornstein,J. *et al.* (2012) Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. *Mol. Cell*, **45**, 483–93.

St-Pierre,F. and Endy,D. (2008) Determination of cell fate selection during phage lambda infection. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 20705–10.

Sulston,J.E. (1983) Neuronal Cell Lineages in the Nematode Caenorhabditis elegans. *Cold Spring Harb. Symp. Quant. Biol.*, **48**, 443–452.

Sulston,J.E. *et al.* (1983) The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.*, **100**, 64–119.

Swain,P.S. *et al.* (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 12795–800.

Tang,F. *et al.* (2011) Deterministic and stochastic allele specific gene expression in single mouse blastomeres. *PLoS One*, **6**, e21208.

Tang,F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–82.

Tang,F., Barbacioru,C., Nordman,E., *et al.* (2010) RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.*, **5**, 516–35.

Tang,F., Barbacioru,C., Bao,S., *et al.* (2010) Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell*, **6**, 468–78.

Tarazona,S. *et al.* (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.*, **21**, 2213–23.

Telford,N.A. *et al.* (1990) Transition from maternal to embryonic control in early mammalian development: a comparison of several species. *Mol. Reprod. Dev.*, **26**, 90–100.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–11.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 5116–21.

Veenman,C.J. *et al.* (2002) A maximum variance cluster algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 1273–1280.

Volfson,D. *et al.* (2006) Origins of extrinsic variability in eukaryotic gene expression. *Nature*, **439**, 861–4.

Wang,Q.T. *et al.* (2004) A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev. Cell*, **6**, 133–44.

Warner,J.R. (1989) Synthesis of ribosomes in Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.*, **53**, 256–271.

Xu,C. and Su,Z. (2014) Identification of genes driving lineage divergence from single-cell gene expression data in *C. elegans. Dev. Biol.*, **393**, 236-44.

Xu,C. and Su,Z. (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics,* **31**, 1974-1980.

Yan,L. *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, **20**, 1131–9.

Zahn,C.T. (1971) Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Trans. Comput.*, **C-20**, 68–86.

Zenklusen,D. *et al.* (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.*, **15**, 1263–1271.

Zhang,S. *et al.* (2009) Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes. *Nucleic Acids Res.*, **37**, e72.

Zhu,J. *et al.* (1997) end-1 encodes an apparent GATA factor that specifies the endoderm precursor in Caenorhabditis elegans embryos. *Genes Dev.*, **11**, 2883–2896.

APPENDIX: SUPPLEMENTAL FILES

The Supplementary files for Chapter 1 can be downloaded from http://www.sciencedirect.com/science/article/pii/S0012160614003455 or http://www.webpages.uncc.edu/cxu3/.

The Supplementary files for Chapter 2 can be downloaded from http://bioinformatics.oxfordjournals.org/content/31/12/1974 or http://www.webpages.uncc.edu/cxu3/.

VITA

Chen Xu was born in 1986 at Xi'an, China. She received her B.S. of Biological Sciences from the University of Science and Technology of China in 2008. She then continued to pursue her M.S. degree in the McGill University in Canada, where she worked on epigenetic-environment interactions in Dr. Sarah Kimmins' lab. Her study of the influence of folate deficiency on sperm epigenome and the consequence of transgenerational epigenetic inheritance on the health of next generations was published in Nature Communications 2013 and was in the News published by over hundred media. She continued her journey in science and received her PhD in Computing & Information Systems and M.S. in Computer Science in 2015, both from the University of North Carolina at Charlotte. Most recently, she worked as a research assistant in Dr. Zhengchang Su's lab in the Department of Bioinformatics at UNCC, where she was exposed to the most cutting edge ideas and projects. She devoted most of her time developing computational methods to analyze high-throughput single-cell data. Her clustering algorithm SNN-Cliq (bioinfo.uncc.edu/SNNCliq/) shows high performance on the high-dimensional noisy single-cell RNA-sequencing (RNA-Seq) data. She also investigated the functions of gene expression noise and how it is regulated in response to environmental stresses in yeast. Her expertise is in machine learning and statistical modeling, especially in applying them to high throughput gene expression data to solve complex biological questions.