

12-1-2009

Statistical methods in microarrays and high-throughput flow cytometry

Osorio Meirelles

Follow this and additional works at: https://digitalrepository.unm.edu/biol_etds

Recommended Citation

Meirelles, Osorio. "Statistical methods in microarrays and high-throughput flow cytometry." (2009).
https://digitalrepository.unm.edu/biol_etds/81

This Dissertation is brought to you for free and open access by the Electronic Theses and Dissertations at UNM Digital Repository. It has been accepted for inclusion in Biology ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact disc@unm.edu.


OSORIO MEIRELLES
Candidate

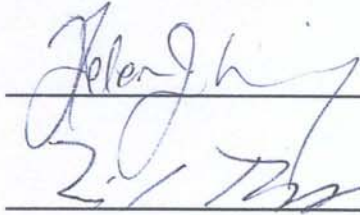
DEPARTMENT OF BIOLOGY
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

, Chairperson





**STATISTICAL METHODS IN MICROARRAYS AND
HIGH-THROUGHPUT FLOW CYTOMETRY**

BY

OSORIO MEIRELLES

MS MATHEMATICS, University of California, Irvine, 1994
BS MATHEMATICS, Pontificia Universidade Catolica, Rio
de Janeiro, Brazil, 1989

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

**Doctor of Philosophy in
Biology**

The University of New Mexico
Albuquerque, New Mexico

December, 2009

©2009, Osorio Meirelles

DEDICATION

I am forever thankful for all the support I have received from my spouse Michele, my son Marcos, my father Osorio, my mother Janet, Tom, my cousins and my friends.

ACKNOWLEDGMENTS

I heartily acknowledge Dr. Margaret Werner-Washburne, my advisor and dissertation chair, for continuing to encourage me through the years of classroom teachings and the long number of months writing and rewriting these chapters. Her guidance and professional style will remain with me as I continue my career.

I also thank my committee members, Dr. Eric Toolson, Dr. Helen Wearing and Dr. Donald Natvig, for their valuable recommendations pertaining to this study and assistance in my professional development.

To my colleges at the Werner-Washburne lab for the enriching discussions: Melissa, Swagata, Harriet, Sushmita, Elaine, Shannon, Philip, Anne, Jason Thomas and Anthony.

To my colleges at UNM Biology, also for the enriching discussions: Annette Evangelisti and George Davidson.

To my colleges at UNM Statistics for our valuable study groups sessions preparing for classes, qualifying and comprehensive exams: Alina, Raisa, Kristina, Tammy, Ed. Graham, Dan and Alvaro.

And finally to my wife, Michele, and my son Marcos, your love is the greatest gift of all.

**STATISTICAL METHODS IN MICROARRAYS AND
HIGH-THROUGHPUT FLOW CYTOMETRY**

BY

OSORIO MEIRELLES

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy in Biology

The University of New Mexico
Albuquerque, New Mexico

December, 2009

STATISTICAL ANALYSIS OF MICROARRAYS AND HIGH-THROUGHPUT FLOW CYTOMETRY

by

OSORIO MEIRELLES

B.S Mathematics, Pontificia Universidade Catolica, Rio de Janeiro, Brazil, 1989

M.S Mathematics, University of California, Irvine, CA, 2004

PhD. Biology, University of New Mexico, Albuquerque, NM, 2009

ABSTRACT-I

Background: Heterogeneous cell populations have previously been described as noisy. However, recent studies have demonstrated that heterogeneity can be biologically significant. We present here an approach for rapid and complete identification of heterogeneous cell populations from high-throughput flow cytometry data. We have developed a novel measure Slope Differentiation Identification (SDI) using flow cytometry-based protein expression, quantifying the rate of change in protein expression between two conditions (exponential and stationary phase) of yeast cells, as a function of cell size or cell granularity.

Results: SDI had superior Gene Ontology enrichment when compared with other approaches such as k-means clustering and an approach based on the bi-modality of the fluorescence intensity distribution. Cell populations were also validated using gradient-separation followed by microscopy, where proteins with high SDI measure showed significant levels of differentiation between high and low density cells.

Conclusion: Overall, our approach has identified novel protein expression patterns that differentiate quiescent and non-quiescent cell populations.

ABSTRACT-II

Background: With the advent of genomics, there has been a rapid increase in the use of two and one-color microarrays, used to measure mRNA abundance for the entire genome. Variability in microarray analysis undermines its utility in identifying the entire subset of differentially expressed mRNAs. Recent microarray studies have shown that, although it is assumed that variances are constant for every hybridized spot within a microarray, variances may differ for each biological sample analyzed (Ritchie, Diyagama et al. 2006). Another common assumption is that log-intensity values for any given gene have a Normal distribution. For many datasets, both assumptions have been shown to be incorrect, resulting in distortions in the significance when testing for differential expression of each gene (Bar-Even, Paulsson et al. 2006; Wentzell, Karakach et al. 2006).

Approach: To overcome the limitations of existing approaches in identifying significant, differentially expressed genes, we have developed a novel unsupervised statistical approach called Calibration Regression Analysis of Microarrays (CRAM) that uses a combination of empirical Bayes and regression calibration. The main novelty of our approach is the modeling of gene expression variances as a function of the log-intensity within each sample. Another version was later developed CRAM-GS in which the association between genes is captured using an adjusted gene correlation measure.

Results: CRAM was compared to four existing approaches for identifying differentially expressed genes. Performance was based on the ability to identify co-regulated genes in the same Gene Ontology process. CRAM exhibited a marginal improvement in GO process enrichment compared with the other approaches. To the original datasets, three more were included in which the later version CRAM-GS, showed a significant improvement compared to CRAM, suggesting a major additional benefit of incorporating gene correlations into the model. All versions of CRAM were two orders of magnitude

faster than the existing approaches. Overall, CRAM provides an adaptive, computationally efficient approach for accurate identification of differentially expressed genes.

TABLE OF CONTENTS

LIST OF FIGURES	xv
LIST OF TABLES	xx
CHAPTER 1 INTRODUCTION	1
Functional Genomics	1
New Technologies	2
Microarrays.....	2
Flow cytometry.....	4
Other technologies	6
Yeast as a model organism	6
Stationary Phase cultures in yeast.....	7
Quiescent and non-quiescent cells.....	9
Computational challenges.....	10
Gene Ontology enrichment.....	12
Concordance	12
Association	13
Dissertation overview	13
CHAPTER 2 REVIEW OF RELATED LITERATURE	18
Microarrays.....	18
Differentially expressed genes.....	18
Reproducibility	20
Ranked differential expression and generation of gene lists	24
Generation of gene lists - algorithm version 1.....	25
Generation of gene lists - algorithm version 2.....	28

Measures of gene list overlaps.....	33
Generation of p-value for Gene Ontology categories.....	33
Measuring GO enrichment - two gene lists	37
Analyzing pair-wise overlaps - examples.....	42
Multi-dimensional overlaps	45
GO enrichment evaluation of gene lists.....	53
Conclusion - Novel biological results.....	56
CHAPTER 3 - SDI.....	58
Abstract.....	58
Introduction.....	59
Methods	61
Generating the data.....	61
Visual identification of two-peak samples	61
Slope Differentiation Identification (SDI).....	62
<i>k</i> -means clustering	63
Average fold change.....	64
Identification of GFP strains and GO process categories.....	64
Results.....	65
SDI and two-peak plots	65
Biological process enrichment.....	69
Marginal enrichment.....	71
Intersection analysis of compared approaches	72
Microscopic examination of gradient-separated cells	73
Discussion.....	76
Conclusion	77

Acknowledgments	78
Supplemental Materials	79
Growth conditions	79
High-throughput flow cytometric screening.....	79
Flow dataset	80
Reproducibility analysis between biological samples	80
Reproducibility analysis between technical replicates	81
SDI algorithm	81
GO Term Finder settings	83
CHAPTER 4 - CRAM	84
Abstract.....	84
Introduction.....	86
Related work.....	89
Methods	90
Datasets.....	90
Method Overview	91
Linear model.....	93
CRAM model: sample-specific variance with different contribution per gene.....	98
Comparison of CRAM against other approaches	100
Other versions of CRAM.....	101
Results.....	109
Normality Assumption.....	109
Enrichment measures - CRAM.....	111
Enrichment measures - CRAM-GC.....	114
Discussion and Conclusion.....	118

Overcoming biological heterogeneity.....	119
Gene enrichment performance.....	120
Computational speed factors.....	122
Future applications.....	123
Acknowledgments	123
Supplemental materials.....	124
Sample variance distortion in biological heterogeneous datasets.....	124
Proof: $P(\boldsymbol{\tau}_i x_{ij})$	125
Proof: $P(\boldsymbol{\tau}_i \mathbf{X})$	127
Weight estimation.....	128
CRAM algorithm.....	129
Estimating the variance of the posterior expectation.....	131
Optimizing the alpha parameter.....	131
Alpha estimation - method 1.....	132
Alpha estimation - method 2.....	133
Other versions of CRAM.....	133
CHAPTER 5 DISCUSSION	141
Key ideas	141
Gene list overlaps	141
Concordance and measures of differential expression	142
Limitations of standard t-statistics.....	142
Concordance and moderated t-statistics	143
Concordance vs. GO enrichment.....	144
Associations and sample weight estimation	145
Linear association between samples.....	145

Concordance between samples	147
Linear association between genes.....	148
CRAM: Future work.....	149
Detection of heterogeneous cell populations - SDI	150
Multi-dimensional fold change.....	151
Modeling SDI using multi-dimensional flow data	151
Modeling SDI with multiple conditions	152
Limitations of concordance between samples	153

REFERENCES	155
-------------------------	------------

LIST OF FIGURES

Chapter 1

Figure 1. Microarray	3
Figure 2. Flow cytometer.....	5
Figure 3. Stationary phase in yeast	8
Figure 4. Quiescent and non-quiescent cell populations.....	4
Figure 5. SDI – scatter plot.....	15
Figure 6. CRAM – scatter plot	16

Chapter 2

Figure 1. Venn Diagram – Example of overlap between two sets	37
Figure 2. Venn Diagram – Example of overlap among three sets.....	38
Figure 3. Venn Diagram – Real example of overlap between two sets.....	39
Figure 4. Venn Diagram – Real example of overlap among three sets.	40
Figure 5. Venn Diagram – Real example of overlap among three sets	43
Figure 6. Venn Diagram – Real example of overlap among three sets	45

Chapter 3

Figure 1. Histogram of fluorescence intensity.....	65
Figure 2. Histograms of flow cytometry output	66
Figure 3. Scatter plot of log side-scatter vs. avg. log fluorescence intensity	67
Figure 4. Reg. scatter plot side-scatter vs. log fold-change in fluorescent intensity	67
Figure 5. Line plot with marginal enrichments	72
Figure 6. Venn Diagram – overlaps using SDI, SPV and SKM.....	75

Chapter 4

Figure 1. Histogram of standardized expected log-intensity –	110
--	-----

LIST OF TABLES

Chapter 1

Table 1. Example of a microarray dataset with log-intensity values	11
Table 2. Example of microarray dataset in which a gene list is selected	11

Chapter 2

Table 1. Description of datasets.....	23
Table 2. P-values and ratios for all pair-wise overlaps	44
Table 3. P-values and ratios for all pair-wise overlaps	45
Table 4. Approximate expectation for all six cases of triple overlaps	48
Table 5. Gene Ontology enrichment for most significant category	54
Table 6. Gene Ontology enrichment for most significant category	54
Table 7. Gene Ontology enrichment for most significant category	55
Table 8. Gene Ontology enrichment for most significant category	55
Table 9. Gene Ontology enrichment for most significant category	56

Chapter 3

Table 1. Gene Ontology biological process enrichment comparison	69
Table 2. Gene Ontology biological process enrichment comparison	69
Table 3. Number of genes in each GO biological process category	70
Table 4. Number of genes in each GO biological process category	70
Table 5. Visual microscopy identification of fluorescence differentiation	75

Chapter 4

Table 1. Description of the three additional datasets – CRAM-GC	103
Table 2. Enrichment Dataset 1 – CRAM.....	111
Table 3. Enrichment Dataset 2 – CRAM	112
Table 4. Enrichment Dataset 3 – CRAM.....	112
Table 5. Enrichment Dataset 4 – CRAM.....	113
Table 6. Enrichment Dataset 1 – CRAM-GC.....	115
Table 7. Enrichment Dataset 2 – CRAM-GC.....	115
Table 8. Enrichment Dataset 3 – CRAM-GC.....	116
Table 9. Enrichment Dataset 4 – CRAM-GC.....	116
Table 10. Enrichment Dataset 5 – CRAM-GC.....	117
Table 11. Enrichment Dataset 6 – CRAM-GC.....	117
Table 12. Enrichment Dataset 7 – CRAM-GC.....	118
Table 13. Enrichment for combined arrays – Q CRAM-GC.....	121
Table 14. Enrichment for combined arrays – NQ CRAM-GC.....	121

Chapter 1 – Introduction

1.1 Functional Genomics

Functional genomics, the study of levels of cellular organization in a whole-genome context, was developed as an outgrowth of genomic sequencing projects. Functional genomics aims to describe gene and protein functions and their interactions, extending basic concepts of genomics and proteomics by describing dynamic aspects such as transcription, translation and protein-protein interactions. In summary, functional genomics can be viewed as a dynamic evolution of the static aspects of genomic information such as DNA sequences. After the completion of the Human Genome Project in 2001, an immediate challenge using functional genomics, was to identify the relations between genes, proteins and the environment that were responsible for the evolution and functioning of dynamic living systems (Sebastiani, Gussoni et al. 2003). Our main goal is to apply quantitative methods in functional genomics in order to better understand the biological processes involved in many diseases such as cancer, aging and stem cells. For this purpose, we developed novel high-throughput statistical methods applied to recent technologies, to study stationary phase cultures in yeast. However, our methods can also be applicable to other types of experiments.

1.2 New technologies

Recent advances in technologies, such as *microarrays* (Fig. 1) and *flow cytometry* (Fig. 2), permit researchers to make inferences on dynamic living systems, by observing relations between thousands of mRNAs or proteins in an organism, under the same experimental conditions. Massive amounts of data resulting from these technological advances soon became available, giving rise to another challenge with the analysis of all this new information. To handle the size and complexity of all the new biological data, sophisticated and computationally intensive data analysis methods had to be developed.

1.2.1 Microarrays

Since the introduction of microarray technology in the 1990's, microarray experiments have been used in molecular biology and in medicine to quantify the abundance of all mRNA in an organism (Fig. 1), and to attempt to infer the relationship between mRNA abundance, biological development, disease and physiology (Eisen and Brown 1999). In the analysis after compiling all the networks and information in a microarray dataset, lists of candidate genes are generated and are often referred to as differentially expressed genes. The overall concept involving lists of candidate genes is the assumption that these genes are interrelated and are part of the same metabolic pathway.

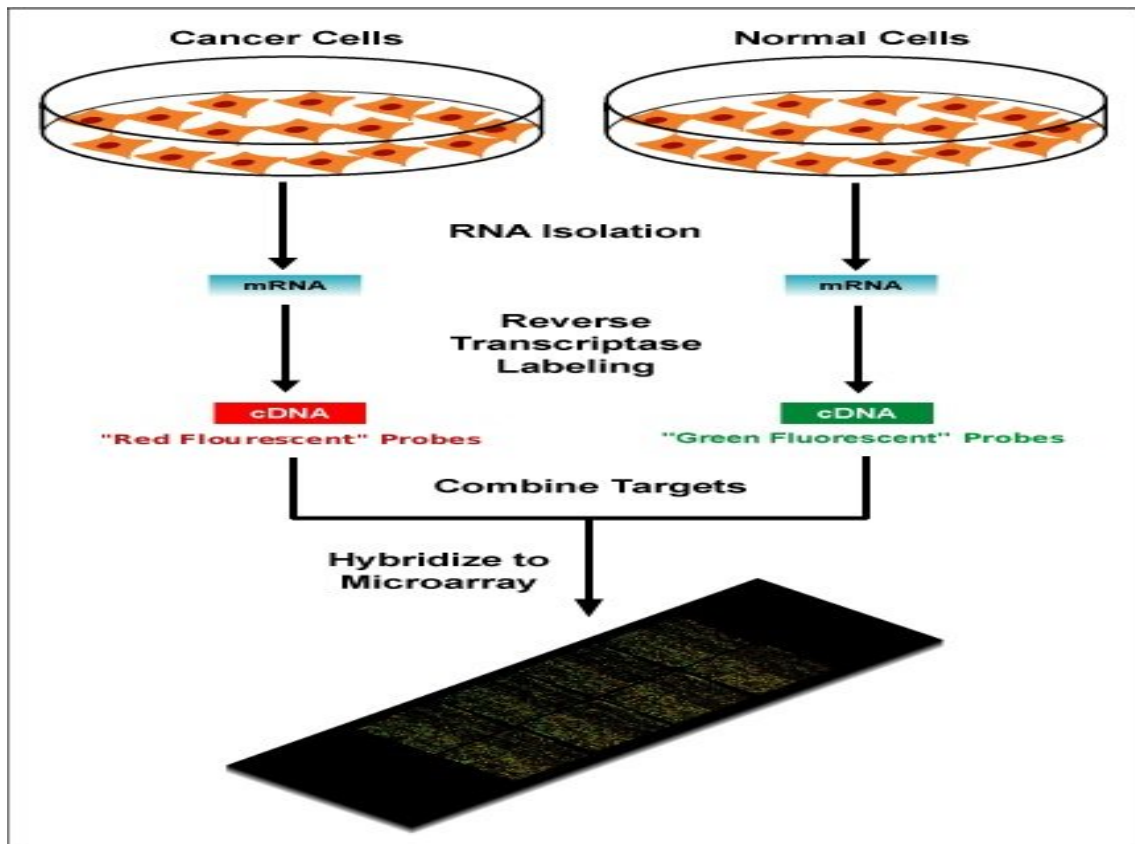


Figure 1 - (<http://www.dnamicroarray.net/>) In spotted microarrays the probes are oligonucleotides, cDNA, or small fragments of PCR products that correspond to mRNAs. Each probe contains a different, characteristic sequence that is specific to a different group of genes under study. These probes are then spotted onto glass substrate to form an array. One common approach uses an array of fine needles controlled by a robotic arm that is dipped into wells containing different DNA probes. Each needle then deposits its probe onto designated locations on the array surface. The probes are then ready to hybridize with complementary cDNA and cRNA targets derived from experimental or clinical samples.

One of the challenges in the analysis of microarray data is to integrate and compare the differential gene lists from multiple experiments for common or unique underlying biological themes (Yi, Mudunuri et al. 2009). One way to approach this challenge is by selecting common genes from these gene lists and then subjecting these genes to enrichment analysis in order to reveal the underlying biology. However, this

approach is highly restricted by the limited gene overlaps shared by datasets from multiple experiments, which could be originated by the complexity of the biological system itself. On the other hand, small gene overlaps can be the result of sub-optimal measures of differential expression, and gene list overlaps can be largely improved by the use of more accurate statistical measures (Shi, Perkins et al. 2008). In the current work, we introduce a novel statistical method, Calibration Regression Analysis of Microarrays (CRAM) and some of its variations, in which microarrays are used to optimally model gene expression in yeast (*Saccharomyces cerevisiae*).

1.2.2 Flow cytometry

Flow cytometry (Fig. 2) is a well established technique in cell biology, first developed in the 1970's for quantifying fluorescence of single cells and other morphological characteristics, such as size and granularity (Watson 1987). Since its introduction, the flow cytometer has rapidly become an essential instrument for biological sciences. In the cytometer, cells are suspended and aspirated into a flow chamber passing one at a time through a focused laser beam (Fig. 2). When the light strikes the cell, it is either scattered or absorbed, resulting in quantitative information for every cell. Since large numbers of cells are analyzed in a short period of time (>30,000/sec), a large amount of valid information about cell populations is quickly

obtained (Riley 2002).

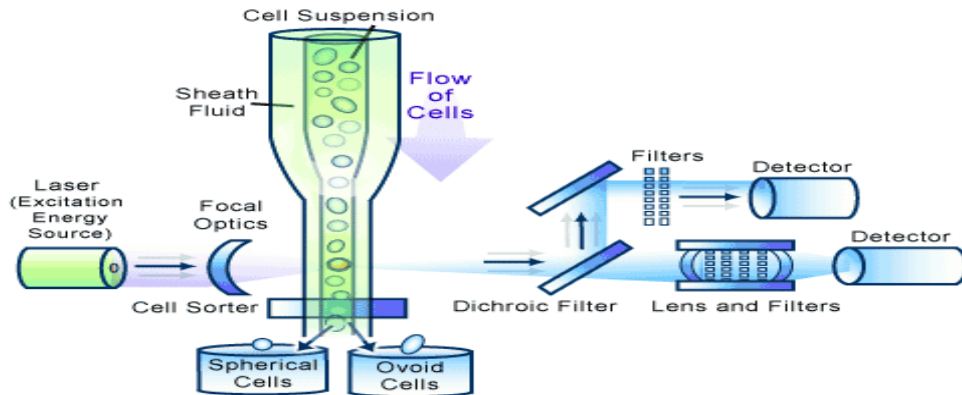


Figure 2 -

<http://www.bioteach.ubc.ca/MolecularBiology/FlowCytometry/flowcytometry.gif>

Detection of fluorescent measurements of cells passing through a laser beam, where many thousands of cells can be measured, counted and sorted.

Flow data has multiple dimensions leading to far greater computational analysis and high information content. These large numbers of multi-dimensional data points are from a statistical point of view, one of flow cytometry's major strengths (Kruzick, Irish et al. 2004). In the current work, a novel high-throughput flow cytometry method called Slope Differentiation Identification (SDI) is developed, based on protein expression measurements in yeast. Similar to microarrays, the primary goal is to generate statistically robust lists of candidate genes that reveal distinct populations of cells in heterogeneous cultures.

1.2.3 Other Technologies

A more recent technology called RNA-Seq, also called "Whole Transcriptome Shotgun Sequencing" (WTSS) is a recently developed approach in transcriptomics that uses deep-sequencing technologies. Studies using RNA-seq have already changed our view of the extent and complexity of eukaryotic transcriptomes (Wang, Gerstein et al. 2009). This method is also more precise in measuring levels of transcripts and their isoforms than other methods. RNA-Seq provides researchers with efficient ways to measure transcriptome data experimentally, allowing them to get information such as how different alleles of a gene are expressed, detect post-transcriptional mutations or identifying gene fusions (Maher, Kumar-Sinha et al. 2009). Although we do not analyze RNA-seq datasets in this dissertation, the same methods developed for microarrays can be easily extended to this new technology.

1.3 Yeast as a model organism

Yeasts (*Saccharomyces cerevisiae*) are eukaryotic, unicellular, microorganisms classified in the kingdom Fungi, with about 1,500 species currently described (Kurtzman 2006), although some species with yeast forms may become multicellular through the formation of a string of connected budding cells known as *pseudohyphae*. Yeast is also one of the most researched eukaryotic microorganisms in modern biology, where researchers have used it to gather information about the biology of eukaryotic cells and ultimately human biology (Ostergaard, Olsson et al. 2000).

Yeast was chosen as a model organism for many reasons. (1) It has low generation time. The average doubling time of a yeast culture is approximately 2 hours at 30 °C, making it suitable for growing cultures in a short amount of time. (2) Can be easily manipulated. It can be easily transformed by either altering genes (addition or deletion) through homologous recombination. The process of generating gene knockout strains is also largely simplified due to its ability to grow as a haploid. (3) DNA is highly conserved as an eukaryote. Yeast has similar complex internal cell structures of plants and animals, without the large amounts of non-coding regions from the DNA in higher organisms.

Yeast has been used to study cell cycle (Spellman, Sherlock et al. 1998), various responses to stress (Gasch 2002; Werner-Washburne, Wylie et al. 2002), and entry into (Gasch, Spellman et al. 2000; Radonjic, Andrau et al. 2005) and exit from stationary phase (Martinez, Roy et al. 2004; Radonjic, Andrau et al. 2005). In addition, yeast has been used to study many human diseases such as cancer (Simon, Szankasi et al. 2000; Marks, Rifkind et al. 2001) as well as the aging process (Ashrafi, Sinclair et al. 1999; Bitterman, Medvedik et al. 2003; Fabrizio and Longo 2003; McMurray and Gottschling 2004; Piper 2006; Kaeberlein, Burtner et al. 2007).

1.4 Stationary Phase cultures in yeast

Stationary phase is an identifiable component of the culture cycle of microorganisms that is functionally defined as the time when there is no further net

increase in cell number (Fig. 3) (Werner-Washburne 1993). When all external sources of carbon have been exhausted, cells enter stationary phase.

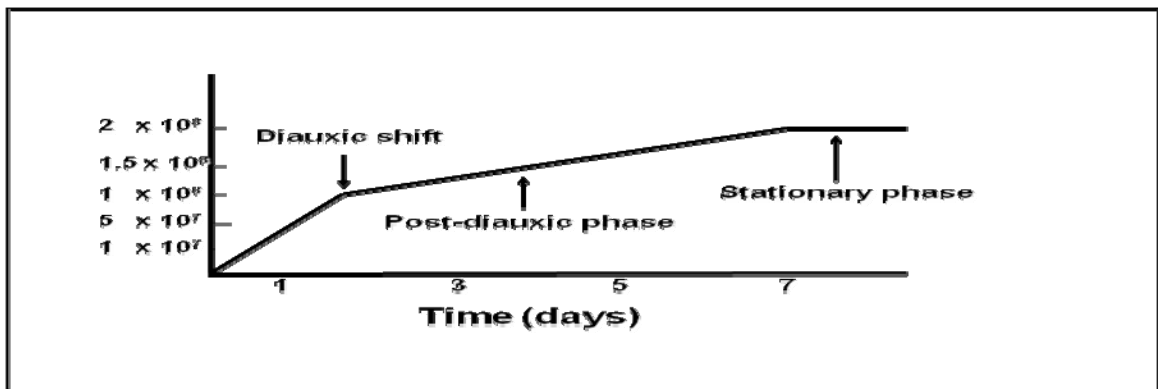


Figure 3 - Time chart with the different phases leading to stationary phase. Cells are initially in a glucose-rich media (YPD). In what is known as exponential phase, cells reproduce rapidly through glucose fermentation until all glucose is totally consumed, reaching what is known as diauxic shift. During post diauxic shift, cells change their metabolism to respiration, consuming ethanol as the primary energy source and reproducing very slowly.

Different phases occur before stationary phase when yeast cells are grown in a glucose rich medium: exponential phase, diauxic shift and post-diauxic shift. When cells in exponential phase have exhausted their sources of glucose, the diauxic shift occurs and they adapt from fermentation to respiration. During the post-diauxic shift cell growth is highly reduced and after approximately seven days cells enter stationary phase as a result of carbon starvation (Lillie and Pringle 1980). Stationary phase cultures have very different properties when compared to exponential phase cultures, such as a rate of translation 300 times slower (Fuge, Braun et al. 1994), a rate of transcription three to five times lower than exponential phase (Paz, Meunier et al. 1999), and are also highly resistant to stress (Werner-Washburne 1993). As cells exhaust glucose and enter stationary phase, they differentiate into quiescent (Q) and non-quiescent (NQ) cells (Fig. 4).

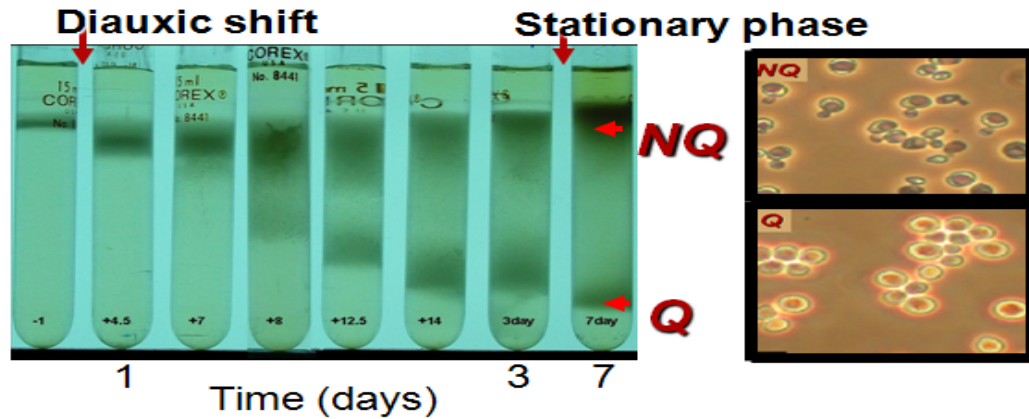


Figure 4 - Differentiation of quiescent and non-quiescent cell populations. After performing gradient centrifugation in stationary phase, two types of heterogeneous populations are observed, an upper (non-quiescent) and a lower band (quiescent).

1.5 Quiescent and non-quiescent cells

Quiescence is the most common cell cycle state on earth (Lewis and Gattie 1991).

Also known as G_0 , quiescence is critical to the survival of all organisms, where the efficiency of entrance and exit from quiescence provides a large selective advantage to microorganisms (Vulic and Kolter 2001) and also long-lived cells such as neurons (Morrison, Kinoshita et al. 2002) and egg cells (Bever and Izadyar 2002). Quiescent cells are also directly involved in tuberculosis (Parrish, Dick et al. 1998), cancer (Gray, Petsko et al. 2004), stem-cell maintenance (Suda, Arai et al. 2005), wound healing (Chang, Yang et al. 2002) and sexual reproduction.

Many phenotypic differences are present between quiescent and non-quiescent cell populations. Quiescent cells are denser, more refractive, have thicker membranes, are stress resistant, 90% are daughters and are synchronous in the cell cycle. On the other hand, non-quiescent cells are lighter, genetically unstable, less than 50% can divide and

show high levels of oxidative stress (Allen, Buttner et al. 2006) . In summary, differences between stationary and exponential phase cultures as well as differences between quiescent and non-quiescent cell populations make the study of stationary phase cultures ideal for high-throughput flow cytometry.

1.6 Computational challenges

With the latest developments in microarray and flow cytometry technologies, we face the ongoing challenges of analyzing large amounts of data, which have been growing at an exponential rate. In order to achieve quick and reliable results, large datasets require optimal algorithms. Although sophisticated computational procedures have been developed in recent years, many of these algorithms require either powerful computers or long processing time. Moreover, powerful computational algorithms often provide questionable benefits in detecting additional biological information when compared to simpler and more efficient ones (Fodor, Tickle et al. 2007). It is our goal to infer precise biological knowledge and at the same time to be computationally efficient.

Unsupervised models, which are characterized by the lack of a dependent variable, present some of the main challenges in quantifying biological knowledge, requiring sophisticated methods to extract biological relationships from experimental data. First we generate for all genes a measure of differential expression based on fold change in intensity (mRNA or protein abundance) and related t-statistics (Table 1).

Table 1 - Example of a microarray dataset with log-intensity values. After averaging the fold change in log-intensity for every gene over all three arrays, an expected log fold change (exp. FC) for each gene is calculated. After calculating the variance in log fold change intensity for every gene, the standard deviation is obtained and combined with the expected log fold change to generate a standard t-statistic.

Gene (ORF)	array1	array2	array3	gene var.	exp.FC	t-stat
YAL001C	0.03	0.24	0.53	0.063	0.27	1.06
YAL002W	0.28	-0.03	0.62	0.106	0.29	0.89
YAL003W	0.25	0.17	0.46	0.022	0.29	1.96
...
YPR203W	-0.47	-0.58	-0.5	0.003	-0.52	-9.09

Next, differentially expressed genes are selected into gene lists based on ranking the measure of differential expression (Table 2). Finally, the significance of gene lists is evaluated using performance measures. Two common performance measures are often used: Gene Ontology (GO) enrichment and gene list overlap (often referred to as “Concordance”).

Table 2 - Example of microarray dataset in which a gene list is selected. After sorting genes in descending order by expected log fold change intensity, the top 100 genes (blue) are selected into a gene list. A similar gene list with the top 100 genes could have been selected based on sorting genes by the t-statistic.

Gene (ORF)	array1	array2	array3	gene var.	exp.FC	t-stat
YIL101C	3.15	2.86	3.09	0.023	3.03	19.82
YML042W	2.92	2.74	2.9	0.010	2.85	28.92
YOL126C	3.31	2.34	2.84	0.235	2.83	5.83
YDR256C	2.63	2.62	3.11	0.078	2.79	9.95
YDR384C	3.02	1.77	3.2	0.607	2.66	3.42
YDR034W-B	2.32	2.93	2.67	0.094	2.64	8.62
...
YJL016W	1.46	1.05	1.37	0.046	1.29	6.00
YLR136C	1.15	1.13	1.6	0.071	1.29	4.87
YDR545W	0.2	1.75	1.88	0.874	1.28	1.37
YJR019C	1.06	1.67	1.09	0.118	1.27	3.70
YPL147W	1.06	1.23	1.51	0.052	1.27	5.57
YBR294W	0.83	1.4	1.57	0.150	1.27	3.27
YPR184W	1.3	1.33	1.15	0.009	1.26	13.07
...
YNL052W	-2.57	-3.63	-2.73	0.327	-2.98	-5.21
YCR021C	-3.1	-3.28	-2.9	0.036	-3.09	-16.27
YBR054W	-3.02	-4.1	-4.64	0.680	-3.92	-4.75

1.6.1 Gene Ontology enrichment

Gene Ontology (GO) enrichment is a powerful tool to infer biological information. The basis of GO enrichment derives from measuring the association between gene lists and known biological knowledge based on Gene Ontology (GO) categories (GeneOntology ; Boyle 2004). Genes are grouped into GO categories according to biological process, biological function or cell localization and then GO enrichment of a gene list is measured based on the number of genes in the list that belong to each GO category. If the number of genes from a specific GO category is significantly higher than expected by chance, we say the gene list is enriched in that particular GO category. More specifically, when a GO category is enriched, it will have a significant p-value, giving us the confidence that our gene list is detecting groups of genes likely to belong to the same pathway.

1.6.2 Concordance

Concordance is a measure of reproducibility, often defined as the percentage of overlap between two gene lists. The assumption is based on statistical measures to generate independent gene lists in which the overlap is significantly high. Under these conditions, we can claim that a high overlap is indicative of major biological information (Lee, Kuo et al. 2000). In other words, when the observed overlap is significantly higher than the expected overlap (where it is assumed no biological relationship exists), it is presumed that genes from the two lists describe similar biological patterns.

1.6.3 Association

Association can be viewed as an extension of the concept of concordance applicable to understanding relations between microarrays or between genes. Rank correlations are often applied as a robust measure of association (Kim, Rha et al. 2004). When using technical or biological replicates, microarrays with high rank correlation will provide a high confidence in the gene expression values. Similarly when there is a high rank correlation between genes, these are more likely to be in the same pathway and therefore are more likely to be in the same GO category.

1.7 Dissertation overview

In microarray data, differences in log-fluorescence intensity between two conditions are calculated and then a measure of differential expression is generated and used to determine differentially expressed genes. Researchers are often interested in the difference between a test and a control group, and similarly in differences between two test groups (different treatments). Under such conditions the concept of differential expression can be further extended for optimal modeling in multi-dimensional datasets such as flow cytometry, where differences in protein expression are measured.

The goal of this dissertation is to describe novel statistical methods based on unsupervised models applied to microarray and flow cytometry datasets. These methods are compared against current methods using exclusively GO enrichment and concordance

as performance measures for different datasets. The present work is divided into 5 chapters.

In chapter 2 we present applications of gene list overlaps between lists of differentially expressed genes and describe some of the main challenges that led the author to the development of more sophisticated algorithms described in chapter 4. Two versions of these algorithms that test for differential expression are presented, followed by the description of an application that combines concordance and GO enrichment for multiple gene lists. In addition, results from five yeast datasets are shown, together with a brief discussion. Moreover, we also describe a method for measuring the significance of multi-dimensional gene list overlaps. Summarizing, chapter 2 aims at describing some introductory ideas in measuring differential expression without getting into much detail.

In chapter 3, a method is described for rapid and complete identification of heterogeneous cell populations from high-throughput flow cytometry data. We present a novel measure, **Slope Differentiation Identification (SDI)** using flow cytometry-based protein expression. SDI is used to quantify the rate of change in protein expression between two conditions (exponential and stationary phase of yeast cells), as a function of size or granularity of cells (Fig. 5).

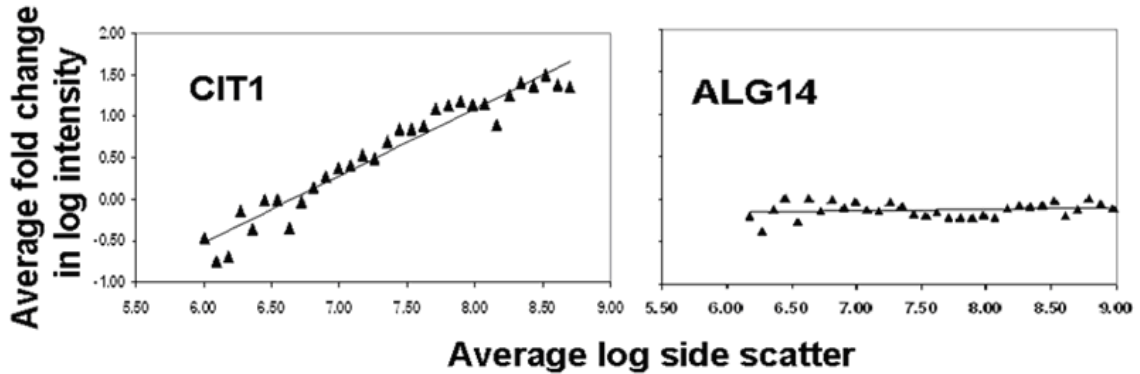


Figure 5 - Scatter plots with slope as a multi-dimensional measure of fold change for two yeast mutants (CIT1 and ALG14). The X axis (log side-scatter) is related to cell granularity, whereas triangles represent the corresponding average fold change in log-intensity between stationary and exponential phase cultures for different levels of side-scatter. On the left hand side, there is a strong indication that the slope is greater than zero for the CIT1 yeast mutant strain. On the right hand side, the slope for the ALG14 yeast is very close to zero.

Results showed SDI had superior GO enrichment performance when compared to other methods such as k-means clustering, average fold change and a method based on the bi-modality of the fluorescence intensity distribution, referred to as “Visual Two Peak Classification”. Cell population differences were also validated using a gradient-separation procedure in stationary phase followed by microscopy, where proteins with high SDI showed significant levels of differentiation between high and low density cells.

In chapter 4 we describe a method that incorporates in a systemic way, the concepts of concordance and association in order to provide more accurate measures of differential expression. In order to overcome the limitations presented by current methods in identifying differentially expressed genes, we developed a novel unsupervised statistical method called **Calibration Regression Analysis of Microarrays (CRAM)**, in which empirical Bayes and regression calibration are systemically conjugated. The main

novelty of CRAM is based on association between microarrays to model variance in gene expression as a function of the intensity levels within each microarray (Fig. 6).

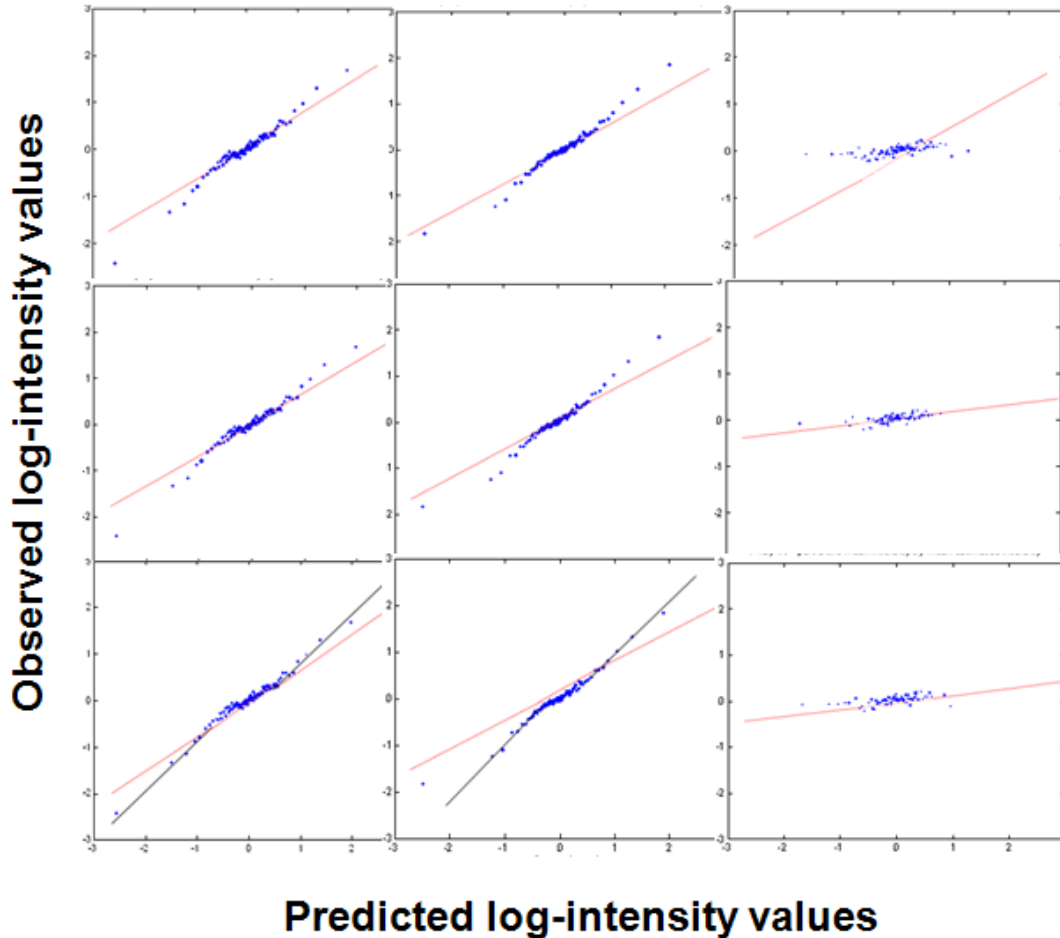


Figure 6 Scatter plots for each microarray under three different assumptions. Association is measured between sample log-fold change intensity and predicted log-fold change intensity from linear regression used by CRAM with remaining microarrays as explanatory variables. Slopes are a transformation inversely proportional to the variance. In the top row, gene variances from every sample are assumed constant and equal to each other (identical slopes). In the middle row, gene variances are assumed constant within each sample but different between microarrays (each microarray has a different slope). In the bottom row, gene variances are assumed to change within each microarray sample (sections within each sample have different slopes).

CRAM gene correlation (CRAM-GC), a more sophisticated version of CRAM that incorporates gene correlation is presented and compared against other known

methods. Results using 7 datasets are presented comparing CRAM and CRAM-GC against other methods. Whereas CRAM showed a marginal improvement when compared to other methods, CRAM-GC had significantly superior performance.

Chapter 5 summarizes results derived from chapters 2, 3 and 4 and describes potential improvements for the various methods presented. We show how the concepts of association and reproducibility are present in all methods. Similarly, these concepts also appear in gene list performance measures (GO enrichment and concordance), followed by a discussion of the relationship between variance in gene expression and concordance. In addition, the benefits of modeling the difference between two conditions and improvements in modeling multi-dimensional and multiple conditions using high-throughput flow data are discussed.

Chapter 2 – Applications

2.1 Microarrays

A DNA microarray is a technology that evolved from Southern blotting and is highly used in medicine and in molecular biology (Kulesh, Clive et al. 1987). Microarrays are generally classified into two types: two-channel (often referred to as cDNA or spotted microarrays) and one-channel (often referred to as oligonucleotide microarrays). Each microarray is made of thousands of spots of short nucleic acid polymers. These polymers can be a short section of a gene or other DNA fragment that are used as probes (usually 100 to 1000 bases long) to hybridize a cDNA (called target) under very specific conditions. In order to detect hybridization of the probe to its complementary target sequence, the probe is labeled with a fluorescent marker. Next, the level of fluorescence is quantified to determine relative mRNA abundance in the target.

2.2 Differentially expressed genes

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. Many steps in the gene expression process can be modulated, including the transcription, RNA splicing, translation, and post-translational modification of a protein. Often these products are proteins, but in non-protein coding genes such as rRNA genes or tRNA genes, the product is a functional RNA (Huttenhofer, Schattner et al. 2005).

In a microarray experiment, several replicates (technical or biological) are generally used to obtain gene-expression measurements for every gene. When technical replicates are used, all samples originate from the same tissue (humans) or the same culture (yeast) and are expected to be biologically identical. Thus, differences in expression can be attributed to measurement error such as array to array variation, reagent variation and dye incorporation.

When biological replicates are used, samples are expected to be biologically similar but not biologically identical. One example of biological replicates in yeast is when each array has a different yeast culture belonging to the same strain, which were grown following the same strict protocol but nevertheless, have an additional source of variation (besides all the sources of variation present in technical replicates) which is attributed to differences between cultures as they grow. We call the data in the experiments using technical or biological replicates as '*homogeneous datasets*'. In addition, some experiments have each sample originated from a different strain (yeast deletion set), which makes biological variability among samples much greater when compared to biological replicates, and thus an additional source of variability (variability between strains) is introduced. We call the data in these types of experiments '*heterogeneous datasets*'.

The primary reason for using replicates is to obtain a good level of significance for a combined measure of differential expression. Once a combined measure of gene-expression is obtained, it is used to generate lists of differentially expressed genes. Lists of differentially expressed genes are then used to understand the corresponding

associations between mRNA abundance with development, disease and physiology (Eisen and Brown 1999).

Ideally, if we performed the same experiment again, we would expect to obtain identical lists of differentially expressed genes (gene lists). However, the many sources of variation present in gene-expression measurements are such that each time a microarray experiment is repeated, a different gene list is almost always generated. A correct understanding of the variations in gene expression measurements in a microarray experiment is one of the main challenges in analyzing microarray data.

2.3 Reproducibility

As a basic requirement in microarray experiments, one must assume that although lists of differentially expressed genes generated by similar experiments are not identical, they should at least have high reproducibility. However, recent studies have shown how unreliable microarray experiments can be: lists of differentially expressed genes, generated by similar experiments, have low overlap between them (Ivanova, Dimos et al. 2002; Ramalho-Santos, Yoon et al. 2002; Tan, Downey et al. 2003).

Parametric models, commonly used to measure differential expression are based in standard t-statistics, in which gene lists are selected by their p-value ranks. Similarly, non-parametric models such as Wilcoxon rank-sum tests are also used to generate measures of differential expression and their corresponding p-values (Shi, Tong et al.

2005; Shi, Reid et al. 2006). In both parametric and non-parametric statistical models, gene lists generated by most methods have often resulted in low level of reproducibility.

Extensive comparisons in recent work have shown that gene lists based on simple measures of differential expression such as fold change are often much more reliable than using more complex methods (Shi, Perkins et al. 2008). Gene lists produced by fold change, are equivalent to gene lists using standard t-statistics, under the simplifying assumption that all genes have the same variance in gene expression. Thus, it comes as a surprise why this simplification would generate much higher rate of reproducibility than the more ‘statistically’ sound standard t-statistics (or other sophisticated statistics which assume different gene expression variances). Many possible reasons such as small number of microarrays (Ein-Dor, Zuk et al. 2006) and sub-optimal standards in the manufacturing processes (Tan, Downey et al. 2003), have been suggested without a final conclusion, as to why this happens.

The complexities involved in measurements of differentially expressed genes make it very likely that most gene expression models are based on incorrect assumptions, and thus incorrect p-values are generated. Ranking genes using fold change, which is equivalent to assuming that the variance in gene expression is constant for every gene (homogenous variances), is also not a perfect assumption. However, results based on the high reproducibility of gene list overlaps have suggested that the assumption of homogeneous variances is valid (Shi, Perkins et al. 2008). In contrast, most methods which assume a different variance for every gene, model correctly the middle of the true distribution of the measure of differential expression, but often generate distortions when

modeling the tails. Since selecting for lists of differentially expressed genes is a decision problem involving the tails, incorrect modeling of the tails will often produce sub-optimal gene lists.

In order to overcome the limitations of standard t-statistics we developed a measure of differential expression based on ranks of gene expression for every microarray. At the time our method was developed, the study by Shi and Perkins was not yet published, yet potential benefits of ranking each microarray based on the expression values had already been demonstrated (Qin, Kerr et al. 2004) .

Our approach was tested on five yeast datasets using two channel microarrays and where RNA transcript abundance (gene expression) was measured and normalized using the software GenePix 6.0. Each dataset consisted of yeast grown in stationary phase, where Q (quiescent) and NQ (non-quiescent) cell populations were separated from stationary phase cultures using density centrifugation.

Our goal was to identify differences in gene-expression between Q and NQ and determine the GO biological processes that differ most between these two cell populations. Since Q and NQ populations are sub-populations of stationary phase cultures, and are originated from the same subject, they are likely to be highly correlated and therefore a paired-design was used, where differences between transformed intensity values were taken for every Q/NQ pair of microarrays. Although our method was tested in paired-design datasets it is also applicable to non-paired designs. Summarizing, for all five datasets, each stationary phase culture gave rise to a pair of microarrays Q and NQ. For each corresponding pair of microarrays, differences in log-intensity between Q and

NQ were taken generating a corresponding array of paired differences in log-intensity (Table 1). Datasets 1, 4 and 5 use biological replicates and therefore are classified as homogeneous, whereas datasets 2 and 3 have different strains (deletion mutant), and therefore are classified as heterogeneous.

Table 1 - Description of datasets, number of arrays, number of paired-difference samples and total microarrays used in this study.

DATASET	Description	# strains	# cultures	# arrays	Q arrays	NQ arrays
1	Biological replicates in SP - BY4742	1	6	12	6	6
2	Different mutants in SP	80	80	160	80	80
3	Different mutants in SP	88	88	176	88	88
4	Biological replicates in SP - BY4742	1	16	32	16	16
5	Biological replicates in SP - S288C	1	10	20	10	10

In dataset 1, auxotrophic parental BY4742 strains were grown and separated into Q and NQ populations using a two-step density-gradient protocol (Allen, Buttner et al. 2006). As a result, 6 biological replicates (strains) from each Q/NQ population were used for a total of 12 microarrays. The remaining 4 datasets (Aragon, Rodriguez et al. 2008) were processed similarly to dataset 1. Due to the paired-design structure of our approach, we only used cell populations that separated into Q and NQ. Dataset 2 had 80 microarrays from the Q population and 80 from NQ, where each microarray measured the gene expression profile of a different yeast deletion mutant strain. Dataset 3, similar to dataset 2, had 176 microarrays from 88 mutant strains. Datasets 4 and 5 had 32 and 20 microarrays respectively, and correspond to 16 auxotrophic parental (BY4742) strains and 10 wild type (S288C) strains (Table 1).

2.4 Ranked differential expression and generation of gene lists

To identify the genes strongly associated between Q and NQ, we calculated for each gene a measure of differential expression between Q and NQ for all microarrays in the dataset. If the measure of differential expression for a gene is positive, this gene is more highly expressed in the Q population. Conversely, if the measure of differential expression is negative, this gene is more highly expressed in the NQ population. Our goal is to obtain gene lists from both Q and NQ and obtain GO enrichment from the *Saccharomyces* Genome Database Gene Ontology (Ontology ; TermFinder) as a measure of performance of gene lists.

Two versions of the ranking method, the original version (version 1) and an improved version (version 2) were developed, with both having superior performance compared to standard t-statistics. Both versions generate p-values based on t-statistics, with version 1 assuming different variances (after rank transformation) for every gene, and version 2 assuming the same variance (after rank transformation) for all genes. Initially we use these two versions to generate overall measures of differential expression, by combining transformed intensity values in all microarray pairs. Next, for each version we generated gene lists of different sizes, and for each size we compared GO enrichment between the corresponding genes lists for each version for Q only (due to the fact that enrichment results for NQ were very unstable since each gene list from a different method produces vary different GO enrichment output, making them hard to compare). Given the many limitations observed in the two versions, we developed a

much more powerful approach, Calibration Regression Analysis of Microarrays (CRAM), which is described in chapter 4.

2.4.1 Generation of gene lists – algorithm version 1

The goal of the rank transformation used by this version is to minimize the effect of outliers, which is done by normalizing the data using a z-statistic transformation for Q similarly, a z-statistic transformation for NQ. From this point on, the differences between the transformed z-statistics for Q and NQ are taken for every pair of arrays. These are then combined to generate a ranked based t-statistic. The following is a step-by-step methodology for the microarray statistical ranking analysis. For this purpose, let the dataset have $2k$ microarrays with k corresponding pairs of arrays of Q and NQ and n be the number of genes in the dataset.

Ranked t-statistics - version 1

For each array,

1. Replace each missing value with the median of the non-missing values in the array. This enables us to manipulate the expression values for every gene and at the same time, minimize the impact of the replacing value (which is close to zero). To minimize the bias in the data, we could also chose to add to the median, a small random error (such

as the mean error under a two way ANOVA), although in this particular case (using the five tested datasets), results were similar with or without the addition of a random error.

2. Rank each array in ascending order, generating a vector \mathbf{r} of size n , where each element r_i contains the rank of gene i within the array. This transforms the data in each array into ranks, thus minimizing the effect of outliers.

3. For each gene i , its rank r_i is transformed to a value p_i between 0 and 1, through the expression

$$p_i = (r_i - 0.5) / n$$

4. Use the inverse standard normal cumulative distribution function Φ^{-1} to transform p_i into a corresponding z-statistic (standard Normal random variable with mean zero and standard deviation one), which is a robust statistic for every gene.

$$z_i = \Phi^{-1}(p_i)$$

For each corresponding pair j of complementary arrays (Q and NQ),

5. Denote \mathbf{zQ}_j the vector of z statistics corresponding to the quiescent array and \mathbf{zNQ}_j the vector of z statistics for the corresponding non-quiescent array.

6. Let $\mathbf{d}_j = \mathbf{zQ}_j - \mathbf{zNQ}_j$, be a vector of differences with components d_{ij} corresponding to each gene i from array pair j . This step generates the vector of differences between the transformed statistics.

For each gene i ,

7. Calculate the average u_i over all elements of \mathbf{d}_i

$$u_i = \frac{1}{k} \sum_{j=1}^k d_{ij}$$

8. Denote s_i , standard deviation of d_{ij} across arrays for each gene i .
9. Let a_i which is assumed to be a t-statistic with $k - 1$ degrees of freedom, be given by

$$a_i = \sqrt{k} u_i / s_i,$$

which is measure of differential expression for gene i .

10. Estimate a two tail p-value for gene i using $T(a_i, k - 1)$, the cumulative distribution function of a t-statistic with $k - 1$ degrees of freedom evaluated at value a_i , given by

$$\text{p-value}_i = 2T(-|a_i|, k - 1)$$

11. At this stage, a p-value _{i} for each gene for the overall difference between Q and NQ has been calculated.

This p-value will be used to determine if a gene is significantly differentially expressed.

12. Let C be the level of significance for every gene, specified by the user.
13. If p-value _{i} < C , and $a_i > 0$, then the gene i is selected such that class _{i} = "Q".

This generates a gene list where genes are differentially more expressed in Q than in NQ.

If $p\text{-value}_i < C$ and $a_i < 0$ then $\text{class}_i = \text{"NQ"}$. This generates a gene list where genes are differentially more expressed in NQ than in Q.

Note: if $p\text{-value} \geq C$, class_i is not classified as "Q" or "NQ", that is $\text{class}_i = \textit{blank}$.

Note: steps 12 and 13 can be replaced if instead of selecting by p-value cutoff, we select for the m most significant p-values, where m is chosen arbitrarily by the user in a similar way as C .

2.4.2 Generation of gene lists – algorithm version 2

In version 1, which is based on a rank-transformed t-statistic measure of differential expression, we have achieved noise levels below those produced by a standard t-statistics. However, in version 1, genes have different variances in the rank-transformed gene expression values. In order to generate an even more robust measure of differential expression, we developed version 2, which is based on the assumption that variances in the transformed expression values are homogenous, and therefore assumes every gene has a variance equal to the average of the variances over all genes. This is equivalent to a fold change of the rank-transformed statistics. We expect that at least under certain conditions (heterogeneous data) version 2 will be more robust than version 1. This superior robustness in heterogeneous data from version 2, comes from the fact

that even with rank-transformed data, gene variances (estimated by version 1) are highly distorted when large biological variability is present among samples.

Ranked t-statistics - version 2

Steps 1-7 as in version 1.

For each gene i ,

8. Let a_i , which is assumed to be approximately a z-statistic, be given by

$$a_i = u_i / k$$

This is equivalent to an expected fold change across array pairs.

9. Estimate a two tail p-value for gene i , Φ the cumulative distribution function of the standard normal distribution is used:

$$\text{p-value}_i = 2 \Phi (-|a_i|)$$

10. At this stage, a p-value _{i} for each gene for the overall difference between Q and NQ has been calculated.

This p-value will be used to determine if a gene is significantly differentially expressed.

11. Let C be the level of significance for every gene, specified by the user.
12. If $\text{p-value}_i < C$, and $a_i > 0$, then the gene i is selected such that $\text{class}_i = \text{“Q”}$.

If $p\text{-value}_i < C$ and $a_i < 0$ then $\text{class}_i = \text{"NQ"}$.

Note: if $p\text{-value} \geq C$, class_i is not classified as "Q" or "NQ".

Note: steps 11 and 12 can be replaced if instead of selecting by p-value cutoff, we select for the m most significant p-values, where m is chosen arbitrarily by the user in a similar way as C .

The choice of C , the level of significance for cutoff purposes, depends on how many false positives one is willing to accept. For example, let us assume our dataset has a total of 6,000 genes and that by using a cutoff level = 0.01, we get a total of 200 genes more highly expressed in Q than in NQ. Under the assumption of a random scenario, the expected number of genes is equal to 60 genes (6000×0.01), which is the number of genes that could have been selected just by chance. Thus, we can say that within the 200 selected genes, we expect to have on average a total of 60 false-positive genes (genes that do not belong to the set of differentially expressed genes), leaving us with only 140 true genes. This corresponds to a false discovery rate equal to $30\% = (60/200)$. To illustrate this example in a more formal way, we define false discovery rate as the probability that a gene is our list is not differentially expressed and we represent it by $P(\text{False} \mid \text{List})$. So let:

F represent the event that the gene is not truly differentially expressed,

T represent the event that the gene is truly differentially expressed,

L represent the event that the gene belongs to the gene list.

Using Bayes theorem, we have:

$$P(F | L) = P(L | F)P(F)/P(L) = P(L | F)P(F)/(P(L | F)P(F) + P(L | T)P(T)).$$

Since we have no prior information about the probability that a gene is be either True or False, we let $P(T) = P(F) = \frac{1}{2}$ and therefore $P(F | L)$ simplifies to

$$P(F | L) = P(L | F)/(P(L | F) + P(L | T)) =$$

$$P(L | F) / P(L) = 1\%/(200 / 6000) = 1\%/(1 / 30) = 30\%$$

Assuming we had prior information that $P(T) = 5\%$, that is, only 5% of the genome is truly differentially expressed, we would have:

$$P(F | L) = (1\%)(95\%)/((1\%)(95\%) + (200/6000)(5\%)) = 85.1\%$$

The main idea is that since we had prior knowledge that it was much more likely for the gene not to be differentially expressed, the false discovery rate (FDR) became much higher. The main limitation when selecting a gene list is to know which genes are the ones truly differentially expressed and which are not. In the example described, we are unable to know which are the true 60 genes and the false 140 genes. In this case, prior knowledge of the probability of a gene being differentially expressed can be highly informative and will generally lead to a smaller gene list (based on a p-value cutoff), when the prior probability of a gene being differentially expressed is low. Since it is common for scientists to work with $FDR < 15\%$ (although this can vary depending on the experiment), and in the example where $P(T) = 5\%$, we had an unacceptable FDR of 85.1%, in order to keep FDR below 15%, we would have to make our p-value cutoff level under $3.1E-4$. A more detailed description of this framework, also referred to as False Discovery Rates (FDR) is found in (Benjamini 1995).

We should also be aware that any statement about false positives is based on the assumption that the p-values from the measure of differential expression are correct. In almost all methods, including versions 1 and 2, there are distortions on the tails of the distribution of the measures of differential expression which leads to distortions of p-value estimates (Fodor, Tickle et al. 2007). Thus, inferences involving false positive rates must be viewed with caution in all differential expression methods. The most common consequence of incorrect estimation of p-values in these tails, is the underestimation of p-values, leading to the generation of gene lists larger than they should be. However, the accuracy of the p-values generated by version 2, based on the suggestion by (Fodor, Tickle et al. 2007; Reid and Fodor 2008) is largely improved, resulting in a better fit for the distribution tails, and thus, generating more accurate gene lists. This improved modeling of the tail of the distribution of differential expression measures, is particularly useful when gene lists are generated based on p-value cutoffs.

2.5 Measures of gene list overlaps

A measure of Gene Ontology category enrichment for a gene list can be generated using the software application *GO Term Finder*, in order to generate p-values for GO categories obtained from a gene list. This software is used for measuring enrichment of a single gene list, which is equivalent to generating a p-value for the overlap between two sets of genes (the gene list originally generated and the set of genes belonging to a GO category). *GO Term Finder* uses a simple statistical method based on the hypergeometric distribution, which is also used to measure the significance of the

overlap between two sets of genes (generated as a result of different treatments). However, valuable biological information can be obtained when three or more sets of genes are observed with their multiple pair-wise overlaps. For this purpose we have developed a multi-dimensional approach to generating p-values for gene list overlaps. An example is described using the three sets of genes, which measures GO category enrichment using two gene lists (GO category set of genes + two gene lists).

2.5.1 Generation of p-value for Gene Ontology (GO) categories

Before describing the improved methodology to measure GO enrichment of a gene list, I will describe the standard methodology currently used by *GO Term Finder*. The goal of measuring GO enrichment is to infer biological knowledge based on the level of concentration of annotated genes present in our gene list. For this purpose, a gene list is submitted to *GO Term Finder* and a p-value is returned for the most significant categories from a set of approximately 2000 annotated categories. Next, I present an example of how the p-value is generated for a specific GO category.

Let T denote the total number of genes (in the population)

Let k be the number of genes of the GO category (the number of successes in the population)

Let m be number of genes in the submitted gene list (sample size)

Let l be the observed overlap, defined as the number of genes present in the overlap between the set of m genes from our submitted gene list and the set of k genes from the GO category. In other words, l is the number of genes present in both the submitted gene list and the GO category.

Next, under the hypergeometric distribution assumption, let X be the random variable representing the unknown number of successes of the sample before the observed outcome of l successes. The p-value after observing l genes belonging to the specific GO category is calculated by the following expression:

$$\text{p-value} = \sum_{i=l}^{\min\{k,m\}} P(X = i) ,$$

where $P(X = i)$, is given by the hypergeometric probability density function

$$P(X = i) = \frac{C_i^k C_{m-i}^{T-k}}{C_m^T}$$

Calculating the p-value under the hypergeometric distribution assumption is not straight forward since we need to use the hypergeometric cdf (cumulative distribution function), which is not present in standard software (such as Microsoft Excel), thus more mathematically sophisticated software is required. Thus, in order to estimate the p-value for the hypergeometric distribution, an approximation is used with very similar practical results. The p-value for the hypergeometric distribution function can be approximated by a standard normal distribution through the following procedure:

$$E[X] = \mu = mk / T \quad \text{and}$$

$$\text{Var}[X] = \sigma^2 = \mu(T - k)(T - m)/(T(T-1))$$

A z-statistic is calculated by

$$z = (l - \mu) / \sigma,$$

and a one-tail p-value is calculated by

$$\text{p-value} = 1 - \Phi(z)$$

where Φ is the standard Normal cumulative distribution function. The numerical example below clarifies the procedure.

Example:

Let $T = 6300$, the total number of genes of the data set.

Let $m = 630$, the number of genes in the gene list.

Let $k = 200$, the number of genes of a specific process, say “protein biosynthesis”.

Let $l = 100$, the number of observed genes in the category “protein biosynthesis”

belonging to our gene list. Then, $\text{p-value} = \sum_{i=100}^{200} P(X = i)$. Since the proportion of the

number of genes m in our gene list with respect to the total number of genes T is $m / T = 630/6300 = 10\%$. Thus, we expect to have by the independence assumption, 10% of the total number of genes from the “protein biosynthesis” category (10% of 200 = 20 genes) that belongs to our gene list.

The ratio $lT/(mk) = (100)(6300)/((630)(200)) = 5$, describes how much more concentrated the genes from the category “protein biosynthesis” are present in our gene list, compared to the expected number of genes from that category in a random gene list of size m . Both measures (p-value and the ratio) should be used together as a way to obtain the enrichment of a gene list. This approach may be extended by analyzing enrichment between overlaps of any pair of gene lists.

A one-tail p-value is used (instead of two-tail) in all calculations of GO enrichment, since researchers are interested in genes from a gene list that are more highly

concentrated (greater than the expected overlap under a random scenario) in a GO category, and are rarely interested in genes from a gene list that are underrepresented in a GO category. Moreover, if a researcher is interested in genes that are underrepresented in a particular GO category, it is unlikely that there will be enough statistical power to have a level of significance < 0.01 . As an example, let m equal the expected number of overlapping genes between our gene list and genes from a specific GO category. We assume two scenarios: we observe a total of $m+a$ overlapping genes in the first scenario, and a total of $m-a$ overlapping genes in the second scenario. The p-value in the first scenario is much more likely to be significant than in the second scenario. This is due to the fact that the probability of a gene from our gene list to belong to a GO category in most cases will be much smaller than the probability of the same gene not belonging to the GO category. Thus, the likelihood of observing an overlap with $m+a$ genes, will be greater than the likelihood of observing $m-a$ genes. Moreover, we can easily adapt the p-value generating algorithm to perform two-tail p-values.

2.5.2 Measuring GO enrichment – two gene lists

Here, we describe a novel method of measuring GO enrichment using two lists and compare it with the standard method using the ‘unknown’ GO category as an example. The reason for choosing the ‘unknown’ GO category is because it is suited to detect novel genes not yet assigned to any particular known GO biological process (Fig. 1 and 2). Although we have chosen to illustrate our examples with the ‘unknown’ GO

category, this enrichment approach using two gene lists can be applied in a single step to all the thousands of GO categories.

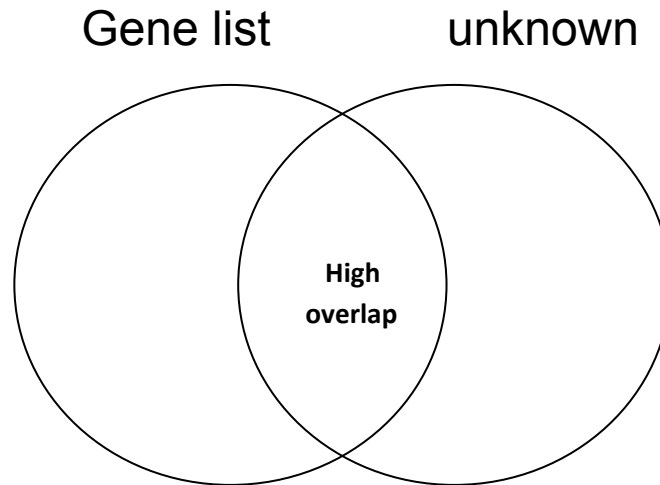


Figure 1- Example of a high overlap between the ‘unknown’ GO category and a single gene list. A high overlap, which is an overlap significantly greater than expected overlap between two sets by the independence assumption (or by chance) is indicative of potential candidate genes for belonging to either a new GO category or an existing one.

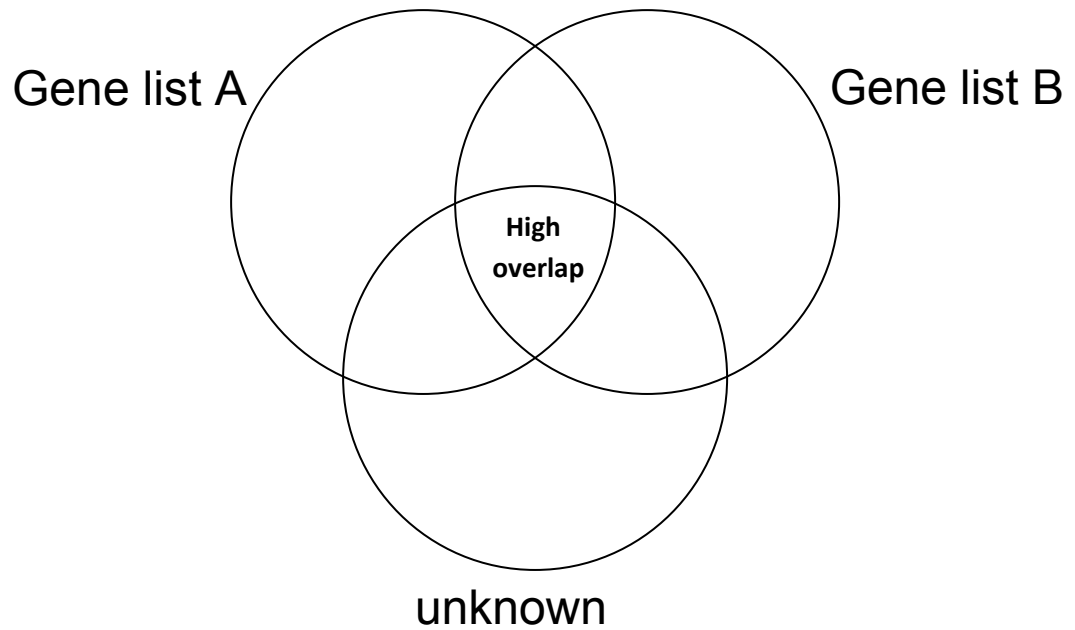


Figure 2 - Example of a high overlap among the ‘unknown’ GO category and two gene lists. A high triple overlap, which is an overlap significantly greater than the expected overlap among 3 sets by the independence assumption (or by chance), is indicative of potential candidate genes for belonging to either a new GO category or an existing one.

To illustrate the method, I will use the gene lists generated from datasets 2 and 3 obtained from (Aragon, Rodriguez et al. 2008). Datasets 2 and 3 generated respective gene lists A and B, of respective sizes 1080 and 1374 for genes classified as ‘Q’ with a total of 683 genes present in the overlap between both gene lists (Fig. 3).

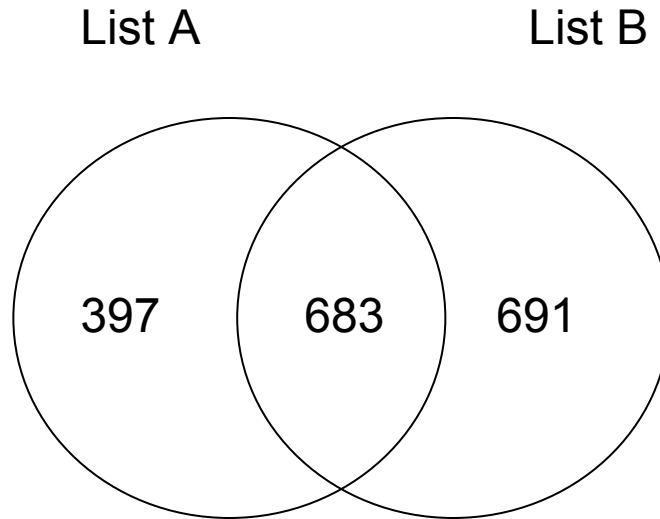


Figure 3 – Venn Diagram for gene lists high in Q relative to NQ, obtained from datasets 2 and 3. The overlap of 683 genes, measures the number of genes present in both lists.

There are 2540 genes in the “unknown” GO category annotated in the *GO database*, of which 232 genes are also present in the overlap of 683 genes between lists A and B. So if we name list U as the list of all 2540 genes in the ‘unknown’ GO category, the set of 232 genes will have a set theory representation as $(A \cap B) \cap U$. Assuming both datasets have a total of 6359 genes each, under the random assumption, then we have $2504 \times (683/6359) = 269$ expected number of genes (within the ‘unknown’ GO category) in the overlap. Generating the ratio between the observed 232 genes and the expected 269 genes, we have a ratio of $232/269 = 0.86 < 1$, which is less than the expected, under a random sampling, resulting in a non significant p-value.

Let us assume that the population is the set of all genes from the ‘unknown’ GO category, corresponding to the lower portion of the Venn diagram (Fig. 4), resulting in a total of 591 genes, representing the overlap between: the overlap of the ‘unknown’

category (2504 genes) and set A (385 genes), with the overlap between the ‘unknown’ category (2504 genes) and set B (438 genes), which is represented by $(A \cap U) \cap (B \cap U)$.

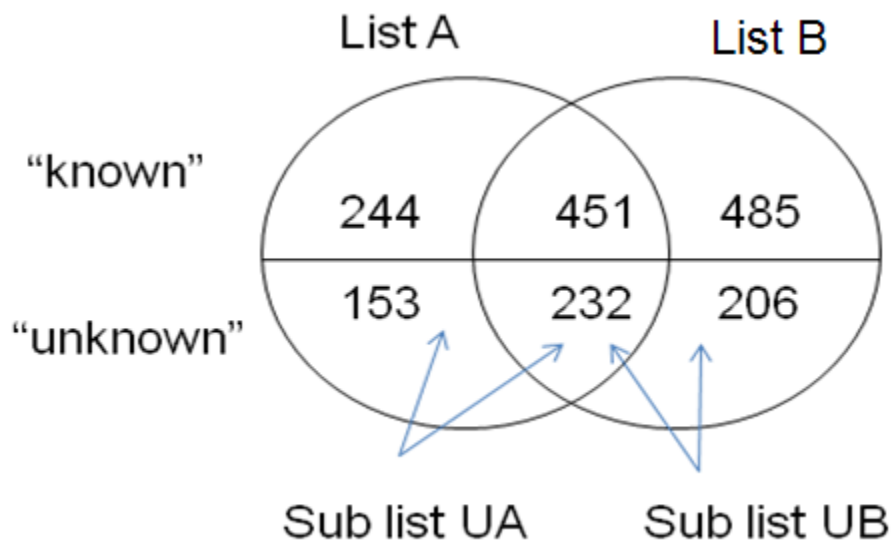


Figure 4 - Venn diagram for gene lists high in Q relative to NQ, obtained from datasets 2 and 3. The overlap of 232 represents the number of genes from the ‘unknown’ GO category present in sub list UA and also in sub list UB.

Thus, sub list UA (385 genes which can be represented by $A \cap U$), containing only genes from the ‘unknown’ GO category, and similarly sub list UB (438 genes which can be represented by $B \cap U$). The overlap between both lists is equal to 232 genes from the “unknown” category. However, the expected number of genes in the “unknown” category in the overlap between lists A and B is equal to $385 \times (438 / 2504) = 67.3$. This generated a p-value that was very significant ($2.8E-100$) and a ratio equal to $232/67.3 = 3.45$. The main interpretation of this result is that a highly significant number of the **same** genes from the “unknown” category are present in both lists A and B, even though

the **total** number of genes (232) in the “unknown” category in the overlap between A and B is less than the expected overlap (269). The representation in set theory of our list of 232 genes in this scenario is given by $(A \cap U) \cap (B \cap U)$. So it is interesting to notice that although the list of 232 genes can be represented by either $(A \cap B) \cap U$ or by $(A \cap U) \cap (B \cap U)$, which are identical, both methods give totally different p-values, as a result of different expected overlaps (269 vs. 67.3), and therefore depend on the order which they are applied using the hypergeometric distribution function. This is a characteristic of triple overlaps as well as higher dimensional overlaps which will be explained in more detail in section 2.5.3.

Consider the following example using a normal approximation for the hypergeometric distribution, is presented in which we estimate the p-value for the overlap between sub list UA and the sub list UB.

$$\mu = mk / T = 385 \times 438 / 2504 = 67.3$$

$$\sigma^2 = 67.3(2504 - 385)(2504 - 438) / (2504(2504 - 1)) = 47.00, \text{ then } \sigma = 6.85$$

$$\text{A z-statistic is calculated from } z = (l - \mu) / \sigma = (232 - 67.3) / 6.85 = 24.02$$

The p-value using the normal approximation to the hypergeometric distribution, which is given by the expression $1 - \Phi(24.02) = 8.5E-128$, when compared to the p-value (2.8E-100) using the true hypergeometric distribution, gives slightly different values, although for practical applications the conclusion is nearly the same, that is, the observed overlap of 232 genes is very significant, making these ‘unknown’ genes, good candidates for being classified into either a new GO category or an existing GO category. In addition, the high significance of the observed overlap, indicate that these genes are likely to be correlated and therefore are good candidates to belong to the same pathway.

We should keep in mind that when using $(A \cap B) \cap U$ to measure the significance of the triple intersection, we are asking the question: how significant is the number of ‘unknown’ genes present in both datasets A and B? On the other hand, when using $(A \cap U) \cap (B \cap U)$, the question is: how significant is the overlap between the ‘unknown’ genes from datasets A, and the ‘unknown’ genes from dataset B? As we can see, these are two different questions, each generating a different expected value for the triple overlap, and therefore, different levels of significance are obtained for the observed overlap of 232 genes.

2.5.3 Analyzing pair-wise and triple overlaps - Examples

In the previous section, we illustrated an example of how the significance of the triple overlap $(A \cap B \cap U)$ gave two different results, depending on the order in which the triple overlap was generated. This is a typical characteristic of triple overlaps. The overlap between three sets which is a natural extension of the commonly used overlaps between two sets, has many applications in biological problems. To better explain applications involving triple overlaps, two practical examples are illustrated using Venn diagrams from data from experiments in (Aragon, Quinones et al. 2006). One of the main goals in these experiments was to determine if stress resistance in stationary cells would protect against oxidative stress (induced by the use of menadione).

In our first example, three treatments were used in stationary phase cultures where each treatment generated a list of differentially expressed genes, corresponding to genes that had ≥ 2 fold increase in transcript abundance relative to its initial condition (before

the treatment). Three treatments were applied, Proteinase K, 1 minute oxidative stress and 30 minutes exposure to high-temperature, in which three gene lists were generated. A Venn diagram with the overlap between three gene lists under the following three different treatments produced the following results (Fig. 5).

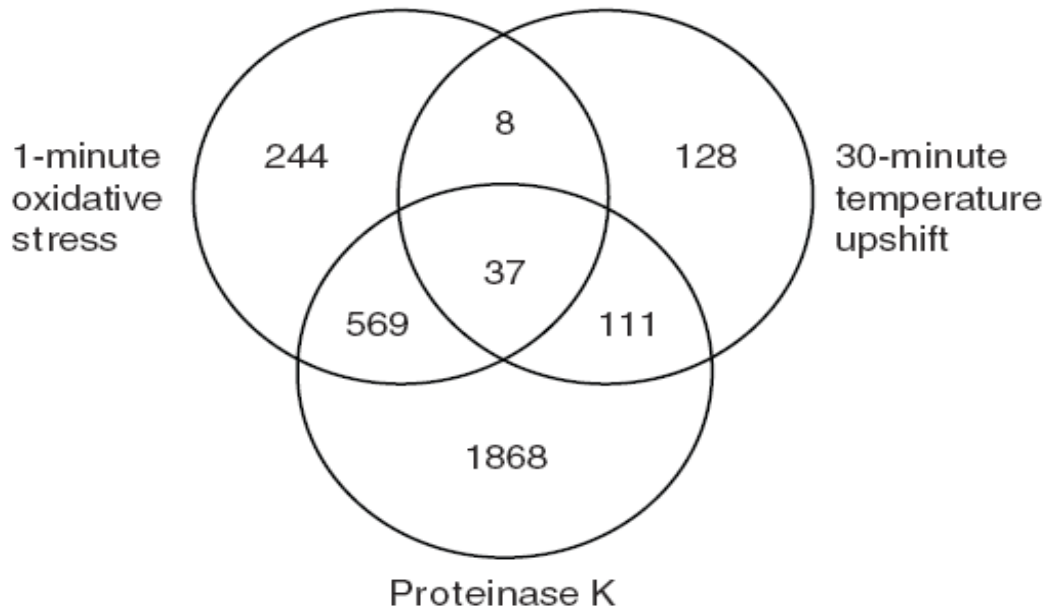


Figure 5 - Venn diagram of transcripts that increased after three treatments were used: oxidative stress, proteinase K or high temperature. Transcripts were evaluated if they had a ≥ 2 fold increase relative to T_0 cell lysates: abundance by 1 minute oxidative stress or 30 minutes after oxidative stress or after proteinase K treatment. Transcripts were also required to have good spots in 80% of the time points.

A significant overlap was detected between the gene list generated using Proteinase K and the gene list generated by 1 minute oxidative stress. Another significant overlap was detected between the genes list generated using Proteinase K and the gene list generated by a 30 minute exposure to increased temperature. There was a significant triple overlap (Table 2).

Table 2 - P-values and ratios for all pair-wise overlaps and the triple overlap between gene lists obtained from the treatments Proteinase K, oxidative stress or high temperature. Ratios are defined as the observed overlap divided by expected overlap. Total genes = 6359.

Overlap	p-value	Overlap	expected	Ratio
(PK vs. OS)	1.5E-164	606	263.18	2.30
(PK vs. HT)	1.1E-15	148	87.11	1.70
(OS vs. HT)	3.1E-01	45	39.29	1.15
(OS vs. HT) vs. PK	6.1E-09	37	18.3	2.02

In another experiment from the same article, a comparison between three other treatments was tested. The three treatments were Proteinase K, 1 minute oxidative stress and 30 minutes of oxidative stress. Similar to our previous examples, the three treatments were used in stationary phase cultures where each treatment generated a list of differentially expressed genes, corresponding to genes that had ≥ 2 fold increase in abundance relative to its initial condition. A Venn diagram showing the overlaps between all three treatments is presented (Fig. 6).

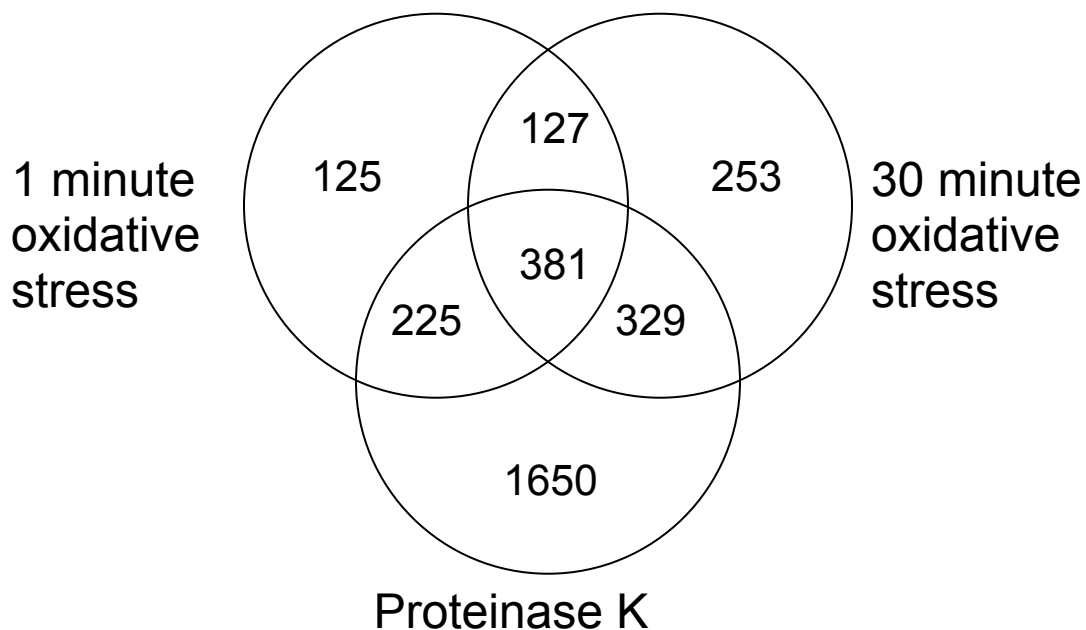


Figure 6 - Venn diagram of transcripts that increased by ≥ 2 fold increase in: abundance by 1 minute oxidative stress or 30 minutes after oxidative stress or after proteinase K treatment. Transcripts were also required to have good spots in 80% of the time points.

In this case, all pair-wise overlaps were very significant. The triple overlap was also significant although this was expected given the significance of all three pair-wise overlaps (Table 3).

Table 3 - P-values and ratios for all pair-wise overlaps and the triple overlap between gene lists obtained from the treatments Proteinase K, 1 minute oxidative stress or 30 minutes of oxidative stress for T₀ cell lysates. Ratios are defined as the observed overlap divided by expected overlap. Total genes = 6359.

Overlap	p-value	overlap	expected	Ratio
(PK vs. 1min OS)	2.4E-59	608	388	1.56
(PK vs. 30 min OS)	5.4E-49	710	493	1.44
(1min OS vs. 30 min OS)	2.9E-231	508	164	3.10
(PK vs. 30 min OS) vs. 1 min OS	1.6E-242	381	95.8	3.98

2.5.4 Multi-dimensional overlaps

We have seen some examples of how overlap analysis of gene lists is not limited to two dimensions and can potentially be extended to multiple dimensions. Since biological systems are such that multiple interactions can occur, overlap analysis can be used to effectively detect some of the higher order interactions. We will illustrate some potential applications of multi-dimensional overlap analysis in detecting higher order interactions.

2.5.4.1 Measuring the significance of a triple overlap

There are many ways of measuring the significance of a triple overlap between given gene lists A, B and C. We can measure the triple overlap by six different ways:

1. $(A \cap B) \cap C$
2. $(A \cap C) \cap B$
3. $(B \cap C) \cap A$
4. $(B \cap A) \cap (C \cap A)$
5. $(C \cap B) \cap (A \cap B)$
6. $(A \cap C) \cap (B \cap C)$

In the first case, we can evaluate the overlap $(A \cap B)$ first and obtain the p-value of the overlap $(A \cap B)$ with C. In the second case we can evaluate the overlap $(A \cap C)$ first and then obtain the p-value for the overlap $(A \cap C)$ with B, and finally, in the third case we can obtain the overlap between $(B \cap C)$ first and obtain the p-value of the overlap between $(B \cap C)$ with A. Similarly, in cases 4, 5 and 6, we take the overlap between pair-wise overlaps. The interesting fact is that all six cases will give

different p-values when measuring the significance of the triple overlap ($A \cap B \cap C$). This happens due to the fact that the expectation for the triple overlap is different in each one of the six cases. In order to define the significance of a triple overlap, we recommend choosing the most significant overlap between the six cases (the case where the expectation is the smallest). We will show in subsection 2.5.4.3, how to optimize the selection of the most significant triple overlap.

2.5.4.2 The non-greedy triple overlap

Let's assume gene lists A, B and C, are such that all pair-wise overlaps ($A \cap B$), ($A \cap C$) and ($B \cap C$) have insignificant overlap but ($A \cap B \cap C$) has a significant overlap. Although a multidimensional hypergeometric method is capable of detecting this type of overlap, they are quite limited since they do not use any knowledge of pair-wise overlaps. This limitation becomes clear in the example used in figure 4 with the 'unknown' GO category. In this example, by assuming independence between sets A, B and U, we have an expected triple overlap equal to $6359 \times (1080/6359) \times (1374/6359) \times (2504/6236) = 92$ which is still greater than the expected overlap ($A \cap U$) \cap ($B \cap U$) which is equal to 67. Moreover, we should also keep in mind that if we want to detect the significance of all overlaps in higher dimensions, we might run into a computationally unfeasible problem. A more realistic approach, applied to three or more dimensions may be found in the example using the greedy triple overlap described next.

2.5.4.3 Greedy overlap analysis

In the greedy overlap approach to evaluating the significance of triple overlaps, we assume that in order for a triple overlap to be significant, there has to be at least one

significant pair-wise overlap. To give a better idea, let us assume gene lists A, B, C are such that overlap $(A \cap B)$ is a significant overlap, but overlaps $(A \cap C)$ and $(B \cap C)$ are not significant, indicating that C does not interact with either A or B. However, if the triple overlap $(A \cap B) \cap C$ is significant, this implies that C interacts with the set $(A \cap B)$. When the number of genes in each set A, B and C is large (>50), we can use a binomial approximation to estimate the expected values of overlaps. We present how the expectation is calculated for each of the six cases previously described (Table 4).

Table 4 – Approximate expectation using the binomial approximation for all six cases of triple overlaps. In the first column, we have each triple overlap. In the second column we have the corresponding formula for calculating the expected overlap ($\#A$, $\#B$, $\#C$ represent the total number of genes from each set; $\#T$ represents the total number of genes used in the experiment; $\#(A \cap B)$, $\#(A \cap C)$, $\#(B \cap C)$ correspond to the number of genes in the pair-wise overlaps). The third column represents the expected overlap with the data from figure 5.

Triple Overlaps	Expected triple overlaps	expectation
1. $(A \cap B) \cap C$	$\#(A \cap B)(\#C/\#T)$	18.3
2. $(A \cap C) \cap B$	$\#(A \cap C)(\#B/\#T)$	27.1
3. $(B \cap C) \cap A$	$\#(B \cap C)(\#A/\#T)$	20.0
4. $(B \cap A) \cap (C \cap A)$	$\#(B \cap A)\#(C \cap A)/\#A$	31.8
5. $(C \cap B) \cap (A \cap B)$	$\#(C \cap B)\#(A \cap B)/\#B$	23.5
6. $(A \cap C) \cap (B \cap C)$	$\#(A \cap C)\#(B \cap C)/\#C$	34.7

Since the smallest expected overlap is 18.3 (from case 1), we select this as the optimal way to measure the significance of the triple overlap, and has a p-value equal to $6.1E-09$. Had we chosen the largest overlap (from case 6), with expected overlap equal to 34.7, the corresponding p-value would have been equal to 0.32, which is not significant. An interesting fact in this example, is that the p-values for the pair-wise overlaps $(A \cap B)$, $(A \cap C)$ and $(B \cap C)$ are respectively equal to 0.59, $1.3E-82$, $1.6E-22$, and at the same time, the overlap $(A \cap B) \cap C$ is the most significant triple overlap. We can say that since the pair-wise overlap $(A \cap B)$ is the least significant, it also has a low expected overlap,

making it a good candidate to also generate a low expected triple overlap. Thus, in a greedy approach, our best candidates to measure the significance of the triple overlap in the example in table 4, would be $(A \cap B) \cap C$ and $(A \cap B) \cap (B \cap C)$, which correspond to cases 1 and 5, $(A \cap B) \cap C$ is equivalent to the least significant pair-wise overlap $(A \cap B)$ overlapped with the remaining set A, and where $(A \cap B) \cap (B \cap C)$ is equivalent to the least significant pair-wise overlap $(A \cap B)$ overlapped with the second least significant pair-wise overlap $(B \cap C)$.

Applying the step-wise procedure to the results obtained in figure 4, where the ‘unknown’ GO category was used, if we take all three pair-wise overlaps $(A \cap B)$, $(A \cap U)$ and $(B \cap U)$, their corresponding p-values are respectively 9.9E-292, 4.9E-3 and 0.35. The least significant of these pair-wise overlaps is $(B \cap U)$, and therefore, this leaves us with two candidate overlaps: $(B \cap U) \cap A$ or $(B \cap U) \cap (A \cap U)$. The corresponding expected overlaps for both $(B \cap U) \cap A$ and $(B \cap U) \cap (A \cap U)$ are equal to 74.4 and 67.3, and therefore since $(B \cap U) \cap (A \cap U)$ has the smallest expected triple overlap, it is the most significant triple overlap.

The procedure of testing all combinations of three overlaps can be easily extended to testing all combinations of higher order overlaps. However, we should remember that the number of overlaps (cases as presented in table 4) increases exponentially as a function of the number of dimensions. Remembering that the step-wise procedure greatly reduces the number of overlaps in which the significance is measured, we can extend the procedure to higher order overlaps but at the same time we can be computationally efficient. This is the goal of the greedy approach.

We can extend this procedure to higher order overlaps using the greedy approach, if we consider only overlaps such as cases 1 to 3 in table 4 (we do not include the more complex overlaps such as cases 4 to 6). To illustrate the algorithm, we will use as an example a 5 dimensional overlap, using sets (gene lists) A, B, C, D and E.

Let #T be the total number of genes in the experiment, and #A,#B,#C,#D and #E, be the total number of genes for corresponding sets A,B,C,D and E. We first need to know if the multiple overlap $A \cap B \cap C \cap D \cap E$ is underrepresented or overrepresented. We define as overrepresented if the number of genes in the multiple overlap

$$\#(A \cap B \cap C \cap D \cap E) > \#A(\#B/\#T)(\#C/\#T)(\#D/\#T)(\#E/\#T),$$

which is the expected overlap assuming all 5 sets are independent. Similarly we say $A \cap B \cap C \cap D \cap E$ is underrepresented if

$$\#(A \cap B \cap C \cap D \cap E) < \#A(\#B/\#T)(\#C/\#T)(\#D/\#T)(\#E/\#T) .$$

Let us assume $A \cap B \cap C \cap D \cap E$ is overrepresented. Then in order to maximize the significance of the multiple overlap, we should *minimize* the expected multiple overlap, which in this case is the expected fifth-order overlap. To minimize the expected fifth-order overlap, we chose the lower-order overlaps such that the expected overlap is the smallest as possible. Let the lift between two sets S1 and S2 be defined as:

$$\text{Lift}(S1,S2) = \#(S1 \cap S2)/(\#S1(\#S2/\#T))$$

For example, if $\#(A \cap B) < \#A(\#B/\#T)$, meaning that the pair-wise overlap $A \cap B$ is underrepresented, then we have $\text{Lift}(A, B) = \#(A \cap B)/(\#A(\#B/\#T)) < 1$. Thus Lift between two sets can also be viewed as a measure of relative expectation. In this example with 5 sets,

$$\#(A \cap B)(\#C/\#T)(\#D/\#T)(\#E/\#T) < \#A(\#B/\#T)(\#C/\#T)(\#D/\#T)(\#E/\#T),$$

which results in the fifth-order overlap $(A \cap B) \cap C \cap D \cap E$ being more significant than the fifth-order overlap $A \cap B \cap C \cap D \cap E$. Thus our goal is to find underrepresented ($\text{lift} < 1$) lower-order overlaps within $A \cap B \cap C \cap D \cap E$, such that the fifth-order expected overlap is the smallest. In this example, (where the fifth-order overlap is overrepresented), we start by selecting from all pair-wise overlaps, the one with the smallest lift.

Say the pair-wise overlap $(A \cap B)$ is such that its lift $\#((A \cap B) / (\#A(\#B/\#T)))$ is the smallest among all pair-wise lifts. For each remaining set, in our case sets C, D and E, we select the triple overlaps with the smallest lift among the lifts:

- 1- $\#((A \cap B) \cap C) / \#(A \cap B)(\#C/\#T)$
- 2- $\#((A \cap B) \cap D) / \#(A \cap B)(\#D/\#T)$
- 3 - $\#((A \cap B) \cap E) / \#(A \cap B)(\#E/\#T)$
- 4- $\#((A \cap B) / (\#A(\#B/\#T)))$

We notice that if the lift in case 4 is selected, it is because that lift is smaller than the lift in cases 1, 2 and 3. This case is necessary since we want any additional grouping of sets to improve (decrease) the lift, thus, if no additional grouping improves the lift, we would chose the lift from the previous step, which is the smallest lift up to this point.

Let case 2, be the overlap with the smallest lift. The fact that $(A \cap B) \cap D$ is chosen, indicates that the triple overlap $(A \cap B) \cap D$ is underrepresented, since $\#((A \cap B) \cap D)$ is less than $\#(A \cap B)(\#D/\#T)$ which assumes that the sets $(A \cap B)$ and D are independent. Next we select the fourth-order overlap with the smallest lift between

- 1 - $\#(((A \cap B) \cap D) \cap C) / (\#((A \cap B) \cap D)(\#C/\#T))$
- 2- $\#(((A \cap B) \cap D) \cap E) / (\#((A \cap B) \cap D)(\#E/\#T))$
- 3- $\#((A \cap B) \cap D) / (\#(A \cap B)(\#D/\#T))$

Once again, case 3 assumes cases 1 and 2 had greater lift than in the previous step. Say the smallest lift was case 2. Finally we chose the last set C to generate the fifth-order overlap, so we chose between

$$1- \#(A \cap B \cap C \cap D \cap E) / (\#((A \cap B) \cap D \cap E))(\#C/\#T)$$

$$2- \#(((A \cap B) \cap D) \cap E) / (\#((A \cap B) \cap D))(\#E/\#T)$$

Let us suppose case 2 (which was the lift from the previous step) was the smallest lift. Thus, we assume $\#(A \cap B \cap C \cap D \cap E) / (\#((A \cap B) \cap D \cap E))(\#C/\#T)$ is the smallest lift for the fifth-order overlap. The main interpretation of this fifth-order overlap is that the fourth-order overlap $(A \cap B \cap D \cap E)$ is underrepresented, such that $\#(A \cap B \cap D \cap E) < \#(A \cap B \cap D)(\#E/\#T) < \#(A \cap B)(\#D/\#T)(\#E/\#T) < \#A(\#B/\#T)(\#D/\#T)(\#E/\#T)$ and therefore $\#(A \cap B \cap D \cap E)(\#C/\#T)$ is the expected fifth-order overlap and also a good candidate for being the smallest expected overlap (we can't be sure it is the smallest since not all possible lower order combinations were considered, but the greedy solution is likely to be to be close to the optimal).

If the observed overlap $A \cap B \cap C \cap D \cap E$ was underrepresented, we would select at each step for the set that *maximized* the lift, such that the fifth-order overlap had the largest expected overlap, and therefore our observed fifth-order overlap would be the furthest from the expected fifth-order overlap, resulting in the most significant fifth-order overlap.

2.5.4.4 Variable gene list size overlaps

A more complex and potentially more powerful extension of overlap analysis results from applying a flexible cutoff for the selection of individual gene lists. As an example,

consider selecting cutoffs c_1 , c_2 and c_3 used to create lists A, B and C, such that the significance of overlaps is optimized. In this case, c_1 , c_2 and c_3 are optimally chosen such that the expected triple overlap is smallest and therefore the significance of the triple overlap is highest. However, due to the multiple choices involving c_1 , c_2 and c_3 , it would be required to adjust for optimistic p-values which occur as a result of testing for multiple cutoffs.

2.6 GO enrichment evaluation of gene lists – five datasets

Enrichment for gene lists based on versions 1 and 2 were generated from datasets described in section 2.2. Both versions were compared by selecting the most significant GO category of version 1 and then observing its performance when applied to version 2 (Tables 5 to 9). This will lead us to expect that in a random setting, version 1 will most often have a superior enrichment in relation to version 2. In (Table 5), derived from a set of biological replicates, version 2 was significantly superior to version 1 for small gene lists (50 and 100), suggesting that the variance is likely to be distorted leading to inflated t-statistics. Corresponding GO enrichment (not on table) for list sizes 50 to 400 for the standard t-statistics showed no GO enrichment.

Table 5 - Gene Ontology enrichment for most significant category in genes significantly higher in Q than in NQ. Two tail p-values for the overlap between the two versions are obtained using the normal approximation of the Hypergeometric distribution. Dataset is homogeneous.

Dataset 1 - Quiescent list					
List size	Go category	version 1	version 2	# in cat	p-value
50	Transposition RNA-mediated	34	43	93	4.6E-13
100	Transposition RNA-mediated	44	69	93	2.1E-45
200	Transposition RNA-mediated	70	73	93	1.1E-01
300	Transposition RNA-mediated	71	77	93	2.3E-02
400	Transposition RNA-mediated	75	79	93	1.2E-01

In (Table 6), version 1 appears marginally better than version 2. Taking into consideration that the GO category was selected based on the lowest p-value from version 1 we can say that both versions have similar enrichment. Corresponding GO enrichment (not on table) for list sizes 50 to 400 for the standard t-statistics showed no GO enrichment.

Table 6 - Gene Ontology enrichment for most significant category in genes significantly higher in Q than in NQ. P-values are obtained using the normal approximation of the Hypergeometric distribution. Dataset is heterogeneous.

Dataset 2 - Quiescent list					
List size	Go category	version 1	version 2	# in cat	p-value
50	<i>not enriched</i>				
100	Response to oxidative stress	7	6	70	2.5E-01
200	Response to oxidative stress	12	8	70	2.6E-02
300	Response to oxidative stress	15	8	70	2.5E-03
400	Response to oxidative stress	16	13	70	1.5E-01

The results in (Table 7) show that version 2 is significantly superior to version 1 for gene list sizes of 200, 300 and 400. Taking into consideration that the GO category in version 1 was selected based on the lowest p-value, we can say that version 2 is superior to version 1 on this dataset. Corresponding GO enrichment (not on table) for list sizes 50 to 400 for the standard t-statistics showed no GO enrichment.

Table 7 - Gene Ontology Gene Ontology enrichment for most significant category in genes significantly higher in Q than in NQ. P-values are obtained using the normal approximation of the Hypergeometric distribution. Dataset is heterogeneous.

Dataset 3 - Quiescent list					
List size	Go category	match v1	match v2	# in cat	p-value
50	Alcohol metabolic processes	10	10	160	5.0E-01
100	Cell biosynthetic process	27	35	1619	9.5E-02
200	Cell biosynthetic process	46	74	1619	5.5E-04
300	Cell biosynthetic process	71	105	1619	5.5E-04
400	Cell biosynthetic process	95	125	1619	6.0E-03

The results in (Table 8) indicate version 1 is superior to version 2, and although GO categories in version 1 were selected based on the most significant GO category, the p-value for the difference between the two versions is very significant, pointing to a clear superior GO enrichment performance for version 1. Corresponding number of genes for each GO category (not on table) for list sizes 100 to 400 for the standard t-statistics were 12, 15, 21 and 24 genes. No GO enrichment was found for the t-statistics for list size 50.

Table 8 - Gene Ontology enrichment for most significant category in genes significantly higher in Q than in NQ. P-values are obtained using the normal approximation of the Hypergeometric distribution. Dataset is homogeneous.

Dataset 4 - Quiescent list					
List size	Go category	version 1	version 2	# in cat	p-value
50	Generation of precursor metabolites	9	5	181	7.9E-03
100	Monocarboxylic acid metabolic process	8	4	128	2.1E-02
200	Monocarboxylic acid metabolic process	16	11	128	3.5E-02
300	Monocarboxylic acid metabolic process	21	13	128	8.6E-03
400	Monocarboxylic acid metabolic process	23	19	128	1.5E-01

The results in (Table 9) indicate that version 1 is marginally superior to version 2.

Taking into consideration that the GO category in version 1 was selected based on the most significant GO category, we can say that version 1 and 2 are roughly equivalent.

Corresponding number of genes for each GO category (not on table) for list sizes 50 to 300 for the standard t-statistics were 9, 67, 118, 149 genes. No enrichment was found for the t-statistic for list size equal to 400.

Table 9 - Gene Ontology enrichment for most significant category in genes significantly higher in Q than in NQ. P-values are obtained using the normal approximation of the Hypergeometric distribution. Dataset is homogeneous.

Dataset 5 - Quiescent list					
List size	Go category	version 1	version 2	# in cat	p-value
50	Glycolysis	9	2	22	2.6E-33
100	Cell biosynthetic process	57	46	1619	3.6E-02
200	Cell biosynthetic process	117	75	1619	4.8E-07
300	Cell biosynthetic process	143	98	1619	7.8E-06
400	Cell biosynthetic process	163	114	1619	2.0E-05

2.7 CONCLUSION – Novel Biological Results

I have presented in this chapter the benefits of measuring the significance of multi-dimensional overlaps. A useful application was described in which the significance of a triple overlap was used to determine the enrichment of a GO category and any two gene lists. Another application measuring significance of triple overlaps involving experiments with 3 different treatments was also presented.

On the topic of Gene Ontology enrichment of gene lists, we have shown the benefits of different approaches based on the hypergeometric distribution applied to Venn diagrams. The main novelty derives from the combination of GO enrichment measures using two gene lists. This approach was described as a variation of the triple overlap, in which we showed that very different p-values could result from the same triple overlap, depending on the order of pair-wise grouping. Next we presented a general framework of how to detect multi-dimensional interactions by using multiple gene lists and described some interpretations of the results. Moreover, a greedy approach was described that can identify multi-dimensional biological interactions in a computationally efficient way.

On the topic of differential expression, different studies have shown conflicting results as to selecting gene lists assuming identical gene variances or selecting gene lists

based on standard t-statistics assuming different variances for every gene. The results presented suggest that the ideal choice between equal or different gene variances is likely to depend on the dataset. To address some of these limitations, we describe in chapter 4 the CRAM algorithm which is far more sophisticated, and of which the latest version, CRAM-GS, incorporates correlations between genes into the model.

In homogeneous datasets, versions 1 and 2 were roughly equivalent, whereas in heterogeneous datasets, version 2 was marginally superior. The main reason version 2 was superior in heterogeneous datasets is primarily due to its assumption of same variance for rank-transformed intensity values, which provides greater robustness than version 1, making it better to model the more noisy data, present in heterogeneous datasets. We should also keep in mind that although versions 1 and 2 had similar GO enrichment performance, version 2 is better to infer measures of false positive rates, since it produces more accurate p-values.

Chapter 3 - SDI

A statistical approach for detection of heterogeneous cell populations in high-throughput flow cytometry data

Osorio Meirelles¹, Sushmita Roy², Ray Joe¹, Phillip Tapia¹, Chris Allen³, Mark B. Carter³, Susan M. Young³, Bruce S. Edwards³, Larry A. Sklar³, Margaret Werner-Washburne¹

¹ Department of Biology, University of New Mexico

² Department of Computer Science, University of New Mexico

³ Cytometry and Department of Pathology, Cancer Research and Treatment Center, University of New Mexico Health Sciences Center

3.0 ABSTRACT

Background. Heterogeneous cell populations have previously been described as noisy. However, recent studies have demonstrated that heterogeneity can be biologically significant. We present here an approach for rapid and complete identification of heterogeneous cell populations from high-throughput flow cytometry data. We have developed a novel measure Slope Differentiation Identification (SDI) using flow cytometry-based protein expression, quantifying the rate of change in protein expression between two conditions (exponential and stationary phase) of yeast cells, as a function of cell size or cell granularity. *Results.* SDI had superior Gene Ontology enrichment when compared with other approaches such as k-means clustering and an approach based on the bi-modality of the fluorescence intensity distribution. Cell populations were also validated using gradient-separation followed by microscopy, where proteins with high SDI

measure showed significant levels of differentiation between high and low density cells. *Conclusion.* Overall, our approach has identified novel protein expression patterns that differentiate quiescent and non-quiescent cell populations.

3.1 INTRODUCTION

Heterogeneous cell populations while sometimes thought of as “noisy” can sometimes result from important differences in cellular function. For example, stationary phase cultures of the yeast *Saccharomyces cerevisiae* are known to be heterogeneous because of the formation of two populations of cells separable by density (Allen, Buttner et al. 2006; Aragon, Rodriguez et al. 2008). Other differences, such as age, cell cycle stage, cellular differentiation, and other non-random intra- and inter-cellular differences can contribute to heterogeneity (Raser and O'Shea 2004; Raser and O'Shea 2005).

Flow cytometry is a technology used to detect fluorescent measurements of cells passing through a laser beam, where many thousands of cells can be measured, counted and selected (HTC). The recent application of high-throughput flow cytometry using the yeast GFP-fusion library (Ghaemmaghami, Huh et al. 2003; Howson, Huh et al. 2005) (4159 strains, each with a green-fluorescent tagged protein), has led to new challenges in analysis of proteomics data. High throughput flow cytometry not only measures thousands of cells in each sample, but potentially hundreds of samples per minute, producing millions of data points per assay. Analysis of these massive flow datasets requires sophisticated computational methods for quantifying protein expression and detecting important population characteristics such as heterogeneity.

We developed a novel approach called Slope Differentiation Identification (SDI) to detect heterogeneous cell populations from high throughput flow cytometry data. Our approach detects heterogeneity by modeling change in fluorescence intensity from two conditions as a function of cell size or granularity.

We applied SDI to detect heterogeneous cell populations in a flow cytometry dataset measuring expression levels of ~4000 yeast GFP-fusion strains in stationary phase. Because stationary phase samples are known to be heterogeneous this dataset served as a good candidate for validation of SDI as well as discovery of novel strains exhibiting heterogeneity.

We compared SDI against other approaches for detecting heterogeneity, including visual inspection and three-dimensional k-means clustering. For each approach we tested if predicted heterogeneous strains were statistically overrepresented in biological process categories (GeneOntology). SDI outperformed these approaches, generating heterogeneous candidate strains that were more overrepresented in biological process categories, than other approaches. Additional validation with stationary phase cultures, showed SDI-identified GFP-fusion strains to be strongly associated with heterogeneous populations identified using gradient-separation and microscopy.

Overall SDI is a computationally efficient approach for analyzing flow cytometry measurements of thousands of proteins, and detecting strains that are statistically overrepresented in several biological processes. SDI-identified strains are also highly likely to form heterogeneous cell populations identifiable by microscopy.

3.2 METHODS

3.2.1 Generating the data

Slope Differentiation Identification (SDI) was applied on two high-throughput flow datasets measuring fluorescence intensity of GFP-fusion strains from stationary and exponential phase cultures. Each dataset contained three technical replicates for each of the 3941 GFP-fusion strains. A HyperCyt® autosampler (Edwards, Oprea et al. 2004; Young, Bologa et al. 2005) controlled by HyperSip software was used to measure fluorescence intensity of approximately 30,000 cells (events) per sample at a sampling rate of approximately 40 samples/min. The software package IDLQuery (IDLQuery) was used to capture and analyze data, generating output flow measurements for every strain. Each sample had approximately 30,000 three-dimensional measurements of fluorescence intensity, forward-scatter (cell size) and side-scatter (cell granularity). Overall, approximately 24,000 samples were analyzed in both datasets.

3.2.2 Visual identification of two-peak samples

The software package IDLQuery was used to generate fluorescence intensity histograms for each GFP sample. Each sample was visually classified as either two-peak, if it had bi-modal distribution of fluorescence intensity, or as one-peak, if the distribution of fluorescence intensity was uni-modal.

3.2.3 Slope Differentiation Identification (SDI) method for unsupervised two-peak detection

SDI can be generated using either one of the two flow-cytometric measurements: side-scatter and forward-scatter. We will describe only the SDI measure using side-scatter since the procedure for forward-scatter is identical. To generate SDI measure, we grouped side scatter measurements from the SP dataset into 100 bins. Side-scatter measurements from exponential phase were grouped into similar 100 bins. Although we chose to group by 100 bins, several different numbers of bins were tested with very similar results for number of bins between 10 and 200. Next, for every sample, profiles of average log intensity and average side scatter were generated, separately for stationary and exponential phase datasets using all events within each bin.

The SP samples were combined with their corresponding exponential samples, forming three sample-pairs* (one pair for each technical replicate). To obtain SDI for stationary phase, we first calculated the fold change in log fluorescence intensity by subtracting average log intensity in exponential phase from average log intensity in stationary phase. This was done for each sample for all bins, generating a profile of fold change as a function of side scatter, in stationary phase. After excluding GFP strain which had bad samples in either SP or exponential phase, the initial GFP library of 4159 strains was reduced to 3941 strains.

*The term “sample” and “sample-pairs” both refer to a strain carrying a specific GFP-fusion protein. A sample can be cells from a stationary or an exponential culture, whereas a sample-pair refers to information from both stationary and exponential phase for a strain carrying the same GFP-fusion protein.

We have regressed fold change in log intensity on the average log side scatter using only bins with ≥ 50 events in both stationary and exponential datasets, in order to assure statistical significance of fold change for each bin (we also tested other cutoff values from 100 to 500 in steps of 50, with similar results). Selecting bins with ≥ 50 events assured statistical confidence in estimations of average log-intensity and average log side-scatter. Regressing fold change in log intensity on the average log side-scatter, generated a slope for each sample-pair per replicate, resulting in three slopes for each GFP-fusion strain. These slopes were shown to have high reproducibility (please see supplemental materials for more details).

The median m_i is calculated for each gene (GFP) i , over the three slopes. An approximate z -statistic is obtained for every sample by subtracting each m_i from its mean over 3941 genes and dividing the difference by the standard deviation of m_i over all 3941 genes. This z -statistic is the numerical value for SDI. Similarly, SDI for exponential phase was obtained, where fold change was calculated by subtracting average log intensity in SP from the average log intensity in exponential phase, followed by linear regression.

3.2.4 k -means clustering

k -means clustering was performed on each dataset using the ratio of average log intensity to average log forward-scatter. The number of clusters for k -means was pre-specified as 20 (the number of clusters was tested from 5 to 100 in increments of 5 with very similar results). The average profile for each cluster was computed, followed by

visual identification of clusters with broad or jagged profiles. These clusters were expected to contain proteins with multiple populations. This analysis identified one cluster of 80 samples from stationary phase, and one cluster of 99 samples from exponential phase. Samples from these clusters were compared against candidate heterogeneous samples from other methods.

3.2.5 Average fold change

Average fold change is frequently used in microarray data. Similar to SDI, but using limited to a single dimension (fluorescence intensity), the average log fluorescence intensity over all cells (events) from each GFP strain was obtained for each sample for both SP and exponential phase. Next, for every GFP, we subtracted the average log intensity from exponential phase, from the average log fluorescence intensity from SP, and generated the average fold change measure, for each sample. We define the average fold change for every gene i as the median over all 3 samples of the average fold change measure.

3.2.6 Identification of GFP strain lists and GO process categories

Four approaches were used to identify proteins with heterogeneous samples: SDI, visual identification, k -means clustering and average fold change. For visual identification of two-peak samples approximately 8000 samples were examined using IDLQuery. Two sample lists were generated: 147 SP, two-peak samples (SPV) and 45 exponential phase two-peak samples (EPV). For k -means clustering two sample lists were generated: (SKM) with 80 SP samples and (EKM), with 99 exponential phase samples. Lists generated by SDI measure (SDI) and average fold change (AFC) were

compared with lists generated by the other two approaches for Gene Ontology (GO) process enrichment. Lists from SDI were generated by sorting according to decreasing SDI measure and then selecting top n samples. n depended on the type of comparison (See Results). Similarly lists from average fold change were obtained.

After sample lists were generated, each list was evaluated using GO Term Finder (www.geneontology.org), available at Saccharomyces Genome Database. For each list, p -values for GO biological process categories were obtained and the most significant categories of each list were selected with their respective p -values (for more details about p -value generation, please see supplemental materials).

3.3 RESULTS

3.3.1 SDI and two-peak plots

The histogram distribution obtained from IDLQuery shows the distribution of fluorescence intensity of two yeast GFP-fusion strains from stationary-phase cultures (Fig. 1).

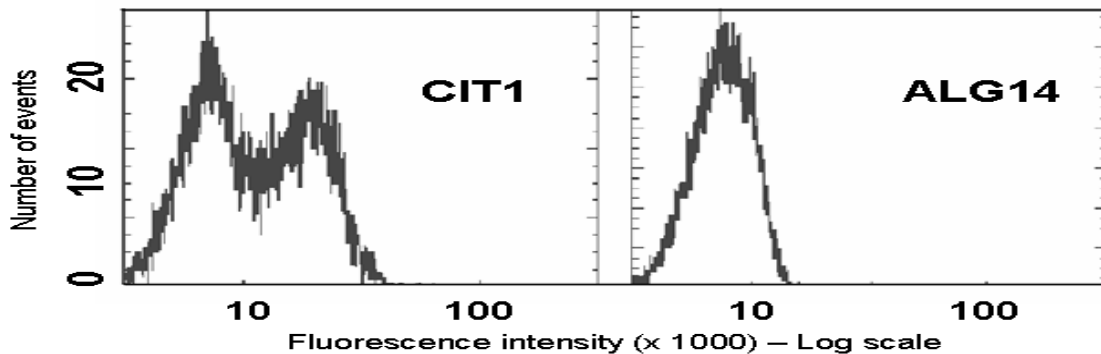


Figure 1 - Histograms of flow cytometry output comparing log fluorescence intensity vs. number of events for two yeast GFP-fusion strains in SP. CIT1 shows a two-peak distribution whereas ALG14 shows a single peak.

Similarly, the histogram distribution of fluorescence intensity of the same two yeast GFP-fusion strains from exponential-phase cultures is shown (Fig. 2).

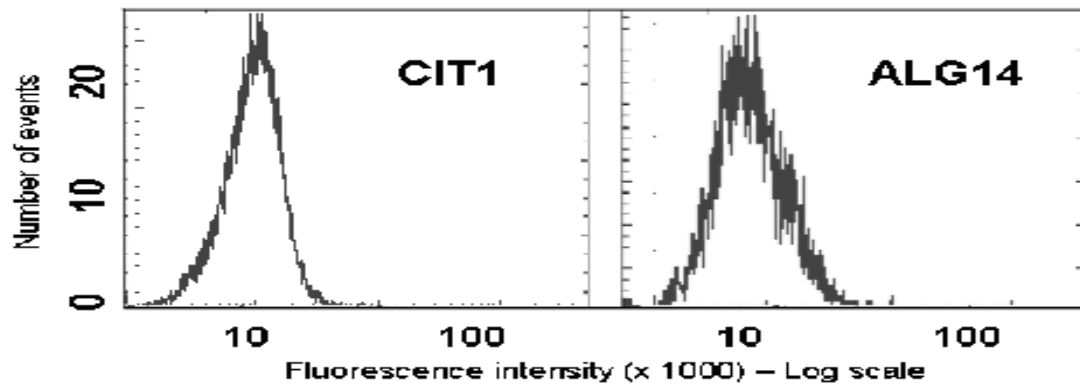


Figure 2 - Histograms of flow cytometry output comparing log fluorescence intensity vs. number of events for two yeast GFP-fusion strains in exponential phase. CIT1 and ALG14 show single peak distributions.

These were the types of outputs used to visually detect samples with two peaks, i.e., bimodal distribution of fluorescence intensity. The same yeast strains from both exponential and stationary phases were compared using two scatter plots: one displaying fluorescence intensity as a function of side scatter (Fig. 3), and the other displaying the fold change in fluorescence intensity as a function of side scatter (Fig. 4). As can be seen, fluorescence intensity of CIT1 changes at a higher rate in SP than in exponential phase. This difference is captured by the SDI measure (Fig. 4), which is high (0.81) for CIT1 and close to 0 for ALG1 (0.02). Through this analysis, we determined that a significant slope is strongly associated with strains that have bimodal distributions of fluorescence intensity. This association between large slope values and bimodal intensity

distributions is exploited by SDI to detect heterogeneous cell populations in a high-throughput fashion.

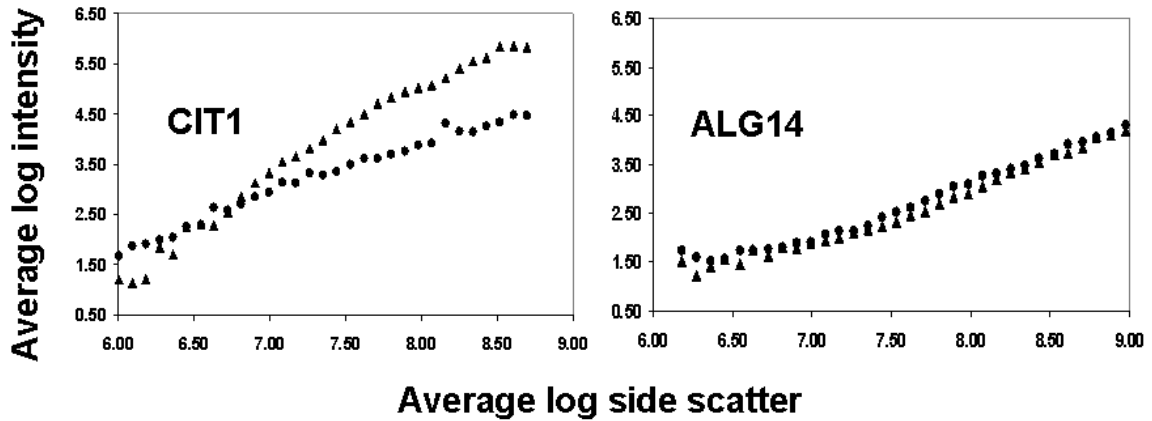


Figure 3 - Scatter plot of log side-scatter vs. log fluorescent intensity. Triangles represent the average log fluorescence intensity in SP and circles represent the average log fluorescence intensity in exponential phase. To assure statistical significance, each bin in the both plots was selected only if it had at least 50 events in both datasets.

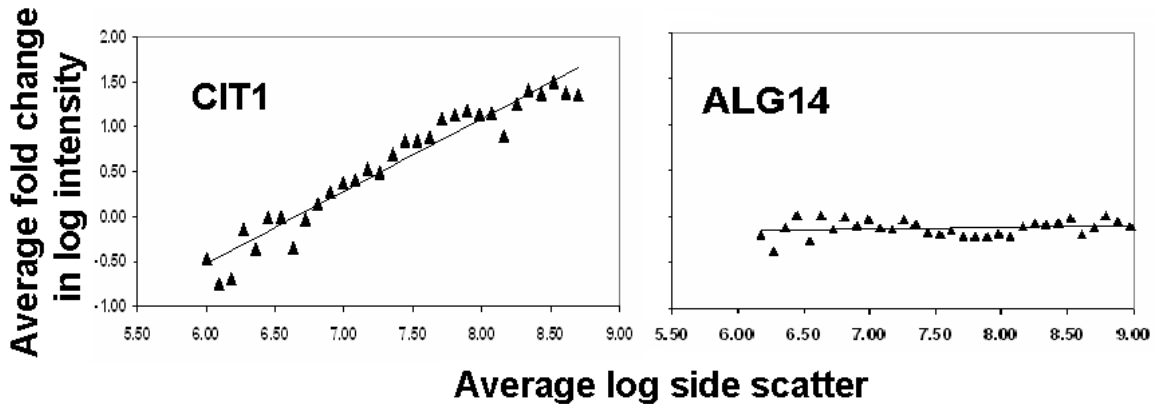


Figure 4 - Regression scatter plot of log side-scatter vs. log fold-change in fluorescent intensity. Triangles represent average fold change in log intensity between stationary and exponential phase. To assure statistical significance, each bin in the both plots was selected only if it had at least 50 events in both datasets.

To determine the significance of the slope in figure 4, a weighted linear regression, where each bin is assigned the weight which is the minimum of the number of cases for that bin, between SP and exponential. Similarly, we estimate the weighted of

the slope, and by divide the slope by its standard deviation, to obtain a t-statistics and its corresponding p-value. However, gene lists produced based on ranking by slopes turned out to have far better enrichment, than gene lists produced by ranking by t-statistics, and thus we chose to rank by slope (which is the same as ranking by t-statistics, but assuming all slopes have the same variance).

The modeling of the fold change for each bin as a function of side-scatter can be further improved by assuming a non-linear association. This non-linear association is suited to model saturation regions which fold change will tend to be constant for high levels of side-scatter, such as the 5 triangles in the extreme right of figure 4. One simple non-linear function to model the relation between log-fold change in intensity and log side-scatter is $f(x) = \mathbf{a}(1 - \exp(-(\mathbf{b}x+\mathbf{c})))$, where x is the log side-scatter which has an upper bound (asymptotic) at \mathbf{a} , and will look like the line $\mathbf{a}(\mathbf{b}x+\mathbf{c})$ near the region in which $\mathbf{b}x+\mathbf{c}$ is close to zero, that is, in the neighborhood where x is close to $-\mathbf{c} / \mathbf{b}$. This function also has an interesting property which is that it also models the local slope (slope in this neighborhood), which is equal to $\mathbf{a}\mathbf{b}$ and is a non-linear version of SDI for every GFP, and would also be used to rank GFPs and produce gene lists. However, in this dataset, genes lists generated by ranking genes by $\mathbf{a}\mathbf{b}$ was very similar to gene lists generated by ranking by the linear SDI, and therefore no additional benefit was observed for using this non-linear model.

3.3.2 Biological process enrichment

Enrichment performance using Gene Ontology (GO) biological process categories was obtained for SP visual two peak list (SPV, 147 GFP-fusion strains) and same sized lists (147 GFP-fusion strains) generated using two SDI measures SDI_SS (which uses side-scatter), SDI_FS (which uses forward-scatter) and average fold change (AFC) for SP (Table 1). Since SDI_FS was very similar to SDI_SS, after table 2 until the end of this chapter, we use only SDI_SS, and for simplicity we will refer to it as SDI. Also for simplicity, all p-values from the tests will be one-tail p-values until the end of this chapter.

Table 1 - Gene Ontology biological process enrichment comparison for stationary visual two-peak list (SPV), SDI_SS (SDI side-scatter), SDI_FS (SDI forward-scatter) and (AFC) of the same size (147 samples).

GO CATEGORY	SPV	AFC	SDI_SS	SDI_FS
generation of precursor metabolites and energy	7.9E-21	6.8E-15	3.1E-32	8.4E-30
oxidative phosphorylation	1.6E-17	4.0E-19	2.5E-24	2.5E-24
cofactor metabolic process	1.4E-13	1.0E-12	2.1E-18	2.0E-15

GO enrichment of SDI_SS and SDI_FS lists was significantly superior compared to SPV and AFC. Similarly, GO enrichment was generated for the *k*-means list (SKM) and same size lists (80 GFP-fusion strains) using SDI and AFC (Table 2). GO enrichment of SDI_SS and SDI_FS lists was significantly superior to SKM and AFC.

Table 2 - Gene Ontology biological process enrichment comparison for stationary *k*-means (SKM), SDI_SS (SDI side-scatter), SDI_FS (SDI forward-scatter) and (AFC) of the same size (80 samples).

GO CATEGORY	SKM	AFC	SDI_SS	SDI_FS
generation of precursor metabolites and energy	1.1E-08	1.6E-11	2.7E-24	5.7E-27
oxidative phosphorylation	1.4E-10	1.4E-13	2.7E-22	2.0E-20
cofactor metabolic process	3.1E-07	7.0E-12	2.2E-17	4.5E-15

GO enrichment of SDI_SS and SDI_FS lists was significantly superior compared to SKM and AFC. In Table 3 we show the significance between the number of genes from SDI, compared to SPV and AFC, for results from Table 1.

Table 3 – Number of genes in each GO biological process category present in the gene list using different approaches (SDI, SPV, AFC). In column 6 we have the p-value under the normal approximation for the hypergeometric distribution that SDI is significantly greater than SPV and similarly in column 7 we have the p-value that SDI is significantly greater than AFC.

GO CATEGORY	GFP in cat	SDI	SPV	AFC	SDI.vs.SPV	SDI vs. AFC
generation of precursor met	140	45	35	29	5.0E-04	6.9E-08
oxidative phosphorylation	30	21	17	18	2.5E-03	1.8E-02
cofactor metabolic process	134	32	27	26	4.6E-02	2.2E-02

In Table 4 we show the significance between the number of genes from SDI, compared to SPV and AFC, for results from Table 2.

Table 4 - Number of genes in each GO biological process category present in the gene list using different approaches (SDI, SKM, AFC). In column 6 we have the p-value under the normal approximation for the hypergeometric distribution that SDI is significantly greater than SKM and similarly in column 7 we have the p-value that SDI is significantly greater than AFC.

GO CATEGORY	GFP in cat	SDI	SKM	AFC	SDI.vs.SKM	SDI vs. AFC
generation of precursor met	140	31	15	19	7.2E-13	5.5E-08
oxidative phosphorylation	30	17	10	12	2.0E-11	1.2E-06
cofactor metabolic process	134	24	14	19	7.7E-06	1.5E-02

We did a similar GO enrichment analysis for lists from exponential phase cultures. However, the enrichment performance for all three lists was similar and much lower than lists from SP. For example, p-value for the best category of visual exponential two-peak list was $> 5.7E-05$, of *k*-means list was $> 5.0E-04$, of average fold change was $> 5.0 E-04$ and for SDI $> 1.8E-4$. These results suggest that heterogeneous cell populations are more likely to occur in stationary than in exponential phase cultures.

3.3.3 Marginal enrichment

Marginal enrichment comparisons between two lists of the same size are used to identify the enrichment of each list after excluding the overlap between them, measuring the exclusive enrichment of each list. Excluding from SDI the samples present in the overlap of SDI and SPV resulted in a list of 52 samples, which we call SDI–SPV. Similarly, excluding the same overlap from SPV resulted in a same size list called SPV–SDI. Next we selected a random list of the same size as a control for enrichment comparisons. The three lists were submitted to GO Term Finder and enrichment for the most significant 20 categories of each list was ranked by $-\log(\text{p-values})$ and compared. Marginal enrichment was also compared between SDI and AFC with lists of size 95 and similarly, comparisons between SDI and SKM with list sizes of 46 samples were performed (Fig. 5). SDI–SPV list is more enriched than SPV–SDI list for all category ranks. SDI–AFC list is more enriched than AFC–SDI for all category ranks. This illustrates the benefits in using two-dimensional fold change, which is SDI, compared to using a single dimensional fold change, which is AFC. Furthermore, SDI–SKM shows a high level of enrichment, whereas SKM–SDI shows enrichment no different from random.

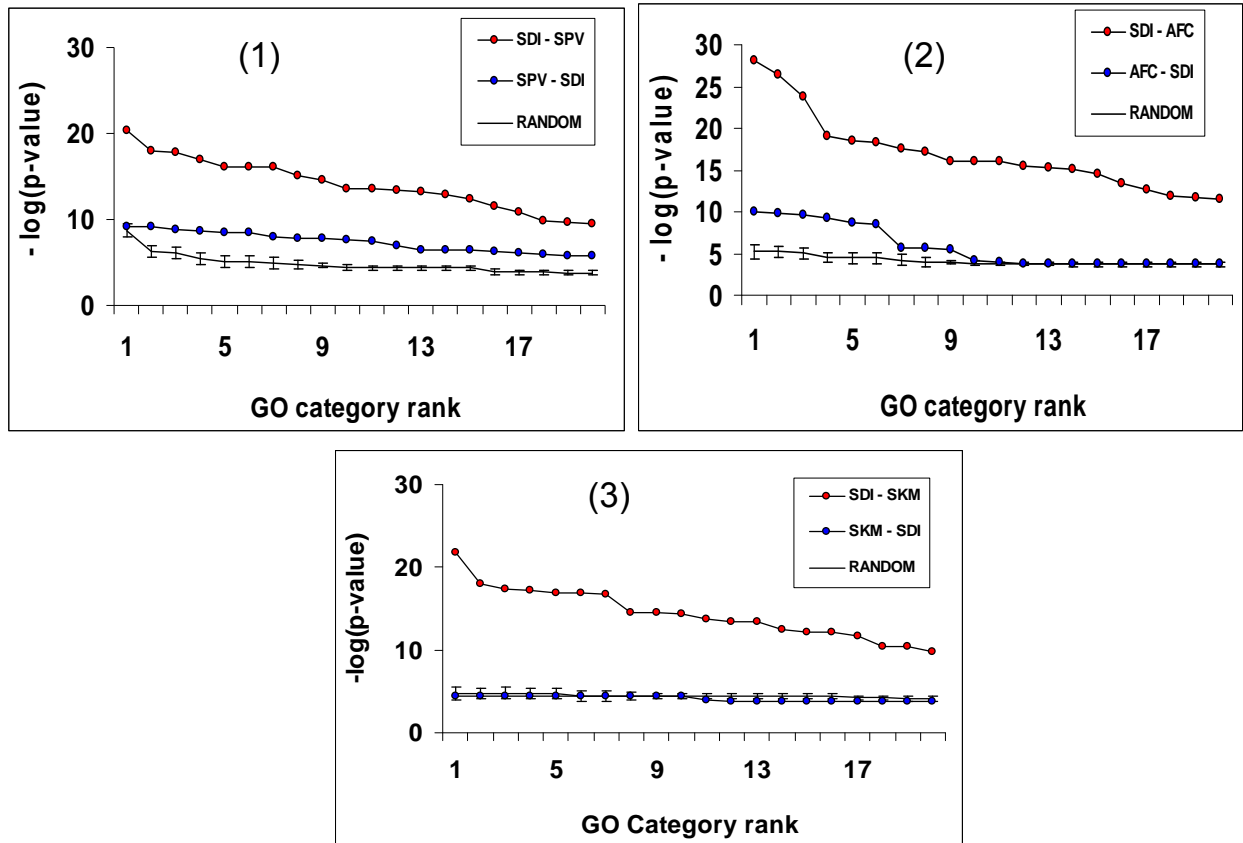


Figure 5 – (1) Line plot comparing SP marginal lists, SDI vs. visual 2 peak list (SPV). Standard errors for the $-\log(p\text{-value})$ from random lists are not shown since they are very small. (2) Line plot comparing SP marginal lists, SDI vs. average fold change (AFC). (3) Line plot comparing SP marginal lists, SDI vs. k -means (SKM). Standard errors for the $-\log(p\text{-value})$ from random lists are not shown since they are very small.

3.3.4 Intersection analysis of compared approaches

We compared the overlap between lists generated by each approach for heterogeneous sample detection. We used a p -value cutoff of 0.01 to generate an SDI list of 78 samples. Next, a triple overlap between this SDI list, stationary-phase two-peak SPV and stationary-phase k -means SKM lists was obtained, which had 33 samples (Fig. 5).

The overlap between SDI and SPV (39) excluding the triple overlap is greater than the overlap between SDI and SKM (1) excluding the triple overlap, and similarly, greater than overlap between SPV and SKM (9) excluding the triple overlap. This suggests a higher similarity between SDI and SPV. Because SPV was generated via visual analysis, and deemed to be of high quality, the high similarity of SDI list further indicates SDI to be a reliable approach of detecting heterogeneous samples.

3.3.5 Microscopic examination of gradient-separated cells

In order to provide a stronger validation of our candidate heterogeneous samples, we performed phenotypic analysis using density separation and microscopy, of 35 high confidence candidates. These 35 samples included the triple overlap (33 samples) of all three approaches and two additional samples corresponding to second and third highest SDI measures (the sample with highest SDI measure was already in the triple overlap).

Density separation of GFP-fusion strains for each of the 35 samples resulted in an upper and lower fraction in stationary phase. For every sample both upper and lower fractions were isolated, giving rise to 70 cultures: 35 cultures containing high-density cells and 35 cultures containing low density cells, corresponding to 35 GFP-fusion strain pairs. For simplicity we will use the term GFP-fusion strains instead of GFP-fusion strain pairs. Next, each GFP-fusion strain was compared with visual microscopy to identify differences in GFP-fusion localization among the high and low density cells. This resulted in 20 GFP-fusion strains with visual differences in fluorescence between

their corresponding high and low density populations. Majority of these 20 strains had a high SDI score (pvalue < 1.1E-4).

In order to provide additional comparison of the approaches, samples that were exclusively identified (Fig. 6) by each approach were analyzed microscopically. Specifically, we selected 5 (out of 37), 5 (out of 67), and 5 (out of 6) samples uniformly at random from two peak visual analysis, k-means and SDI, respectively. After visual inspection, we used the following classification for each GFP-fusion strain: 1 if there was clear visual fluorescence difference, and 0 if there was no difference (Table 5). SDI showed differences in fluorescence in all 5 GFP-fusion strains, followed by SPV with 3 differentiated GFP-fusion strains out of 5. None of the samples from SKM had differences in fluorescence.

Inspection of GO categories for the list of 35 samples identified ‘Carboxylic acid metabolic process’ to be one of the highly significant categories. Interestingly, all strains from our list that were annotated with Carboxylic acid metabolic process (9 out of 35) had visual differentiation in fluorescence intensity. This was highly significant (p-value equal to 1.4E-8) and suggests that proteins involved in Carboxylic acid metabolic process are highly likely to form heterogeneous populations.

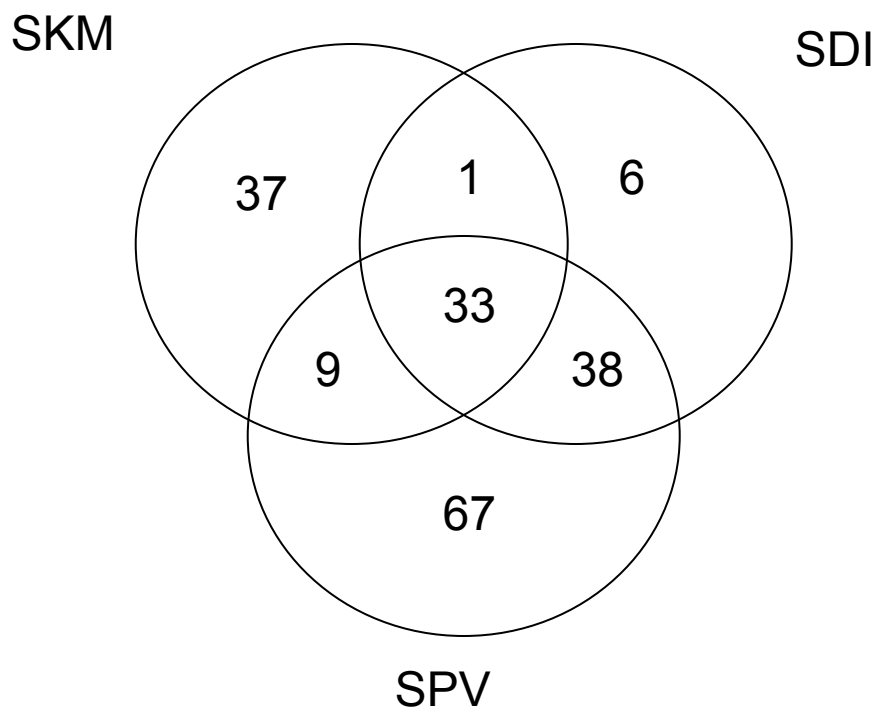


Figure 6 - Venn diagram of the overlaps using SDI, SPV and SKM. Overlap between SPV and SDI is much higher than overlaps with SKM.

Table 5 - Visual microscopy identification of fluorescence differentiation of non-overlapping GFP-fusion strains. A '1' represents visual differentiation in fluorescence intensity and '0' represents no differentiation.

Random sample of non-overlapping genes (GFP-fusion)					
SDI	Diff	SPV	Diff	SKM	Diff
TEF1	1	ATP1	1	MDV2	0
URA1	1	MES1	0	UBC1	0
TDH3	1	PRT1	0	NMD4	0
ATP10	1	ABF2	1	PET18	0
SDH4	1	PRE3	1	ENP1	0

3.4 DISCUSSION

Heterogeneous cell populations can be detected using different methods, with every method having its own limitations. For example, gradient separation can identify heterogeneous cells populations based on density, but not if heterogeneous populations cannot be separated on the basis of density. Visual detection of two-peak samples can detect heterogeneity from a single condition, but it is not clear how to detect heterogeneity across two conditions. Further, this approach requires manual inspection, which does not scale when data from multiple conditions are available.

SDI identifies heterogeneous cell populations, by incorporating the relative change in fluorescence intensity between two conditions. This makes SDI suited to detect heterogeneity between two conditions. The strength of SDI relies on the assumption that different subpopulations within a heterogeneous population exhibit different relations between fluorescence intensity and side-scatter (forward-scatter). Based on this assumption, SDI uses a linear regression as a computationally efficient way to detect differences in these relations.

SDI requires data from two conditions. However, if heterogeneity is much larger in one of the two conditions, SDI is likely to work as well as an approach that looks at a single condition at a time (two-peak analysis or k-means). This was true for our setup, where most of the meaningful heterogeneous candidates were in SP and not in exponential phase. The marginal enrichment analysis showed that most of the enrichment of SPV lists was due to the overlap between SDI and SPV lists.

3.5 CONCLUSION

In this paper, we have described a scalable approach for detecting heterogeneous populations from high-throughput flow cytometry data. Sample lists obtained from SDI measures had superior enrichment compared to lists obtained from visual two-peak distributions, average fold change and k -means clustering. The superior enrichment of SDI was also supported by our marginal enrichment analysis, where most of the enrichment of other approaches was due to the overlap between these lists and SDI list. Moreover, gradient-separation followed by visual microscopy, showed that samples identified by SDI were highly likely to have differences in GFP localization in high and low density cells.

SDI currently performs linear regression of the fold change in log intensity to side scatter. However, for some of the GFP-fusions strains, this linear relationship does not hold. Extending SDI to perform a piece-wise linear or non-linear regression is an important direction of future research.

As high throughput flow cytometry becomes more routine with many thousands of measurements per minute, approaches that allow rapid characterization of samples will become increasingly important. The SDI approach provides a simple, scalable way to identify strain heterogeneity that can identify important biological differences in high throughput data that might not otherwise be accessible for evaluation.

At this point, we do not completely understand all the factors that govern heterogeneity within a cell population. A comprehensive study involving more complex

experimental designs over many conditions in concert with approaches like SDI will be instrumental in improving our understanding of the cause and benefit of heterogeneous cell populations.

3.6 ACKNOWLEDGMENTS

We specially thank Swagata Chakraborty and Melissa Wilson for their valuable contribution and discussions, as well as the other members of the laboratory. This work was supported by National Science Foundation (NSF) grant MCB-0645854 and National Institutes of Health (NIH) grant GM-67593 (to M.W.W) and NIH grant 1U54MH084690-01 (to L.S). P.H.T was supported by NIH/IMSD grant GM-060201. R.J was supported by NIH grant GM-075149.

3.7. SUPPLEMENTAL MATERIALS

3.7.1 Growth conditions

Individual strains from the Yeast GFP Collection that were constructed from the base strain ATCC 201388: *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0* (S288C) (Brachmann, Davies et al. 1998) were replicated into 96 well plates containing YPD + A (2% yeast extract, 1% peptone, 2% glucose, 0.04 mg/mL adenine, and 50 μg/ml ampicillin; (Rose 1990) using pin tools. The plates were covered with Breathe Easy sealing membranes (Sigma Aldrich cat #380059) and the strains were cultured at 30°C with aeration either overnight (for exponential growth) or for 7 days (for stationary-phase growth).

3.7.2 High-throughput flow cytometric screening

Three steps were used to prepare the samples for high throughput screening. First, dilution plates were prepared by transferring 90μL of peptide dilution flow buffer (30mM HEPES*1/2 Na, 110mM NaCl, 10mM KCl, 1mM MgCl₂*6H₂O) into each well of the 384-well plates (Greiner Bio-one Cat #781280) using the Biomek NX_{MC} (Beckman Coulter, Fullerton, CA.) liquid handling robot. Second, 10μL of each yeast strain were transferred from the 96-well growth plates into 3 adjacent wells of the 384-well dilution plates using the Biomek NX_{S8} (Beckman Coulter) liquid handling robot. This step created a 1:10 dilution and generated three technical replicates for each sample. The 4th, 8th, 12th, 16th, 20th, and 24th columns of the dilution plates do not contain samples, just buffer alone. These columns serve as a wash well in between different samples to minimize any sample carryover. Third, the cells were sampled with a HyperCyt®

(Edwards, Oprea et al. 2004; Young, Bologna et al. 2005) autosampler controlled by HyperSip software and interrogated for GFP fluorescence with a CYAN ADP (Dako Cytomation, Ft. Collins, CO) flow cytometer using excitation at 488 nm and collection of fluorescent emissions with a 530/40 nm filter set. The data were processed using IDLQuery software and the median channel fluorescence for each sample was calculated and used for subsequent analyses.

3.7.3 Flow dataset

Approximately 4000 GFP were used in both SPand exponential phase generating two datasets. In each dataset, three technical replicates were generated for every GFP where each GFP has approximately 30,000 events. After applying a filter in which we excluded GFP with missing data in either exponential or stationary phase, there were a total of 3941 GFP for used in both stationary and exponential datasets.

3.7.4 Reproducibility analysis between biological samples

In order to make an evaluation on the quality of the data, an additional biological sample was generated approximately 4 weeks after the completion of the initial experiment, by selecting at random four plates of 96 GFP each. This additional biological sample also had 3 technical replicates. For both stationary and exponential, the two biological samples were joined having a total of 384 GFP, 3 technical replicates from the first sample and 3 technical replicates from the second sample. Each technical

replicate contained for each GFP, the average log fluorescence intensity over thousands of events, which were then correlated between biological samples. All the Spearman correlation coefficients were over 0.90 and thus we can state the high reproducibility of the average fold change.

3.7.5 Reproducibility analysis between technical replicates

Correlations using slopes from SDI were performed between the three technical replicates to show the reproducibility of both side-scatter slopes and forward-scatter slopes over the total 3941 GFP. Respective Spearman correlations for slopes from each pair of replicates were 0.966(rep1, rep2), 0.971(rep1, rep3) and 0.967(rep2, rep3). Given that all Spearman correlations were above 0.90, we can state the high reproducibility of the slope used in SDI.

3.7.6 SDI Algorithm

Generating the groups:

1. Sort the exponential- and stationary-phase datasets by side-scatter and define k , e.g.,
 $k = 50$, equally populated groups, defined over the set of all side-scatter events.
2. Assign all events from exponential- and stationary-phase samples to their corresponding groups.

For every sample j (GFP-strain):

3. Calculate FIS_{ij} as the average of \log_2 of the fluorescent intensity of each event from the SP set, for group j and sample j , and let FIE_{ij} be the analogous value for the exponential phase set.
4. Combine corresponding samples from SP sample j and exponential phase sample j into sample-pairs, referred to sample-pair j .
5. Three sample-pairs, one per technical replicate. For every sample-pair, the fold change in log-intensity was calculated for each group by taking the difference in average log intensity between them.
6. Calculate $\Delta FC_{ij} = FIS_{ij} - FIE_{ij}$, the average fold change for group j from sample-pair j .
7. For every sample-pair i , select groups with both number of events ≥ 50 in SP sample i and number of events ≥ 50 in exponential phase sample i . In the end, each sample-pair i will have a total of k'_{i1} selected groups, where $k'_{i1} \leq k$.
8. Calculate the average \log_2 in side-scatter for each corresponding selected groups, denoted as $x_{i1}, x_{i2}, \dots, x_{ik}$. Let array X denote the set containing these values.
9. Let $\Delta FC_{i1}, \Delta FC_{i2}, \dots, \Delta FC_{ik}$ be the array Y of fold-change values for the selected groups for sample j .
10. Regress Y on X and generate slope $_i$, the slope for sample j .
Combining the sample-pairs for the three replicates from each GFP-strain i :
11. Three slopes are generated, one for each j replicate sample-pair.
12. Calculate the median of the slopes over all three sample-pairs m_{ij} .
13. An approximate z-statistic is obtained for every sample by subtracting each m_{ij} by its mean over 3941 samples and dividing the difference by the standard deviation

of m_{ij} over all 3941 samples. This z-statistic is the numerical value for GFP-strain i which is called SDI_i .

3.7.7 GO Term Finder settings

The main settings used for GO term finder are: ORF's only, no 'dubious' categorized genes, 'manually annotated' with a background set of genes being the set of 3941 ORF's corresponding to their respective GFP (<http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.pl>). When calculating the p-value the option 'no Bonferroni adjustment' was chosen. Ontology was set to 'process' and the cut-off level for the category p-values was 0.01(default value).

Chapter 4 - CRAM

Calibration Regression Analysis of Microarrays

Osorio Meirelles¹, Sushmita Roy² and Margaret Werner-Washburne¹

4.0 ABSTRACT

Background: With the advent of genomics, there has been a rapid increase in the use of two and one-color microarrays, used to measure mRNA abundance for the entire genome. Variability in microarray analysis undermines its utility in identifying the entire subset of differentially expressed mRNAs. Recent microarray studies have shown that, although it is assumed that variances are constant for every hybridized spot within a microarray, variances may differ for each biological sample analyzed (Ritchie, Diyagama et al. 2006). Another common assumption is that log-intensity values for any given gene have a Normal distribution. For many datasets, both assumptions have been shown to be incorrect, resulting in distortions in the significance when testing for differential expression of each gene (Bar-Even, Paulsson et al. 2006; Wentzell, Karakach et al. 2006).

Approach: To overcome the limitations of existing approaches in identifying significant, differentially expressed genes, we have developed a novel unsupervised statistical approach called Calibration Regression Analysis of Microarrays (CRAM) that uses a combination of empirical Bayes and regression calibration. The main novelty of our

approach is the modeling of gene expression variances as a function of the log-intensity within each sample. Another version was later developed CRAM-GS in which the association between genes is captured using an adjusted gene correlation measure.

Results: CRAM was compared to four existing approaches for identifying differentially expressed genes. Performance was based on the ability to identify co-regulated genes in the same Gene Ontology process. CRAM exhibited a marginal improvement in GO process enrichment compared with the other approaches. To the original datasets, three more were included in which the later version CRAM-GS, showed a significant improvement compared to CRAM, suggesting a major additional benefit of incorporating gene correlations into the model. All versions of CRAM were two orders of magnitude faster than the existing approaches. Overall, CRAM provides an adaptive, computationally efficient approach for accurate identification of differentially expressed genes.

Keywords — Empirical Bayes, microarrays, regression, measurement error, calibration.

4.1 INTRODUCTION

Accurate identification of differentially expressed genes, detected as changes in mRNA abundance, is crucial for extracting biological relevance from microarray data. Methods for identification of differentially expressed genes typically use some type of **t-statistic**, requiring the estimation of both expected log-intensity values across samples for expression of every gene, and the variance of the expected log-intensity for the same genes.

In this paper, we define gene-expression variance for one microarray spot as a product of **sample-specific** variance and **gene-specific** variance (Ritchie, Diyagama et al. 2006). Sample-specific variance results from technical or measurement error. Gene-specific variance is described as the variance contribution attributed to biological expression of each gene or the deviations from the expected log-intensity across samples.

Different methods make different assumptions for estimating the gene-specific variance and sample-specific variance. Gene-specific variance is commonly estimated using log ratios for each gene, across all samples. However, when datasets have a small number of arrays, gene-specific variance is often underestimated, resulting in inflated t-statistics (Smith 2004).

Many approaches improving gene-specific variance estimators have been developed to address the problem of distorted t-statistics (Efron B 2001; Tusher, Tibshirani et al. 2001; Lonnstedt and Speed 2002; Broberg 2003; Smith 2004; Kristiansson, Sjogren et al. 2005; Kristiansson, Sjogren et al. 2006; Ritchie, Diyagama et al. 2006; Astrand, Mostad et al. 2007; Sjogren, Kristiansson et al. 2007; Astrand, Mostad et al. 2008). These approaches assume that the expected expression values have a Normal distribution. However, in microarray data from heterogeneous biological samples, where each sample corresponds to a different subject, this is unlikely to be true. In heterogeneous samples, the Normality assumption is typically violated because of large variations between samples, producing inaccurate gene-specific variance estimates, and thus, distortions in t-statistics estimates.

Sample-specific variance, although usually distinct for each sample, is usually modeled by assuming it is constant for every gene within each sample. Unfortunately, sample-specific variance is often dependent on the intensity level of each gene (Bar-Even, Paulsson et al. 2006; Wentzell, Karakach et al. 2006). To address the limitations of both sample-specific variance and gene-specific variance assumptions, we have developed a novel statistical algorithm. Calibration Regression Analysis of Microarrays (CRAM) models both sample- and gene-specific variance based on analysis of the data, by combining empirical Bayes (Baldi and Long 2001; Lonnstedt and Speed 2002; Smith 2004; Kristiansson, Sjogren et al. 2005; Kristiansson, Sjogren et al. 2006; Sartor, Tomlinson et al. 2006; Astrand, Mostad et al. 2007; Sjogren A 2007; Astrand, Mostad et al. 2008) and regression calibration (Spiegelman, McDermott et al. 1997; Schneeweis B 2005) to accurately identify differentially expressed genes.

CRAM was compared with four existing approaches: **Locally moderated weighted t-statistics** (LMW) (Astrand, Mostad et al. 2008), **Weighted moderated t-statistic** (WAME) (Smith 2004; Kristiansson, Sjogren et al. 2005; Kristiansson, Sjogren et al. 2006; Astrand, Mostad et al. 2007; Sjogren A 2007), **fold change** (FC), and ordinary **t-statistic** (t), on four yeast microarray datasets (Aragon, Rodriguez et al. 2008). Performance was measured using Gene Ontology (GO) process enrichment(GeneOntology ; GOTermFinder). CRAM showed a marginal enrichment improvement compared to other approaches. Additionally, CRAM is highly computationally efficient compared with other methods, scoring a dataset of 88 samples in less than one second as compared with several minutes for other approaches.

In addition to capturing associations between microarray samples, the more recent CRAM-GS captures associations between genes. The underlying assumption is that since every gene belongs to a pathway containing one or more genes, a differentially expressed gene should have a strong level of association with at least some other gene. Similarly to correlating samples, a gene highly correlated with another gene is an indication of confidence in the expression values of that gene. For this reason, the quantification of gene correlations will prove to be a significant improvement over CRAM. CRAM-GC is also computationally very efficient, making it applicable to large microarray datasets.

4.2 RELATED WORK

A basic model for identifying differentially expressed genes estimates the expected log-intensity by calculating the average of log-intensity across all samples. Similarly, gene-specific variance is estimated by calculating the variance of log-intensity values across all samples, generating a standard t-statistic for every gene. This model often underestimates variances in log intensity when the number of samples is small, leading to overestimated t-statistics. To address the issue of inflated t-statistics, many methods have been developed in recent years.

Penalized t-statistics type approaches, add a constant to the gene standard deviation across samples, whereas the **posterior odds t-statistic**, also known as B-statistics, adds a constant to the gene expression variance, providing a better solution (Efron B 2001; Tusher, Tibshirani et al. 2001; Lonnstedt and Speed 2002; Broberg 2003). The posterior

odds t-statistic method was later extended into the **moderated t-statistic** (Smith 2004). However, moderated t-statistics does not account for sample-specific variance.

Moderated t-statistic was extended to **Linear models for microarray data**, LIMMA, with the assumption that each array had a constant but distinct sample-specific variance (Ritchie, Diyagama et al. 2006). LIMMA is also a software application part of the Bioconductor Projects web page (LIMMA). LIMMA however, assumes measurements from biological samples are independent, which is not true for many datasets (Kristiansson, Sjogren et al. 2005).

Weighted moderated t-statistic (WAME), overcomes the measurement independence assumption by introducing a correlation structure between samples.

Locally moderated weighted t-statistic (LMW), also part of the PLW software application, is an improved version of WAME that incorporates the modeling of gene-specific variance as a function of the expected intensity level of each gene (PLW-Astrand 2008).

All three approaches, LIMMA, WAME and LMW use some form of moderated t-statistic based on the estimation of gene expression variance using an independence assumption between gene-specific variance and sample-specific variance. With this independence assumption, gene expression variance can be defined as a product of gene-specific variance and sample-specific variance. Similarly, our method estimates gene expression variance under the same independence assumption, however, we allow the sample-specific variance to vary within each sample as a function of intensity levels for

every gene in the sample. The independence assumption between gene-specific variance and sample-specific variance is also used in CRAM-GS.

4.3 METHODS

4.3.1 Datasets

A total of four yeast cDNA microarray datasets were used to validate our approach (Aragon, Rodriguez et al. 2008). Arrays are assumed to have undergone a standard microarray normalization process, followed by natural logarithm transformation on all expression values, generating a log-intensity value for every gene in every array. Each dataset measures gene expression change from two yeast cell populations, quiescent and non-quiescent, separated from stationary phase cultures using density centrifugation.

The datasets used are the same as used in chapter 2, with the main difference being that the datasets described and referred to as 1 to 4, correspond to the datasets in chapter 2 from 2 to 5. Dataset 1 has 80 microarrays from the quiescent population and 80 from non-quiescent, where each microarray measures the gene expression profile of a single yeast deletion mutant. Dataset 2 is similar to dataset 1 with 176 microarrays from 88 mutants. Datasets 3 and 4 have 32 and 20 microarrays respectively, corresponding to 16 auxotrophic parental (BY4742) strains and 10 wild type (S288C) strains.

Datasets 1 and 2 are said to be heterogeneous since they have very dissimilar samples due to the genetic differences, resulting in high biological variability. Datasets 3 and 4 are said to be homogeneous datasets in which samples are biological replicates and

therefore are expected to have lower biological variability than datasets 1 and 2. We excluded genes with >80% missing values in any dataset producing a total of 5649 genes in all datasets.

For each biological sample we subtracted the log intensity measurements in the non-quiescent microarray from the quiescent microarray. Next, this difference in log-intensity was adjusted by subtracting the mean log-intensity difference over all genes, such that the mean of log-intensity difference was equal to zero. We will refer to differences in log-intensity as **delta log-intensity** throughout this article.

4.3.2 Method Overview

Differentially expressed genes are typically identified using t-statistics, which require the estimation of the variance of every gene. Gene expression variance is a product of two entities: gene-specific variance and sample-specific variance (Ritchie, Diyagama et al. 2006). Gene-specific variance is defined as the variance contribution attributed to each individual gene, and sample-specific variance is defined as the variance contribution of each individual sample.

Calibration Regression Analysis of Microarrays (CRAM) is a novel approach for identifying differentially expressed genes that is based on three themes:

(a) gene-specific variance is treated as a weighted average between the variance estimate of a gene and the average of these variance estimates over all genes,

(b) sample contribution of each to the gene expression variance is weighted according to the quality of each sample, and

(c) sample-specific variance is not constant, but rather depends on the intensity level of the genes within the sample.

CRAM first uses a linear regression model to predict a sample using the remaining samples. Next, the predicted sample is regressed using the original sample where a slope is generated. Using an Empirical Bayes approach, slopes are transformed into weight parameters, inversely proportional to the sample-specific variances. We extend this weight estimating procedure to gene subsets within each sample, increasing the precision of the weights and therefore the accuracy of expected delta log-intensity.

4.3.2.1 Notation

Let k represent the number of biological samples in a dataset and n the number of genes.

Denote $\mathbf{X}_j = \{x_{1j}, x_{2j}, \dots, x_{nj}\}$ as the set of delta log-intensity values for every gene i in sample j . A dataset with k samples will be represented as the set $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k\}$.

4.3.3 Linear model

Let \mathbf{x}_{ij} , the observed $\Delta \log$ -intensity value for gene i in sample j be written as

$$\mathbf{x}_{ij} = \boldsymbol{\tau}_i + \mathbf{e}_{ij}, \quad \text{for } i = 1, 2, \dots, n \text{ and } j = 1, 2, \dots, k \quad (4.1)$$

where $\boldsymbol{\tau}_i$ is $N(\mu_i, \sigma_i^2)$ -distributed random variable, representing the unknown measure of the expected $\Delta \log$ -intensity for gene i , μ_i is a known prior expected $\Delta \log$ -intensity and σ_i^2 is the unknown variance of $\Delta \log$ -intensity for gene i . The random variable \mathbf{e}_{ij} is an unknown measurement error for gene i in sample j , assumed to be $N(0, \sigma_i^2/w_j)$ -distributed, which variance is proportional to the variance of $\boldsymbol{\tau}_i$ by a factor equal to $1/w_j$. We also assume that $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \dots, \boldsymbol{\tau}_n$ are independent, that the \mathbf{e}_{ij} are independent for all i and j , and that the $\boldsymbol{\tau}_i$ and \mathbf{e}_{ij} are independent for all i and j . Thus, since the sum of normally distributed random variables is itself normally distributed, we see that the random variable \mathbf{x}_{ij} is $N(\mu_i, \sigma_i^2 + \sigma_i^2/w_j)$ -distributed, or equivalently, $N(\mu_i, \sigma_i^2(1 + w_j)/w_j)$ -distributed.

The identification of differentially expressed genes requires us to perform a hypothesis test for every gene i :

$$H_0: \boldsymbol{\tau}_i = 0$$

vs.

$$H_1: \boldsymbol{\tau}_i \neq 0$$

If H_1 is true, gene i is said to be differentially expressed, and if H_0 is true gene i is said not to be differentially expressed. For example, in datasets 1 to 4, if H_1 is true, a gene i is said to be differentially expressed in relation to Q vs. NQ. Similarly, the same idea applies to any differences between two conditions. Summarizing, we assume random variables τ_i and e_{ij} to have the prior distributions

$$\tau_i \sim N(\mu_i, \sigma_i^2) \quad \text{and} \quad e_{ij} \sim N(0, \sigma_i^2 / w_j), \quad (4.2)$$

where τ_i and e_{ij} are mutually independent, e_{ij} is independent for gene i and sample j , $\sigma_i^2 \geq 0$ is the unknown variance of τ_i , μ_i is the known prior expectation for τ_i , and $w_j > 0$ is an unknown positive parameter for sample j . The parameter σ_i^2 is defined as the gene-specific variance, whereas the sample-specific variance is defined as $1/w_j$. Thus, the variance of e_{ij} is σ_i^2 / w_j , the product of gene-specific and sample-specific variances. In addition, the set of unknown expected Δ log-intensity values $\tau_1, \tau_2, \dots, \tau_n$ are assumed to be independent, which is equivalent to assuming uncorrelated genes or more formally, that the knowledge of parameters $\tau_{i'}, \sigma_{i'}^2$ for any gene i' does not influence (or change) the conditional distribution of τ_i for gene i . Although this is a simplistic assumption since it is known that many genes are often correlated, this assumption is still often used in most of the current models. We now need to estimate the gene parameters τ_i, σ_i^2 , and the sample parameters w_1, w_2, \dots, w_k , which we describe in the next subsection.

4.3.3.1 Estimating τ_i

From (4.1) and (4.2):

$$P(x_{ij} | \tau_i = \lambda_i) \sim N(\lambda_i, \sigma_i^2/w_j). \quad (4.3)$$

The posterior distribution $\tau_i | w_j, x_{ij}$ is also a Normal distribution

$$P(\tau_i | x_{ij}) \sim N((\mu_i + w_j x_{ij})/(1+ w_j), \sigma_i^2/(1+ w_j)). \quad (4.4)$$

For proof please see supplemental materials 4.7.2.

From the expression above, if we do not know the value for the prior μ_i , we set it equal to zero, which is the mean of the normalized expression values x_{ij} over sample j . When w_j is equal to zero, the posterior variance is equal to σ_i^2 (the prior variance of τ_i) implying that the value x_{ij} has no useful information since the posterior mean is shrunk to μ_i (the prior expectation of τ_i). When w_j becomes large, the posterior variance becomes close to zero, $w_j/(1+ w_j)$ becomes closer to one and thus the posterior mean becomes close to x_{ij} , indicating a large confidence in the value x_{ij} as an estimate of τ_i and thus, the influence of μ_i in estimating τ_i becomes negligible. So when w_j is small (large sample-specific variance), there is low confidence in the observed x_{ij} values in sample j whereas when w_j is large, there is high confidence in these values. Denoting $W = (w_1, w_2, \dots, w_k)$ and $X_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, after observing all samples and assuming independence among samples X_j , we generalize the posterior distribution

$P(\tau_i | X_i)$ which has a Normal distribution with mean u_i , where

$$\mathbf{u}_i = E(\boldsymbol{\tau}_i | \mathbf{X}_i) = (\boldsymbol{\mu}_i + \sum_{j=1}^k w_j \mathbf{x}_{ij}) / (1 + \sum_{j=1}^k w_j) \quad (4.5)$$

is also the maximum likelihood estimator of $\boldsymbol{\tau}_i$. Furthermore, $\boldsymbol{\mu}_i$ parameters can be very useful if they were obtained as the output (a measure of expected gene expression) from a previous experiment performed under very similar conditions, which could potentially largely improve the precision of the current model. Considering that in our datasets we do not know the prior $\boldsymbol{\mu}_i$ parameters, we set all of them equal to zero, in all datasets.

However, we chose to keep the $\boldsymbol{\mu}_i$ parameters in most equations, in order to describe the most general case. Setting all $\boldsymbol{\mu}_i$ equal to zero, equation (4.5) simplifies to

$$\mathbf{u}_i = E(\boldsymbol{\tau}_i | \mathbf{X}_i) = (\sum_{j=1}^k w_j \mathbf{x}_{ij}) / (1 + \sum_{j=1}^k w_j) .$$

The conditional variance of $\boldsymbol{\tau}_i | \mathbf{X}_i$ is equal to

$$\text{Var}(\boldsymbol{\tau}_i | \mathbf{X}_i) = \boldsymbol{\sigma}_i^2 / (1 + \sum_{j=1}^k w_j). \quad (4.6)$$

See supplemental materials 4.7.3 for more details.

4.3.3.2 Estimating w_j

To estimate the parameter w_j we use a linear regression to estimate sample \mathbf{X}_j by regressing it on the remaining samples, generating a sample vector \mathbf{Y}_j of predicted values of \mathbf{X}_j . Next, we use an Empirical Bayes approach and equate the posterior expectation $E[\boldsymbol{\tau}_i | \mathbf{X}_j] = \mathbf{Y}_j$. By minimizing the sum

$$\sum_{i=1}^n (E[\tau_i | w_j, x_{ij}] - y_{ij})^2 = \sum_{i=1}^n \left((\mu_i + \sum_{j=1}^k w_j x_{ij}) / (1 + \sum_{j=1}^k w_j) - y_{ij} \right)^2, \text{ we}$$

find the optimal weight parameter w_j , given by:

$$w_j = \left(\sum_{i=1}^n (y_{ij} - \mu_i)(x_{ij} - \mu_i) \right) / \left(\sum_{i=1}^n (x_{ij} - y_{ij})(x_{ij} - \mu_i) \right) \quad (4.7)$$

For the proof see supplemental materials in 4.7.4.

4.3.3.3 Testing for differential expression

To test for differential expression we calculate the variance of the posterior expectation

u_i :

$$\text{Var}(u_i) = \sigma_i^2 \sum_{j=1}^k w_j(1+w_j) / \left(1 + \sum_{j=1}^k w_j\right)^2. \quad (4.8)$$

For more details on the proof see supplemental materials.

Dividing the expectation u_i in (4.6) by the square root of (4.8) we get a z-statistic for

gene i ,

$$a_i = (\mu_i + \sum_{j=1}^k w_j x_{ij}) / (\sigma_i (\sum_{j=1}^k w_j (1+w_j))^{1/2}) \sim N(0,1). \quad (4.9)$$

4.3.4 CRAM model: sample-specific variance with different contribution per gene

CRAM allows genes with different intensity values to have different weights. The Δ log-intensity values x_{ij} in sample j , are first categorized into smaller groups. In order to obtain more stable weight estimates, genes are categorized into three groups. Genes within a sample can be categorized into any number of groups but the influence of outliers on the weight parameters increases with the number of groups. Empirical analysis with different number of groups showed that three groups were optimal in our setup.

A sample X_j is sorted such that the highest one third of its Δ log-intensity values are assigned to group 1, the middle one third are assigned to group 2 and the lowest one third are assigned to group 3. Similar to the procedure in section 4.3.3, genes in each group g from sample j are regressed with their corresponding genes in other samples and a weight W_{gj} for group g is estimated. After estimating W_{gj} for each group, genes in sample j are assigned a weight from the group to which they belong $w_{ij} = W_{gj}$. This procedure is repeated for the other samples where all weights are estimated. Although modeling the weight for each group in a sample, implies that weights and expected Δ log-intensity are not independent, we assume a local-independence in which the weight assigned to a group is independent from the Δ log-intensity values in that group.

4.3.4.1 Estimating the gene-specific variance σ_i^2

The first step to estimate the gene-specific variance σ_i^2 is to calculate s_i^2 , the pre-moderated gene-specific variance across samples given by:

$$s_i^2 = \frac{\sum_{j=1}^k w_{ij}(x_{ij}-u_i)^2}{\left(\sum_{j=1}^k w_{ij}\right)^2 - \sum_{j=1}^k w_{ij}^2} / \sum_{j=1}^k w_{ij}$$

Next, we calculate V^2 , the average of s_i^2 over n genes,

$$V^2 = \sum_{i=1}^n s_i^2 / n$$

A gene-specific variance moderating factor α , a parameter between zero and one is used such that the estimate for the moderated variance is given by:

$$\hat{\sigma}_i^2 = \alpha s_i^2 + (1 - \alpha) V^2 \quad (4.10)$$

For more details in how to estimate α , please see supplemental materials.

4.3.4.2 Estimating the CRAM z-statistic

Extending (4.9) to different weights within samples, a **CRAM z-statistic** is estimated for gene i , under the local-independence assumption as introduced in 4.3.4,

$$\hat{a}_i = \left(\mu_i + \sum_{j=1}^k w_{ij} x_{ij}\right) / \left(\hat{\sigma}_i \left(\sum_{j=1}^k w_{ij} (1+w_{ij})\right)^{1/2}\right) \quad (4.11)$$

The statistic above is only a standard Normal distribution in ideal situations, when all underlying assumptions are true. An additional transformation is often required to

generate more accurate p-values as suggested by (Eaves, Wicker et al. 2002). For more accurate p-values, we calculate a standardized transformation generating the CRAM z-statistic, given by:

$$z_i = (\hat{a}_i - \text{average}(\hat{a}_1, \hat{a}_2, \dots, \hat{a}_n)) / \text{stdev}(\hat{a}_1, \hat{a}_2, \dots, \hat{a}_n)$$

and the two tail p-value is given by

$$\text{p-value} = 2 \Phi(-|z_i|) \quad (4.12)$$

where Φ is the standard normal cumulative distribution function.

4.3.5 Comparison of CRAM against other approaches

We compare CRAM against all other methods using the biological significance of gene lists identified by each approach. To measure the biological significance of gene lists, we use enrichment of Gene Ontology (GO) biological process categories (GeneOntology ; GOTermFinder). The primary assumption is that a more accurate gene list will have on average, more significant GO categories than a less accurate gene list.

We used an iterative approach to assess biological significance of gene lists from each approach. In each iteration a subset of k' samples ($k' = 4$ in this paper) is randomly selected and submitted to each method, which produces a candidate list of differentially expressed genes of a pre-specified size m . Gene lists are generated by sorting in descending order by the measure of differential expression of each approach (z-statistic in CRAM). Next the top m genes are selected into a gene list. We consider gene lists of

sizes $m = 200, 400$ and 800 genes. A total of 1000 iterations are performed for each dataset and for each gene list size.

For each gene list, a p-value is obtained for each GO category using an algorithm based on GO Term Finder (Boyle 2004). The most significant GO category (smallest p-value) is selected and its corresponding p-value is stored for each approach and for each gene list size. After all iterations are performed, the average of the $-\log(\text{p-value})$ for each approach and each gene list size is computed. The average $-\log(\text{p-value})$ over all iterations is used as the measure of biological significance.

4.3.6 Other Versions of CRAM

4.3.6.1 CRAM-binary

This is a robust version of CRAM and was developed to handle datasets with a high incidence of outliers. Instead of using its original values, all samples X_j are ranked and categorized as a dichotomous variable. For a detailed description of the algorithm, please see supplemental materials.

4.3.6.2 CRAM expectation maximization (CRAM-EM)

This is the fastest version of CRAM. It replaces the regression step as in 4.3.3.2 by assigning initial weights $w_1, w_2, \dots, w_k = 1$. In this version, X_j is also estimated using the remaining samples but instead of using a regression, X_j is estimated using a iterative

procedure where a weighted average is used. Then \mathbf{Y}_j , which is the estimate of \mathbf{X}_j is given by

$$\mathbf{Y}_j = \left(\sum_{l \neq j=1}^k w_l \mathbf{X}_j \right) / \left(\sum_{l \neq j=1}^k w_l \right) \text{ where } y_{ij} = \left(\sum_{l \neq j=1}^k w_l x_{ij} \right) / \left(\sum_{l \neq j=1}^k w_l \right).$$

For a detailed description of the algorithm, please see supplemental materials in 4.7.4.

4.3.6.3 CRAM gene correlation (CRAM-GC)

In addition to capturing associations between microarray samples, this version also captures associations between genes. The underlying assumption is that since every gene belongs to the same pathway of another gene, then a differentially expressed gene should have a strong level of association with at least some other gene. Similarly to correlating samples, a gene highly correlated with another gene is an indication of confidence in the expression values of that gene. For this reason, the quantification of gene correlations will prove to be a significant improvement over all previous versions of CRAM. CRAM-GC is also computationally very efficient, making it applicable to large microarray datasets.

4.3.6.3.1 Additional Datasets:

To test the CRAM-GC approach, we used the 4 datasets described in the beginning of this chapter, and added 3 additional datasets (Gasch, Spellman et al.

2000; Tu, Kudlicki et al. 2005; Hu, Killion et al. 2007) , named respectively datasets 5, 6 and 7 (Table 1).

Table 1 – Description of the three additional datasets used to measure the performance of CRAM-GC.

DATASET	Description	ARRAY TYPE	# microarrays
5	Yeast mutant strains	Two colors non-paired	174
6	Yeast CEN.PK strain	One color non-paired time course	36
7	Yeast mutant strains (transcription factor deletion)	Two colors non-paired	269

All of the additional datasets were non-paired, meaning that each sample corresponds to a specific microarray. Dataset 5 with 174 cDNA microarrays (two colors) is a dataset in which responses were measured for many different stresses. Dataset 6 with 36 microarrays is an Affymetrix (Affymetrix) type microarray (one color) with a time course covering three complete cell cycles. In dataset 7 with 269 cDNA microarrays, each microarray corresponds to a transcription factor deletion (263 non-essential and 6 essential). Dataset 5 and 7 can be considered heterogeneous datasets, whereas dataset 6 is not clearly defined, given that each sample is a biological replicate but at a different time point, and therefore has higher heterogeneity than a typical biological replicate experiment, but lower heterogeneity than datasets 5 and 7. An overall description of the CRAM-GC version is presented next, followed by the algorithm.

4.3.6.3.2 CRAM-GC linear model:

Let $x_{ij} = \tau_{ij} + e_{ij}$, where $e_{ij} \sim N(0, \sigma_i^2/w_{ij})$, and x_{ij} is the gene expression (say log-intensity or Δ log-intensity) for gene i and sample j , $w_{ij} > 0$ is a weight (or precision) parameter of gene i and sample j , for a total of k samples and n genes and τ_{ij} is an unknown random variable for gene i and sample j , with distribution $N(\mu_{ij}, \sigma_i^2)$, where μ_{ij} is a known prior estimate of τ_{ij} for gene i and sample j .

4.3.6.3.3 Posterior expectation

As a result, the posterior expectation $E[\tau_{ij} | x_{ij}]$ is a Gaussian distribution with mean equal to $(\mu_{ij} + w_{ij} x_{ij})/(1 + w_{ij})$ and variance equal to $\sigma_i^2 / (1 + w_{ij})$.

4.3.6.3.4 Homogeneous sample weight assumption

In this version, we assume that the weight for every sample is constant within each sample. A non-constant array (sample) weight can also be used, but since only a minor improvement in GO enrichment was observed, we will use this assumption for simplicity. In this version of CRAM, the sample weight for sample j is denoted as a_j .

4.3.6.3.5 Sample-gene weight independence assumption

The cell (gene-sample combination) weight w_{ij} can be written as $w_{ij} = a_j g_i$, where g_i is defined as the gene weight for gene i , under the assumption that the sample weight a_j is independent from the gene weight g_i .

4.3.6.3.6 Estimating sample weights a_j

In order to increase computational performance with a very minor loss in precision, we estimate the sample weights based on the highest correlated sample with each sample. We first generate a Spearman correlation matrix for all k samples and select for each sample j the maximum absolute value of the correlation coefficient over all remaining $k-1$ samples. Next, we let ρ_j be the maximum correlation coefficient of sample X_j and let X_j' be the sample with maximum correlation with X_j . Next we generate a predictive value Y_j by performing the regression $Y_j = c X_j'$, where c minimizes the loss function L for sample j , given by sum of the squared residuals $L = \sum_{i=1}^n (x_{ij} - c x_{ij}')^2$.

Similar to 4.3.3.2, if we let $\boldsymbol{\mu}_j$ be the vector of prior parameters for sample j denoted by $(\mu_{1j}, \mu_{2j}, \dots, \mu_{nj})$ and letting $\mathbf{T}_j = \{\boldsymbol{\tau}_{1j}, \boldsymbol{\tau}_{2j}, \dots, \boldsymbol{\tau}_{nj}\}$ be the vector of unknown expectations for sample j the weight a_j is estimated by equating the posterior expectation $E[\mathbf{T}_j | X_j] =$

$(\mu_j + X_j a_j)/(1 + a_j) = Y_j$, where $a_j > 0$ is the weight parameter for sample j , we

estimate a_j that minimizes the sum $\sum_{i=1}^n (E[T_j | X_j] - Y_j)^2$, and similar to (4.7) we get

$$a_j = \left(\sum_{i=1}^n (y_{ij} - \mu_{ij})(x_{ij} - \mu_i) \right) / \left(\sum_{i=1}^n (x_{ij} - y_{ij})(x_{ij} - \mu_i) \right).$$

4.3.6.3.7 Estimating gene weights g_i .

The next step is to generate a Spearman correlated matrix for all n genes and select for each gene i the maximum absolute value of the correlation coefficient over all remaining $n - 1$ genes and call ρ_i the corresponding correlation coefficient. In datasets in which the number of samples k are small, given that the number of gene n is large, a maximum absolute correlation can be a high number just by chance, that is, under a random scenario.

For example, under a random scenario in which expression values are permuted within each gene, the maximum correlation between genes can be highly inflated by chance, leading to the expectation $E[\rho_i] > 0$ but most of all, $E[\rho_i]$ can become close to one. In order to correct for optimistic gene correlations for datasets of all sizes, we need to deflate each ρ_i , generating an adjusted ρ_i which we will call ρ_i' .

The basis of this approach is to test if $|\rho_i|$ is greater than $|\rho_{iR}|$ the maximum correlation between genes in a random scenario, obtain a p-value, and then use the p-

value to deflate $|\rho_i|$. We should keep in mind that $|\rho_{iR}|$ is a random variable, whereas $|\rho_i|$ is the observed maximum correlation between gene i and the remaining genes.

4.3.6.3.8 Fisher's transformation of Spearman's correlation coefficient

Applying Fisher's transformation to ρ_i , which we will call $f(\rho_i)$, we get

$$z_i = f(\rho_i) = 0.5(k-3)^{1/2} \ln \left(\frac{1+|\rho_i|}{1-|\rho_i|} \right)$$

where z_i is approximately the absolute value of a $N(0,1)$ under a random scenario.

Let α be the level of significance such that

$$\alpha = 1/n \quad \text{and} \quad z_{0i} = \Phi^{-1}(1-\alpha),$$

where Φ^{-1} represents the inverse of the Normal cumulative distribution function and z_{0i} represents the standardized z-statistic corresponding to the maximum correlation under a random scenario with significance level equal to α . Let f^{-1} be the inverse function of the Fisher transformation such that

$$f^{-1}(z_{0i}) = (\exp\{2z_{0i}/(k-3)^{1/2}\} - 1) / (\exp\{2z_{0i}/(k-3)^{1/2}\} + 1),$$

and

$$z_{1i} = z_{0i} + f^{-1}(z_{0i}) (k-3)^{1/2} / 2(k-1)$$

where z_{1i} is the mean of the transformed maximum correlation coefficients

(mean of $f(|\rho_{iR}|)$), generated under the random scenario. Finally let $z_i' = z_i - z_{1i}$, which is the distance between z_i (the transformed $|\rho_i|$) and z_{1i} (the transformed $|\rho_{iR}|$).

Let \mathbf{R} be the random variable $N(z_{1i}, 1)$. Then, the likelihood of observing z_i in \mathbf{R} , is the density of the conditional distribution $P(\mathbf{R} | z_i)$ which is $N(z_i', 1)$. In order to avoid negative values for z_i' , we set the constraint that $z_i' = 0$, if $z_i' < 0$. By using the inverse Fisher transformation on z_i' , we get $|\rho_i'| = f^{-1}(z_i')$, where $|\rho_i'|$ is the adjusted value for $|\rho_i|$, the absolute value of the observed correlation coefficient. After generating the adjusted correlation $|\rho_i'|$ for every gene i , we find the gene weight following similar steps to estimating the sample weights a_j . We regress $\mathbf{X}_i = \{x_{i1}, x_{i2}, \dots, x_{ik}\}$ with the gene with the highest correlation $\mathbf{X}_i' = \{x_{i1}', x_{i2}', \dots, x_{ik}'\}$ such that $\mathbf{X}_i = c_i \mathbf{X}_i'$, where the slope c_i minimizes the sum of the squared residuals. Without any adjustment, c_i is given by $\rho_i (\sigma_x' / \sigma_x)$, thus the adjusted c_i is given by $\rho_i' (\sigma_x' / \sigma_x)$, which is equal to the original c_i multiplied by the shrinking factor $|\rho_i'| / |\rho_i|$ and thus each y_{ij} from (4.7) can be written as $c_i |\rho_i'| / |\rho_i| x_{ij}$, thus, the weight for gene i is given by

$$g_i = \left(\sum_{j=1}^k (c_i |\rho_i'| / |\rho_i| x_{ij} - \mu_{ij})(x_{ij} - \mu_i) \right) / \left(\sum_{j=1}^k (x_{ij} - c_i |\rho_i'| / |\rho_i| x_{ij})(x_{ij} - \mu_i) \right).$$

Comment: the adjustment for the sample correlations (estimating a_j) was not used since the number of samples ($k = 300$) is small compared to the number of genes n (≈ 6000 for yeast) and therefore it is not necessary to adjust for sample correlation coefficients. In summary, the main reasons for not adjusting the sample correlation

coefficients is **(a)** there are a much smaller number of correlations between samples in which the maximum is selected, as compared to correlating genes and **(b)** the sample vectors are much larger than the gene vectors (n vs. k), making the sample correlation coefficients much more stable.

4.3.6.3.9 Expected log-intensity of a gene-sample combination.

Since $w_{ij} = a_j g_i$, after all sample weights and gene weights are estimated, we have the expected value of the posterior

$$E[\tau_{ij} | x_{ij}] = (\mu_{ij} + x_{ij} w_{ij}) / (1 + w_{ij}). \quad (4.13)$$

So if either $a_j = 0$ or $g_i = 0$, this would imply $w_{ij} = 0$, indicating we have no confidence in our value x_{ij} and therefore $E[\tau_{ij} | x_{ij}] = \mu_{ij}$, the prior expectation for gene i .

4.4 RESULTS

4.4.1 Normality assumption

We first assess the extent to which each dataset satisfies the normality assumption. For all four datasets, a standardized expected Δ log-intensity value was generated for every gene by using equation (4.9) which assumes weights are constant within each sample. Next a histogram of the standardized values was plotted for each dataset (Fig. 1). In addition, a Kolmogorov-Smirnoff test for normality was used. Dataset 1 had a p-

value $< E-300$, dataset 2 had a p-value equal to $7.6E-242$, dataset 3 had a p-value equal to $1.7E-04$ and dataset 4 had a p-value equal to $3.9E-128$. These p-values also suggest data in all four datasets are not normally distributed, although in dataset 3 this can be attributed to the left tail, since the right tail looks close to Normal.

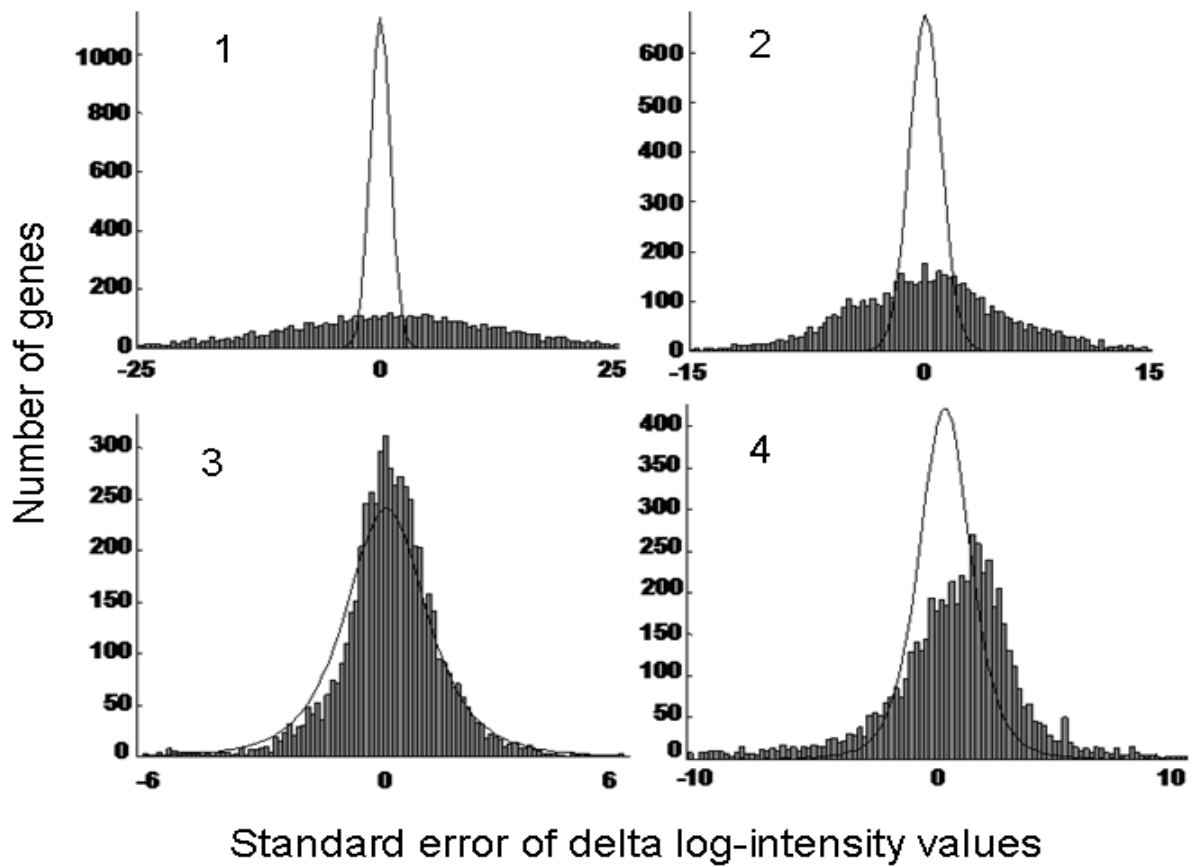


Figure 1 – Histogram of standardized expected $\Delta \log$ -intensity ($u_i/\text{stdev}(u_i)$) for each of the four datasets. Histograms for datasets 1 and 2 are very different from the continuous line (standard Normal distribution). The distribution of standardized values

in dataset 3 looks Normal, whereas in dataset 4 it has clearly a thicker tail than a Normal distribution and its histogram is skewed to the right. The result of long tails in datasets 1 and 2 and to a lesser extent in dataset 4, suggest inflated t-statistics, originating from the underestimation of gene-specific variances. Also, heavy tails on datasets 1 and 2 indicate high biological heterogeneity between samples.

4.4.2 Enrichment measures – CRAM

We compared the enrichment of CRAM to four other approaches of identifying differentially expressed genes. These approaches were two state of the art algorithms for measuring differential expression, **locally moderated t-statistic** (LMW), **weighted moderated t-statistic** (WAME) and two general algorithms **fold change** (FC) and **t-statistic** (t). We also compare against an approach, RANDOM, which generates random gene lists as a baseline for comparison.

We first considered up-regulated genes that were high in the quiescent population compared to non-quiescent. CRAM had the highest enrichment for gene list sizes 400 and 800 in dataset 1 (Table 2) and had equal performance, compared to WAME for gene list size 200.

Table 2 – Enrichment (average $-\log$ p-value) for gene lists of different size for each method. Highest enrichment for each sample size in bold. Standard errors are approximately 0.25 for all methods excluding RANDOM with SE = 0.04. If the highest and second highest enrichment are not significantly different, the two enrichment values are made bold.

DATASET 1 - GO ENRICHMENT 1000 simulations						
list size	CRAM	WAME	LMW	FC	t	RANDOM
200	24.84	24.89	22.78	23.3	10.91	6.93
400	37.97	36.97	31.68	31.89	12.52	7.12
800	32.28	31.00	31.23	28.97	16.04	6.88

Dataset 2 showed CRAM and WAME as the most enriched and equivalent for all sample sizes (Table 3).

Table 3 – Enrichment (average -log p-value) for gene lists of different size for each method. Highest enrichment for each sample size in bold. Standard errors are approximately 0.25 for all methods excluding RANDOM with SE = 0.04. If the highest and second highest enrichment are not significantly different, the two enrichment values are made bold.

DATASET 2 - GO ENRICHMENT 1000 simulations						
list size	CRAM	WAME	LMW	FC	t	RANDOM
200	24.11	24.22	23.85	23.49	15.15	6.92
400	103.71	103.91	85.24	84.5	36.25	6.87
800	84.81	84.65	83.71	76.8	44.95	6.96

In dataset 3, CRAM had the highest enrichment for all gene list sizes (Table 4).

Table 4 – Enrichment (average -log p-value) for gene lists of different size for each method. Highest enrichment for each sample size in bold. Standard errors are approximately 0.25 for all methods excluding RANDOM with SE = 0.04. If the highest and second highest enrichment are not significantly different, the two enrichment values are made bold.

DATASET 3 - GO ENRICHMENT 1000 simulations						
list size	CRAM	WAME	LMW	FC	t	RANDOM
200	13.76	11.56	11.5	10.02	13.04	6.91
400	17.8	15.07	15.91	11.2	14.62	6.97
800	21.61	18.67	20.11	13.4	15.7	7.01

In dataset 4 CRAM had the highest enrichment for gene list sizes 800 and had similar

enrichment to standard t statistics for sample size 400 (Table 5).

Table 5 – Enrichment (average -log p-value) for gene lists of different size for each method. Highest enrichment for each sample size in bold. Standard errors are approximately 0.25 for all methods excluding RANDOM with SE = 0.04. If the highest and second highest enrichment are not significantly different, the two enrichment values are made bold.

DATASET 4 - GO ENRICHMENT 1000 simulations						
list size	CRAM	WAME	LMW	FC	T	RANDOM
200	70.0	53.37	53.13	26.43	92.5	6.53
400	84.75	74.7	70.41	22.16	85.0	6.97
800	68.59	63.2	61.77	24.07	62.06	6.88

4.4.3 Enrichment measures – CRAM-GC

To measure the results, each sample had its gene enrichment evaluated separately. Given that some tested datasets are heterogeneous (biologically dissimilar) in which case, combining samples is of questionable use. Our goal is to measure the improvement in GO enrichment from every individual sample.

In heterogeneous datasets, standard methods for evaluating the gene-specific variance σ_i^2 , assume a gene-specific variance moderation parameter equal to zero, meaning the gene-specific variance is constant for all genes. This constant variance assumption has been shown to produce higher overlap between samples even in homogeneous datasets. In this sense, we chose to compare GO enrichment of every array using the original fold change values, compared to CRAM-GC expected fold change using (4.13). Also, the

sample-specific variance from other approaches is assumed constant within the sample, meaning that it does not influence the rank order of the genes when selecting for a gene list using a single sample. Thus, ranking genes from a single sample by their original fold change values is the equivalent of ranking them by using other common approaches.

The p-value for the most significant GO category for every sample comparing the original fold change log intensity value with CRAM-GC expected fold change was calculated, where the corresponding $-\log(\text{p-value})$ for the two approaches was averaged for all samples for each dataset. A one tail p-value for differences of means of between the two averages of $-\log(\text{p-value})$ for CRAM-GC and original fold change was generated for every gene list sizes of 200, 400 and 800. In addition, a measure of percentage of times CRAM-GS had superior enrichment compared to the original value was performed (tables 6 to 12). In all tables CRAM-GC had a significant p-value (< 0.05) for all genes list sizes. When comparing the number of times CRAM-GC had superior enrichment, CRAM-GC was marginally lower in (Table 6) (gene list size 200) and (Table 7) (gene list size 400), but superior in all other cases.

Table 6 – Enrichment (average $-\log$ p-value) for gene lists of different sizes for CRAM-GC and the original fold change (columns 1 and 2). P-values for differences of means (over 80 samples) between columns 1 and 2 were generated (column 3). In column 4 we present the % of times CRAM-GC had superior enrichment to the original fold change value (in case there were ties, a 0.5 count was assigned to both Original FC and CRAM-GC).

DATASET 1 – 80 paired samples				
List size	Original-FC	CRAM-GC	p-value diff	% Cram > Orig.
200	18.57	19.15	2.0E-02	45%
400	26.35	34.15	8.7E-10	72%
800	23.24	28.22	2.4E-08	69%

Table 7 – Enrichment (average $-\log$ p-value) for gene lists of different sizes for CRAM-GC and the original fold change (columns 1 and 2). P-values for differences of means (over 88 samples) between columns 1 and 2 were generated (column 3). In column 4 we present the % of times CRAM-GC had superior enrichment to the original fold change value (in case there were ties, a 0.5 count was assigned to both Original FC and CRAM-GC).

DATASET 2 – 88 paired samples				
List size	Original-FC	CRAM-GC	p-value diff	% Cram > Orig.
200	18.50	19.17	6.8E-03	57%
400	27.62	33.86	2.0E-07	49%
800	26.61	30.66	2.4E-08	56%

Table 8 – Enrichment (average $-\log$ p-value) for gene lists of different sizes for CRAM-GC and the original fold change (columns 1 and 2). P-values for differences of means (over 16 samples) between columns 1 and 2 were generated (column 3). In column 4 we present the % of times CRAM-GC had superior enrichment to the original fold change value (in case there were ties, a 0.5 count was assigned to both Original FC and CRAM-GC).

DATASET 3 – 16 paired samples				
List size	Original-FC	CRAM-GC	p-value diff	% Cram > Orig.
200	13.35	33.82	1.7E-03	89%
400	13.39	34.37	8.7E-04	98%
800	13.03	28.90	7.6E-04	73%

Table 9 – Enrichment (average $-\log$ p-value) for gene lists of different sizes for CRAM-GC and the original fold change (columns 1 and 2). P-values for differences of means (over 10 samples) between columns 1 and 2 were generated (column 3). In column 4 we present the % of times CRAM-GC had superior enrichment to the original fold change value (in case there were ties, a 0.5 count was assigned to both Original FC and CRAM-GC).

DATASET 4 – 10 paired samples				
List size	Original-FC	CRAM-GC	p-value diff	% Cram > Orig.
200	6.60	8.84	2.8E-06	90%
400	7.91	9.22	1.2E-03	78%
800	8.07	9.16	3.4E-02	58%

Table 10 – Enrichment (average -log p-value) for gene lists of different sizes for CRAM-GC and the original fold change (columns 1 and 2). P-values for differences of means (over 174 samples) between columns 1 and 2 were generated (column 3). In column 4 we present the % of times CRAM-GC had superior enrichment to the original fold change value (in case there were ties, a 0.5 count was assigned to both Original FC and CRAM-GC).

DATASET 5 – 174 unpaired samples				
List size	Original-FC	CRAM-GC	p-value diff	% Cram > Orig.
200	38.19	50.18	1.2E-07	62%
400	44.54	59.41	1.5E-11	77%
800	50.86	65.47	2.1E-13	79%

Table 11 – Enrichment (average -log p-value) for gene lists of different sizes for CRAM-GC and the original fold change (columns 1 and 2). P-values for differences of means (over 36 samples) between columns 1 and 2 were generated (column 3). In column 4 we present the % of times CRAM-GC had superior enrichment to the original fold change value (in case there were ties, a 0.5 count was assigned to both Original FC and CRAM-GC).

DATASET 6 – 36 unpaired time course samples				
List size	Original-FC	CRAM-GC	p-value diff	% Cram > Orig.
200	91.76	93.06	3.6E-02	65%
400	132.50	136.68	1.5E-05	67%
800	138.67	142.17	1.5E-08	79%

Table 12 – Enrichment (average $-\log p$ -value) for gene lists of different sizes for CRAM-GC and the original fold change (columns 1 and 2). P-values for differences of means (over 269 samples) between columns 1 and 2 were generated (column 3). In column 4 we present the % of times CRAM-GC had superior enrichment to the original fold change value (in case there were ties, a 0.5 count was assigned to both Original FC and CRAM-GC).

DATASET 7 – 269 unpaired samples				
List size	Original-FC	CRAM-GC	p-value diff	% Cram > Orig.
200	31.00	63.62	4.0E-47	93%
400	29.71	55.71	7.8E-42	91%
800	35.93	59.03	1.1E-53	88%

4.5 DISCUSSION AND CONCLUSION

We have presented the Calibration Regression Analysis of Microarrays (CRAM) and shown that CRAM had superior performance compared to other approaches. Similar to existing approaches, CRAM models gene-specific variance as a function of intensity (Baldi and Long 2001; Sartor, Tomlinson et al. 2006; Fodor, Tickle et al. 2007; Astrand, Mostad et al. 2008; PLW-Astrand 2008). CRAM improves on existing approaches by explicitly modeling the dependency between intensity level and sample-specific variance. The estimation of different sample-specific variances for every gene in each sample was shown to improve the accuracy in testing for differential expression. In addition, CRAM-GC was shown to use effectively the correlation between genes, producing a major improvement in GO enrichment compared to the standard CRAM model. CRAM-CG was also shown to be applicable to heterogeneous datasets.

4.5.1 Overcoming biological heterogeneity

We described in section 4.4.1, the distortions in estimating gene-specific variance in datasets with high heterogeneity. In datasets 1, 2 and 4, extreme values for the standardized values of expected Δ log-intensity indicate a strong underestimation of gene-specific variance. Our results show CRAM overcomes some limitations of distortions in gene-specific variance estimates, which are more extreme in heterogeneous datasets, by finding that the optimal gene-specific variance moderating factor $\alpha = 0$, which is equivalent to assuming a constant gene-specific variance. However, estimating gene-specific variances in heterogeneous datasets based on the expected log-intensity, will generate underestimated variances resulting in inflated t-statistics, such that gene lists based on these t-statistics will have inferior enrichment.

Studies have shown that different genes have different variances in expression in homogeneous datasets (Bar-Even, Paulsson et al. 2006; Wentzell, Karakach et al. 2006). On the other hand, recent studies have shown that in homogeneous datasets, a higher concordance between samples is achieved using fold change measures (Shi, Perkins et al. 2008). Curiously, when estimating t-statistics measures in homogeneous datasets, modeling gene-specific variances based on optimizing for the highest concordance between samples, often results in sub-optimal GO enrichment. This is still a highly debatable topic in current literature.

4.5.2 Gene enrichment performance

Although LMW is an extension of WAME, modeling gene-specific variance as a function of intensity, WAME performed better than LMW most of the time (Kristiansson, Sjogren et al. 2005; Bar-Even, Paulsson et al. 2006; Kristiansson, Sjogren et al. 2006; Wentzell, Karakach et al. 2006; Astrand, Mostad et al. 2007; Sjogren, Kristiansson et al. 2007; Astrand, Mostad et al. 2008). This behavior can be attributed to over fitting of the gene-specific variance by LMW, due to the non-normality of the distribution of Δ log-intensity values. CRAM, which is similar to WAME with the main difference being the modeling of sample-specific variance within each sample, is also robust to over fitting of the gene-specific variance. In general, modeling gene-specific variance as a function of intensity by CRAM (using groups) proved to be more robust than LMW and at the same time, generated higher enrichment than WAME, making CRAM biologically accurate with heterogeneous and homogeneous data.

The enrichment performance of CRAM gene correlation (CRAM-GC) was superior in every tested dataset, showing that the untransformed fold change values were significantly less enriched than when gene correlations between genes were taken into account. This superior enrichment was shown in paired and non-paired designs, in two and one-color microarrays, in single time point and time course data. However, in order to fully achieve the benefits of using gene-correlation, it is necessary to have a datasets with number of samples greater or equal to 10. The primary consequence of using a sample dataset too small is that correlations will be too unstable with high correlations being detected as a result of chance (even after Fisher's transformation adjustment),

generating a large number of false positive genes, thus resulting in reduced GO enrichment.

CRAM-applied was also applied to datasets 1 to 4, where most significant GO enrichment categories for genes higher in Q and similarly for genes higher in NQ (tables 13 and 14) were obtained by combining all samples.

Table 13 – Most significant GO categories for gene lists significantly more expressed in Q than in NQ (p-value < E-6). An additional list was obtained for a sample of size 200 (due to the fact that the gene list based on p-value was not enriched). On the column named ‘p-value’ we have the p-value for the column ‘# of genes’. On the right hand side we have the ratio between observed number of genes for GO category and expected number of genes under the random assumption.

Q Gene lists – Biological Process GO enrichment						
DATASETS	GO category	# genes	total in category	sample	p-value	ratio
Dataset 1	Ethanol Metabolic Process	5	11	59	8.6 E-06	40
	Fermentation	5	16	59	7.8E-05	30
	Monocarboxylic Acid MP	22	132	200	1.9E-07	5
Dataset 2	Monocarboxylic Acid MP	14	132	60	4.9E-09	10
	Amine Catabolic Process	4	7	60	5.8E-05	65
Dataset 3	Glucose Catabolic Process	6	39	64	5.5E-04	15
Dataset 4	Pyruvate Metabolic Process	13	40	68	7.4E-15	30
	Glycolysis	9	22	68	6.2E-11	35

Table 14 – Most significant GO categories for gene lists significantly more expressed in NQ than in Q (p-value < E-6). An additional list was obtained of size 200. An additional list was obtained for a sample of size 200 (due to the fact that the gene list based on p-value was not enriched). On the column named ‘p-value’ we have the p-value for the column ‘# of genes’. On the right hand side we have the ratio between observed number of genes for GO category and expected number of genes under the random assumption.

NQ Gene lists – Biological Process GO enrichment						
DATASETS	GO category	# genes	total in category	sample	p-value	ratio
Dataset 1	Hexose Transport	6	25	62	2.50E-05	25
Dataset 2	Hexose Transport	14	132	200	4.9E-09	10
Dataset 3	Asparagin MP	6	39	58	5.5E-04	15
Dataset 4	Carboxylic Acid MP	13	40	65	7.4E-15	30

Both datasets 1 and 2 shared similar GO categories for both Q (“Monocarboxylic Acid Metabolic Process”). Similarly for NQ, both datasets shared the same GO categories (“Hexose Transport”). Dataset 4 most significant GO categories for Q (“Pyruvate MP” and “Glycolysis”) are subsets of “Monocarboxylic Acid MP”. Curiously, for NQ, dataset 4 most significant GO category was Carboxylic Acid Metabolic Process, which is a parent category of “Monocarboxylic Acid MP”, suggesting that some genes from the category Carboxylic Acid may be more highly expressed in Q and others more highly expressed in the NQ cell population.

4.5.3 Computational speed factors

CRAM was written in MATLAB, and was shown to be highly computationally efficient, taking an order of magnitude less time, to score the datasets on the same computer, as compared to other sample-specific variance modeling approaches. The treatment of each sample as an individual dataset using the framework in (Ibrahim and Chen 2000), combined with an empirical Bayes approach and measurement error is the main reason for this efficiency (HerbertR. 1956; Fuller 1987).

4.5.4 Future applications

We intend to apply CRAM to time series microarray data, which can be highly heterogeneous depending upon the temporal dynamics of the system under study. To accurately model log-intensity from time series data, CRAM needs to be enhanced to have gene-specific variance σ_{ij}^2 for every sample j and for every gene i as opposed to

using a constant gene-specific variance σ_i^2 for every sample for gene i . More robust estimation of the weights can be developed where each sample is classified into more groups. Instead of grouping genes within each sample based on their transformed log intensity levels, it is possible to group genes using GO biological process categories, which is likely to better model interactions among genes, leading to improved performance. Finally, similar experiments can be combined by using each gene's expected log-intensity (or Δ log-intensity) from a previous experiment, as the prior expected log-intensity in the current experiment.

4.6 ACKNOWLEDGMENTS

We thank three unknown reviewers for their insightful comments. We thank Diego Martinez and Michele Guindani for their comments on an earlier version of this work.

4.7 SUPPLEMENTAL MATERIALS

4.7.1 Sample variance distortion in biological heterogeneous datasets

Estimation based on gene sample variance in heterogeneous datasets, can highly distort the estimation of gene-specific variance, and thus distorts gene-expression variance estimates. The assumption of a log-intensity expectation τ_i for each gene i in estimating gene sample variances, does not hold in the presence of high biological heterogeneity where each gene i for each sample j , has a more specific (to gene i and sample j) expectation τ_{ij} . When assuming a fixed τ_i for a gene i , the error component for gene i in sample j , is based on the deviation between τ_i and x_{ij} , whereas the true error should be based on the deviations between τ_{ij} and x_{ij} . As a result, assuming samples have the same weights, when estimating the gene sample variance s_i^2 , many approaches assume the true unknown sum of errors across arrays to be $\sum_{j=1}^k (x_{ij} - \tau_i)^2$ when it should

be $\sum_{j=1}^k (x_{ij} - \tau_{ij})^2$. We can write $\sum_{j=1}^k (x_{ij} - \tau_i)^2$ as the sum

$$\sum_{j=1}^k (x_{ij} - \tau_{ij})^2 + \sum_{j=1}^k (\tau_i - \tau_{ij})^2 - 2 \sum_{j=1}^k (x_{ij} - \tau_{ij})(\tau_i - \tau_{ij}).$$
 When true errors are small,

that is, when x_{ij} is close to τ_{ij} , the third term becomes very small and we have

$$\sum_{j=1}^k (x_{ij} - \tau_i)^2 \geq \sum_{j=1}^k (x_{ij} - \tau_{ij})^2,$$
 showing an overestimation of the sample variance when

using an overall τ_i . On the other hand, when

$\sum_{j=1}^k (\boldsymbol{\tau}_i - \boldsymbol{\tau}_{ij})^2 < 2 \sum_{j=1}^k (x_{ij} - \boldsymbol{\tau}_{ij}) (\boldsymbol{\tau}_i - \boldsymbol{\tau}_{ij})$, we have an underestimation of the variance.

In Fig.1, the observed heterogeneity of datasets 1, 2 and 4, sample variance is underestimated, suggesting the presence of large errors, where x_{ij} is far from $\boldsymbol{\tau}_{ij}$ relative to the distance between $\boldsymbol{\tau}_i$ and $\boldsymbol{\tau}_{ij}$. In datasets 1 and 2, underestimation of gene sample variance was extreme to the point that it was more accurate in some cases, to assume a homogeneous (constant) gene-specific variance, than modeling gene-specific variance based on the gene sample variance s_i^2 .

4.7.2 Proof: $P(\boldsymbol{\tau}_i | x_{ij}) \sim N((\mu_i + w_j x_{ij}) / (1 + w_j), \sigma_i^2 / (1 + w_j))$.

We assume that every gene i for an sample j , has a corresponding random variable $\boldsymbol{\tau}_i$ with distribution $P(\boldsymbol{\tau}_i) \sim N(\mu_i, \sigma_i^2)$ which has the known prior expectation μ_i for gene i and where $w_j > 0$, is an unknown parameter which is constant within sample j , representing the level of confidence we have in this sample. The parameter w_j is also the inverse of the sample-specific variance defined as $1/w_j$. Given this, we can write

$$P(x_{ij} | \boldsymbol{\tau}_i = \lambda_i) = P(\boldsymbol{e}_{ij} + \boldsymbol{\tau}_i | \boldsymbol{\tau}_i = \lambda_i) = P(\boldsymbol{e}_{ij} + \lambda_i | \boldsymbol{\tau}_i = \lambda_i) \sim N(\lambda_i, \sigma_i^2 / w_j),$$

since \boldsymbol{e}_{ij} is $N(0, \sigma_i^2 / w_j)$ -distributed.

We pause to recall a convention from Bayesian statistics. If one has two random variables \mathbf{X} and \mathbf{Y} , one can write the conditional distribution $P(\mathbf{Y}|\mathbf{X})$ as

$$P(\mathbf{Y}|\mathbf{X}) = P(\mathbf{X},\mathbf{Y})/P(\mathbf{X}) = P(\mathbf{Y})P(\mathbf{X}|\mathbf{Y})/P(\mathbf{X}).$$

This is a form of Bayes' Theorem. Now, we consider $1/P(\mathbf{X})$ as a constant of proportionality since \mathbf{X} is given and fixed in the conditional probability $P(\mathbf{Y}|\mathbf{X})$. Thus, we can write

$$P(\mathbf{Y}|\mathbf{X}) \propto P(\mathbf{Y})P(\mathbf{X}|\mathbf{Y})$$

where \propto is the symbol for proportionality (Lee 1997).

By the above discussion, we then have $P(\boldsymbol{\tau}_i | \mathbf{x}_{ij}) \propto P(\mathbf{x}_{ij} | \boldsymbol{\tau}_i) P(\boldsymbol{\tau}_i)$.

Let \mathbf{Z} be a random variable that has the distribution

$$P(\mathbf{x}_{ij} | \boldsymbol{\tau}_i = \lambda_i) P(\boldsymbol{\tau}_i) \sim N(\lambda_i, \sigma_i^2/w_j) N(\mu_i, \sigma_i^2)$$

Then

$P(\mathbf{Z} = z) = \left(\frac{1}{2} 2\pi \sigma_i^2 w^{1/2} \right) \exp\left(-\frac{1}{2} \{w_j(z - x_{ij})^2 + (z - \mu_i)^2\} / \sigma_i^2 \right)$ which is proportional to

$$\exp\left(-\frac{1}{2} \{w_j z^2 - 2w_j x_{ij} z + w_j x_{ij}^2 + z^2 - 2z\mu_i + \mu_i^2\} / \sigma_i^2 \right) =$$

$$\exp\left(-\frac{1}{2} \{ (w_j+1) z^2 - 2z (w_j x_{ij} + \mu_i) + w_j x_{ij}^2 + \mu_i^2 \} / \sigma_i^2 \right) =$$

$\exp(-\frac{1}{2} \{ (w_j+1)[z^2 + 2(w_j x_{ij} + \mu_i)/(1+ w_j) z+ w_j x_{ij}^2/(1+ w_j) + \mu_i^2/(1+ w_j)]\}/ \sigma_i^2)$ and by completing the square by adding and subtracting the term $(x_{ij}w_j + \mu_i)^2/(1+ w_j)^2$ in the part of the equation between the brackets, we have

$$\exp(-\frac{1}{2} [z- (x_{ij} w_j + \mu_i)/(1+ w_j)]^2 /(\sigma_i^2/(1+ w_j)))\exp(-\frac{1}{2} [w_j x_{ij}^2 - (x_{ij}w_j + \mu_i)^2/(1+ w_j)]/ \sigma_i^2).$$

Since the second exponential term does not have the term z , it is just a proportionality factor, thus, we can say the previous expression is proportional to the first exponential term

$$\exp(-\frac{1}{2} [z- (x_{ij} w_j + \mu_i)/(1+ w_j)]^2/ (\sigma_i^2/(1+ w_j)))$$

which is proportional to a normal distribution with mean $(\mu_i + w_j x_{ij})/(1+ w_j)$ and variance $\sigma_i^2 / (1+ w_j)$. Thus we can write

$$P(\tau_i | x_{ij}) \propto N((\mu_i + w_j x_{ij})/(1+ w_j) , \sigma_i^2 / (1+ w_j))$$

and we are done.

4.7.3 Proof: $P(\tau_i | X) \sim N((\mu_i + \sum_{j=1}^k w_j x_{ij})/(1+ \sum_{j=1}^k w_j) , \sigma_i^2 / (1+ \sum_{j=1}^k w_j))$.

$P(\tau_i | x_{i1}, x_{i2}, \dots, x_{ik}, w_1, w_2, \dots, w_k)$ is proportional to

$P(x_{i1}, x_{i2}, \dots, x_{ik} | w_1, w_2, \dots, w_k, \mu_i) P(\boldsymbol{\tau}_i)$ and by independence between samples,

$P(x_{i1}, x_{i2}, \dots, x_{ik} | w_1, w_2, \dots, w_k, \boldsymbol{\tau}_i)$ is proportional to

$P(x_{i1} | w_1, \boldsymbol{\tau}_i) P(x_{i2} | w_2, \boldsymbol{\tau}_i) \dots P(x_{ik} | w_k, \boldsymbol{\tau}_i)$ and therefore

$P(\boldsymbol{\tau}_i | x_{i1}, x_{i2}, \dots, x_{ik}, w_1, w_2, \dots, w_k)$ is proportional to

$P(x_{i1} | w_1, \boldsymbol{\tau}_i) P(x_{i2} | w_2, \boldsymbol{\tau}_i) \dots P(x_{ik} | w_k, \boldsymbol{\tau}_i) P(\boldsymbol{\tau}_i)$ which is the kernel of a Normal distribution with mean $(\mu_i + \sum_{j=1}^k w_j x_{ij}) / (1 + \sum_{j=1}^k w_j)$ and variance $\sigma_i^2 / (1 + \sum_{j=1}^k w_j)$.

This last step can be shown in greater detail if we follow the same steps as in 4.7.2 where we complete the square and get an expression proportional to

$$\exp\left(-\frac{1}{2} \left[(\boldsymbol{\tau}_i - (\mu_i + \sum_{j=1}^k w_j x_{ij}) / (1 + \sum_{j=1}^k w_j))^2 / (\sigma_i^2 / (1 + \sum_{j=1}^k w_j)) \right]\right).$$

4.7.4 Estimating w_j , the weight parameter of a sample j .

Remembering that $\mathbf{Y}_j = \{y_{1j}, y_{2j}, \dots, y_{nj}\}$ is an estimate of $\mathbf{X}_j = \{x_{1j}, x_{2j}, \dots, x_{nj}\}$, let the vector $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_n\}$ be the vector of known prior expectations and denoting

$\mathbf{T} = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \dots, \boldsymbol{\tau}_n\}$ the vector of unknown expectations, from (4.4) or 4.7.2, we have

$$E[\mathbf{T} | \mathbf{X}_j] = (\boldsymbol{\mu} + w_j \mathbf{X}_j) / (1 + w_j).$$

By performing the regression $Y_j = E[T | X_j]$, we estimate w_j which minimizes the loss function L_j , denoted as the sum $\sum_{i=1}^n ((\mu_i + w_j x_{ij})/(1 + w_j) - y_{ij})^2$, we set the derivative

$$\partial L_j / \partial w_j = 0, \text{ and thus } \sum_{i=1}^n 2((\mu_i + w_j x_{ij})/(1 + w_j) - y_{ij})(x_{ij} - \mu_i)/(1 + w_j)^2 = 0 .$$

Since the term $(1+w_j)^2$ is constant throughout sample j , we can simplify the equation by

$$\text{making } \sum_{i=1}^n ((\mu_i + w_j x_{ij})/(1 + w_j) - y_{ij})(x_{ij} - \mu_i) = 0 .$$

Thus, solving for w_j we get

$$w_j = \left(\sum_{i=1}^n (y_{ij} - \mu_i)(x_{ij} - \mu_i) \right) / \left(\sum_{i=1}^n (x_{ij} - y_{ij})(x_{ij} - \mu_i) \right). \quad (4.7)$$

4.7.5 CRAM algorithm

We present the CRAM algorithm where the CRAM *z-statistic* is calculated.

/*Calculation of weights w_{ij} , for all n genes, k samples and g groups per sample.*/

for each sample j :

- 1- Group genes into three groups using their intensity values, by sorting in descending order by x_{ij}

for each group g :

a. Perform a linear regression where $\mathbf{X}_j = \{x_{1j}, x_{2j}, \dots, x_{nj}\}$ is estimated using the remaining set of \mathbf{X}_l where $l \neq j$, generating a set of $\mathbf{Y}_j = \{y_{1j}, y_{2j}, \dots, y_{nj}\}$ predicted values.

b. Calculate $W_{gj} = \left(\sum_{i=1}^n I_g(y_{ij} - \mu_i)(x_{ij} - \mu_i) \right) / \left(\sum_{i=1}^n I_g(x_{ij} - y_{ij})(x_{ij} - \mu_i) \right)$,

where I_g is the indicator function for a gene belonging to group g .

end loop

c. For every gene i in array vector j , assign the weight $w_{ij} = W_{gj}$.

end loop

/*Calculation of z-statistic*/

for each gene i :

1- Get the estimate u_i for the unknown expectation τ_i :

$$u_i = \left(\mu_i + \sum_{j=1}^k w_{ij} x_{ij} \right) / \left(1 + \sum_{j=1}^k w_{ij} \right)$$

2- Calculate the weighted sample variance:

$$s_i^2 = \sum_{j=1}^k w_{ij} (x_{ij} - u_i)^2 / \left(\left(\sum_{j=1}^k w_{ij} \right)^2 - \sum_{j=1}^k w_{ij}^2 \right) / \sum_{j=1}^k w_{ij}$$

3- Calculate the average of all s_i^2 :

$$V^2 = \sum_{i=1}^n s_i^2 / n$$

4- Use optimal α and estimate the gene-specific variance:

$$\hat{\sigma}_i^2 = \alpha s_i^2 + (1 - \alpha) V^2$$

5- Generate the estimate for \hat{a}_i

$$\hat{a}_i = (\mu_i + \sum_{j=1}^k w_{ij} x_{ij}) / (\hat{\sigma}_i \sum_{j=1}^k w_{ij} (1 + w_{ij}))^{1/2}$$

6- Calculate $\hat{a}_{i_m} = \sum_{i=1}^n \hat{a}_i / n$ and $S = (\sum_{i=1}^n (\hat{a}_i - \hat{a}_{i_m})^2 / (n-1))^{1/2}$

7- Generate the CRAM *z*-statistic, which is approximately $N(0,1)$:

$$z_i = (\hat{a}_i - \hat{a}_{i_m}) / S$$

A two tail p-value for each gene can be generated by $2\Phi(-|z_i|)$, where Φ is the standard Normal cumulative distribution function.

end loop

4.7.6 Estimating the variance of the posterior expectation u_i

To calculate the variance of the estimated expectation u_i , we have

$$\text{Var}(u_i) = \text{Var}((\mu_i + \sum_{j=1}^k w_j x_{ij}) / (1 + \sum_{j=1}^k w_j)) = \sum_{j=1}^k w_j^2 \text{Var}(x_{ij}) / (1 + \sum_{j=1}^k w_j)^2.$$

By (4.1) and (4.2) variance of x_{ij} is given by

$\text{Var}(x_{ij}) = \text{Var}(\tau_i) + \text{Var}(e_{ij}) = \sigma_i^2 + \sigma_i^2 / w_j = \sigma_i^2 (1 + w_j) / w_j$ and thus

$$\text{Var}\left(\frac{\mu_i + \sum_{j=1}^k w_j x_{ij}}{1 + \sum_{j=1}^k w_j} \right) = \sigma_i^2 \frac{\sum_{j=1}^k w_j (1 + w_j)}{\left(1 + \sum_{j=1}^k w_j\right)^2}$$

(4.8)

4.7.7 Optimizing the α parameter

In this paper, we will set $\alpha = 0.9$ for homogeneous datasets (datasets 3 and 4) and $\alpha = 0$ for heterogeneous datasets (datasets 1 and 2). This criterion for selecting α is used to generate the CRAM z-statistic in (4.11) which is used for generating lists of size 200, 400 and 800. We present two methods to estimate α .

4.7.8 α estimation for the whole dataset (Method 1)

The main assumption is that the α which best models the upper tail of a true $N(0,1)$, is the α that best moderates the gene-specific variance. We chose the 90th percentile ($q = 0.90$) to define the upper tail, although other values could be used as well. An individual α is optimized for each dataset.

For every simulated $\alpha = 0, 0.1, 0.2, \dots, 0.9, 1.0$.

1. Estimate u_i for every gene i , using equation (6)
2. Estimate σ_i^2 for every gene i using equation (10).
3. Generate the z-statistic a_i by (9).

4. Generate $a'_i = a_i - \{\text{mean of } a_i \text{ over all } n \text{ genes}\}$.
5. Sort in descending order by a'_i .
6. Select a'_i corresponding to the 90th percentile. Call this a'_{90th} .
7. Generate $p_{90th} = \Phi(-|a'_{90th}|)$ where Φ is the Standard Normal cdf.
8. Calculate $\Delta(\alpha) = |p_{90th} - (1 - q)| = |p_{90th} - 0.10|$.

End loop

Select α that generates the smallest $\Delta(\alpha)$ for all simulated values of α .

4.7.9 α estimation for each simulated sample (Method 2):

If the research conducting the microarray experiment has a small dataset (small number of samples), this method will be more appropriate than method 1. In order to estimate α , the dataset is divided into test data and hold out data. The main assumption is that the optimal alpha is the one that has the highest overlap between the test data (Test) and the hold out data (Ho).

For every simulated $l = 1, \dots, 11$

Let $\alpha = 0.1(l-1)$

For each sample $j = 1, \dots, k$

1. Select array j as the hold data. Select the remaining arrays as test data.

2. Use the weight w_j of sample j , to estimate the expectation of the hold out data.

$$\text{Ho_u}_i = (\mu_i + w_j x_{ij}) / (1 + w_j)$$

3. Use the weights excluding w_j to estimate the expectation of the test data.

$$\text{Test_u}_i = (\mu_i + \sum_{l=1 \neq j}^k w_{il} x_{il}) / (1 + \sum_{j=1}^k w_{il})$$

4. Estimate s_i^2 for every gene i in the test data.

$$s_i^2 = \sum_{l=1 \neq j}^k w_{il} (x_{il} - \text{Test_u}_i)^2 / ((\sum_{l=1}^k w_{il})^2 - \sum_{l=1}^k w_{il}^2) / \sum_{l=1}^k w_{il}$$

5. Calculate the average of all s_i^2 in the test data:

$$V^2 = \sum_{i=1}^n s_i^2 / n$$

6. Estimate σ_i^2 for every gene i in the test data using the simulated α .

$$\hat{\sigma}_i^2 = \alpha s_i^2 + (1 - \alpha) V^2$$

7. Standardize the expectations of the hold out data and test data by dividing by $\hat{\sigma}_i$.

$$\text{Test_t}_i = \text{Test_u}_i / \hat{\sigma}_i$$

$$\text{Ho_t}_i = \text{Ho_u}_i / \hat{\sigma}_i$$

8. Sort both Test_t and Ho_t in descending order.

9. In both sorted Test_t and Ho_t, select for the top C genes.

10. Count the number of overlapping genes after selection between Test_t and Ho_t, and call this Pre_overlap(*l*, *j*).

End loop

11. Get the mean of the Pre_overlap(*l*, *j*) over all *j*. Call this Overlap(*l*).

End loop

12. Set *pos* = Argmax(Overlap).

13. Optimal α is equal to $0.1(pos-1)$.

4.7.10 Other versions of CRAM

4.7.10.1 CRAM-binary

This is a robust version of CRAM and was developed to handle datasets with a high incidence of outliers. Instead of using its original values, all samples X_j are ranked and categorized as a dichotomous variable.

4.7.10.2 CRAM expectation maximization (CRAM-EM)

This is the fastest version of CRAM. It replaces the regression step in 4.3.3.2 by assigning initial weights $w_1, w_2, \dots, w_k = 1$. In this version, X_j is also estimated using

the remaining samples but instead of using a regression, X_j is estimated using an iterative procedure where a weighted average is used. Thus Y_j , the estimate of X_j is the vector denoted by $(y_{1j}, y_{2j}, \dots, y_{nj})$, where $y_{ij} = \left(\sum_{l \neq j=1}^k w_l x_{ij} \right) / \left(\sum_{l \neq j=1}^k w_l \right)$.

4.7.11.2 CRAM-EM Algorithm

1- Assign initial weights $w_1, w_2, \dots, w_k = 1$.

For every iteration until convergence of the weights

For every sample j

2- Estimate Y_j using the current weights, such that

$$y_{ij} = \left(\sum_{l \neq j=1}^k w_l x_{ij} \right) / \left(\sum_{l \neq j=1}^k w_l \right).$$

3- Estimate a new weight for sample j such that

$$w_j = \left(\sum_{i=1}^n (y_{ij} - \mu_i)(x_{ij} - \mu_i) \right) / \left(\sum_{i=1}^n (x_{ij} - y_{ij})(x_{ij} - \mu_i) \right).$$

Assuming negative weights are not possible, if $w_j < 0$ we make $w_j = 0$.

End loop

4 - Update all weights w_1, w_2, \dots, w_k .

5- If convergence then end.

End loop

Once the weights are defined, the rest of the procedure is identical to CRAM. The as the number of samples increase, the EM procedure becomes much faster than using a multiple linear regression, with the same performance.

4.7.11.3 CRAM-binary algorithm

Define γ , a cutoff value such that $0 \leq \gamma \leq 1$.

For every sample j

Sort each X_j in ascending order and define $V_j = \{v_{1j}, v_{2j}, \dots, v_{nj}\}$ where

$v_{ij} = 1$ if x_{ij} is in the top γ proportion of the total number of genes and $v_{ij} = 0$

otherwise.

End Loop

As a result V_j will have approximately γ values equal to 1 and $1 - \gamma$ values equal to 0.

1- Assign initial weights $w_1, w_2, \dots, w_k = 1$.

For every iteration until convergence of the weights

For every sample j

2- Estimate Y_j using the current weights Y_j such that

$$Y_j = (1 + \sum_{l \neq j=1}^k w_l V_j) / (1/\gamma + \sum_{l \neq j=1}^k w_l).$$

3- Estimate a new weight w_j for sample j such that

$$w_j = \frac{1-\gamma}{\gamma} \text{Cov}(\mathbf{Y}_j, \mathbf{X}_j) / (\text{Var}(\mathbf{X}_j) - \text{Cov}(\mathbf{Y}_j, \mathbf{X}_j)).$$

Assuming negative weights are not possible, if $w_j < 0$ we set $w_j = 0$.

End loop

4 - Update all weights w_1, w_2, \dots, w_k .

5- If convergence then end.

End loop

6- Generate the final estimate u_i for every gene given by

$$u_i = (1 + \sum_{j=1}^k w_j v_{ij}) / (1/\gamma + \sum_{j=1}^k w_j). \text{ Note: if prior probabilities } \pi_i \text{ are known,}$$

$$\text{then } u_i = (1 + \sum_{j=1}^k w_j v_{ij}) / (1/\pi_i + \sum_{j=1}^k w_j).$$

4.7.11.4 CRAM-GC algorithm

1. Generate a Spearman correlation matrix with all samples.

For every sample j

2. Select the maximum correlation ρ_j among all the remaining samples.

3. Generate the sample weight a_j is given by

$$a_j = \left(\sum_{i=1}^n (y_{ij} - \mu_{ij})(x_{ij} - \mu_i) \right) / \left(\sum_{i=1}^n (x_{ij} - y_{ij})(x_{ij} - \mu_i) \right).$$

End loop

For every gene i

4. Generate a Spearman correlation matrix with all genes.
5. Select the maximum absolute correlation $|\rho_i|$ among all the remaining genes.
6. Use the Fisher's transformation to get a corresponding z_i statistics

$$z_i = f(\rho_i) = 0.5(k-3)^{1/2} \ln \left(\frac{1+|\rho_i|}{1-|\rho_i|} \right)$$

7. Let $\alpha = 1/n$ be the level of significance and let $z_{0i} = \Phi^{-1}(1-\alpha)$, where Φ^{-1} represents the inverse of the Normal cumulative distribution function.
8. Let $f^{-1}(z_{0i}) = (\exp\{2z_{0i}/(k-3)^{1/2}\} - 1) / (\exp\{2z_{0i}/(k-3)^{1/2}\} + 1)$
9. Let $z_{1i} = z_{0i} + f^{-1}(z_{0i}) (k-3)^{1/2} / 2(k-1)$.
10. Let $z_i' = z_i - z_{1i}$. If $z_i' < 0$ then set $z_i' = 0$.
11. By using the inverse Fisher transformation on z_i' , generate $\rho_i' = f^{-1}(z_i')$.

12. Generate the gene weight given by

$$g_i = \left(\sum_{j=1}^k (c_i |\rho_i'| / |\rho_i| x_{ij} - \mu_{ij})(x_{ij} - \mu_i) \right) / \left(\sum_{j=1}^k (x_{ij} - c_i |\rho_i'| / |\rho_i| x_{ij})(x_{ij} - \mu_i) \right).$$

End loop

For every array j

For every gene i

13. Generate the sample-gene weight $w_{ij} = a_j g_i$.

14. Generate $E[\tau_{ij} | x_{ij}] = (\mu_{ij} + w_{ij} x_{ij}) / (1 + w_{ij})$.

End loop

End loop

Chapter 5 - DISCUSSION

5.1 KEY IDEAS

The studies presented in the previous chapters describe applications in genomics of three key ideas: concordance, association and multi-dimensional differential expression. For each of them we have presented applications that identify differentially expressed genes in a computationally efficient way. In chapter 2 a general framework was described in which concordance was used to obtain the significance of multi-dimensional gene list overlaps. In chapter 4, CRAM models were developed where concordance and association concepts were combined to generate superior measures of differential expression in unsupervised data. In these models, variance in gene expression was estimated within and between samples, and was applied to heterogeneous and homogeneous datasets. In chapter 3 we described the SDI approach, which models differences in fluorescence intensity in multiple dimensions in order to detect heterogeneous cell populations in a high-throughput flow cytometry experiment.

5.2 GENE LIST OVERLAPS

The novel method to estimate the significance of gene list overlaps described in chapter 2 expands on the currently used GO term Finder application (GOTermFinder ; Boyle 2004) by estimating the significance of overlaps between multiple sets of genes. An example was described in figure 4, where two gene lists were combined with the set

of genes belonging to the ‘unknown’ GO category, in which the triple overlap was highly significant. One immediate application of this method is in the identification and classification of candidate genes into either a new GO category or a previously known one. In addition, we have shown that when measuring the significance of overlaps among three or more sets of genes, the order in which successive pairwise overlaps are obtained, is critical, resulting in a different p-value for each different order. Another way to measure the significance of higher order overlaps would be to apply multidimensional hypergeometric distributions (Kerov 2005), however, we have shown the limitations of this method since it limited to a single type of higher-order overlap (assuming independence among all sets), and thus, it is likely to miss some important lower-order overlaps, that are key to identify the optimal overlap. Based on the preliminary results in chapter 2, we hypothesize that important biological interactions can be detected and easily interpreted by using the method we developed.

5.3 CONCORDANCE AND MEASURES OF DIFFERENTIAL EXPRESSION

5.3.1 Limitations of standard t-statistics

Standard t-statistics applied to microarrays are based on assumptions that are often violated. The normality assumption of the expected fold change for each gene often does not hold due to large distortions in the tails of the distribution in most datasets (Fodor, Tickle et al. 2007). In order to improve the standard t-statistics, several variations of moderated t-statistics have been developed in which the variance in intensity

level for each gene is smoothed using highly sophisticated computational procedures applied to supervised data (Smith 2004; Kristiansson, Sjogren et al. 2005; Kristiansson, Sjogren et al. 2006; Astrand, Mostad et al. 2007; Astrand, Mostad et al. 2008) . Nevertheless, even moderated t-statistics often produce gene lists with lower reproducibility compared to genes lists generated using fold change. In order to maximize concordance between samples, we described a procedure in which t-statistics is moderated so that concordance between gene lists is maximized.

5.3.2 Concordance and moderated t-statistics

Most of the currently used moderated t-statistics moderation algorithms have been developed for supervised datasets where a dependent variable is present. In these algorithms a variance moderation parameter is optimized in such a way that the moderated t-statistics best predicts the dependent variable. Since our datasets are all unsupervised, we had to find alternative ways to moderate gene-specific variances.

In chapter 4 we have presented an algorithm that moderates the variance such that the average concordance between samples is optimized. As expected, the optimal variance moderating parameter was equal to zero for all heterogeneous datasets tested, which implies that the moderated t-statistics has the same rank ordering as fold change. On the other hand, even in the homogeneous datasets (3 and 4), the variance moderating parameter was equal to zero, confirming the assumption that optimal concordance is reached by using fold change measures (Shi, Perkins et al. 2008).

5.3.3 Concordance vs. GO enrichment

Fold change and t-statistics (moderated or not) are the two most common methods to evaluate differential expression. As described in detail by (Tusher, Tibshirani et al. 2001; Shi, Tong et al. 2005; Tu, Kudlicki et al. 2005; Shi, Reid et al. 2006; Shi, Perkins et al. 2008) when using biological replicates to measure concordance between gene lists, fold change is almost always superior to standard t-statistics and similarly to moderated t-statistics. However, although fold change is more likely to generate a higher concordance than a t-statistics, we should ask the question: Does a higher concordance imply in a higher GO enrichment? Preliminary results show that although a hypothetical gene list A can have a lower concordance than a gene list B, gene list A can have higher GO enrichment than gene list B. An example using microarray data, can be described in chapter 4 (tables 4 and 5) where biological replicates were used. In both tables, superior enrichment was observed for gene lists generated using standard t-statistics when compared to fold change, although when using t-statistics, concordance between samples was inferior to fold change. Another example was observed using flow cytometry data, where the gene list generated using SDI had a much higher GO enrichment than the gene list generated using average fold change, even though SDI samples had slightly lower concordance. Based on these findings, we hypothesize that concordance is more suited to detect measurement or experimental variability rather than biological variability. The fundamental idea behind this is that if biological differences exist between samples, we can always increase concordance by using a transformation which smooth's the original data. However, we should also keep in mind that by performing this type of smoothing transformation, we may also be losing valuable biological information. This is the case

when extreme values in gene expression are biologically accurate, and smoothing the data will have the effect of down weighting these values, resulting in inferior GO enrichment.

5.4 ASSOCIATION AND SAMPLE WEIGHT ESTIMATION

Another concept of concordance is based on the idea that gene expression measurements in a sample are more reliable when the sample has a high reproducibility with respect to at least one of other remaining samples in a set. In chapter 4, in addition to the previously described methods to estimate gene-specific variances, we also described some of the challenges of modeling sample-specific variances in unsupervised datasets by estimating sample weights. To estimate sample weights two important concepts were used: linear association and concordance.

5.4.1 Linear association between samples – CRAM model

Sample weights were estimated using the CRAM algorithm, in which association between samples was identified. To measure these associations, samples were regressed with remaining samples, generating a predicted log-intensity for every sample. A sample weight is thus, a measure of association between observed and predicted log-intensity for

that sample. In CRAM, the weight of a sample was estimated by applying a transformation such that the sum of square of the differences between predicted log-intensity and transformed observed log-intensity was minimized. This concept of estimating sample weights based on a linear regression was further extended to sub-groups of genes within each sample. A general framework was also presented assuming a known prior expected log-intensity for every gene.

Results in heterogeneous datasets 1 and 2, showed that the WAME method, which has constant weights within each sample but different weights among samples and is supervised method similar to the basic version of CRAM, had significant improvement in GO enrichment when compared to fold change (which assumes all sample weights are equal). This underlines the importance of modeling sample-specific variances as a way to detect more accurate biological patterns. Moreover, when WAME was compared to CRAM, results in datasets 1, 3 and 4 showed CRAM with a marginal but significant improvement in GO enrichment, stressing the importance of modeling sample-specific variances within each sample, which is a main characteristic of the CRAM method.

Estimating optimal sample weights is highly dependent on the dataset. Our results show that sample weight estimation is highly dependent on the level of homogeneity between samples and will tend to perform better in homogeneous than in heterogeneous data. The primary reason for this is that when large biological differences are present among samples, sample weights will usually be underestimated, since the CRAM model will interpret incorrectly that biological discrepancies among samples, are the result of noise (measurement error), rather than biological variability. Nevertheless,

in practice, when CRAM was applied to heterogeneous datasets and datasets using biological replicates (homogeneous datasets), it generated superior GO enrichment to all other methods in datasets 1, 3 and 4, and was at least as good as WAME in dataset 2.

Based on these findings, we hypothesize that CRAM is more accurate from a theoretical perspective, in experiments in which technical replicates are used, since biological replicates from homogeneous datasets can have high biological variability to the point that weights can be distorted just as in heterogeneous datasets. Thus, in order to optimally model biological variability in both heterogeneous and homogeneous datasets, we developed CRAM-GS, which incorporates correlations between genes into the measure of differential expression.

5.4.2 Concordance between samples – CRAM BINARY model

A robust version of CRAM (CRAM BINARY) was developed where concordance was used to estimate the weight of each sample. Samples that had a high concordance with at least some other sample in the dataset were more heavily weighted whereas samples that had low concordance with all remaining samples were down weighted. This use of concordance for weight estimation provides a robust measure of reproducibility between samples even when log-intensity values deviate largely from a Normal distribution. Preliminary results comparing CRAM BINARY to fold change, generated superior GO enrichment in most cases.

5.5 LINEAR ASSOCIATION BETWEEN GENES

Another novel model called CRAM gene correlation (CRAM-GC), was introduced in chapter 4, implementing an adjusted Spearman rank correlation between genes and improving on a recent work where gene correlations were modeled, but without any adjustment (Leek and Storey 2008). The main assumption is based on the fact that genes more highly correlated with other genes are likely to be enriched, compared to uncorrelated genes. One of the main reasons for improvements in GO enrichment in CRAM-GC derives from the adjustment of rank correlations over all genes in order to correct for optimistic correlations.

CRAM-GC when compared to fold change showed significant improvement in GO enrichment in both homogeneous and heterogeneous datasets, in one and two-channel microarrays, and in paired and non-paired designs. In CRAM-GC, rank correlation between genes was used as an extension of the concept of reproducibility, with the main advantage that any two genes do not need to be very similar in gene expression (log-intensity) in order to have a high association, and only need to have a high rank correlation. Gene weights and sample weights were combined in the CRAM-GC model, such that if a gene was not highly correlated with any other gene, the weight for that gene would be close to zero. Thus, the expected fold change for that gene would be close to zero, resulting in the gene not being differentially expressed. This down weighting of genes that had low correlation with all remaining genes is hypothesized as the main reason for the major improvement in GO enrichment. We further hypothesize, that this improvement in GO enrichment, results from the fact that if a gene belongs to a particular GO category, then it is very likely to be highly correlated with at least another

gene in the same GO category. This makes sense from a biological perspective, since a gene in the same GO category is very likely to belong to the same pathway of at least another gene in the same GO category, and thus is expected to have some level of statistical dependence, resulting in both genes being significantly correlated.

5.6 CRAM: FUTURE WORK

CRAM gene correlation (CRAM-GC) is the most powerful version of all the CRAM models, due to the high level of biological information obtained from incorporating gene correlations into the estimation of measures of differential expression. One potential improvement of CRAM-GC would be to group genes by GO category and perform gene-correlations restricted to genes within each GO category, generating a measure of differential expression for every gene and for every GO category. This potential improvement of CRAM-GC would be feasible as long as GO categories have a reasonable size (≥ 20 genes). This type of restriction by GO category, would generate much smaller gene correlation matrices, and would potentially increase biological accuracy by reducing the number the false positive genes resulted from high correlations between genes, which sometimes appear even after the adjustments based on Fisher's transformation.

The flexibility of CRAM models makes them highly applicable to some of the latest technologies in transcriptomics such as “Whole Transcriptome Shotgun Sequencing”, which aims at measuring RNA content (Morin, Bainbridge et al. 2008). In

microarrays, CRAM models can also be easily adapted to one-channel datasets in which multiple probes are used for every gene. Moreover, it can be also applied in flow cytometry datasets such as SDI measures. Finally, CRAM can use prior knowledge for every gene, generated as outputs from similar experiments, together with the data from the current experiment in order to improve GO enrichment results.

5.7 DETECTION OF HETEROGENEOUS CELL POPULATIONS – SLOPE DIFFERENTIAL IDENTIFICATION (SDI)

In order to identify heterogeneous cell populations, we have described in chapter 3 the method Slope Differentiation Identification (SDI), which was shown to have superior performance in GO enrichment compared to other methods. SDI also exhibits a high overlap with our gold standard gene list (gene lists obtained from stationary two peak distributions). Validation via microscopy also indicated that SDI was superior in differentiating between quiescent (Q) and non-quiescent (NQ) cell populations in stationary phase cultures. Moreover, SDI was shown to be highly computationally efficient taking only a few seconds to score the whole Green fluorescence protein (GFP) dataset.

5.7.1 Multi-dimensional fold change

SDI is an extension of the concept of fold change in log-fluorescence intensity between two conditions, in a sense that it is applied to multiple dimensions. In the flow experiment described in chapter 3, SDI had two additional dimensions besides

fluorescence intensity: side-scatter and forward-scatter. These two additional dimensions were used as control variables, where averages of fold change of log-fluorescence intensity were evaluated for different levels of side-scatter and forward-scatter.

Although recent studies have used clustering methods to model multi-dimensional flow datasets, these are not optimized to model differences between two conditions in multiple dimensions (Zeng, Pratt et al. 2007; Pedreira, Costa et al. 2008). The modeling of the differences in log-fluorescence intensity between stationary and exponential phase cultures, controlled either by side-scatter or forward-scatter, was shown to add discriminatory power in accurately detecting heterogeneous cell populations. One of the advantages in modeling the differences in log-fluorescence intensity between two conditions rather than modeling each condition separately is that differences in log-intensity provide a highly simplified model for every GFP-fusion strain. Another reason in modeling differences in fluorescence intensity comes from the strong association between conditions, for a given GFP-strain, and therefore noise levels of the predicted values from the model, are largely reduced. This same procedure can be extended to more dimensions, in which case, each dimension would be used as a control variable whereas the differences in log-fluorescence intensity would be the dependent (or explanatory) variable.

5.7.2 Modeling SDI using multi-dimensional flow data

Defining the correct binning of a control variable is a crucial step for modeling multi-dimensional differences in log-fluorescence intensity. Although the fluorescence

intensity is measured for every cell (event), differences in log-fluorescence intensity can only be measured for a group of events. In this case, groups are defined based on different ranges of a control variable (side-scatter or forward-scatter in our experiment). This grouping procedure may be applied simultaneously to a set of multiple control variables and their respective ranges, where the *average* of the difference in log-fluorescence intensity between two conditions is estimated for each group.

An improvement of SDI to handle multi-dimensional data would be to generate a regression tree for every GFP, combining the most important control variables in which branches and nodes would be selected based on the most significant differences in average log-fluorescence intensity between two conditions. The output of such trees for every GFP would produce nodes with various levels in differences in log-intensity (fold change), where large variations in fold change between nodes, such as high average fold change for some nodes and low average fold change for other nodes, which would be indicative of the presence of heterogeneous cell populations.

5.7.3 Modeling SDI with multiple conditions

In chapter 3 it was shown that the pairing between stationary and exponential phase was necessary in order to calculate the differences in log-fluorescence intensity, controlled by either side-scatter or forward-scatter. This pairing requirement is essential when testing for multiple conditions, where fold change between each pair of conditions is modeled. This approach may be applied to time course experiment in which every two consecutive time points are treated as two different conditions. Moreover, we can extend

the concept of differences between two time points using experimental design contrasts in which three or more time points are used. This would allow higher order differences to be calculated, such as acceleration in fold change.

5.7.4 Limitations of concordance between samples

Although concordance between samples in a dataset is a strong indicator of confidence in data reproducibility, it is not always an indication of optimal enrichment. For example, in the flow dataset, average fold change had an average concordance between samples equal to 73% vs. 63% for SDI, whereas the enrichment for SDI was clearly superior to average fold change. As an example, the overlap between AFC list and the gold standard stationary SPV list was only 31% (46/147), whereas the overlap between SDI list and SPV list was 65% (96/147). Recent studies suggest that in order for biological knowledge to be inferred from the data it is necessary that we have reproducible samples (Shi, Perkins et al. 2008). However, as our examples show, having optimal reproducibility between samples does not imply that the approach is optimal to detect the main biological relationships we are interested in, which was also the case for SDI. Therefore, the most accurate statistical approaches are likely to have both: highly reproducible samples and significant GO enrichment.

With the large increase in the amounts of data generated from recent technologies such as flow cytometry, we expect that future improvements in SDI will be suited for

disease studies such as Cancer, and to better understand important biological relations, which result from identifying heterogeneous cell populations. In addition, we expect that further development of the CRAM methodology, will combine datasets from similar experiments to generate more biological accuracy. Finally, given the increase in number of dimensions in flow data (approximately 30 dimensions in the next few years) and further improvement in gene annotations to GO categories, we believe there is a huge potential for developing methods that optimally model biological relations.

5.8 REFERENCES

- Affymetrix "<http://www.affymetrix.com>."
- Allen, C., S. Buttner, et al. (2006). "Isolation of quiescent and nonquiescent cells from yeast stationary-phase cultures." Journal of Cell Biology **174**(1): 89-100.
- Aragon, A. D., G. A. Quinones, et al. (2006). "Release of extraction-resistant mRNA in stationary phase *Saccharomyces cerevisiae* produces a massive increase in transcript abundance in response to stress." Genome Biol **7**(2): R9.
- Aragon, A. D., A. L. Rodriguez, et al. (2008). "Characterization of differentiated quiescent and nonquiescent cells in yeast stationary-phase cultures." Molecular Biology of the Cell **19**(3): 1271-1280.
- Ashrafi, K., D. Sinclair, et al. (1999). "Passage through stationary phase advances replicative aging in *Saccharomyces cerevisiae*." Proceedings of the National Academy of Sciences of the United States of America **96**(16): 9100-9105.
- Astrand, M., P. Mostad, et al. (2007). "Improved covariance matrix estimators for weighted analysis of microarray data." Journal of Computational Biology **14**(10): 1353-1367.
- Astrand, M., P. Mostad, et al. (2008). "Empirical Bayes models for multiple probe type microarrays at the probe level." Bmc Bioinformatics **9**: -.
- Baldi, P. and A. D. Long (2001). "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes." Bioinformatics **17**(6): 509-519.
- Bar-Even, A., J. Paulsson, et al. (2006). "Noise in protein expression scales with natural protein abundance." Nature Genetics **38**(6): 636-643.
- Benjamini, Y. a. H., Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing
" Journal of the Royal Statistical Society Series B **57** (1): 289--300.
- Bevers, M. M. and F. Izadyar (2002). "Role of growth hormone and growth hormone receptor in oocyte maturation." Molecular and Cellular Endocrinology **197**(1-2): 173-178.
- Bitterman, K. J., O. Medvedik, et al. (2003). "Longevity regulation in *Saccharomyces cerevisiae*: Linking metabolism, genome stability, and heterochromatin." Microbiology and Molecular Biology Reviews **67**(3): 376-+.
- Boyle, E., Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G (2004). "GO Term Finder -- open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes." Bioinformatics **20**(18): 3710-5.
- Brachmann, C. B., A. Davies, et al. (1998). "Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications." Yeast **14**(2): 115-132.
- Broberg, P. (2003). "Statistical methods for ranking differentially expressed genes." Genome Biology **4**(6): -.
- Chang, E., J. W. Yang, et al. (2002). "Aging and survival of cutaneous microvasculature." Journal of Investigative Dermatology **118**(5): 752-758.
- Eaves, I. A., L. S. Wicker, et al. (2002). "Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: The NOD model of type 1 diabetes." Genome Research **12**(2): 232-243.

- Edwards, B. S., T. Oprea, et al. (2004). "Flow cytometry for high-throughput, high-content screening." Current Opinion in Chemical Biology **8**(4): 392-398.
- Efron B, T. R., Storey JD, Tusher V (2001). "Empirical Bayes ANalysis of a Microarray Experiment." J. Amer. Statist. Assoc. **96**: 1151-1160.
- Ein-Dor, L., O. Zuk, et al. (2006). "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer." Proceedings of the National Academy of Sciences of the United States of America **103**(15): 5923-5928.
- Eisen, M. B. and P. O. Brown (1999). "DNA arrays for analysis of gene expression." Cdna Preparation and Characterization **303**: 179-205.
- Fabrizio, P. and V. D. Longo (2003). "The chronological life span of *Saccharomyces cerevisiae*." Aging Cell **2**(2): 73-81.
- Fodor, A. A., T. L. Tickle, et al. (2007). "Towards the uniform distribution of null p-values on Affymetrix microarrays." Genome Biology **8**(5): -.
- Fuge, E. K., E. L. Braun, et al. (1994). "Protein synthesis in long-term stationary phase cultures of *Saccharomyces cerevisiae*." Journal of Bacteriology (176): 5802-5813
- Fuller, W. A. (1987). "Measurement Error Models." Wiley, New York.
- Gasch, A. P. (2002). "Yeast genomic expression studies using DNA microarrays." Guide to Yeast Genetics and Molecular and Cell Biology, Pt B **350**: 393-414.
- Gasch, A. P., P. T. Spellman, et al. (2000). "Genomic expression programs in the response of yeast cells to environmental changes." Molecular Biology of the Cell **11**(12): 4241-4257.
- GeneOntology "Gene Ontology <<http://www.geneontology.org/>>."
- Ghaemmaghami, S., W. Huh, et al. (2003). "Global analysis of protein expression in yeast." Nature **425**(6959): 737-741.
- GOTermFinder "GO Term Finder<<http://go.princeton.edu/index.html>>."
- Gray, J. V., G. A. Petsko, et al. (2004). ""Sleeping beauty": Quiescence in *Saccharomyces cerevisiae*." Microbiology and Molecular Biology Reviews **68**(2): 187-+.
- Herbert R. (1956). "An empirical Bayes approach to statistics." Proceesings of the thrid Berkely Symposium on Mathematical Statistics and Probability **1**: 157-163.
- Howson, R., W. K. Huh, et al. (2005). "Construction, verification and experimental use of two epitope-tagged collections budding yeast strains." Comparative and Functional Genomics **6**(1-2): 2-16.
- HTC "<http://www.cytologyproject.info/>."
- <http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.pl>.
- Hu, Z. Z., P. J. Killion, et al. (2007). "Genetic reconstruction of a functional transcriptional regulatory network." Nature Genetics **39**(5): 683-687.
- Huttenhofer, A., P. Schattner, et al. (2005). "Non-coding RNAs: hope or hype?" Trends Genet **21**(5): 289-97.
- Ibrahim, J. G. and M. H. Chen (2000). "Power prior distributions for regression models." Statistical Science **15**(1): 46-60.
- IDLQuery "<http://www.nrl.navy.mil/tira/Projects/HIRAAS4/downloads.htm>."
- Ivanova, N. B., J. T. Dimos, et al. (2002). "A stem cell molecular signature." Science **298**(5593): 601-4.
- Kaeberlein, M., C. R. Burtner, et al. (2007). "Recent developments in yeast aging." Plos Genetics **3**(5): 655-660.
- Kerov, S. V. (2005). "Multidimensional Hypergeometric Distribution and Characters of the Unitary Group." Journal of Mathematical Sciences **129**(2): 3697-3729.

- Kim, B. S., S. Y. Rha, et al. (2004). "Spearman's footrule as a measure of cDNA microarray reproducibility." Genomics **84**(2): 441-448.
- Kristiansson, E., A. Sjogren, et al. (2005). "Weighted analysis of paired microarray experiments." Statistical Applications in Genetics and Molecular Biology **4**: -.
- Kristiansson, E., A. Sjogren, et al. (2006). "Quality optimised analysis of general paired microarray experiments." Statistical Applications in Genetics and Molecular Biology **5**: -.
- Krutzik, P. O., J. M. Irish, et al. (2004). "Analysis of protein phosphorylation and cellular signaling events by flow cytometry: techniques and clinical applications." Clinical Immunology **110**(3): 206-221.
- Kulesh, D. A., D. R. Clive, et al. (1987). "Identification of Interferon-Modulated Proliferation-Related Cdna Sequences." Proceedings of the National Academy of Sciences of the United States of America **84**(23): 8453-8457.
- Kurtzman, C. P., Fell, J.W. (2006). "Yeast Systematics and Phylogeny-Implications of Molecular Identification Methods for Studies in Ecology." Biodiversity and Ecophysiology of Yeasts, The Yeast Handbook, Springer.
- Lee, M. L. T., F. C. Kuo, et al. (2000). "Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations." Proceedings of the National Academy of Sciences of the United States of America **97**(18): 9834-9839.
- Lee, P. M. (1997). "Bayesian Statistics." Arnold Publishing, New York(2nd Edition): 17.
- Leek, J. T. and J. D. Storey (2008). "A general framework for multiple testing dependence." Proceedings of the National Academy of Sciences of the United States of America **105**(48): 18718-18723.
- Lewis, D. L. and D. K. Gattie (1991). "Predicting Chemical Concentration Effects on Transformation Rates of Dissolved Organics by Complex Microbial Assemblages." Ecological Modelling **55**(1-2): 27-46.
- Lillie, S. H. and J. R. Pringle (1980). "Reserve Carbohydrate-Metabolism in Saccharomyces-Cerevisiae - Responses to Nutrient Limitation." Journal of Bacteriology **143**(3): 1384-1394.
- LIMMA "<<http://www.bioconductor.org/packages/release/bioc/html/limma.html>>."
- Lonnstedt, I. and T. Speed (2002). "Replicated microarray data." Statistica Sinica **12**(1): 31-46.
- Maher, C. A., C. Kumar-Sinha, et al. (2009). "Transcriptome sequencing to detect gene fusions in cancer." Nature **458**(7234): 97-U9.
- Marks, P. A., R. A. Rifkind, et al. (2001). "Histone deacetylases and cancer: Causes and therapies." Nature Reviews Cancer **1**(3): 194-202.
- Martinez, M. J., S. Roy, et al. (2004). "Genomic analysis of stationary-phase and exit in Saccharomyces cerevisiae: Gene expression and identification of novel essential genes." Molecular Biology of the Cell **15**(12): 5295-5305.
- McMurray, M. A. and D. E. Gottschling (2004). "Aging and genetic instability in yeast." Current Opinion in Microbiology **7**(6): 673-679.
- Morin, R. D., M. Bainbridge, et al. (2008). "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing." Biotechniques **45**(1): 81-+.
- Morrison, R. S., Y. Kinoshita, et al. (2002). "Neuronal survival and cell death signaling pathways." Molecular and Cellular Biology of Neuroprotection in the Cns **513**: 41-86.
- Ontology, G. <http://www.geneontology.org/>.

- Ostergaard, S., L. Olsson, et al. (2000). "Metabolic engineering of *Saccharomyces cerevisiae*." Microbiology and Molecular Biology Reviews **64**(1): 34-+.
- Parrish, N. M., J. D. Dick, et al. (1998). "Mechanisms of latency in *Mycobacterium tuberculosis*." Trends in Microbiology **6**(3): 107-112.
- Paz, I., J. R. Meunier, et al. (1999). "Monitoring dynamics of gene expression in yeast during stationary phase." Gene **236**(1): 33-42.
- Pedreira, C. E., E. S. Costa, et al. (2008). "A multidimensional classification approach for the automated analysis of flow cytometry data." Ieee Transactions on Biomedical Engineering **55**(3): 1155-1162.
- Piper, P. W. (2006). "Long-lived yeast as a model for ageing research." Yeast **23**(3): 215-226.
- PLW-Astrand (2008). "PLW: An R implementation of Probe level Locally moderated Weighted median-t (PLW) and Locally Moderated Weighted-t (LMW)."
- Qin, L. X., K. F. Kerr, et al. (2004). "Empirical evaluation of data transformations and ranking statistics for microarray analysis (vol 32, pg 5471, 2004)." Nucleic Acids Research **32**(19): -.
- Radonjic, M., J. C. Andrau, et al. (2005). "Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S-cerevisiae* stationary phase exit." Molecular Cell **18**(2): 171-183.
- Ramalho-Santos, M., S. Yoon, et al. (2002). "'Stemness': transcriptional profiling of embryonic and adult stem cells." Science **298**(5593): 597-600.
- Raser, J. M. and E. K. O'Shea (2004). "Control of stochasticity in eukaryotic gene expression." Science **304**(5678): 1811-4.
- Raser, J. M. and E. K. O'Shea (2005). "Noise in gene expression: Origins, consequences, and control." Science **309**(5743): 2010-2013.
- Reid, R. W. and A. A. Fodor (2008). "Determining gene expression on a single pair of microarrays." Bmc Bioinformatics **9**: -.
- Riley, R. S. (2002). "Preface - Flow cytometry and its applications in hematology and oncology." Hematology-Oncology Clinics of North America **16**(2): Xi-Xii.
- Ritchie, M. E., D. Diyagama, et al. (2006). "Empirical array quality weights in the analysis of microarray data." Bmc Bioinformatics **7**: -.
- Rose, M. D., F. Winston, and P. Hieter. (1990). "Methods in Yeast Genetics: A Laboratory Course Manual." Cold Spring Harbor Laboratory Press, Cold.
- Sartor, M. A., C. R. Tomlinson, et al. (2006). "Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments." Bmc Bioinformatics **7**: -.
- Schneeweis B, H. a. A., Thomas (2005). "Some recent advances in measurement error models and methods." Collaborative Research Center 386, Discussion Paper 452.
- Sebastiani, P., E. Gussoni, et al. (2003). "Statistical challenges in functional genomics." Statistical Science **18**(1): 33-60.
- Shi, L., R. G. Perkins, et al. (2008). "Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential." Curr Opin Biotechnol **19**(1): 10-8.
- Shi, L., L. H. Reid, et al. (2006). "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." Nat Biotechnol **24**(9): 1151-61.

- Shi, L., W. Tong, et al. (2005). "Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential." Bmc Bioinformatics **6 Suppl 2**: S12.
- Simon, J. A., P. Szankasi, et al. (2000). "Differential toxicities of anticancer agents among DNA repair and checkpoint mutants of *Saccharomyces cerevisiae*." Cancer Research **60**(2): 328-333.
- Sjogren A, K. E., Rudemo M, Nerman O (2007). "Weighted analysis of general microarray experiments." BMC Bioinformatics **8**(387).
- Sjogren, A., E. Kristiansson, et al. (2007). "Weighted analysis of general microarray experiments." Bmc Bioinformatics **8**: -.
- Smith, G. (2004). "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments." Stat. Appl. Genet. Mol. Biol. **2004** **3**:article **3**.
- Spellman, P. T., G. Sherlock, et al. (1998). "Identification of cell cycle regulated genes in yeast by DNA microarray hybridization." Molecular Biology of the Cell **9**: 371a-371a.
- Spiegelman, D., A. McDermott, et al. (1997). "Regression calibration method for correcting measurement-error bias in nutritional epidemiology." American Journal of Clinical Nutrition **65**(4): S1179-S1186.
- Suda, T., F. Arai, et al. (2005). "Hematopoietic stem cells and their niche." Trends in Immunology **26**(8): 426-433.
- Tan, P. K., T. J. Downey, et al. (2003). "Evaluation of gene expression measurements from commercial microarray platforms." Nucleic Acids Res **31**(19): 5676-84.
- TermFinder, G. <http://yeastgenome.org/cgi-bin/GO/goTermFinder.pl>.
- Tu, B. P., A. Kudlicki, et al. (2005). "Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes." Science **310**(5751): 1152-1158.
- Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response (vol 98, pg 5116, 2001)." Proceedings of the National Academy of Sciences of the United States of America **98**(18): 10515-10515.
- Vulic, M. and R. Kolter (2001). "Evolutionary cheating in *Escherichia coli* stationary phase cultures." Genetics **158**(2): 519-526.
- Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics **10**(1): 57-63.
- Watson, J. V. (1987). "Flow-Cytometry in Biomedical Science." Nature **325**(6106): 741-742.
- Wentzell, P. D., T. K. Karakach, et al. (2006). "Multivariate curve resolution of time course microarray data." Bmc Bioinformatics **7**: -.
- Werner-Washburne, M. (1993). "**Stationary Phase in the Yeast *Saccharomyces cerevisiae*.**" Microbiology Reviews: 383-401.
- Werner-Washburne, M., B. Wylie, et al. (2002). "Comparative analysis of multiple genome-scale data sets." Genome Research **12**(10): 1564-1573.
- www.geneontology.org.
- Yi, M., U. Mudunuri, et al. (2009). "Seeking unique and common biological themes in multiple gene lists or datasets: pathway pattern extraction pipeline for pathway-level comparative analysis." Bmc Bioinformatics **10**: -.
- Young, S. M., C. Bologna, et al. (2005). "High-throughput screening with HyperCyt (R) flow cytometry to detect small molecule formylpeptide receptor ligands." Journal of Biomolecular Screening **10**(4): 374-382.
- Zeng, Q. T., J. P. Pratt, et al. (2007). "Feature-guided clustering of multi-dimensional flow cytometry datasets." Journal of Biomedical Informatics **40**(3): 325-331.