Statistical Analysis of NGS Data from N=1 Studies

Contents

[Abstract 1](#_Toc503642715)

[1.0 Introduction 2](#_Toc503642716)

[2.0 Methods 2](#_Toc503642717)

[2.1 Data 2](#_Toc503642718)

[2.1.1 Dataset1: db/db model of obesity-induced type 2 diabetes: neuroretina 2](#_Toc503642719)

[2.1.2 Dataset2: Transcriptomic analysis of effects of histone methyltransferase Setd7 knockdown and isothiocyanate PEITC in human prostate cancer cell by microarray 2](#_Toc503642720)

[2.2 In vivo vs. in vitro samples and sources of variability 2](#_Toc503642721)

[2.3 Signal-To-Noise Ratio Approach 2](#_Toc503642722)

[2.2 Standard Deviation with LOESS of Moving Window (LMW) 4](#_Toc503642723)

[2.3 Standard Deviation with Smoothing Spline of Absolute Differences (SSAD) 5](#_Toc503642724)

[5](#_Toc503642725)

[2.4 The Agreement between the Two Methods 5](#_Toc503642726)

[2.4 Test Statistic Distribution and Cut-Offs 6](#_Toc503642727)

[3.0 Results 6](#_Toc503642728)

[4.0 Discussion 6](#_Toc503642729)

[5.0 Acknowledgements 6](#_Toc503642730)

[References to add 6](#_Toc503642731)

[References 6](#_Toc503642732)

# Abstract

Recent advances in genomics and sharp decreases in gene sequencing price created a flood of data generated by laboratories of all sizes. However, many of the laboratories are operating on strictly constrained budgets and cannot afford large number of samples sequenced for a single project, rerouting resources saved on replication into testing more compounds and models instead. Very often, the results produced by a genome-wide analysis are the first step in screening for genes of interest and are usually validated further by qPCR or pyrosequencing. This may result in a large number of false positives, very quickly diluting the usefulness of such first-pass screening. The purpose of this paper is to describe a statistical approach for evaluation of differential expression of genes in experiment with small number of replicates, including an extreme example of single samples per treatment. Our method is based on the relationship between the differences and the means of gene expressions. This relationship allows estimation of standard errors by borrowing strength across all genes in the samples. The results indicate that even in experiments with no replicates, useful test statistics can be calculated and applied for filtering of significantly differentially expressed genes. This method does not address the technical variability and assumes that the experimenters’ techniques were reasonably good and no other major technical factor was contributing to heterogeneity of the data. This assumption holds for the data used in this paper that was generated by Dr. T. Kong’s laboratory, as we were able to confirm our findings for single replicates using a study with four replicates per group (Yue’s curcumin data reference here).

# 1.0 Introduction

# 2.0 Methods

## 2.1 Data

Multiple data sources were used to test the methods described in this paper.

### 2.1.1 Dataset1: db/db model of obesity-induced type 2 diabetes: neuroretina

This dataset is publically available through the National Center for Biotechnology Information (NCBI) website[[1]](#footnote-1). Bogdanov et al analyzed neuroretina from 8-week C57BLKsJ-db/db diabetic mice. These leptin-receptor-deficient db/db mice have been shown to exhibit early features of diabetic retinopathy (DR) (Bogdanov P, 2014) (more details about microarray data here).

### 2.1.2 Dataset2: Transcriptomic analysis of effects of histone methyltransferase Setd7 knockdown and isothiocyanate PEITC in human prostate cancer cell by microarray

This dataset was generated from a microarray experiment that examined differential expression in genes of wild type (WT) and Sedt7 knock-down (KD) mice treated with phenethyl isothiocyanate (PEICT) compared to controls (Chao’s LNCaP reference and details here).

### 2.1.3 Dataset3: Skin UVB experiment

In this experiment, samples were harvested from mice at 2, 15 and 25 weeks. The mice were subjected to one of 3 treatments – UVB only (positive control), UVB + UAA, and UVB + SFN. A group of untreated animals was used as a negative control. Two mice per group were used. Both, RNA expression and DNA methylation data was collected for each mouse. We compared existing methods - DESeq2 (differential gene expression analysis based on the negative binomial) for RNA expression and DSS (dispersion shrinkage for sequencing data) for the methylation, with our method.

## 2.2 In vivo vs. in vitro samples and sources of variability

There are two main sources of variability in biological experiments: technical and biological. Technical variability can be estimated with technical replicates. However, if the organisms are much diverged biologically, samples from multiple animals or human subjects must be collected for each treatment group so inference about a wider group of organisms with shared trades (both, genotype and phenotype) can be made. Biological variability, in general, is much higher in in vivo experiments compared to in vitro, especially if the latter is conducted using cells from the same passage (i.e. same generation of cells). Sometimes, a workaround is applied when only a limited number of samples can be analyzed but biological variability is assumed to be relatively high: samples from multiple animals, tissues or Petri dishes can be pooled together and analyzed as one [REFERENCE NEEDED]. This type of sample pooling attempts to create an “average” sample with a hope that differences between the animals in the same group will cancel each other out and the effect of the treatment will be amplified above the background noise.

## 2.3 Signal-To-Noise Ratio Approach

The simplest approach for filtering differentially expressed genes is using fold-change threshold, i.e. if the ratio of the average expressions of the same gene from two different groups is above a certain value, the difference is *biologically significant*. However, as explained by Amaratunga et al, group mean expressions are the estimates of the true but unknown means, and the quality of these estimates will depend on how variable each gene is. Therefore, for genes with higher variability observing a large fold-change is more likely than it is for the low-variability genes (Amaraunga & Cabrera, 2013).

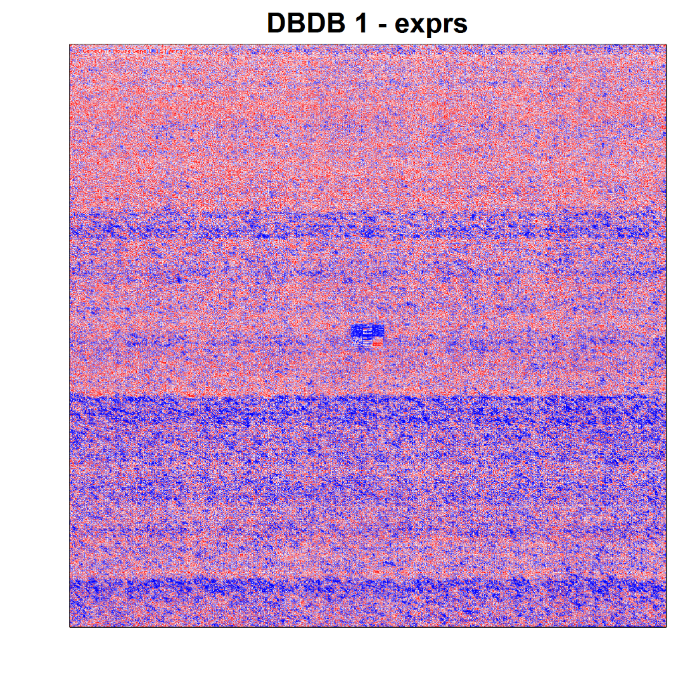
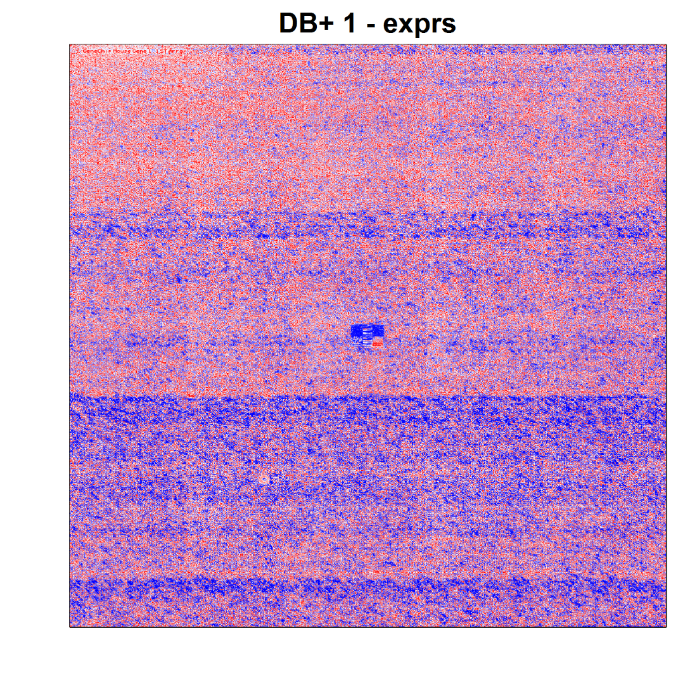
To account for the differences in variability, the estimate of the mean differences is divided by the estimate of variability (e.g. standard error). This is referred to as *signal-to-noise ratio*.

When the number of samples in each group is small or there are no replicates, the usual method for estimation of variability can be very inaccurate or impossible. Instead of trying to estimate variability of each gene separately, we can borrow strength across many genes. There are several approaches to borrowing strength. The simplest is to assume that the variability is the same for all genes but, as we already discussed above, the more variable genes will produce higher differences much more often. An alternative is to exploit the relationship of differences and the means of gene expressions. For the remainder of this paper, we will only discuss the extreme cases of having data with no replicates (i.e. single sample per treatment group) which eliminates any possibility of estimating standard errors for any single gene but the method can be extended to experiments with any number of replicates.

For each of the data sets analyzed in this paper, microarray data was first preprocessed and normalized. We used *robust multichip average* (RMA) algorithm for data normalization (Irizarry, 2003). We then calculated all pairwise differences and means between each two treatments for each gene. The data at this point was on the log2 scale so the differences represented number of two-fold changes (e.g. log2 of 3 is 3 two-fold changes, or 2^3 = 8-fold change).

C:\git_local\single.repl.stat\tmp\dbdb_raw_vs_rma.norm.tiffOur first data set (db/db diabetes model) consisted of 8 samples obtained from 4 diabetic (db/db) and four non-diabetic (db/+) C57BL/KsJ male mice. The data was imported into R using *read.celfiles* function from *oligo* package (Carvalho, 2010). The CEL data files contained raw gene expression values derived from the microarray images, and annotated using R package *pd.mogene.1.1.st.v1*. Next, pseudo-images of the microarray chips were reconstructed and visually examined (Fig 1). Since no obvious defects were found in any of the pseudo-images, we proceeded to the next step: normalization.

Figure 1: Pseudo-mages of a db/+(left) and a db/db (right) macroarrays



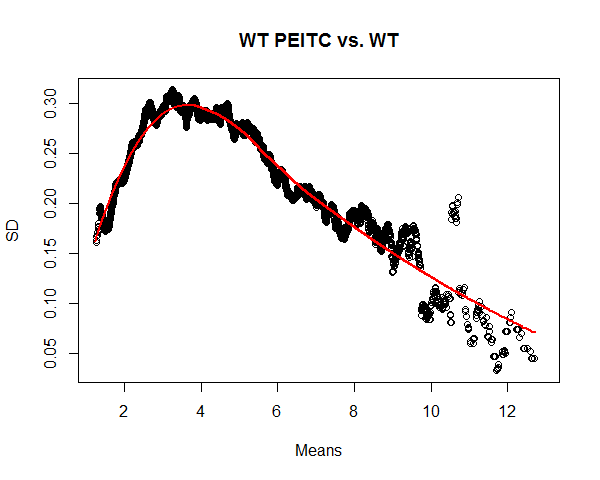
We used RMA algorithm (function *rma* in R package *oligo*) for background correction, quantile normalization and summarization (via median-polish). The results are shown in Figure 2.

Once the data was preprocessed, we moved to testing the differences in gene expressions between the groups (db/db vs. db/+). To simulate a single-sample situation, we calculated pairwise differences and means between all 8 samples.

Figure 2: Raw values (left) vs. RMA-normalized gene expressions (right)

## 2.2 Estimating Standard Deviation with Local Polynomial Regression Fitting

Local Polynomial Regression (LOESS) is a nonparametric method combining k-nearest-neighbor (KNN) approach with regression.



## 2.3 Standard Deviation with Smoothing Spline of Absolute Differences (SSAD)

## 

## 

## 

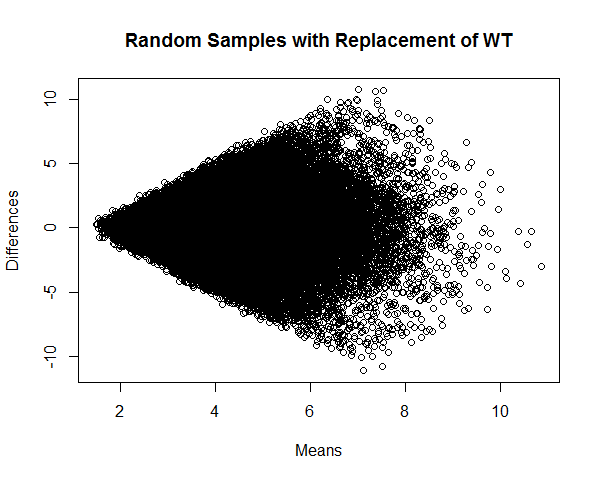
## 2.4 The Agreement between the Two Methods

LMW has a higher estimate of variability than SSAD. The differences decrease toward the average SD values.

## 2.5 Estimating Standard Deviation in DNA Methylation Data

C:\Users\dsargsy\OneDrive - JNJ\skin.uvb\tmp\skin_uvb_maplot_ctrl_w15-w02.tiffSimilarly, to RNA expression data, DNA methylation data can be visualized using MA-type plot. When difference in methylation is plotted against the mean of any two groups for each gene, the relationship between these two statistics can be modeled and used to estimate the variability of each gene conditioned on the mean. Unlike RNA expression, the shape is not a funnel but close to oval. When DSS is applied to the data, the genes at with means at the extremes (0 and 1) have lower variability hence the difference is significant at relatively small difference compared to the genes with the means closer to 0.5 (Figure XX below). Since DSS cannot be applied to N=1 studies, we used this relationship to estimate variability for each gene when there are no replicates.

## 2.6 Test Statistic Distribution and Cut-Offs

NOTE: Prof. Cabrera proposed sampling from controls and estimating SD with SSAD for the controls only. However, there is only 1 sample of control. I cannot randomly sample and reassign expression values as each gene most likely has a range. Otherwise, with random sampling, I completely distort the data structure.

# 3.0 Results

# 4.0 Discussion

# 5.0 Acknowledgements

Everybody in Kong’s lab who is not a coauthor; J&J bioinformaticians that gave me the original idea, Javier and Dhammika, Christian from RUCRD, Cardiovascular Institute.

# References to add

1. [Question: Microarray Data Without Replicates. Any Hope To Get Something Out Of It?](https://www.biostars.org/p/14130/)
2. [How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC115224/)
3. [The Power of Replicates by Illumina](https://www.illumina.com/Documents/products/technotes/technote_power_replicates.pdf)

# References

Amaraunga, D., & Cabrera, J. (2013). *Exploration and Analysis of DNA Microarray and Other High-Dimentional Data.*

Bogdanov P, C. L. (2014). The db/db mouse: a useful model for the study of diabetic retinal neurodegeneration. *PLoS One*.

1. https://www.ncbi.nlm.nih.gov/sites/GDSbrowser#details [↑](#footnote-ref-1)