arXiv:1103.3434v2 [q-bio.QM] 28 Aug 2013

Volcano Plots in Analyzing Differential Expressions with mRNA Microarrays

Wentian Li

The Robert S. Boas Center for Genomics and Human Genetics

The Feinstein Institute for Medical Research, North Shore LIJ Health System

Manhasset, 350 Community Drive, NY 11030, USA.

ABSTRACT:

Volcano plot displays unstandardized signal (e.g. log-fold-change) against noise-adjusted/standardized signal (e.g. t-statistic or $-\log_{10}(p\text{-value})$ from the t test). We review the basic and an interactive use of the volcano plot, and its crucial role in understanding the regularized t-statistic. The joint filtering gene selection criterion based on regularized statistics has a curved discriminant line in the volcano plot, as compared to the two perpendicular lines for the "double filtering" criterion. This review attempts to provide an unifying framework for discussions on alternative measures of differential expression, improved methods for estimating variance, and visual display of a microarray analysis result. We also discuss the possibility to apply volcano plots to other fields beyond microarray.

KEYWORDS:

microarray; volcano plot; signal-to-noise ratio; regularization

1 Introduction

The microarray technology allows simultaneous measurements of messenger RNA level of thousands of genes, and its adoption dramatic changes the way biological and biomedical research is carried out [1–7]. In particular, the more labor-extensive real-time PCR can be

replaced by microarray profiling in a preliminary round, as the general agreement between the two methods is considered to be good [8–10]. As an emerging technology, there are still many issues to be worked out, such as the consistency among different platforms [11–19] as well as their integration [20], batch effect [21–26], level, source, and distribution of noise [27–35], limit of dynamic range [36], etc. However, with better probe design [7, 37], better data quality control [38–40], better data reporting requirement [41, 42], better normalization scheme [43–50], and better understanding of the study goals, these are not insurmountable problems.

Analyzing large amount of expression data from microarray experiments was thought as a major challenge in early days, but this problem was over-estimated. First, the amount of data from thousands of genes and a hundred or so samples is still much smaller than, e.g., the data generated by whole-genome association studies [51] or next generation sequencing [52], and a moderately sized computer might handle the data without problems. Second, no brand-new statistical learning methods have to be invented and existing machine learning techniques [53] could already extract meaningful information from the data. Third, the problem of larger number of false positives due to the large number of genes being profiled has been addressed and properly handled [54–58]. Fourth, in using multiple genes in constructing classifiers, the well known "large p, small n" problem (large number of variables with small number of sample size) can be solved by the variable/subset/feature/model selection techniques [59–66]

One of the most common applications of microarrays is "differential expression" profiling: identifying mRNAs/genes whose expression level is very different under two conditions, e.g., with disease and being healthy. Not only could differentially expressed genes provide insight into the biological processes involved in disease etiology, but also these can be used as biomarkers for diagnosis [67–71] or prognosis [72–75]. The phrase "differential expression" means that the *averaged* expression level of a mRNA/gene in one phenotype-specific group is much *larger* or *smaller* than that in another group. However, the terms "*average*" and "*larger/smaller*" are up to various interpretations.

There are at least two definitions of average: arithmetic mean ($E[x] = \frac{1}{n}\sum_{i=1}^{n} x_i$) or geometric mean ($G[x] = (x_1 x_2 \cdots x_n)^{1/n}$). For fluorescence-light-intensity based microarray data x, it is a common practice to logarithmically transform the data $x' = \log_{10}(x)$, because x' fits better than x to a normal distribution (without losing generality, the base of the logarithmic func-

tion is chosen at 10 in this review). Then $E[x'] = \frac{1}{n}\sum_{i=1}^{n}\log_{10}(x_i) = \log_{10}(x_1 x_2 \cdots x_n)^{1/n} = \log_{10} G[x]$, connecting the two means. Yet another measure of average is the median, being unaffected by log-transformation, which has been used in [76].

Deciding "how large one group's average is compared to the other" is no less trivial. Fold-change and t-statistic are the two main choices for measuring differential expression. In microarray analysis field, these two measures have been in and out of favor at various time. Fold-change had been commonly used before it was pointed out that it did not take the noise into account [77, 78]. t-statistic enjoyed its acceptance until another round of papers suggesting that genes selected by fold-change are more consistent among different microarray platforms than those selected by t-statistics [39, 79, 80]. This result led to more discussion on the relationship between reproducibility and accuracy [81–83], and between biological and statistical signal [84].

Despite development of sophisticated methods for microarray analysis, one question we analysts hear the most from the end-users is "should I use fold-change or t-statistic?". The problem with fold-change is that the same fold-change value will be less impressive if the variance is large. Although t-statistic aims at taking the noise level into account, the practical problem is that the variance may not be estimated reliably, especially when the sample size is small. An answer provided by this review is basically "use both": the volcano plot is exactly such a visual tool to display both fold-change and t-statistic.

This review is organized as follows: Section 2 establishes a relationship between the fold-change and t-statistic; Section 3 introduces volcano plots and its basic usage; Section 4 summarizes the idea of "moderated", "regularized", "penalized" statistics by adding an extra positive term to the sample-based variance or standard error; Section 5 discusses the regularized statistics in the context of volcano plot; Section 6 surveys the software packages in Bioconductor that are relevant to this review; Section 7 introduces the idea of stratified volcano plots; and the final Section is the discussion and conclusion section.

All plots in this paper use the same published dataset containing 37 case/patient samples and 18 control samples, with 48804 probesets in Illumina platform, normalized by "quantile normalization".

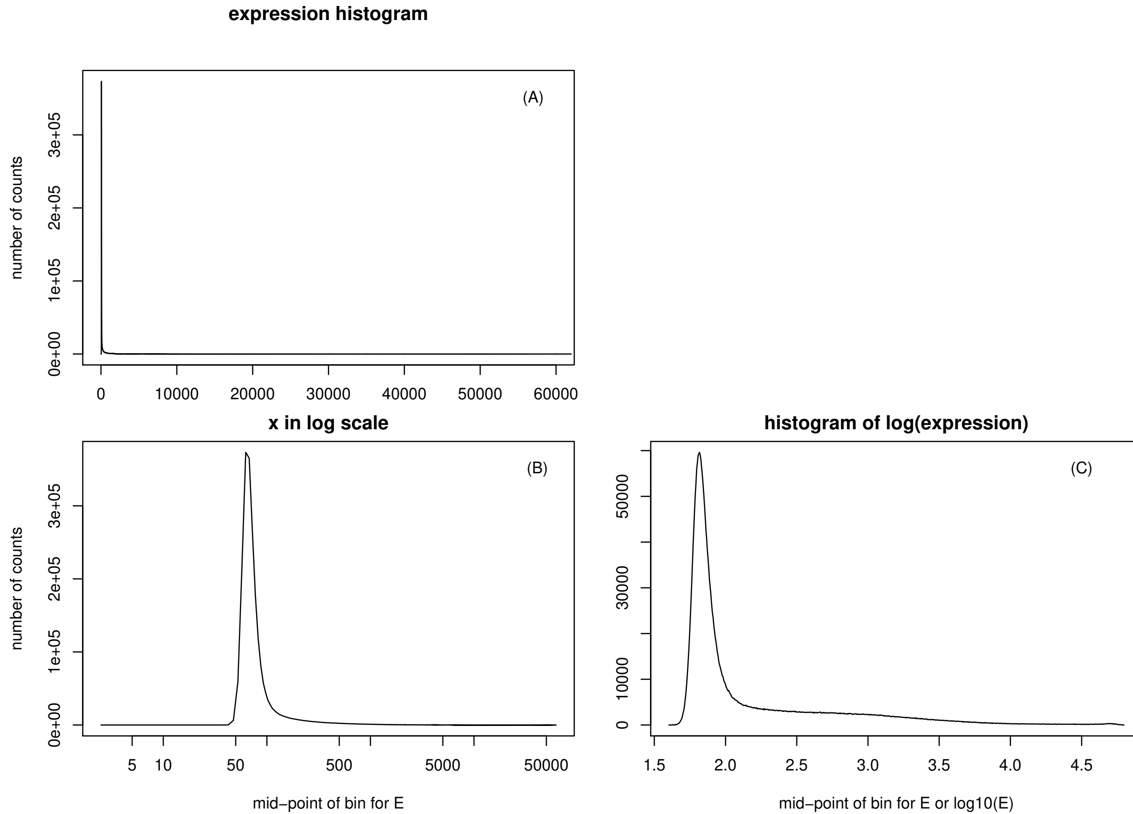

Figure 1: Histogram of expression levels of a microarray experiment: (A) in linear scale. (B) x-axis in a log scale. (C) for log-transformed expression.

2 Fold-change and t-statistic: signal and signal-to-noise ratio

Fold-change (FC) and t-statistic seem to be two very different quantities: one is intuitive and a straightforward measure of differences, another is rooted deeply in the field of statistics. However, with logarithm transformation there is a relationship between the two.

The need for logarithmic transformation can be illustrated by Fig.1. Fig.1 shows the three histograms of fluorescence-light intensity E of a microarray experiment which is indicative of the number of mRNA copies hybridized to the probe, thus a measure of mRNA expression level: (A) in regular scale, (B) in log-transformed x-axis scale, and (C) of $\log_{10}(E)$ itself. Without the logarithmic transformation, the distribution of E is very long-tailed, and very skewed (asymmetric). With the log transformation (or other similar transformations in a recognition that log transformation cannot handle zero level [85]), even though the distribution is still not

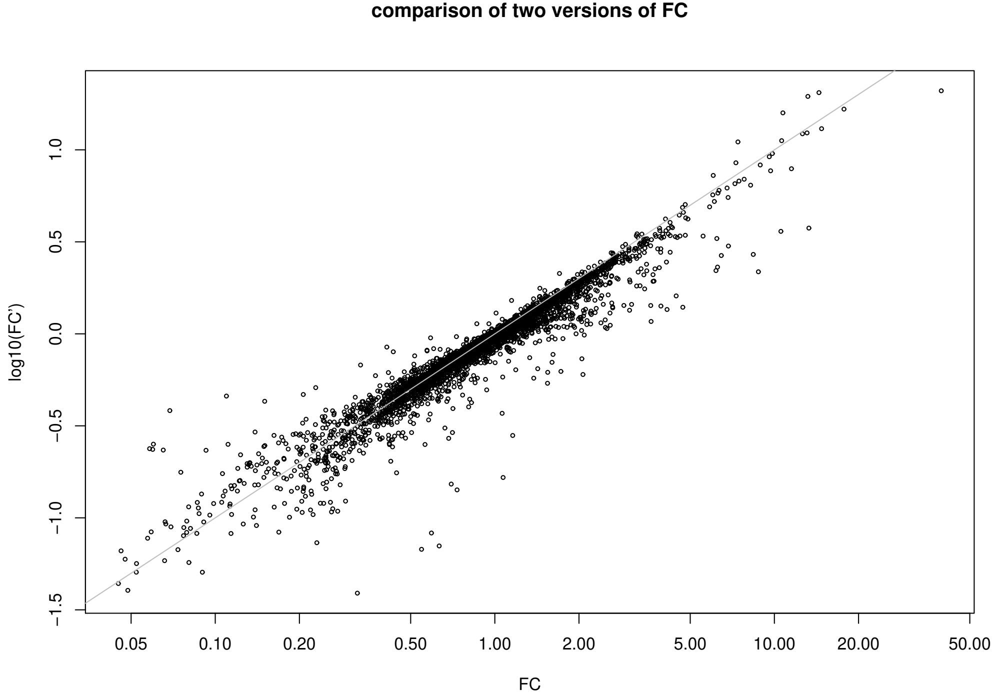

Figure 2: Comparison of two definitions of fold-changes. The x is FC $= \langle E_1 \rangle / \langle E_0 \rangle$ in log scale. The y is the $\log_{10}(FC') = \langle \log_{10} E_1 \rangle - \langle \log_{10} E_0 \rangle$ (Eq.(1)).

a perfect normal distribution, it is much more "normal-like".

There are other advantages of a log transformation, e.g. variance is more stablized and does not tend to increase with the mean; it is consistent with a psycho-physics law relating human sensation to the logarithm of the stimulus level [86]. Note that for non-fluorescence-light-density-based technologies for measuring expression level, such as digital expression and RNA-seq [87] we lose this ground for justifying log-transformation. The decision on whether to use a transformation to become a normal distribution, or whether to model the data by another distribution completely, such as the Poisson distribution, is empirically based on the histogram of the data [88–99]. However, we also notice that Poisson distribution is approximately a normal distribution when its mean is large.

The simplest definition of FC is: $FC = \langle E_1 \rangle / \langle E_0 \rangle$, where the arithmetic average is over the fluorescence-light intensity of samples in group 1 (e.g. diseased group) and group 0 (e.g. control group). The logarithm of FC is: $\log_{10}(FC) = \log_{10}\langle E_1 \rangle / \langle E_0 \rangle = \log_{10}\langle E_1 \rangle - \log_{10}\langle E_0 \rangle \approx \langle \log_{10} E_1 \rangle - \langle \log_{10} E_0 \rangle$. Reversing the order of averaging and log-transformation operations

usually does not lead to identical values, so the above expression is only an approximation. We can have a second definition of FC called FC':

$$\log_{10}(FC') = \langle \log_{10} E_1 \rangle - \langle \log_{10} E_0 \rangle \tag{1}$$

Fig.2 shows that FC is mostly similar to FC' and we do not distinguish the two definitions. The same conclusion is also reached in [84].

The t-test is an example of statistical testing whose goal is to compare any observed result with chance events. The statistic used in t-test (e.g. [100]) is the difference of arithmetic means in two groups divided ("standardized") by the estimated standard deviation of that difference. Standard deviation of parameters (e.g., sample mean, sample variance) is often called "standard error" (SE) [100]. One requirement for using t-test is that values in two groups roughly follow normal distributions. As discussed above, we need to log transform the fluorescence light intensity E to have a normal-like distribution, so t-statistic is:

$$t = \frac{\langle \log_{10} E_1 \rangle - \langle \log_{10} E_0 \rangle}{SE_{\langle \log_{10} E_1 \rangle - \langle \log_{10} E_0 \rangle}} = \frac{\langle \log_{10} E_1 \rangle - \langle \log_{10} E_0 \rangle}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}} \tag{2}$$

where the second formula was due to Welsh [101], who assumed different variances in group 1 and group 0 and provided an estimation of SE (s_1^2 and s_0^2 are the estimated variances (of $\log_{10}(E)$) of group 1 and 0, and n_1, n_0 are number of samples in the two groups).

Comparing Eq.(1) and Eq.(2), we establish a relationship between $\log_{10}(FC)$ and t-statistic: t is $\log_{10}(FC)$ standardized by the noise level as measured by the pooled standard error. There are parallel contrasts of measures in other fields, such as the *signal-to-noise ratio* (vs. signal by itself) in engineering, *standardized effect size* (vs unstandardized effect size) in statistical behavioral science, quantitative psychology, epidemiology, and meta-analysis [102]. The exact relationships between them, however, require more careful examination; for example, t-statistic increases with sample n by the factor of $\sqrt{n}$ when it is not zero, whereas standardized effect size does not change with the sample size.

3 Volcano plot and its basic use

If the noise level is known or can be reliably estimated, it is of course preferable to measure

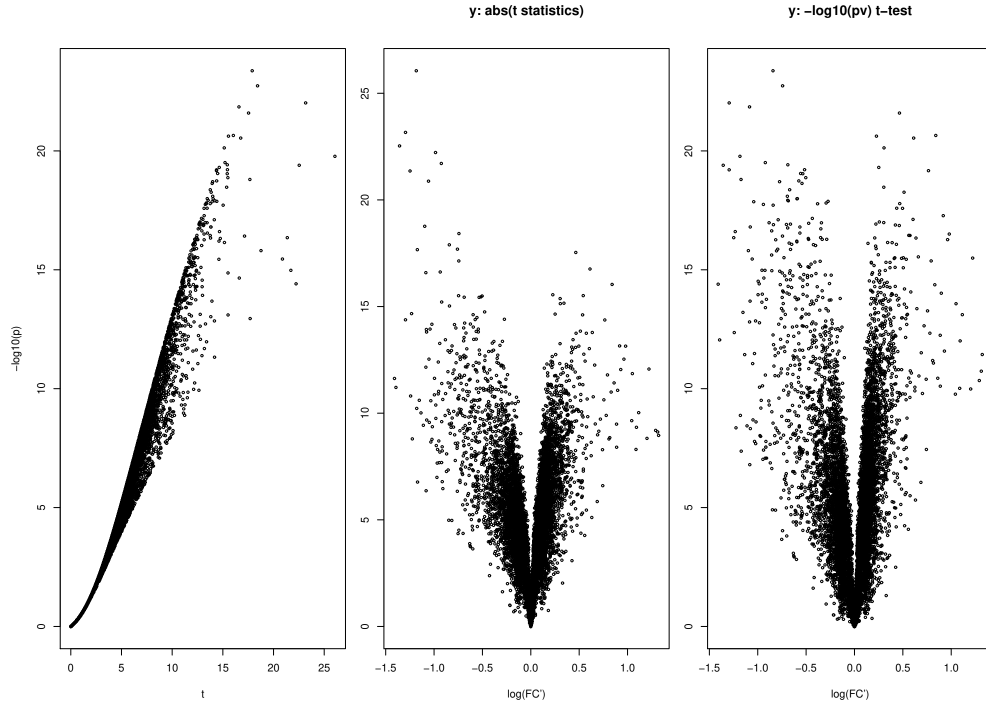

Figure 3: (A) x-axis: t-statistic, y-axis: $-\log_{10}$(p-value) of t-test. (B) Volcano plot using t-statistic as the y-axis (x-axis is $\log_{10}$FC). (C) Volcano plot using $-\log_{10}$(p-value) as the y-axis.

differential expression that takes the noise level into account, such as t-statistic. In reality, not only is smaller sample sizes an issue for variance estimation, but also, if systematic error exists, we may not improve the situation by increasing the sample size. For example, it is observed that noise level during the hybridization stage is much higher than that during the sample preparation or amplification stage [103]. If a probe sequence for an mRNA is highly represented in the genome, cross-hybridization can be a cause of error and variation, and the probability of this error does not seem to decrease with large sample sizes.

Facing this reality, we might just display and use both FC and t-statistic, and this is what the volcano plot does. Volcano plot most often refers to the scatter-plot with $-\log_{10}$(p-value) from the t-test as the y-axis and ($\log_{10}$)FC as the x-axis [104–106]. However, t-statistic and $-\log_{10}$(p-value) are highly correlated (see Fig.3(A)), and whether the t (Fig.3(B)) or $-\log_{10}$(p-value) (Fig.3(C)) is used in the y-axis, the outcome is very similar. The reason why t and p-value from t-test is not one-to-one corresponding (Fig.3(A)) is because in determining p-value, Welsh's t distribution has a degree of freedom parameter which also depends on the

data [107].

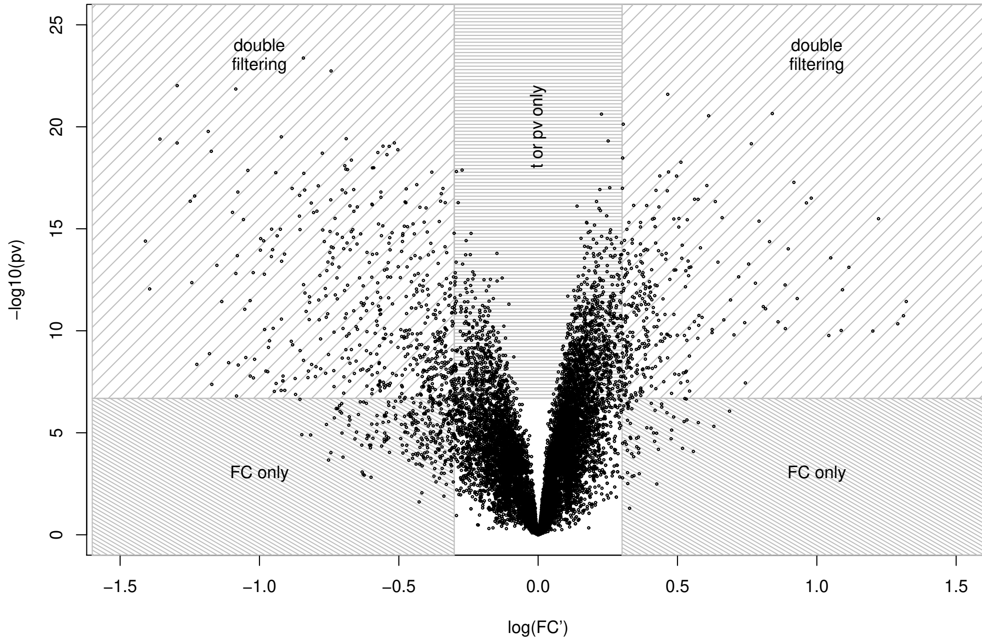

Figure 4: Illustration of the double filtering criterion (upper-left and upper-right corners shaded by sparse lines), FC-only single-gene criterion (lower-left and lower-right corners shaded by dense lines), and t-test-only single-gene criterion ("football goalpost" in the middle shaded by dense horizontal lines).

The basic use of volcano plots is to survey genes that could be selected by one differential expression criterion but not the other. The familiar "double filtering" [108] used by many groups is to set the gene selection criterion by: (i) $|\log_{10} FC| > \log_{10} FC_0$; and (ii) $t > t_0$. Equivalently, it can be defined as (i) $|\log_{10} FC| > \log_{10} FC_0$; and (ii) $p-$value $< p_0$. FC_0, t_0, p_0 are preset threshold values for fold-change, t-statistic, and t-test p-value. The double filtering criterion corresponds to a cutting out of two rectangular corners away from the origin (Fig.4). The single filtering criterion corresponds to delineating (away from origin) regions by horizontal and vertical lines. Then genes chosen by the single *but not* by the double filtering criterion are in the three disjointed regions shaded in Fig.4.

These genes in the shaded area are often not selected for good reasons: (i) genes with large fold-change but nevertheless insignificant test result may be caused by a few outliers with very large values in one group. (ii) genes with significant test result (large t's and small t-test p-values) but low fold change could be false signal due to low variance, which can be caused

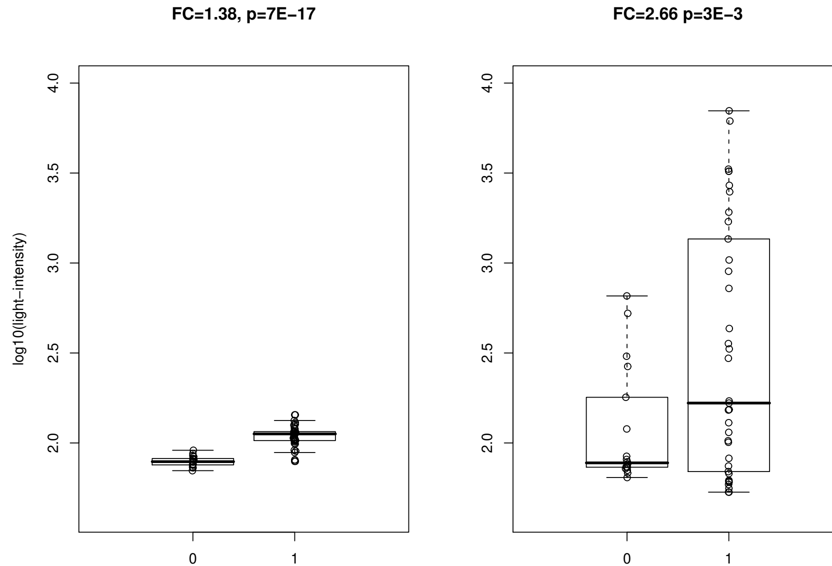

Figure 5: (A) a gene with a significant t-test result (p-value $= 7.7 \times 10^{-17}$) but only moderate fold-change (FC=1.38). (B) a gene with large fold-change (FC=2.66) but weaker t-test significance (p-value= 3×10^{-3}).

by batch effect [109], or low expression level (to be discussed later). A volcano plot allows us to pick some genes from the shaded regions in Fig.4 for further examination.

To understand better the difference between the two single-gene filtering criteria (horizontal and vertical lines in Fig.4), we show two examples of genes selected by the two single filtering criteria in Fig.5. Fig.5(A) is a gene selected by t-test p-value only ($p = 7.7 \times 10^{-17}$) while FC is lower than 2 (FC=1.379). If the true variance is indeed low and we estimated it correctly from 17 control samples, then we trust that this gene is significantly differentially expressed.

On the other hand, The gene in Fig.5(B) is selected by FC only (FC=2.66) whereas the p-value is only 3×10^{-3}. This gene can still be a significantly differential-expression if the large variance in the case group is due to something else, e.g. sub-disease types. Statistical test alone should not be the only foundation for selecting potentially relevant genes, and volcano plot is a way to pick genes that may not lead to the smallest p-values.

Interactively selecting genes in a volcano plot can be done in the statistical package R (*http://www.r-project.org/*). The R function for this purpose is *identify*, which identifies the closest point in a scatter plot to the position clicked by the mouse button. Then information about that point can be printed on screen or in an R session window. Because volcano plot

```
FC <- c(1.03, 2.4, 0.49, 0.6, 1.3, 0.9)
pv <- c(0.5, 3E-7, 2E-4, 5E-3, 0.08, 4E-4)
gname <- c("AAC","ARG1", "CCDC4", "DEFB4", "EIF1", "GNAQ")
x <- log10(FC)
y <- -log10(pv)
plot(x,y)
identify(x,y, n=6, labels=gname)
```

Figure 6: An *R* script illustrating the use of interactive plotting function *identify* in volcano plots. Suppose there are 6 genes whose fold-changes (*FC*) and t-test p-values (*pv*) are given, and the gene names are in *gname*. After the logarithmic transformations, the volcano plot is drawn by *plot(x,y)*.

is usually crowded already, one would prefer to avoid printing long character strings to the screen – a gene name should be often appropriate (human gene names are standardized by HUGO gene nomenclature committee: *http://www.genenames.org/*). An illustrative *R* script for using *identify* is included in Fig.6.

Volcano plot does not show the average expression level of a gene, thought this information can be added using colors (X Hua, X Yan, S Yancopoulos, Y Yang, W Li, "STRAT-VOL: stratified volcano plot for microarray expression analysis", unpublished draft). Nevertheless, the relative magnitude of standard deviation of a gene is provided by the volcano plot, as it is proportional to the tangent of the angle between the point-to-origin line and the y-axis (see Section 5 for more details).

4 Robust variance estimation and regularization

The essential difference between FC and t-statistic is the consideration of statistical noise (variance), but the challenge behind it is how to estimate the variance from a small number of samples [110]. Since variance is calculated around the mean which is also estimated, one idea for robust variance estimation is to iteratively remove outliers then calculate mean and variance [111]. The drawback of this approach is that the number of samples used is further reduced. Artificially increasing the sample size by resampling (Bootstrapping) has been considered [112]. Yet another approach is to use non-parametric tests in place of the t-test (e.g. Mann-Whitney-

Wilcoxon test), so that the variance estimation is not required.

The line of thoughts we pursue for a robust variance estimation is motivated by the typical "large p small n" situation for a microarray experiment [60]. Though the sample size n could be small, the number of genes p is nevertheless large, and that large number of genes make it possible for a reliable estimation of common variance cross all genes [76, 107, 113, 114], at least for the control group.

One main worry about variance estimation is that its value can be low due to the low expression level. To avoid the estimated variance being too low, we may add a constant "penalty" term s_0 to the sample-estimated standard deviation [115] (under a not-so-informative name "SAM" for significance analysis of microarrays):

$$t_{sam} = \frac{\langle \log_{10} E_1 \rangle - \langle \log_{10} E_0 \rangle}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}} + s_0}. \tag{3}$$

The penalty is also called "regularization", reflecting the prior belief (in the Bayesian framework) that variance estimation across different genes should exhibit certain smooth behavior [53, 78].

A popular software package called SAM (Significance Analysis of Microarrays) [116] (*http://www-stat.stanfor* version 4.0, July 2010) is based on Eq.(3). Another *R* implementation of the same idea, *siggenes* [117], is available at
http://www.bioconductor.org/packages/release/bioc/html/siggenes.html. In SAM [116], the s_0 value is chosen to minimize the variability of t_{sam} with respect to the gene-specific standard error term of $\sqrt{s_1^2/n_1 + s_0^2/n_0}$. In [118], s_0 is set at the 90% percentile of standard errors of all genes. In practice, any small value of s_0 can stabilize the variance estimation.

A Bayesian derivation of the extra term in variance estimation is derived in [78]. In this framework, mean, variance of a normal distribution (of $\log_{10}(x) = x'$) has a prior distribution, as well as a posterior distribution after data are observed. For convenience, the inverse Gamma distribution for the variance parameter and the normal distribution for the mean parameter is chosen to ensure both prior and posterior distribution to have the same functional form. It can be shown that (the mean of) posterior variance is a weighted sum of prior variance (σ_0^2)

and the sample-estimated of variance s^2 [78]:

$$E[\sigma^2_{posterior}] = ws^2 + (1-w)\sigma_0^2 \tag{4}$$

where weight w (n is the sample size, ν_0 is the prior degree of freedom for the inverse Gammar distribution):

$$w = \frac{n-1}{\nu_0 + n - 2} \tag{5}$$

tend to close to 1 for larger sample size.

The moderated or regularized variance $\sigma^2_{posterior}$ in Eq.(4) has the effect of drawing gene-specific variance towards the middle, since its change from the sample estimated variance:

$$\sigma^2_{posterior} - s^2 = ws^2 + (1-w)\sigma_0^2 - s^2 = -(1-w)(s^2 - \sigma_0^2), \tag{6}$$

is negative when $s^2 > \sigma_0^2$ and positive when $s^2 < \sigma_0^2$. Note that in Eq.(4), it is the variance that is additive, whereas it is standard error that is additive in the denominator of Eq.(3). However, the idea of moderation/regularization by adding an extra positive and constant term to the sample-estimated one is the same.

In fact, there is a second extra term in variance estimation if the sample-estimated mean is not a good estimate of the true mean [78]. For this reason, it is reasonable to consider removing outliers to make sure the mean is estimated robustly [111].

5 Regularized t-statistic as a joint filtering criterion

What is the relationship between robust variance estimation or regularization discussed in the last section and the volcano plot? If FC can be considered to be the special case when variances of all genes are equal, t-statistic of course contains gene-specific variance, then t_{sam} in Eq.(3) is somewhere in-between [108]. Rewrite $|\langle \log_{10} E_1 \rangle - \langle \log_{10} E_0 \rangle|$ as δ (log-fold-change), $\sqrt{s_1^2/n_1 + s_0^2/n_0}$ as s (standard error), the regularized t-statistic in Eq.(3) can be split into two terms [108]:

$$t_{sam} = \frac{\delta}{s+s_0} = \frac{1}{2(s+s_0)} \cdot \delta + \frac{s}{2(s+s_0)} \cdot \frac{\delta}{s} \tag{7}$$

In other words, t_{sam} is a weighted sum of $\log_{10}(FC')$ and t-statistic, $t_{sam} = a\delta + b(\delta/s)$, where $a = 0.5/(s+s_0)$, $b = 0.5s/(s+s_0)$.

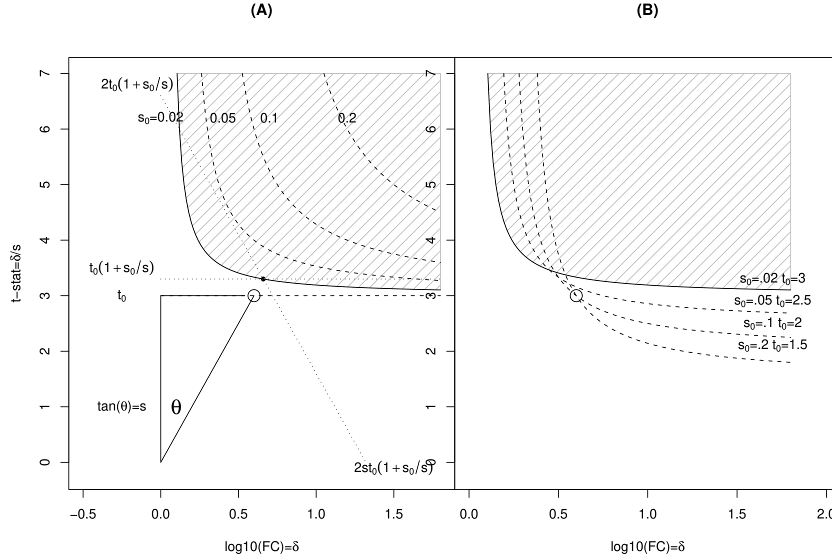

Figure 7: Illustration of the regularized t-statistic (t_{sam}) in volcano plot. For a gene with $x = \delta = \log_{10}(FC') = 0.6$ and $y = \delta/s = t = 3$ (or $s = 0.2$, $\theta = 5.655^o$), the forced linear line (under conflicting limits) for $s_0 = 0.02$ is shown. Also shown are the discriminant lines $t_{sam} \geq t_0 = 3$ at $s_0 = 0.02, 0.05, 0.1, 0.2$. (B) Decreasing t_0 when s_0 is increased: $s_0 = 0.02, t_0 = 3$, $s_0 = 0.05, t_0 = 2.5$, $s_0 = 0.1, t_0 = 2$, and $s_0 = 0.2, t_0 = 1.5$,

Eq.(7) might suggest that t_{sam} is a linear combination of $\log_{10}(FC')$ and t, and the gene filtering criterion $t_{sam} \geq t_0$ discriminant line is a straight line in the volcano plot. However, this geometric interpretation is incorrect. The first hint comes from the fact that the split of t_{sam} into two terms in Eq.(7) can also be carried out for t itself: $t = (1/2s)\delta + (1/2)t$. This is apparently paradoxical as $t \geq t_0$ without regularization should be the plane above the line of $y = t_0$, without a contribution from the x-axis. The second hint is from the observation that the coefficients of "linear function" (a and b) are not constants, but function of the variables themselves.

The third hint can be seen if you want to draw an actual discriminant straight line: the y-intercept is obtained in the limit of $\delta \to 0$, $s \to 0$, but $\delta/s > 0$. Since $x/y = \delta/(\delta/s) = s$, the standard error s has a simple geometric meaning as $\tan(\theta)$ where θ is the angle between the y-axis and the line linking the point and the origin. The above $s \to 0$ limit corresponds to the point to move closer to the y-axis. Similarly, in order to obtain the x-intercept, the limits to be taken are $\delta/s \to 0$, $\delta > 0$, and $s \to \infty$. This is the limit for the point to move away from the y-axis to infinity. Interestingly, under these two conflicting limits, both y- and x-intercept

can be obtained: y-intercept equal to $2t_0(1+s_0/s)$, x-intercept equal to $2st_0(1+s_0/s)$ (see Fig.7(A)).

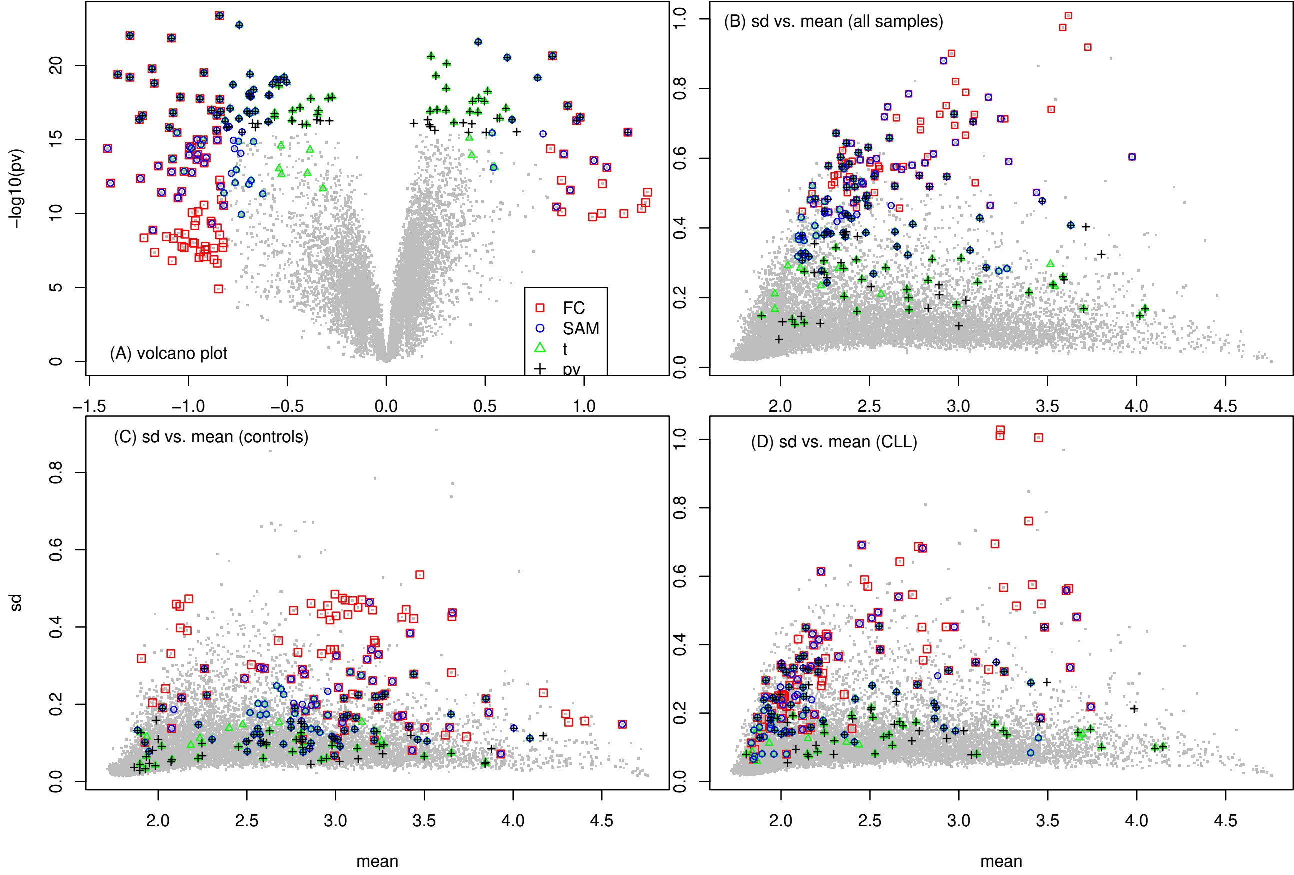

Figure 8: Green, red, green, black dots are the top 100 probes/genes selected by t_{sam}, FC', t-statistic, and p-value of t-test. (A) on volcano plot, x: $\log_{10}(FC')$, y: $-\log_{10}(p\text{-value})$. (B) x: mean of all samples, y: standard deviation of all samples. (C) x: mean of control samples, y: standard deviation of control samples. (D) x: mean of diseased samples, y: standard deviation of diseased samples.

The correct decomposition of t_{sam} splits it into t and s:

$$t_{sam} = \frac{\delta}{s+s_0} = \frac{\delta}{s}(1+\frac{s_0}{s})^{-1} \geq t_0$$

$$\text{or,} \quad t \geq t_0(1+\frac{s_0}{s}) \tag{8}$$

In other words, the discriminant line is a curve which moves up for smaller s's (smaller angles, smaller FC's). A large t-statistic but a small FC (the t-filtering only area in Fig.4) is more

difficult to pass the filtering in Eq.(8). And a large FC without a minimum t-statistic (t_0) (the FC-filtering only area in Fig.4) would not pass the filtering no matter what. These are very different conclusions when compared to the linear discriminant line illustrated in Fig.7(A) where a large FC but a small t may still be selected.

Fig.7 also illustrates the effect of s_0. Besides the discriminant lines at $s_0 = 0.02$, three more lines are shown at $s_0 = 0.05, 0.1$, and 0.2, or 20%, 50%, and 100% of s. We may increase s_0 while decrease t_0 at the same time so that these lines are similar, as shown in Fig.7(B). Under the condition that the same number of top-ranking genes are selected, the exact value of s_0 is less important than the fact that this term is added ($s_0 > 0$), though Fig.7(B) does show that with a larger s_0 value, more genes with less significance but larger FCs are selected.

Fig.8(A) compares the top 100 genes selected by SAM (regularized t) (blue) with those selected by FC (red), t-test p-value (black), and t-statistic itself (green). Although there are certain overlaps among different selection criteria, SAM is able to pick up genes that are not selected by either FC or t-test p-value alone. To address the question on whether t-test criterion tends to select genes with low variance and low expression level. Fig.8(B)(C)(D) show the standard deviation (y-axis) vs. mean (x-axis) for all samples, control samples only, and diseased (CLL) samples only. Indeed, FC-based criterion tend to select genes with high variances, t-test based criterion selects relatively low variance genes, and SAM achieves a balance between the two criteria, selecting genes with intermediate variance values. On the other hand, there is no strong evidence that any selection criterion tends to select low expression level genes.

6 Relevant Bioconductor programs

There are many commercial microarray data analysis programs that include volcano plots. There are also many general graphic packages that intend to handle large number of points, such as *ggplot2* [119]. To make our discussion managable, we limit our summary to Bioconductor programs. Bioconductor site (version 2.10) is a major repository of microarray analysis softwares written in *R* (*http://www.r-project.org/*) [120, 121]. Table 1 lists packages that are relevant to the discussions in this review, roughly grouped into three types:

- *volcano plots:* These are straightforward implementations of the scatter plot, with x-axis usually the log-fold-change and y-axis any other measure of differential expression. We list packages not only for analyzing fluorescence-light intensity-based mRNA expression data (Affymetrix, Illumina, etc.), but also for mRNA expression level based on count data (RNA-seq), and protein expression levels, etc.

- *alternative measures of differential expression:* The fold-change and t-statistic (or $-\log_{10}(p$-value), or the regularized t-statistic (SAM), are not the only measures of differential expression. One large group of alternative measures is the Bayesian calculation of the posterior probability that a gene belong to the differential expression subset (e.g. empirical Bayes analysis of microarray (EBAM)). The packages *DEDS* [122] and *GeneSelector* [83], in particular, include large number of these measures (F-statistic, B-statistic, moderated-F, moderated-t, shrinkage-t, etc.).

- *improvement on error/variance estimation:* Robust and reliable variance estimations are at the heart of the dichotomy choices between t and FC. Some functions in Bioconductor packages directly address this issue, and are listed in Table 1. For example, using variation among replicated samples, using variance between similar probesets, pooling errors, removing outliers, etc.

7 Stratified volcano plots by external information

Volcano plot is a 2-dimensional graphic tool, with potentially interesting genes scattered outward away from the origin. We can make volcano plots even more useful by coloring points with external information. If that external piece of information is relevant to differential expression, we can easily recognize the fact by a visual impression of the plot. This coloring of a volcano plot can be called “stratified volcano plot”. One example is to label all probes/genes that belong to a particular pathway, cellular component, function, or process coded in gene ontology (GO) categories [124].

Fig.9 illustrates a stratified volcano plot by marking 1614 probes/genes that are located on chromosome 6 (red), and 31 probes/genes whose annotation contains the word “cytokine”.

package	functions	comments
a4[131]	volcanoPlot, topTable, limmaTwoLevels	A4 for "automatic Affymetrix array analysis"
ABarray	doPlotFCT, doLPE	AB for "Applied Biosystems"
		FCT for "fold-change and t-statistic"
		DEDS for "differential expression via distance synthesis"
ArrayTools	selectSigGene	double filtering criterion
baySeq[91]	plotPosteriors	Bayesian, RNA-seq
cummeRbund	csVolcano	
DEDS[122]	deds.stat/deds.stat.linkC	include: t, F, FC, SAM, modt, modF, B
		DEDS for "differential expression via distance synthesis"
		SAM for "significance analysis of microarrays"
DEGseq[92]	samWrapper	RNA-seq
DESeq[90]	estimateDispersion, nbinomTest	RNA-seq
diffGeneAnalysis[132]	biasAdjust	
GeneSelector[83]	RankingBaldiLong/Ebam/FC/FoxDimmic,	EBAM for "empirical Bayes analysis of microarrays"
	FoxDimmic/Limma/Permutation/Sam,	the function names preceded by "Ranking"
	ShrinkageT/SoftthresholdT/Tstat	
	WelchT/WilcEbam/Wilcoxon,	
limma[133]	lmFit, eBayes, volcanoplot, topTable	Bayesian
maanova[134]	volcano	
maDB	drawVolcanoPlot	
nudge[135]	nudge1	Bayesian
oneChannelGUI	dfMAPlot	draw from limma's topTable
pickgene[136]	pickgene	
plgem[137]	plgem.obsStn, plgem.deg	PLGEM for "power law global error model"
		STN for "signal-to-noise (ratio)"
		DEG for "differentially expressed genes"
PLPE[138]	lpe.paired	protein level. PLPE for "paired local pooled error"
plw[139]	plw, topRankSummary	PLW for "probe-level locally-moderated weighted (t-test)"
puma[140]	pumaDE, calculateLimma, topGenes	PU for "propagating uncertainty"
RankProd[141]	RP, RPadvance, topGene	RP for "rank product"
SAGx[142]	samrocN	
samr		SAM. not distributed through bioconductor
siggenes[117]	sam, d.stat, ebam	SAM, Bayesian
XDE[143]	xde, calculateBayesianEffectSize	Bayesian
xps[17]	plotVolcano	XPS for "eXpression Profiling System"

Table 1:

From the stratified volcano plot, we can easily identify interesting candidate genes involving cytokines such as CLCF1 (cardiotrophin-like cytokine factor 1, p-value= 1.4×10^{-16}, FC'=0.22), SOCS2 (suppressor of cytokine signaling 2, FC'= 0.11, p-value= 3.8×10^{-8}), SOCS3 (suppressor of cytokine signaling 3, FC'=0.28, p-value= 6.2×10^{-8}), etc. The visual impression immediately shows the top-ranking cytokine-linked genes are all down-regulated instead of up-regulated.

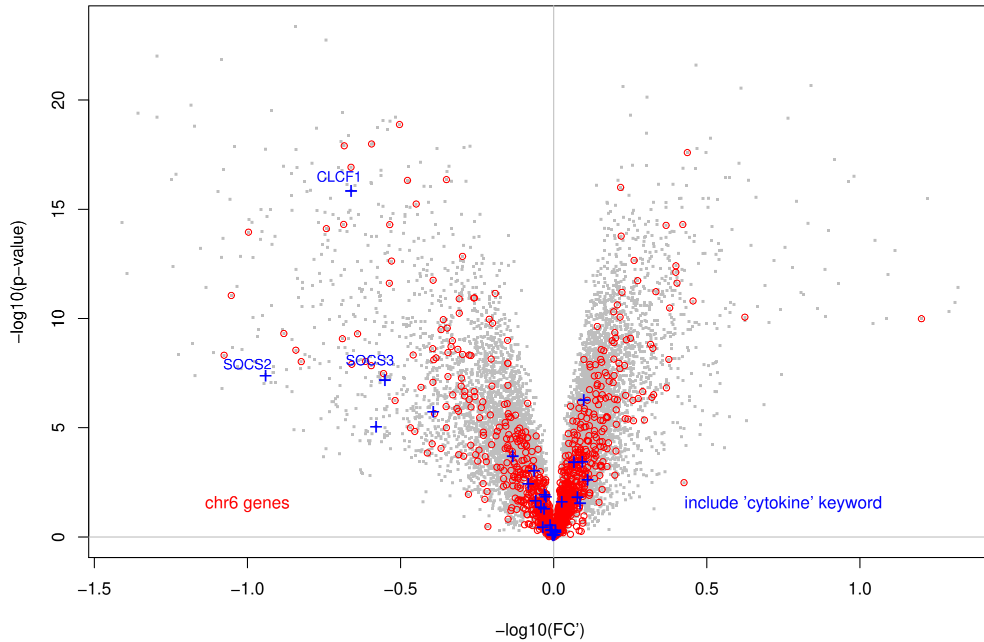

Figure 9: Stratified volcano plot: probes/genes on chromosome 6 are marked by red, and those with "cytokine" in gene annotation is marked by blue.

8 Discussion and conclusions

Finding effect measure of differential expression remains an active research topic [123]. However, the idea of regularization (adding a small positive term to the observed standard error to standardize the differential expression signal) has already successfully combined the two most well known quantities, log-fold-change and t-statistic, in gene filtering. And volcano plot is a natural visual tool to illustrate this procedure.

Simultaneous displaying of noise-level-standardized signal and unstandardized one can also be useful beyond the field of microarray. In genetic association studies, the association signal

of a single-nucleotide polymorphism (SNP) is usually measured by two quantities. One is the odds-ratio (OR) of the 2-by-2 count table with disease status as row and two alleles as column. OR is not standardized by the noise level or sample size, though the 95% confidence interval of OR does become narrower for larger sample sizes thus lower level of chance events [125]. On the other hand, the chi-square statistic or the p-value of the chi-square (χ^2) test strongly dependent on sample size, thus chance event probability. In fact, the chi-square statistics is proportional to the total number of samples for a SNP that contains association signals.

Besides using OR in x-axis (in log scale), another choice is to use the allele frequency difference in case and control group. Denote the four counts in the 2-by-2 table (row for case control status, columns for presence of absence of a particular allele/genotype) in case-control association analysis are a, b, c, d, $\log_{10}$OR is $\log_{10}(ad) - \log_{10}(bc)$, whereas allele frequency difference is $a/(a+b) - c/(c+d) = (ad - bc)(a+b)^{-1}(c+d)^{-1}$. In other words, the difference between the two choices is whether ad and bc are compared in the logarithmic or regular scale.

It is rare for volcano plots being applied to genetic association studies (some examples can be found in [126, 127]). We believe that many extensions and applications of volcano plots in microarray analysis can be equally useful in genetic association analysis. For example, the joint filtering criterion, the stratified volcano plot coloring external pieces of information, and uncovering of systematic patterns when points are colored by other information. We have found that the location of a SNP on the volcano plot is intrinsically related to its minor allele frequency. This will provide further insight on how one should balance the chi-square test result and odds-ratio in selecting genetically associated genes.

In conclusion, volcano plot, together with heatmaps [128], MA plots [129], and cluster/PCA plots [109, 130], is among the most useful and most frequently used visual tools in microarray analysis, Volcano plots display both noise-level-standardized and unstandardized signal concerning differential expression of mRNA levels. Regularized test statistic and joint filtering have an intuitive geometric interpretation in volcano plot, and its advantage over double filter criterion of genes can be easily understood. As a scattering plot, volcano plot can incorporate other external information, such as gene annotation, to aid the hypothesis generating process concerning a disease or phenotype.

Acknowledgements

I would like to thank Frank Batliwalla, Percio Gulko, Max Brenner, Peter Gregersen for asking about differences between genes selected by fold-change and those by t-test, which provided the initial incentive for writing this review, Joy Yan, Sophia Yancopoulos for discussion on volcano plots and allowing me to use the CLL data, Yaning Yang, Xing Hua for collaboration on software development, and Michaela Oswald for reading a draft of the paper.

References

[1] M. Schena, R.A. Heller, T.P. Theriault, K. Konrad, E. Lachenmerier, R.W. Davis "Microarrays: biotechnology's discovery platform for functional genomics", *Trends in Biotech.* **16**,301-306 (1998).

[2] R.A. Young, "Biomedical discovery with DNA arrays", *Cell* **102**, 9-15 (2000).

[3] A. Butte , "The use and analysis of microarray data", *Nat. Rev. Drug Discovery* **1**, 951-960 (2002).

[4] D.K. Slonim , "From patterns to pathways: gene expression data analysis comes of age", *Nat. Genet.* **32(suppl)**, 502-508 (2002).

[5] R.B. Stoughton , "Application of DNA microarrays in biology", *Ann. Rev. Biochem.* **74**, 53-82 (2005).

[6] V. Trevino, F. Falciani, H.A. Barrera-Saldana, "DNA Microarrays: a powerful genomic tool for biomedical and clinical research", *Mol. Med.* **13**, 527-541 (2007).

[7] A.J. Trachtenberg, J.H. Robert, A.E. Abdalla, A. Fraser, S.Y. He, J.N. Lacy, C. Rivas-Morello, A. Truong, G. Hardiman, L. Ohno-Machado, F. Liu, E. Hovig, W.P. Kuo, "A primer on the current state of microarray technologies", *Methods in Mol. Biol.* **802**, 3-17 (2012).

[8] W. Etienne, M.H. Meyer, J. Peppers, R.A. Meyer Jr., "Comparison of mRNA gene expression by RT-PCA and DNA microarray" *BioTechniques* **36**, 618-626 (2004).

[9] P.B. Dallas, N.G. Gottardo, M.J. Firth, A.H. Beesley, K. Hoffmann, P.A. Terry, J.R. Freitas, J.M. Boag, A.J. Cummings, U.R. Kees, "Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR – how well do they correlate?", *BMC Genomics* **6**, 59 (2005).

[10] J.S. Morey, J.C. Ryan, F.M. van Dolah, "Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR", *Biological Procedures Online* **8**, 175-193 (2006).

[11] P.J. Park, Y.A. Cao, S.Y. Lee, J.W. Kim, M.S. Chang, R. Hart, S. Choi, "Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference", *J. Biotech.* **112**, 225-245 (2004).

[12] J.E. Larkin, B.C. Frank, H. Gavras, R. Sultana, J. Quackenbush, "Independence and reproducibility across microarray platforms", *Nat. Methods* **2**, 337-344 (2005).

[13] R.A. Irizarry, D. Warren, F. Spencer, et al., "Multiple-laboratory comparison of microarray platforms", *Nat. Methods* **2**, 345-350 (2005).

[14] S. Draghici, P. Khatri, A.C. Eklund, Z. Szallasi, "Reliability and reproducibility issues in DNA microarray measurements", *Trends in Genet.* **22**, 101-109 (2006).

[15] W.P. Kuo, F. Liu, J. Trimarchi, et al., "A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies", *Nat. Biotech.*, **24**, 832-840 (2006).

[16] T.A. Patterson, E.K. Lobenhofer, S.B. Fulmer-Smentek, et al., "Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project", *Nat. Biotech.* **24**, 1140-1150 (2006).

[17] J.J. Chen, H.M. Hsueh, R.R. Delongchamp, C.J. Lin, C.A. Tsai, "Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data", *BMC Bioinf.* **8**, 412 (2007).

[18] Z. Wen, Z. Su, J. Liu, B. Ning, L. Guo, W. Tong, L. Shi, "The MicroArray Quality Control (MAQC) project and cross-platform analysis of microarray data", in *Handbook of Statistical Bioinformatics*, eds. H Horng-Shing, B Schölkopf, H Zhao, **2011**, pp.171-192 (2011).

[19] Z. Li, J.C. Kwekel, T. Chen , "Functional comparison of microarray data across multiple platforms using the method of percentage of overlapping functions", *Methods in Mol. Biol.* **802**, 123-139 (2012).

[20] J.D. Allen, S. Wang, M. Chen, L. Girard, J.D. Minna, Y. Xie, G. Xiao, "Probe mapping across multiple microarray platforms", *Brief. Bioinf.*, in press (2012).

[21] G.A. Churchill, "Fundamentals of experimental design for cDNA microarrays", *Nat. Genet.* **32**, 490-495 (2002).

[22] C. Li, A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods", *Biostat.* **8**, 118-127 (2007).

[23] K.A. Baggerly, K.R. Coombes, E.S. Neeley, "Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer" *J. Clinical Oncology* **26**, 1186-1187 (2008).

[24] R.R. Kitchen, V.S. Sabine, A.H. Sims, E.J. Macaskill, L. Renshaw, J.S. Thomas, J.I. van Hemert, J.M. Dixon, J.M.S. Bartlett, "Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles", *BMC Genomics* **11**, 134 (2010).

[25] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, C. Liu, "Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods", *PLoS ONE* **6**, e17238 (2011).

[26] J.A. Gagnon-Bartsch, T.P. Speed, "Using control genes to correct for unwanted variation in microarray data", *Biostat.* **13**, 539-552 (2012).

[27] J.P.A. Ioannidis, "Microarrays and molecular research: noise discovery?", *Lancet* **365**, 454-455 (2005).

[28] J.M. Raser, E.K. O'Shea, "Noise in gene expression: origins, consequences, and control", *Science* **309**, 2010-2013 (2005).

[29] L. Ein-Dor, O. Zuk, E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer", *Proc. Nat. Acad. Sci.* **103**, 5923-5928 (2006).

[30] N. Maheshri, E.K. O'Shea, "Living with noisy genes: how cells function reliably with inherent variability in gene expression", *Annu. Rev. Biophy. Biomol. Struct.* **36**, 413-434 (2007).

[31] A. Zeisel, A. Amir, WJ Köstler, E. Domany, "Intensity dependent estimation of noise in microarrays improves detection of differentially expressed genes", *BMC Bioinf.* **11**, 400 (2010).

[32] R. Thomas, L. de la Torre, X. Chang, S. Mehrotra, "Validation and characterization of DNA microarray gene expression data distribution and associated moments", *BMC Bioinf.* **11**, 576 (2010).

[33] RR Kitchen, VS Sabine, AA Simen, J.M. Dixon, J.M.S. Bartlett, A.H. Sims, "Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments", *BMC Genomics* **12**, 589 (2011).

[34] A. Posekany, K. Felsenstein, P. Sykacek, "Biological assessment of robust noise models in microarray data analysis", *Bioinf.* **27**, 807-814 (2011).

[35] V.T.Y. Tang, H. Yan, "Noise reduction in microarray gene expression data based on spectral analysis", *Int. J. Machine Learning and Cybernetics* **3**, 51-57 (2012).

[36] V. Sharov, K.Y. Kwong, B. Frank, E. Chen, J. Hasseman, R. Gaspard, Y. Yu, I. Yang, J. Quackenbush, "The limits of log-ratios", *BMC Biotech.* **4**, 3 (2004).

[37] Y.H. Yang, T. Speed , "Design issues for cDNA microarray experiments", *Nat. Rev. Genet.* **3**, 579-588 (2002).

[38] L. Shi, W. Tong, F. Goodsaid, F.W. Frueh, H. Fang, T. Han, J.C. Fuscoe, D.A. Casciano, "QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies", *Expert Rev. Mol. Diagnostics* **4**, 761-777 (2004).

[39] L. Shi, L.H. Reid, W.D. Jones, et al. MAQC Consortium (2006), "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements", *Nat. Biotech.* **24**, 1151-1161 (2006).

[40] M.N. McCall, P.N. Murakami, M. Lukk, W. Huber, R.A. Irizarry, "Assessing affymetrix GeneChip microarray quality", *BMC Bioinf.* **12**, 137 (2011).

[41] A. Brazma, P. Hingamp, J. Quackenbush, et al., "Minimum information about a microarray experiment (MIAME) – toward standards for microarray data", *Nat. Genet.* **29**, 365-371 (2001).

[42] J.P.A. Ioannidis, D.B. Allison, C.A. Ball, et al., "Repeatability of published microarray gene expression analyses", *Nature Genet.* **41**, 149-155 (2009).

[43] J. Quackenbush, "Microarray data normalization and transformation", *Nat. Genet.* **32**, 496-501 (2002).

[44] J. Vandesompele, K de Preter, F. Pattyn, B. Poppe, N. van Roy, A. de Paepe, F. Speleman, "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes", *Genome Biol.* **3**, 34 (2002).

[45] A. Fujita, J.R. Sato, L. de Oliveira Rodrigues, C.E. Ferreira, M.C. Sogayar, "Evaluating different methods of microarray data normalization", *BMC Bioinf.* **7**, 469 (2006).

[46] C. Steinhoff, M. Vingron, "Normalization and quantification of differential expression in gene expression microarrays", *Brief. Bioinf.* **7**, 166-177 (2006).

[47] P. Stafford, ed. *Methods in Microarray Normalization* (CRC Press, 2008).

[48] R. Autio, S. Kilpinen, M. Saarela, O. Kallioniemi, S. Hautaniemi, J. Astola, "Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations", *BMC Bioinf.* **10(suppl 1)**, S24 (2009).

[49] J. Önskog, E. Freyhult, M. Landfors, P. Rydén, T.R. Hvidsten, "Classification of microarrays; synergistic effects between normalization, gene selection and machine learning", *BMC Bioinf.* **12**, 390 (2011).

[50] R.L. Halpert, S. Sanga, "Robust unattended microarray analysis", Ingenuity Technical Report No. 2011-1 (Ingenuity Systems, Inc. 2012).

[51] K Estrada, A Abuseiris, FG Grosveld, A.G. Uitterlindern, T.A. Knoch, F. Rivadeneira, "GRIMP: a web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data", *Bioinf.* **25**, 2750-2752 (2009).

[52] E.E. Schadt, M.D. Linderman, J. Sorenson, L. Lee, G.P. Nolan, "Computational solutions to large-scale data management and analysis", *Nat. Rev. Genet.* **11**, 647-657 (2010).

[53] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, 2001).

[54] J.D. Storey, R. Tibshirani, "Statistical significance for genomewide studies", *Proc. Nat. Acad. Sci.* **100**, 9440-9445 (2003).

[55] J.D. Storey, "The positive false discovery rate: a Bayesian interpretation and the q-value", *Annals of Statistics* **31**, 2013-2035 (2003).

[56] A. Reiner, D. Yekutieli, Y. Benjamini, "Identifying differentially expressed genes using false discovery rate controlling procedures", *Bioinf.* **19**, 368-375 (2003).

[57] Y. Pawitan, S. Michiels, S. Koscielny, A. Gusnanto, A. Ploner, "False discovery rate, sensitivity and sample size for microarray studies", *Bioinf.* **21**, 3017-3024 (2005).

[58] A. Schwwartzman, X. Lin, "The effect of correlation in false discovery rate estimation", *Biometrika* **98**, 199-214 (2011).

[59] E.P. Xing, M.I. Jordan, R.M. Karp, "Feature selection for high-dimensional genomic microarray data", in *Proc. Eighteenth Int. Conf. Machine Learning (ICML 2001)*, eds. CE Brodley, AP Danyluk, pp. 601-608 (Morgan Kauffmann, 2001).

[60] W. Li, Y. Yang, "How many genes are needed for a discriminant microarray data analysis" in *Methods of Microarray Data Analysis* eds. SM Lin and KF Johnson, pp.137-150 (Kluwer Academic, 2002).

[61] C. Ambroise, G.J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data", *Proc. Nat. Acad. Sci.* **99**, 6562-6566 (2002).

[62] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection", *J. Machine Learning Res.* **3**, 1157-1182 (2003).

[63] H. Peng, F. Long, C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", *IEEE Trans. Pattern Analysis and Machine Learning* **27**, 1226-1238 (2005).

[64] W. Li, "The-more-the-better and the-less-the-better", *Bioinf.* **22**, 2187-2188 (2006).

[65] J.G. Liao. K.V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large p and small n case", *Bioinf.* **23**, 1945-1951 (2007).

[66] C. Zhao, M.L. Bittner, R.S. Chapkin, E.R. Dougherty, "Characterization of the effectiveness of reporting lists of small feature sets relative to the accuracy of the prior biological knowledge", *Cancer Informatics* **9**, 49-60 (2010).

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

[68] I. Hedenfalk, D. Duggan, Y. Chen, et al. "Gene-expression profiles in hereditary breast cancer", *New England J. Medicine* **344**, 539-548 (2001).

[69] S.M. Dhanasekaran, T.R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K.J. Pienta, M.A. Rubin, A.M. Chinnaiyan, "Delineation of prognostic biomarkers in prostate cancer", *Nature* **412**, 822-826 (2001).

[70] T.R. Adib, S. Henderson, C. Perrett, D. Hewitt, D. Bourmpoulia, J. Ledermann, C. Boshoff, "Predicting biomarkers for ovarian cancer using gene-expression microarrays", *Brit. J. Cancer* **90**, 686-692 (2004).

[71] T.J. Yeatman, "Predictive biomarkers: identification and verification", *J. Clinical Oncology* **27**, 2743-2744 (2009).

[72] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, et al. "Prediction of central nervous system embryonal tumour outcome based on gene expression", *Nature* **415**, 436-442 (2002).

[73] M.J. van de Vijver, Y.D. He, L.J. van't Veer, et al., "A gene-expression signature as a predictor of survival in breast cancer", *New England J. Medicine* **347**, 1999-2009 (2002).

[74] H. Colman, L. Zhang, E.P. Sulman, et al. "A multigene predictor of outcome in glioblastoma", *Neuro-Oncology* **12**, 49-57 (2010).

[75] C. Kim, S. Paik, "Gene-expression-based prognostic assays for breast cancer", *Nat. Rev. Clinical Oncology* **7**, 340-347 (2010).

[76] N. Jain, J. Thatte, T. Braciale, K. Ley, M. O'Connell, J.K. Lee, "Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays", *Bioinf.* **19**, 1945-1951 (2003).

[77] Y. Chen, E.R. Dougherty, M.L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images" *J. Biomedical Optics* **2**, 364-274 (1997).

[78] P. Baldi, A.D. Long , "A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes", *Bioinf.* **17**, 509-519 (2001).

[79] L. Shi, W. Tong, H. Fang, U. Scherf, J. Han, R.K. Puri, F.W. Frueh, F.M. Goodsaid, L. Guo, Z. Su, T. Han, J.C. Fuscoe, Z.A. Xu, T.A. Patterson, H. Hong, Q. Xie, R.G. Perkins, J.J. Chen, D.A. Casciano, "Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedures are essential", *BMC Bioinf.* **6(suppl 2)**, S12 (2005).

[80] L. Guo, E.K. Lobenhofer, C. Wang, R. Shippy, S.C. Harris, L. Zhang, N. Mei, T. Chen, D. Herman, F.M. Goodsaid, P. Hurban, K.L. Phillips, J. Xu, X. Deng, Y.A. Sun, W. Tong, Y.P. Dragan, L. Shi, "Rat toxicogenomic study reveals analytical consistency across microarray platforms", *Nat. Biotech.* **24**, 1162-1169 (2006).

[81] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, Y. Li, "Apparently low reproducibility of true differential expression discoveries in microarray studies", *Bioinf.* **24**, 2057-2063 (2008).

[82] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, Q. Liu, J. Wang, D. Wang, C. Wang, Z. Guo, "Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes", *Bioinf.* **25**, 1662-1668 (2009).

[83] A.L. Boulesteix, M. Slawski, "Stability and aggregation of ranked gene lists", *Brief. Bioinf.* **10**, 556-568 (2009).

[84] D.M. Witten, R. Tibshirani, "A comparison of fold-change and the t-statistic for microarray data analysis", Department of Statistics, Stanford University technical report (2007).

[85] R.B. O'Hara, D.J. Kotze, "Do not log-transform count data", *Math. Ecol. Evol.* **1**, 118-122 (2010).

[86] G.T. Fechner, *Elemente der Psychophysik* (Leipzig: Breitkopf und Härtel, 1860).

[87] P.L. Auer, S. Srivastava, R.W. Doerge, "Differential expression – the next generation and beyond", *Brief. Bioinf.* **11**, 57-62 (2011).

[88] M.D. Robinson, D.J. McCarthy, G.K. Smyth, "`edgeR`: a Bioconductor package for differential expression analysis of digital gene expression data", *Bioinf.* **26**, 139-140 (2010).

[89] A. Oshlack, M.D. Robinson, M.D. Young, "From RNA-seq reads to differential expression results", *Genome Biol.* **11**, 220 (2010).

[90] S. Anders, W. Huber, "Differential expression analysis for sequence count data", *Genome Biol.* **11**, R106 (2010).

[91] T.J. Hardcastle, K.A. Kelly, "baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data", *BMC Bioinf.* **11**, 422 (2010).

[92] L. Wang, Z. Feng, X. Wang, X. Wang, X. Zhang, "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data", *Bioinf.* **26**, 136-138 (2010).

[93] J.S. Cumbie, J.A. Kimbrel, Y. Di, D.W Schafer, L.J. Wilhelm, S.E. Fox, C.M. Sullivan, A.D. Curzon, J.C. Carrington, T.C. Mockler, J.H. Chang "GENE-Counter: a computational pipeline for the analysis of RNA-seq data for gene expression differences", *PLoS ONE* **6**, e25279 (2011).

[94] J. Lee, Y. Ji, S. Liang, G. Cai, P. Müller, "On differential gene expression using RNA-Seq data", *Cancer Inform.* **10**, 205-215 (2011).

[95] Z. Chen, J. Liu, H.K.T. Ng, S. Nadarajah, H.L. Kaufman, J.Y. Yang, Y. Deng, "Statistical methods on detecting differentially expressed genes for RNA-seq data", *BMC Sys. Biol.* **5(suppl 3)**, S1 (2011).

[96] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, A. Conesa, "Differential expression in RNA-seq: a matter of depth", *Genome Res.* **21**, 2213-2223 (2011).

[97] D.M. Witten, "Classification and clustering of sequencing data using a Poisson model", *Ann. Appl. Stat.* **5**, 2493-2518 (2011).

[98] V.M. Kvam, P. Liu, Y. Si, "A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data", *Am. J. Botany* **99**, 248-256 (2012).

[99] J. Li, D.M. Witten, I.M. Johnsone, R. Tibshirani, "Normalization, testing, and false discovery rate estimation for RNA-sequencing data", *Biostat.* **13**, 523-538 (2010).

[100] G.W. Snedecor and W.G. Cochran, *Statistical Methods*, eighth edition (Iowa State University Press: Ames, IW). *BMC Biotech.* **4**, 3 (1989).

[101] B.L. Welsh, "The generalization of 'Student's' problem when several different population variances are involved", *Biometrika* **34**, 28-35 (1947).

[102] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition (Lawrence Erlbaum Associates, Inc. 1988).

[103] Y. Tu, G. Stolovitzky, U. Klein, "Quantitative noise analysis for gene expression microarray experiments", *Proc. Nat. Acad. Sci.* **99**, 14031-14036 (2002).

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

[105] W. Jin, R.M. Riley, R.D. Wolfinger, K.P. White, G. Passador-Gurgel, G. Gibson, "The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster", *Nat. Genet.* **29**, 389-395 (2001).

[106] W.G. Alvord, J.A. Roayaei, O.A. Quiñones, K.T. Schneider, "A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R", *Brief. Bioinf.* **8**, 1-13 (2007).

[107] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments", *Bioinf.* **18**, 546-554 (2002).

[108] S. Zhang, J. Cao, "A close examination of double filtering with fold change and t test in microarray analysis", *BMC Bioinf.* **10**, 402 (2009).

[109] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, R.A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data", *Nat. Rev. Genet.* **11**, 733-739 (2010).

[110] Z. Chen, Q. Liu, M. McGee, M. Kong, X. Huang, Y. Deng, "A gene selection method for GeneChip array data with small sample sizes", *BMC Genomics*, 12(suppl 5):S7 (2011).

[111] I. Dozmorov, I. Lefkovits , "Internal standard-based analysis of microarray data. Part 1: analysis of differential gene expressions", *Nucleic Acids Res.* **37**, 6323-6339 (2009).

[112] U.M. Braga-Neto, E.R. Dougherty , "Is cross-validation valid for small-sample microarray classification?", *Bioinf.* **20**, 374-380 (2004).

[113] X. Cui, J.T.G. Hwang, J. Qiu, N.J. Blades, G.A. Churchill, "Improved statistical tests for differential gene expression by shrinking variance components estimates", *Biostat.* **6**, 59-75 (2005).

[114] G.W. Wright, R.M. Simon, "A random variance model for detection of differential gene expression in small microarray experiments", *Bioinf.* **19**, 2448-2455 (2003).

[115] V.G. Tusher, R. Tibshirani, G. Chu , "Significance analysis of microarrays applied to the ionizing radiation response", *Proc. Nat. Acad. Sci.* **98**, 5116-5121 (2001).

[116] G. Chu, B. Narasimhan, R. Tibshirani, V. Tusher, *SAM: "significance analysis of microarrays", users guide and technical document*, v.3.0 (2007).

[117] H. Schwender, A. Krause, K. Ickstadt, "Identifying interesting genes with siggenes", *RNews* **6**, 45-50 (2006).

[118] B. Efron, R. Tibshirani, J.D. Storey, V. Tusher, "Empirical Bayes analysis of a microarray experiment", *J. Am. Stat. Assoc.* **96**, 1151-1160 (2001).

[119] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).

[120] R.C. Gentleman, V.J. Carey, D.M. Bates, et al., "Bioconductor: open software development for computational biology and bioinformatics", *Genome Biol.* **5**, R80 (2004).

[121] eds. R. Gentleman, W. Huber, V.J. Carey, R.A. Irizarry, S. Dudoit, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (Springer, 2005).

[122] Y.H. Yang, Y. Xiao, M.R. Segal, "Identifying differentially expressed genes from microarray experiments via statistic synthesis", *Bioinf.* **21**, 1084-1093 (2005).

[123] Y. Xiao, T.H. Hsiao, U. Suresh, H.I.H. Chen, X. Wu, S.E. Wolf, Y. Chen, "A novel significance score for gene selection and ranking", *Bioinf.*, in press (2012).

[124] M. Ashburner, C.A. Ball, J.A. Blake, et al., "Gene Ontology: tool for the unification of biology" *Nat. Genet.* **25**, 25-29 (2000).

[125] W. Li, "Three lectures on case-control genetic association analysis", *Brief. Bioinf.* **9**, 1-13 (2008).

[126] M. Sirota, M.A. Schaub, S. Batzoglou, W.H. Robinson, A.J. Butte, "Autoimmune disease classification by inverse association with SNP alleles", *PLoS Genet.* **5**, e1000792 (2009).

[127] K Miclaus, M Chierici, C. Lambert, L. Zhang, S. Vega, H. Hong, S. Yin, C. Furlanello, R. Wolfinger, F. Goodsaid, "Variability in GWAS analysis: the impact of genotype calling algorithm inconsistencies", *The Pharmacogenomics J.* **10**, 324-335 (2010).

[128] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proc. Nat. Acad. Sci.* **95**, 14863-14868 (1998).

[129] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngal, T.P. Speed, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation", *Nucleic Acids Res.* **30**, e15 (2002).

[130] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C.M. Perou, J.S. Marron, "Adjustment of systematic microarray data biases", *Bioinf.* **20**, 105-114 (2004).

[131] H. Göhlmann, W. Talloen, *Gene Expression Studies Using Affymetrix Microarrays* (Chapman & Hall/CRC, 2009).

[132] I. Dozmorov, M. Centola, "An associative analysis of gene expression array data", *Bioinf.* **19**, 204-211 (2003).

[133] G.K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments", *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).

[134] H. Wu, M. Kerr, X. Cui, G. Churchill, "MAANOVA: a software package for the analysis of spotted cDNA microarray experiments", in *The Analysis of Gene Expression Data Methods and Software* eds. G Parmigiani, E Garrett, R Irizarry, S Zeger, pp.313-341 (Springer, 2003).

[135] N. Dean, A.E. Raftery, "Normal uniform mixture differential gene expression detection for cDNA microarrays", *BMC Bioinf.* **6**, 173 (2005).

[136] Y. Lin, S.T. Nadler, H. Lan, A.D. Attie, B.S. Yandell, "Adaptive geen picking with microarray data: detecting important low abundance signals", in *The Analysis of Gene Expression Data Methods and Software* eds. G Parmigiani, E Garrett, R Irizarry, S Zeger, pp.291-312 (Springer, 2003).

[137] N. Pavelka, M. Pelizzola, C. Vizzardelli, M. Capozzoli, A. Splendiani, F. Granucci, P. Ricciardi-Castagnoli, "A power law global error model for the identification of differentially expressed genes in microarray data", *BMC Bioinf.* **5**, 203 (2004).

[138] H Cho, DM Smalley, MM Ross, D. Theodorescu, K. Ley, J.K. Lee, "Statistical identification of differentially labelled peptides from liquid chromatography tandem mass spectrometry", *Proteomics* **7**, 3681-3692 (2007).

[139] M. Åstrand, *Normalization and Differential Gene Expression Analysis of Microarray Data* (Ph.D Thesis, Department of Mathematics, Chalmers University of Technology and Göteborg University, 2008).

[140] R.D. Pearson, X. Liu, G. Sanguinetti, M. Milo, N.D. Lawrence, M. Rattray, "puma: a Bioconductor package for propagating uncertainty in microarray analysis", *BMC Bioinf.* **10**, 211 (2009).

[141] R. Breitling, P. Armengaul, A. Amtmann, P. Herzyk, "Rank products: a simple, yet powerful new method to detect differentially regulated genes in replicated microarray experiments", *FEBS Lett.* **573**, 83-92 (2004).

[142] P. Broberg, "Statistical methods for ranking differentially expressed genes", *Genome Biol.* **4**, R41 (2004).

[143] R.B. Scharpf, H. Tjelmeland, G. Parmigiani, A.B. Nobel, "A Bayesian model for cross-study differential gene expression", *J Am. Stat. Assoc.* **104**, 1295-1310 (2009).