Using non-parametric methods in the context of multiple testing to determine differentially expressed genes

Grant G.R.*, Manduchi E., Stoeckert C.J., Jr.
Center for Bioinformatics, University of Pennsylvania (PCBI)

Abstract: Our focus is on the Golub *et al.* ALL/AML oligo-nucleotide array data set [Golub *et al.*, (1999)] with regard to the question of determining differentially expressed genes between pairs of sample types. We use this data set to analyze methods of determining genes which are likely to be differentially expressed between ALL T-cells and ALL B-cells. To this end, we employ non-parametric methods, in the context of multiple testing, for attaching statistical measures of confidence to genes predicted to be differentially expressed. In particular, we apply the method of using t-statistics, with p-values calculated through permutations, and with the Westfall and Young step-down approach to correct for multiple testing, developed by Dudoit *et al.* [Dudoit *et al.*, (2000)]. We also use PaGE [Manduchi *et al.*, (2000)], developed at PCBI, for assigning confidence to predictions by calculating false-positive rates directly from empirical "gene-independent" distributions. We compare the performance of these methods on the Golub *et al.* data. We exploit the large sample size to analyze the effect of the number of observations on a variety of issues relating to the prediction of differential expression, in particular to the reproducibility of results. In addition, we investigate the concept of using shifted intensities for data such as the Golub *et al.* data set. We also investigate the usage of "absent calls" in oligo-nucleotide array data.

INTRODUCTION

In Golub *et al.* [Golub *et al.*, (1999)], Affymetrix data [Lockhart *et al.*, (1996)] have been generated for leukemic myeloid and lymphoblastic cells. The data contain a relatively large number of observations of the expression levels of several cell types. In particular the data contain many biological replicates; that is, the observations were taken from different cells and different individuals. The data also represent a heterogeneous collection of cell types. As such, this data set presents an opportunity to study the effect of the sample size on the results obtained by the current methods which are aimed at giving the least conservative measures of significance for predictions of differential expression. We use the term "prediction" here because array data give only indications of which genes are likely to be differentially expressed. These predictions may then be corroborated by other experimental approaches.

Much of the classical theory that might be used to tackle this problem depends on assumptions about the forms of the distributions of the

*Contact: ggrant@pcbi.upenn.edu

gene intensities. In particular, standard techniques such as the two-sample t-test, depend on these distributions being normal (see [Mood *et al.*, (1963)] for example). The Golub *et al.* data give us many observations, each observation being one array experiment. Examining the gene intensity distributions over the observations shows that many of them are not normal. Some have apparently uniform distributions, some unimodal, others bimodal, or even trimodal. Moreover, the inclusion of absent calls introduces further irregularity into the distributions. Such heterogeneity precludes the possibility of transforming the distributions into standard distributions and the use of non-parametric methods becomes necessary. By non-parametric methods we mean those that make minimal assumptions about the form of the distributions involved. In highly parallel gene expression data, the challenge of having irregular distributions is compounded by issues of multiple testing. Non-parametric approaches for the problem of assigning confidence measures to the prediction of differential expression, in the context of multiple testing, have been proposed using standard and non-standard methods ([Dudoit *et al.*, (2000)], [Manduchi *et al.*, (2000)]). These methods take as input multiple observations for each of two (or more) sample types. We use the Golub *et al.* data as a pool and sample from it, each time applying these algorithms and tabulating the results. This gives information about how reproducibility depends on the sample size.

We also address a complementary issue. One might hope be able to maintain a low false-positive rate when using only a few observations. However, it is key, regardless of which method is used to make predictions, that the data represent sufficiently the variability of the sample types. Variability is generally of two types: experimental, and biological. With a general definition of sample type, such as "Acute Lymphoblastic Leukemia" (ALL) or "Acute Myeloid Leukemia" (AML), biological variability will be the overwhelming factor between the two. AML cells, for example, can be divided into many subcategories, such as M1, M2, etc., male/female, peripheral blood/bone marrow, etc. ALL cells can be either T-cells or B-cells, and each of these divides into respective subcategories. It is also expected, since they are cancerous cells, that they will fall into unknown subcategories of tumor types, perhaps related to the success/failure response to treatment. If the sample type is represented by a few experiments only, then the probability of them falling into one or more subcategories becomes almost a certainty. The Golub *et al.* data give us a chance to see the effect of this clinical heterogeneity on predictions of differential expression, as a function of sample size. Instead of focusing on the ALL and AML classification, we controlled for the

heterogeneity of the sample type as much as possible by focusing on the B-cell and the T-cell experiments within the ALL group. This gives us two reasonably homogeneous sample types, for which we still have many observations.

TYPES OF DIFFERENTIAL EXPRESSION

Generally for two sample types A and B of interest, most genes will not be differentially expressed between them (see [Sagerstrom *et al.*, (1997)]). No gene however is expressed at exactly the same level in a given sample type every time it is measured. Instead each gene will have a distribution of intensities, with respect to observations of the sample type. A sample type can be general, such as "hematopoietic cells," or specific, such as "B-cell lymphoblasts at a particular stage of development, under particular conditions." The distributions will depend heavily on this definition of sample type.

Genes will be differentially expressed in a non-deterministic manner, meaning that the two distributions overlap to some degree, though they still have different means. A deterministically differentially expressed gene, if it existed, would never be incorrectly predicted, regardless of the sample size for each sample type. Figure 1 shows an example of a gene that appears to be nearly deterministically differentially expressed between ALL B-cells and ALL T-cells. This is the exception, however. Between ALL B-cells and ALL T-cells, almost all genes which are differentially expressed appear to be non-deterministic. Figure 2 shows such a gene.

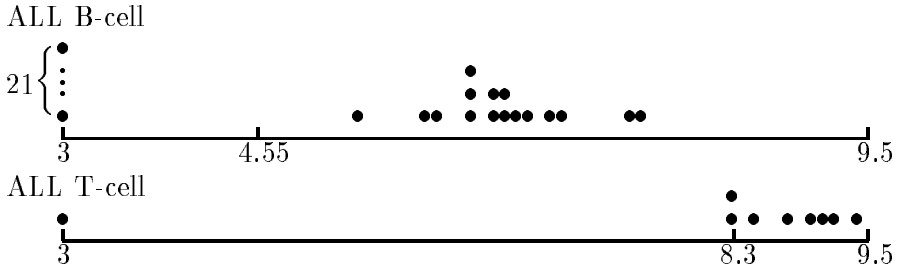

Figure 1: GenBank Accession: U23852, T-lymphocyte specific protein tyrosine kinase. Data points are taken from the Golub *et al.* data set. The top plot is of the expression in the 37 ALL B-cells used, 21 of which were absent calls, indicated on the left. The lower plot is of the expression of the same tag in the 9 ALL T-cells, with one absent call. Graphs are in *ln* scale.

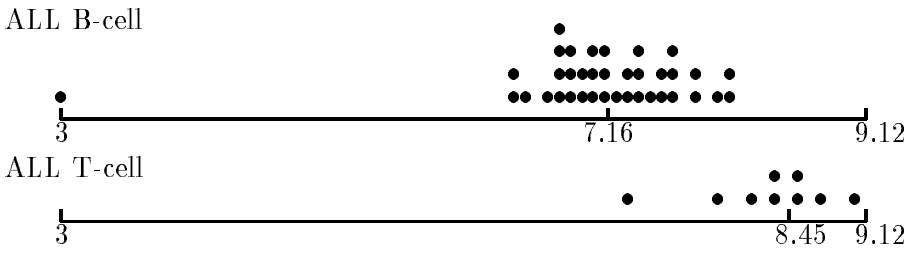

Figure 2: GenBank Accession: M23323 T-CELL SURFACE GLYCOPROTEIN CD3 EPSILON CHAIN PRECURSOR. Data points are taken from the Golub *et al.* data set. Notation is as in Figure 1. Gene displays non-deterministic differential expression.

ABSENT CALLS

When performing statistical analyses of oligo-nucleotide data, one must decide how to handle the absent calls. Absent calls present a complication, in that if we set them to a minimal value, then they tend to create highly bimodal distributions, such as in Figure 3, where the "29" indicates that there are 29 absent calls, which were set to the minimal intensity (3 on this graph, which is in ln scale). The decision to include absent calls was based on a comparison of expected results when we both do and do not include them. In the T-CELL ANTIGEN CD7 PRECURSOR ID:M37271 (see Figure 3), there is negligible difference in the means of the intensities of the present calls between the B-cells and the T-cells. When absent calls are included, however, the means separate clearly (4.04 versus 8.03).

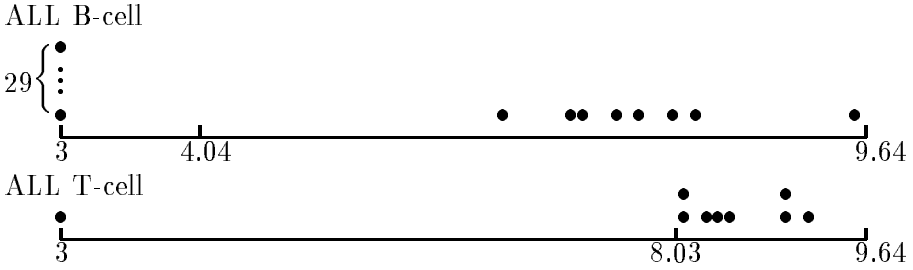

Figure 3: GenBank Accession: D00749, T-CELL ANTIGEN CD7 PRECURSOR. Data points are taken from the Golub *et al.* data set. Notation is as in Figure 1. The twenty-nine intensities at level 3 are absent calls, (set to the minimum intensity $\ln(20) \approx 3$). The mean of the ALL B-cell distribution with the absent calls included is 4.04. The distribution for the same gene, over the nine ALL T-cell experiments, has one absent call and mean 8.03.

Among the B-cell experiments, there are roughly 80% absent calls for

this gene, whereas among the T-cell experiments there are roughly 10%. It is only the absent calls that are differentiating this gene between B- and T-cells. This phenomena was consistent across many T-cell specific genes in the data set. Based on this, absent calls were included for all genes that had at least one non-absent call intensity, by setting them to the minimum value of 20. Genes that had absent calls in all experiments, for both cell types, were eliminated from further analysis, leaving approximately 5000 genes.

A consequence of the inclusion of absent calls is that it dramatically increases the non-deterministic nature of the differential expression, as a high percentage of genes were absent in one or more experiments. As a result, some genes which gain inordinately large variance may not be detected after the inclusion of absent calls. We compared predictions with and without absent calls. With PaGE [Manduchi *et al.*, (2000)], one T-cell specific gene (see Figure 4) was picked up when absent calls were not included that was not picked up when they were included. In this case, the absent calls did not sufficiently separate the means to overcome the large variance present in the expression of this gene. On the other hand, five known differentially expressed genes were detected when absent calls were included, that were not seen without absent calls, including the one in Figure 3.

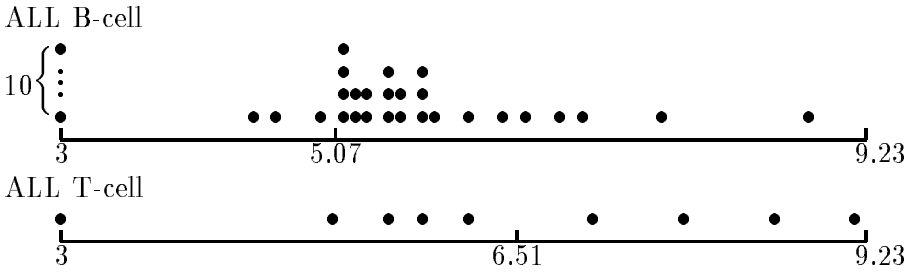

Figure 4: GenBank Accession: M30894, TCRG T cell receptor gamma chain. Data points are from the Golub *et al.* data set. Notation is as in Figure 1.

Some T-cell genes were not detectable by any methods we used. Figure 5 shows such a gene. In this case the differential expression is too subtle to detect, given that approximately 5000 simultaneous tests are performed. If this were the only gene under investigation, then the available number of observations might be sufficient to detect it, however, when looking at thousands of genes simultaneously, this kind of separation is likely to occur just by chance for many genes that are not differentially expressed.

The less specific is the definition of the sample type, the more biologi-

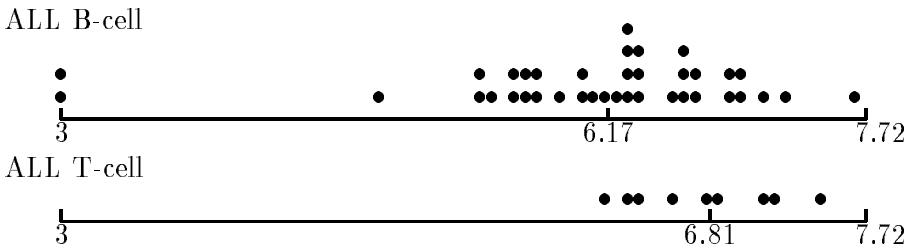

Figure 5: GenBank Accession: X94232, Novel T-cell activation protein. Data points are taken from the Golub *et al.* data set. Notation is as in Figure 1. Differential expression of this gene is too subtle to detect in the context of the thousands of genes on the array.

cal variability we expect to find for more genes. Correspondingly, the less specific is the definition of the sample type, the greater percentage of the differentially expressed genes will have highly overlapping distributions. These issues should be considered when designing array experiments to search for differentially expressed genes. The more non-deterministic is the gene expression, the more observations will be necessary to make accurate predictions.

There are two general approaches to controlling the reliability of the predictions. The classical approach is to control the *experiment-wise* Type I error. In this case the probability of making any false predictions at all is controlled (see [Ewens *et al.*, (2001)] Section 12.4).

An alternative approach is to control the *false-discovery rate*. In this case if predictions of upregulation are being made for type A versus type B, and if M predictions are made with V of them being incorrect, then the false-discovery rate is the expected value of V/M, denoted by $E(V/M)$.

Successfully controlling the experiment-wise Type I error instead of the false-discovery rate gives a set of high confidence results, but at the possible cost of a higher false-negative rate. For example it might be preferable to tolerate 5-10% of false positives among the predictions if it means finding many more true positives.

The two methods we investigate below take these two different approaches. We start with PaGE which controls the false-discovery rate.

THE PaGE APPROACH

PaGE (Patterns from Gene Expression [Manduchi *et al.*, (2000)] is part of a larger software package developed at PCBI for analyzing gene expression data. One of its features is that it reports estimated confidence

measures on predictions of differential expression. The confidence measure reported is one minus the false-discovery rate. PaGE essentially controls for the false-positive rate and then basically translates that into the false-discovery rate using Bayes theorem.

Given any criteria for predicting differential expression, the false-positive rate of this criteria is the percentage of the set of non-differentially expressed genes consisting of genes which satisfies the criteria. When trying to find a small set of differentially expressed genes from a large pool, a seemingly reasonable false-positive rate, say .05, can lead to a set of predictions with a high proportion of false positives among them. For example, if 50 out of 1000 genes are differentially expressed, a false-positive rate of .05 will give $950(.05) = 47.5$ false positives, on average. There can be at most 50 true positives, so if there are 47.5 false positives there will be at best $50/(50 + 47.5)\% = 51.28\%$ confidence in any given prediction, and that will be achieved only if the false-negative rate is zero.

For each gene g, let $X_{g,A}$ and $X_{g,B}$ be the random variables giving the intensities of gene g when measured in a sample of types A and B, respectively. Let $\overline{X}_{g,A}$ (respectively $\overline{X}_{g,B}$) be the average of n_A (respectively n_B) random variables each having the distribution of $X_{g,A}$ (respectively $X_{g,B}$). For a given gene g, we say that it is up-regulated at group B as compared to group A if

$$\frac{\mu_{g,B}}{\mu_{g,A}} > 1,$$

where $\mu_{g,A}$ (respectively $\mu_{g,B}$) denotes the (unknown) true mean of $X_{g,A}$ (respectively $X_{g,B}$). A brief summary of the PaGE approach follows.

Let $s\%$ be the desired false-positive rate. We want to determine C (the "upper cut-ratio") such that, if we say that a gene g is up-regulated at group B as compared to group A when

$$\frac{\overline{x}_{g,B}}{\overline{x}_{g,A}} > C,$$

then the false-positive rate is expected to be no greater than $s\%$, where $\overline{x}_{g,A}$ (respectively $\overline{x}_{g,B}$) are the observed values of $\overline{X}_{g,A}$ (respectively $\overline{X}_{g,B}$).

The false-positive rate is the probability that, if a gene is chosen at random from the set of genes which are true negatives, then the ratio of its sample means between the two groups, $\frac{\overline{X}_{g,B}}{\overline{X}_{g,A}}$, is greater than C. Since the set of genes which are true negatives is the set of genes for which

$\mu_{g,B}/\mu_{g,A} \leq 1$, this probability is bounded above by the probability that, if a gene is chosen at random from the set of genes which are true negatives, then

$$\frac{\frac{\overline{X}_{g,B}}{\mu_{g,B}}}{\frac{\overline{X}_{g,A}}{\mu_{g,A}}} > C. \tag{1}$$

Since only a small percentage of genes are expected to be true positives, we approximate this probability by the (unconditional) probability that (1) holds.

Key ingredient: we approximate the unknown distribution of $\frac{\overline{X}_{g,A}}{\mu_{g,A}}$ by that of

$$\frac{\frac{X_{g,A}}{\overline{X}_{g,A}} - 1}{\sqrt{n_A - 1}} + 1. \tag{2}$$

Similarly for group B. In [Manduchi *et al.*, (2000)] we show that this approximation holds in the case of normally distributed gene intensities that are sufficiently bounded away from zero. Generalizations of this result to other distributions have been verified by simulation (see [Manduchi *et al.*, (2000)]). Simulations were also conducted to justify using the approximation on the Golub *et al.* data. Further empirical tests of this heuristic will appear elsewhere.

We use (2) to approximate the distribution of $\frac{\overline{X}_{g,B}/\mu_{g,B}}{\overline{X}_{g,A}/\mu_{g,A}}$, with the empirical distribution from the data. The random quantities in this expression are g, $\overline{X}_{g,A}$, and $\overline{X}_{g,B}$. This allows us to use (1) to calculate a C by numerical integration of this empirical distribution. This gives us the desired false-positive rate as a function of C.

Approximating (2) with empirical distributions from the data, C is solved for in (1) by numerical integration.

Finally, the confidence in the predictions is related to the false-positive rate via Bayes Theorem.

For down-regulation, we proceed in a similar fashion to determine c (the "lower cut-ratio") such that, if we say that a gene g is down-regulated at group B as compared to group A when

$$\frac{\overline{x}_{g,B}}{\overline{x}_{g,A}} < c,$$

then the false-positive rate is expected to be no greater than $s\%$

The theory above gives no indication of the power of the test. We are investigating this issue using a simulated data study, to appear elsewhere.

We now turn to a more classical approach, but which uses a different measure of confidence, an *experiment-wise* measure.

THE *t*-STATISTIC STEP-DOWN APPROACH

The other method employed has been developed by Dudoit *et al.* [Dudoit *et al.*, (2000)]. This method controls the experiment-wise Type I error. It uses t-statistics, with p-values calculated through permutations, and with the Westfall and Young [Westfall *et al.*, (1993)] step-down approach to correct for multiple testing. We used the one-sided version of the method, to test specifically for up-regulation in ALL T-cells versus ALL B-cells (software to implement this can be downloaded at `http://www.cbil.upenn.edu/tpWY`). In estimating the null hypothesis distribution for the t-statistic, entire experiments (columns) are permuted, so as to preserve the dependencies between the genes. This method achieves a chosen experiment-wise Type I error. This means that if the desired Type I error is set at .05, then there is a 95% chance of *no* false positives. This is in contrast to PaGE, which achieves a desired confidence in the set of predictions. So that if PaGE reports a confidence of 95%, then 95% of the predictions should be true positives.

We now apply these two methods to repeated samplings from the Golub *et al.* data.

REPLICATE STUDY RESULTS

We first used the analyses of [Dudoit *et al.*, (2000)] to find genes which are up-regulated in ALL T-cells versus ALL B-cells. We used 37 B-cell experiments[†] and the nine T-cell experiments from the Golub *et al.* data set. For various values of n we conducted the analysis 100 times, each time choosing a random sampling of n of the B-cell experiments, and testing them against all nine of the T-cell experiments. This was done so as to control for the effect of variation from all but one source, that being the number of B-cell experiments. In practice, variability will come from both sources, so that the true sample size needed to obtain reliable predictions may be even greater than our results indicate. The fraction of times a gene was predicted as up-regulated among the 100

[†]Experiment 72 was accidentally omitted by one of our parsers, so this experiment ended up being eliminated from the study.

experiments was recorded. A portion of the results is given in Table 1. The full table has 106 rows.

36	25	20	15	10	9	8	7	6	5	4	gene information
1	1	1	1	1	1	.99	.6	0	0	0	CD3G CD3G antigen, gamma polypeptide
1	1	1	1	.77	.64	.6	.71	0	0	0	T-CELL DIFFERENTIATION ANTIGEN
1	1	1	.96	.69	.61	.45	.4	.22	.29	0	PTPN7 Protein tyrosine phosphatase
1	1	.79	.73	.67	.75	.48	.62	.52	.02	0	T-CELL ANTIGEN CD7 PRECURSOR
1	1	1	.96	.78	.58	.4	.27	.38	0	0	TCF7 T-cell specific Trans. factor 7
1	.92	.79	.59	.31	.26	.14	.11	.04	.08	0	SMT3A protein
1	.98	.9	.6	.11	.08	.04	0	0	.01	.03	Chromosome 17q21 mRNA clone LF113
1	.62	.41	.3	.09	.11	.03	.02	.03	0	.06	Trophinin mRNA
.01	.35	.44	.46	.34	.34	.1	.09	0	.01	0	T-CELL SURFACE GLYCOPROTEIN CD3 ...
0	.06	.13	.24	.2	.21	.15	.22	.18	.08	0	T-CELL ANTIGEN CD7 PRECURSOR
0	.07	.17	.25	.21	.14	.13	.15	.06	0	0	M-PHASE INDUCER PHOSPHATASE 2
.32	.15	.09	.05	.01	.03	.03	0	.02	.01	.03	GB DEF = Splicing factor, ...
0	0	.02	.06	.05	.11	.03	.05	.11	.13	0	Protein tyrosine kinase related ...
0	.14	.19	.1	.02	.03	0	.01	.01	.01	.01	MACH-alpha-2 protein
0	0	.04	.14	.08	.07	.09	.08	.01	0	0	ANX1 Annexin I (lipocortin I)
0	0	.03	.03	.07	.09	.05	.03	.04	.04	0	CD47 CD47 antigen (Rh-related ...
0	0	0	.03	.08	.04	.02	.04	.07	.08	0	KIAA0050 gene
0	.04	.07	.05	.06	.07	0	.02	.01	.04	0	RNA-BINDING PROTEIN FUS/TLS
0	.06	.05	.08	.03	.03	.02	.01	0	.02	.01	DNA-BINDING PROTEIN NEFA PRECURSOR
0	0	.04	.06	.08	.06	.05	.01	0	0	0	Lactate dehydrogenase B gene ...
0	0	0	.04	.05	.07	.08	.03	0	0	0	MXS1 Membrane component, X ...
0	.05	.07	.03	.01	.03	.01	.01	0	.01	.03	Lipid-activated protein kinase
0	0	0	.02	.05	.04	.05	.09	0	0	0	HIG-1 mRNA
0	0	.01	.06	.02	.05	.03	.01	.01	.04	0	GB DEF = T-lymphocyte specific ...
0	0	0	.02	.06	.05	.03	.04	0	0	0	Signal transducer and activatoin ...
0	0	.01	.05	.02	.05	.02	.01	.02	.02	0	GB DEF = Karyopherin beta 3 mRNA
0	0	0	.03	.03	.04	.02	.02	.04	0	0	GATA3 GATA-binding protein 3
0	.01	.02	.07	.05	.01	.02	0	0	0	0	TXN Thioredoxin
0	0	0	.01	.02	.05	.04	.02	.04	0	0	TYROSINE-PROTEIN KINASE ITK/TSK
0	.03	.04	.01	.02	0	.01	0	.01	0	.04	GB DEF = Inducible nitric oxide ...
0	0	.01	.02	.03	.07	0	0	.02	.01	0	GB DEF = T-cell antigen receptor ...

Table 1: Column labeled n gives the fraction of times the gene was predicted by the methods of [Dudoit *et al.*, (2000)] as up-regulated in ALL T-cells versus ALL B-cells, out of 100 comparisons between n randomly chosen B-cells and all 9 T-cell experiments, from the Golub *et al.* data set. Intensities were preprocessed by taking natural logs. The (experiment-wise) Type I error was taken to be 0.1.

A similar table has been generated using PaGE. A portion of the results is given in Table 2. The complete table contains 81 rows.

It is apparent from these tables that there are not many genes which will be predictably differentially expressed without many observations. Notice that even the gene shown in Figure 1 is not reliably detected using the step-down approach, regardless of the sample size. This is because with this many observations the differential expression is not strong enough to have been statistically significant every time, in the context of the other approximately 5000 genes being analyzed simultaneously. This gene was, however, picked up by virtually all PaGE runs with more than two B-cell observations. The table shows how the dependability of the results drops dramatically as the sample size is reduced.

36	25	20	15	10	9	8	7	6	5	4	3	2	gene information
1	1	1	1	1	1	1	1	1	1	1	1	.88	TCRB T-cell receptor, beta cluster
1	1	1	1	1	1	1	1	1	1	1	1	.73	TCRB T-cell receptor, beta cluster
1	1	1	1	1	1	1	1	1	1	1	1	.73	GB DEF = MAL gene exon 4
1	1	1	1	1	1	1	1	1	1	1	.99	.64	GB DEF=T-lymphocyte specific protein ...
1	1	1	1	1	1	1	1	1	1	1	.95	.57	GB DEF=T-cell antigen receptor gene ...
1	1	1	1	1	1	1	1	1	1	1	.93	.43	Protein tyrosine phosphatase PTPCAAX2...
1	1	1	1	1	1	1	1	1	1	1	.92	.4	CHIT1 Chitinase 1
1	1	1	1	1	1	1	1	1	1	1	.95	.35	Na,K-ATPase gamma subunit mRNA
1	1	1	1	1	1	1	1	1	1	.97	.88	.38	NATURAL KILLER CELLS PROTEIN 4 PRECURSOR
1	1	1	1	1	.99	.96	.93	.89	.87	.95	.81	.43	T-CELL ANTIGEN CD7 PRECURSOR
1	1	1	1	1	1	1	1	.98	.98	.88	.53	.15	GB DEF = Selenoprotein W (selW) mRNA
1	1	1	1	1	1	1	1	1	1	.93	.48	.1	CD1B CD1b antigen (thymocyte antigen)
1	1	1	1	1	1	.99	.99	.95	.89	.84	.64	.18	ADA Adenosine deaminase
1	1	1	1	1	1	1	1	1	.98	.87	.38	.09	TCF7 T-cell specific Trans. factor 7
1	1	1	1	1	1	1	.99	.97	.93	.82	.48	.11	T-CELL SURFACE GLYCOPROTEIN CD3 ...
1	1	1	.99	.99	.99	.96	.96	.89	.85	.7	.52	.24	MIC2 Antigen identified by monoclonal...
1	1	1	1	.75	.72	.77	.88	.86	.87	.9	.5	.12	TCRD T-cell receptor, delta
1	1	.99	.96	.86	.83	.75	.78	.76	.75	.76	.58	.34	PSAP Sulfated glycoprotein 1
1	1	1	.99	.96	.93	.85	.76	.64	.46	.22	.07	.01	GB DEF = CD1 R2 gene for MHC-related ...
1	.86	.83	.77	.76	.77	.62	.68	.53	.51	.47	.31	.1	mRNA fragment encoding beta-tubulin. ...
1	.63	.58	.64	.7	.72	.77	.88	.81	.66	.32	.14	.01	T-CELL ANTIGEN CD7 PRECURSOR
1	.71	.7	.72	.59	.58	.51	.53	.55	.48	.39	.26	.14	Macrophage migration inhibitory fact ...
.98	.59	.61	.53	.55	.59	.59	.58	.51	.51	.4	.32	.07	SM22-ALPHA HOMOLOG
.07	.29	.4	.47	.6	.5	.51	.54	.62	.51	.45	.32	.07	MPO Myeloperoxidase
.13	.26	.44	.47	.48	.48	.47	.5	.42	.45	.38	.12	.02	Lactate dehydrogenase B gene exon 1 ...
.03	.08	.34	.33	.54	.5	.36	.45	.39	.26	.17	.05	.01	TCRG T cell receptor gamma chain
0	.11	.32	.32	.36	.36	.3	.43	.35	.35	.2	.07	.01	KIAA0050 gene
0	.03	.13	.22	.29	.33	.27	.42	.41	.37	.27	.24	.07	RPS3 Ribosomal protein S3
0	.06	.12	.22	.3	.4	.35	.41	.39	.28	.24	.11	.05	LTB Lymphotoxin-beta
0	.11	.25	.26	.33	.36	.34	.38	.23	.22	.16	.09	.04	Homolog Drosophila enhancer of split ...
0	.06	.21	.25	.37	.38	.34	.41	.29	.27	.11	.06	.01	CD2 antigen (p50), sheep red blood ...
0	.07	.16	.3	.32	.31	.24	.34	.22	.25	.18	.09	.06	Put. HMG-17 protein gene extracted ...
0	.04	.12	.23	.21	.28	.27	.3	.29	.28	.2	.19	.07	Histone H1x
0	.05	.14	.18	.28	.31	.25	.31	.27	.26	.19	.06	.04	H4/g gene for H4 histone
0	.04	.18	.2	.25	.29	.23	.24	.24	.25	.12	.13	.05	PTMA gene extracted from Human ...
0	.04	.14	.19	.31	.23	.21	.33	.26	.29	.13	.07	.01	GLUL Glutamate-ammonia ligase ...
0	.03	.14	.19	.29	.18	.25	.3	.21	.27	.19	.12	.04	GB DEF=Polyadenylate binding protein ...

Table 2: Column labeled n gives the fraction of times the gene was predicted by PaGE as up-regulated in ALL T-cells versus ALL B-cells, out of 100 comparisons between n randomly chosen B-cells and all 9 T-cell experiments, from the Golub *et al.* data set. Confidence was set at 90%, and a shift of 5000 was used (see section SHIFTS).

It is interesting to compare the results of the two tables. Of the genes that were picked up reliably with many observations (the top 8 in Table 1, and the top 23 in Table 2), only two are in both tables, the "T-CELL ANTIGEN CD7 PRECURSOR" and the "TCF7 Transcription factor 7 (T-cell specific)." Based on this, it is recommended to use both methods, and to combine the results. For the two genes that do appear on both tables, the T-cell antigen is consistently predicted by PaGE with 4 observations, whereas with the step-down method nearly 25 observations were necessary. For the TCF7 Transcription factor, PaGE consistently predicted it with 5 observations, whereas 15 were required for the step-down method.

We take the genes with 1's in the first column (36) to be our most reliable predictions. At the bottom half of Table 2 we see that many genes which have 0's in the first column are predicted as differentially expressed

fairly often with fewer observations. This is due to clinical heterogeneity. When the cell type is heterogeneous, such as "ALL B-cells," there are many subclasses that the cell type can be classified into, so that if we choose only a small sample of experiments to represent the class, we are likely to have over-represented some subclasses. We investigate this issue further below where we gauge the sample size necessary to overcome this problem. First, however, it is necessary to investigate the nature of the data at a more fundamental level.

SHIFTS

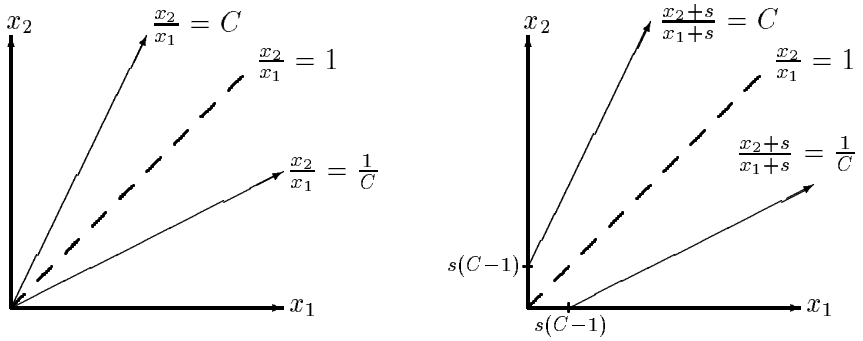

Figure 6: Lines bounding the region of up- and down-regulation expression for the case of no shift (left) and a shift equal to s (right). For simplicity we are assuming the lower cut-ratio c equals the reciprocal of the upper cut-ratio C.

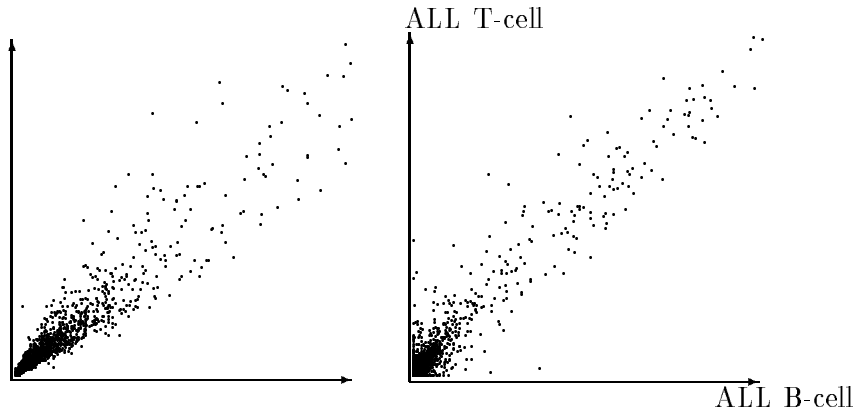

Figure 7: Idealized scatter plot (left) with spread proportional to intensity. Scatter plot of ALL B-cell versus ALL T-cell (right), from the Golub *et al.* data. Axes represent raw unlogged intensities.

PaGE predictions of differential expression are based on ratios of

mean intensities, as opposed to basing them on differences. The use of ratios allows for the convenient introduction of a "shift" parameter. In particular, we investigate the effect of measuring differential expression by the ratio of the shifted intensities $(x_2 + s)/(x_1 + s)$, where x_1 and x_2 are the respective intensities in two different sample types.

Varying the shift allows for different regions of the data to be emphasized. No shift corresponds to a region between the lines through the origin, as seen in the graph on the left in Figure 6. This corresponds to a criteria such as taking a fixed cutoff, say 2-fold, and predicting anything over 2-fold (or under $\frac{1}{2}$-fold) to be differentially expressed. Such a uniform criteria would be justified if the data looked, for example, like that in scatter plot on the left in Figure 7, where the spread is proportional to the magnitude of expression. A graph of the actual ALL B-cell vs. ALL T-cell data is given in the graph on the right in Figure 7. As can be seen from this graph, spread of the data is much greater at the low intensities. This is caused largely by the inclusion of the absent calls. It is therefore clear that a shift is necessary. The shift concept has been used in other methods, e.g. [Newton *et al.*, (2001)].

PaGE results were generated by varying the shift, comparing the 9 T-cell to the 37 B-cell experiments used. The results are shown in Table 3. The set of predictions varies as the shift varies, as can also be seen from the graph in Figure 8.

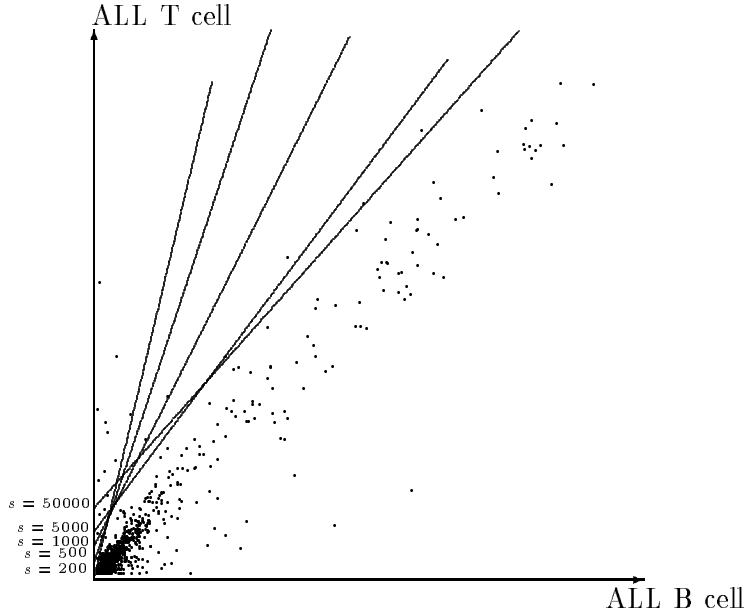

Figure 8: Lines show cutoffs for predictions of up-regulation in ALL T-cells versus ALL B-cells, for varying shifts.

	shift					
ID	200	500	1000	5000	50000	IDinfo
V00599					*	mRNA fragment encoding beta-tubulin. ...
M13792				*	*	ADA Adenosine deaminase
M16279				*	*	MIC2 Antigen identified by monoclonal antibodies...
M23323			*			T-CELL SURFACE GLYCOPROTEIN CD3 EPSILON CHAIN ...
X03934	*	*	*	*	*	GB DEF = T-cell antigen receptor gene T3-delta
X00437	*	*	*	*	*	TCRB T-cell receptor, beta cluster
X76223	*	*	*	*	*	GB DEF = MAL gene exon 4
U14603	*	*	*	*	*	Protein tyrosine phosphatase PTPCAAX2 (hPTPCAAX2)...
U23852	*	*	*	*	*	GB DEF = T-lymphocyte specific protein tyrosine ...
M12886	*	*	*	*	*	TCRB T-cell receptor, beta cluster
M59807	*	*	*	*		NATURAL KILLER CELLS PROTEIN 4 PRECURSOR
M28826	*	*	*	*		CD1B CD1b antigen (thymocyte antigen)
U50743	*	*	*	*		Na,K-ATPase gamma subunit mRNA
M59807	*	*	*	*		NATURAL KILLER CELLS PROTEIN 4 PRECURSOR
U49835	*	*	*	*		CHIT1 Chitinase 1
U67171	*	*	*	*		GB DEF = Selenoprotein W (selW) mRNA
D00749	*	*	*	*		T-CELL ANTIGEN CD7 PRECURSOR
M21624	*	*	*			TCRD T-cell receptor, delta
M16336	*	*	*			CD2 CD2 antigen (p50), sheep red blood cell receptor
M37271	*	*	*			T-CELL ANTIGEN CD7 PRECURSOR
X59871	*	*	*			TCF7 Transcription factor 7 (T-cell specific)
X14975	*	*	*			GB DEF = CD1 R2 gene for MHC-related antigen
S78187	*					M-PHASE INDUCER PHOSPHATASE 2
L10373	*					MXS1 Membrane component, X chromosome, ...
M28825	*					CD1A CD1a antigen (thymocyte antigen)
X62891	*					Mutant coseg gene for vasopressin-neurophysin ...
S56151	*					HMFG
J04132	*					CD3Z CD3Z antigen, zeta polypeptide (TiT3 complex)
L40386	*					DP2 (Humdp2) mRNA
X04145	*					CD3G CD3G antigen, gamma polypeptide (TiT3 complex)
Z19002	*					ZINC FINGER PROTEIN PLZF
L05148	*					Protein tyrosine kinase related mRNA sequence

Table 3: Genes predicted by PaGE to be up-regulated in T-cells (9 samples) versus B-cells (37 samples). Columns correspond to different shifts of the data. Confidence set to 95%. An asterisk indicates that the gene was predicted as up-regulated in T-cells versus B-cells. A shift of 1000 is considered optimal for this dataset, based on empirical studies of the PaGE estimate (2). See also Figure 8.

Some genes are predicted as up-regulated in T-cells, regardless of shift, whereas others are predicted for only a range of shifts. PaGE relies on the estimation (2), which can become inaccurate as the shift decreases to zero. Empirical tests of this approximation using the Golub *et al.* data show that a shift of 1000 achieves the best conservative approximation for comparing 37 to 9 experiments having the type of variation present in these data.[‡] Even if this PaGE approximation were perfect, a moderate shift would be preferred. This can be seen on the graph in Figure 8. With little or no shift, a very large slope must be used to avoid the false positives at the low intensities. The result is that true positives are missed in every intensity range except the lowest. Likewise, a shift that

[‡]In the case of Table 2 we used a uniform shift of 5000, which gave the best approximation, according to the simulations, simultaneously for all values of n used.

is too large might tend to pick up false positives from the higher intensity genes. A review of the literature on the 19 genes that were predicted up-regulated in T-cells, using the shift value of 1000, showed no clear false positives, with most being known to be up-regulated. This gives some empirical verification of the confidence measures in PaGE. On the other hand, of the 10 genes that were predicted to be upregulated with a shift of 200, but were not predicted with higher shifts, there were only four true positives, and five apparent false negatives, with one being unclear. Thus for such a low shift, the confidence appears to have been reduced to around 50%, as might be expected.

THE EFFECT OF CLINICAL HETEROGENEITY

If the sample types under question are not sufficiently homogeneous, then using a relatively small sample size gives a high probability that one or more biological subclasses are over-represented by the experiments. If this happens, the answers obtained might not be answers to the questions being asked. For example, suppose one is looking for genes that are differentially expressed between B-cells and T-cells, using 10 observations of each type. Consider a gene that is not differentially expressed between B-cells and T-cells, but nonetheless is differentially expressed in differing stages of cell cycle or development. Since there are many genes on the chip, it is not unlikely that the majority of the B-cell experiments are in one state, and the majority of the T-cell experiments in another state, with respect to some of the genes. In this case, we may falsely predict the genes that are differentially expressed between these two states to be differentially expressed between B-cells and T-cells. The statistical methods are correctly predicting differential expression, however between different sample types than expected, which can lead to false conclusions.

To measure the effect of this clinical heterogeneity when using the prediction methods of [Dudoit *et al.*, (2000)] and [Manduchi *et al.*, (2000)] on data of a similar nature to that of the Golub *et al.* data, we generated virtual data, based on their ALL B-cell data. To have data that are as variable as theirs, we used the B-cell empirical distributions themselves to generate the virtual data. We sampled randomly 1000 times (with replacement) from the gene tags on the chip. For each of these 1000 gene tags, we sampled from the empirical B-cell distribution of that gene 200 times. This created 200 virtual observations of 1000-gene experiments, containing comparable noise to the actual Golub *et al.* data. The 1000 genes were sampled independently from each other, so that

	number of independent genes	
Observations	1000	3000
5	0.39	0.44
10	0.15	0.19
20	0.10	0.11
30	0.06	0.07
40	0.06	0.02
50	0.03	0.04

Table 4: Fraction of times, out of 100 runs, that any predictions of up-regulation were made, comparing a virtual sample to itself, for varying numbers of observations. All such predictions represent false-positives.

we have created data with 1000 independent genes. We repeated this making 200 virtual experiments with 3000 independent genes. Certainly the approximately 7000 genes on the Affymetrix chip are not expressed independently in B-cells, however it seems reasonable that there might be 1000 to 3000 independent genes. For each data set, we compared random selections of 5, 10, 20, 30, 40, and 50 experiments against each other. Since all experiments were generated from the same distributions, any predictions of up-regulation are false-positives.

The results are given in Table 4. The effect of clinical heterogeneity is clear when only 5 or 10 observations are used. After 20 observations, the effect effectively disappears, and the percentage of false-positives levels off. This percentage of false-positives is not expected to converge to zero, because the PaGE confidence measure is not 100% (it was 90% in this case).

A similar test was conducted with the step-down method. This method was not susceptible to the problem. This is most likely due to the fact that with only 5 observations, the step-down method, which relies on permutations, makes highly conservative predictions. There are not enough experiments to permute very many times. As this problem diminishes, after roughly 10 observations, the effect of clinical heterogeneity has already diminished. It is interesting to note that in the experiments, the step-down method made false-positive predictions roughly 5-8% of the time, with an experiment-wise Type I error of 0.1. This gives some empirical verification that this reported Type I error is close to correct, and to the extent that they are off, they are conservative.

CONCLUSIONS

We have used the Golub *et al.* data to investigate the issue of the impact of variability when differential expression predictions are obtained from a small sample size for each sample type. When the desired confidence measures relate to very homogeneous classes of cells, then as few as two or three observations might suffice. However, often heterogeneous sample types are compared, such as "all lymphoblasts" versus "all myeloid cells," to find any discernable differential expression. This study shows that the non-deterministic effect is greater than might be expected. It is necessary to perform further studies characterizing this effect, as many investigators are currently designing array based analyses of their specific systems of interest, and one of the most basic decisions they have to make is how many and what type of experiments to perform. It is not a simple issue, and to address it, it is necessary to have case studies with results that are verified by a large sample. We have used the Golub *et al.* data for this purpose.

We have also used the data to investigate the performance of two currently existing tools for making predictions of differential expression. At issue is whether to use an experiment-wise Type I error, or the false-discovery rate. The results show that the latter approach allows for greater sensitivity, particularly with few replicates, with the trade-off being the introduction of a small percent of false positives. Each technique predicted some genes that the other technique did not. Therefore it is advisable to use both methods and combine the results.

ACKNOWLEDGEMENTS

We thank Warren Ewens and Sandrine Dudoit for helpful discussions on the step-down method and the false discovery rate, Joan Mazzareli for the literature review to validate the predictions, Jonathan Schug for helpful discussions, and Brian Brunk and Georgi Kostov for help with the E2K cluster. We also thank the rest of the members of the Universtity of Pennsylvania Computational Biology and Informatics Laboratory (CBIL) for their support. This work was supported in part by NIH grants R24-DK56947 and K25-HG-00052-01A1.

REFERENCES

Dudoit S., Yang Y.H., Callow M., Speed T. (2000). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments." UC Berkeley, Technical report #578.
`http://www.stat.berkeley.edu/users/terry/zarray/Html/matt.html`

Ewens W.J., Grant G.R. (2001). "Statistical Methods in Bioinformatics." *Springer-Verlag* New York.

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science* **286(5439)**: 531-537.

Lockhart D.J. *et al.* (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." *Nature Biotechnology* **14**: 1675-1680.

Manduchi E., Grant G.R., McKenzie S.E., Overton G.C., Surrey S., Stoeckert C.J. Jr. (2000). "Generation of patterns from gene expression data by assigning confidence to differentially expressed genes." *Bioinformatics*, **16(8)**: 685-698.
`http://www.cbil.upenn.edu/PaGE`

Mood A.M., Graybell F.A. (1963). "Introduction to the Theory of Statistics." *McGraw-Hill* New York.

Newton M.A., Kendziorski C.M., Richmond C.S., Blattner F.R., Tsui K.W. (2001) "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data." *J. Comput. Biol.*, **8(1)**:37-52.
`http://www.biostat.wisc.edu/geda/eba.html`

Sagerstrom C.G., Sun B.I., Sive H.L. (1997). "Subtractive cloning: past, present, and future." *Ann. Rev. Biochem.*, **66**: 751-783.

Westfall P.H. and Young S.S. (1993). "Resampling-based multiple testing: examples and methods for p-value adjustment." Wiley series in probability and mathematical statistics, Wiley.