OPEN ACCESS Freely available online

PLoS one

Should We Abandon the *t*-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies

Marine Jeanmougin[1,2,3,4]*, Aurelien de Reynies[1], Laetitia Marisa[1], Caroline Paccard[2], Gregory Nuel[3], Mickael Guedj[1,2]

1 Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, Paris, France, **2** Department of Biostatistics, Pharnext, Paris, France, **3** Department of Applied Mathematics (MAP5) UMR CNRS 8145, Paris Descartes University, Paris, France, **4** Statistics and Genome Laboratory UMR CNRS 8071, University of Evry, Evry, France

Abstract

High-throughput post-genomic studies are now routinely and promisingly investigated in biological and biomedical research. The main statistical approach to select genes differentially expressed between two groups is to apply a *t*-test, which is subject of criticism in the literature. Numerous alternatives have been developed based on different and innovative variance modeling strategies. However, a critical issue is that selecting a different test usually leads to a different gene list. In this context and given the current tendency to apply the *t*-test, identifying the most efficient approach in practice remains crucial. To provide elements to answer, we conduct a comparison of eight tests representative of variance modeling strategies in gene expression data: Welch's *t*-test, ANOVA [1], Wilcoxon's test, SAM [2], RVM [3], limma [4], VarMixt [5] and SMVar [6]. Our comparison process relies on four steps (gene list analysis, simulations, spike-in data and re-sampling) to formulate comprehensive and robust conclusions about test performance, in terms of statistical power, false-positive rate, execution time and ease of use. Our results raise concerns about the ability of some methods to control the expected number of false positives at a desirable level. Besides, two tests (limma and VarMixt) show significant improvement compared to the *t*-test, in particular to deal with small sample sizes. In addition limma presents several practical advantages, so we advocate its application to analyze gene expression data.

Citation: Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, et al. (2010) Should We Abandon the *t*-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. PLoS ONE 5(9): e12336. doi:10.1371/journal.pone.0012336

Editor: Kerby Shedden, University of Michigan, United States of America

Received April 8, 2010; **Accepted** June 24, 2010; **Published** September 3, 2010

Copyright: © 2010 Jeanmougin et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by the Ligue Nationale Contre le Cancer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: marine.jeanmougin@genopole.cnrs.fr

Introduction

During the last decade, advances in Molecular Biology and substantial improvements in microarray technology have led biologists toward high-throughput genomic studies. In particular, the simultaneous measurement of the expression levels of tens of thousands of genes has become a mainstay of biological and biomedical research.

The use of microarrays to discover genes differentially expressed between two or more groups (patients *versus* controls for instance) has found many applications. These include the identification of disease biomarkers that may be important in the diagnosis of the different types and subtypes of diseases, with several implications in terms of prognostic and therapy [7,8].

A first approach to identify differentially expressed genes is known as the Fold-Change estimation (FC). It evaluates the average log-ratio between two groups and considers as differentially expressed all genes that differ by more than an arbitrary cut-off. So defined, FC lacks of a solid statistical footing [9]: it does not take the variance of the samples into account. This point is especially problematic since variability in gene expression measurements is partially gene-specific, even after the variance has been stabilized by data transformation [10,11].

Rather than applying a FC cutoff, one should prefer statistical tests: they standardize differential expression by considering their variance [9,12]. Furthermore, corresponding effect sizes, confidence intervals and *p*-values are essential information for the control of false-positives [13] and meta-analysis [14].

The *t*-test is certainly the most popular test and has been matter of discussion. Computing a *t*-statistic can be problematic because the variance estimates can be skewed by genes having a very low variance. These genes are associated to a large *t*-statistic and falsely selected as differentially expressed [2]. Another drawback comes from its application on small sample sizes which implies low statistical power [12]. Consequently, the efficacy of a *t*-test along with the importance of variance modeling have been seriously called into question [15]. It has led to the development of many innovative alternatives, with hope of improved variance estimation accuracy and power.

These alternatives appear very diverse at a first sight, but fall into few nested categories relying on both statistical and biological hypotheses: parametric or non-parametric modeling, frequentist or Bayesian framework, homoscedastic hypothesis (same variance between groups of samples) and gene-by-gene variance estimation. Further propositions come from the field of machine-learning for instance [16], but lie beyond the scope of our study.

A disadvantage of having so many alternatives is that selecting a different test usually identifies a different list of significant genes since each strategy operates under specific assumptions [17]. Moreover, despite the wealth of available methods, the *t*-test remains widely used in gene-expression studies, presumably because of its simplicity and interpretability. Given the tendency to use this method, identifying which approach is the most appropriate to analyze gene expression data remains a crucial issue. Nevertheless, if the development of new methodologies is still an active topic of publication, only few studies have addressed their comparison. This is probably due to the difficulty to implement a realistic framework of comparison for which the differentially expressed genes are known in advance.

In order to sidestep many problems, comparisons frequently rely on the analysis of gene lists resulting from the application of several methods [18] and simulations for which truly differentially expressed genes are known [6]. More empirical alternatives include the use of re-sampling methods (to compare genes from small subsets of samples and those from the full dataset) [3,19], and the use of spike-in data for which a set of genes are differentially expressed by design [12,20]. Finally Jeffery et al. [18] explore an indirect approach by assessing classification performance obtained with genes resulting from the application of the methods to compare. The heterogeneity of the strategies adopted in the literature and the diversity of tests investigated make the formulation of general conclusions difficult. In addition, to our knowledge, no study has focused on the direct comparison of a wide range of variance modeling strategies.

Consequently, we conduct a comparison study of eight tests representative of variance modeling strategies in gene expression data: Welch's *t*-test, ANOVA [1], Wilcoxon's test, SAM [2], RVM [3], limma [4], VarMixt [5] and SMVar [6]. The comparison process relies on four steps: gene list analysis, simulations, spike-in data and re-sampling. Our aim is to benefit from the specificity of each strategy, to make our results comparable to previous studies and to ease the formulation of general, robust and reproducible conclusions.

So defined, we follow a standard statistical framework. First, our main focus concerns the issue of data reduction which relies on the form of the test statistic and impact directly the resulting power. A separate but important issue is calibration (i.e. the accuracy of *p*-values) which can impact the false-positive rate (α). So at each step of the process, tests are compared in terms of statistical power assessed at the same false-positive rate. Control of the false-positive rate to the desired value is checked for each test which is, to our opinion, too rarely considered in the literature. Eventually, in addition to an efficacy comparison, we find relevant to confront each test in terms of practical consideration such as execution time and ease of use.

Methods

Statistical background

Differential analysis consists in testing the null hypothesis (H_0) that the expected values of expression for a given gene are equal between two groups of interest (1 and 2), against the alternative hypothesis (H_1) that they differ. Let Y_{gcr} the level of expression observed for gene g, replicate r, under group c; the general model is then given by:

$$\mathbb{E}(Y_{gcr}) = \mu_{gc} \quad \text{and} \quad Var(Y_{gcr}) = \sigma^2_{gc}$$

So defined, the null hypothesis to test comes down to:

$$\begin{cases} H_0: & \mu_{g1} = \mu_{g2} \\ H_1: & \mu_{g1} \neq \mu_{g2} \end{cases}$$

Given a statistical test, type-I error-rate α (resp. type-II error-rate β) commonly refers to the probability to reject (resp. accept) H_0, H_0 being true (resp. false). The statistical power of the test is then defined as the ability to reject H_0 when it is actually false:

$$\begin{aligned} \text{Power}(\alpha) &= \mathbb{P}_{H_1}(H_0 \text{ rejected at the } \alpha \text{ level}) \\ &= 1 - \mathbb{P}_{H_1}(H_0 \text{ not rejected at the } \alpha \text{ level}) \\ &= 1 - \beta \end{aligned}$$

Type-I and II errors are inversely related: the smaller the risk of one, the higher the risk of the other. Consequently the power depends directly on α, and a valid comparison of several tests has to be driven at the same type-I error-rate to overcome the issue of calibration.

The type-I error-rate is often referred to as false-positive rate. It differs from the false-discovery rate (FDR) in the sense that it represents the rate that truly null features are called significant whereas the FDR is the rate that significant features are truly null [21].

Selection of the eight tests

This selection has focused on tests broadly applied in the literature and representative of different variance modeling strategies. The eight tests selected are described in detail in Methods S1 and re-implemented in R to simplify their application. The package is available on demand.

Briefly, most of the eight tests are parametric and estimate a gene-by-gene variance: ANOVA (homoscedastic), Welch's *t*-test (heteroscedastic), RVM (homoscedastic), limma (homoscedastic and based on a Bayesian framework) and SMVar (heteroscedastic and based on structural model); we also select two non-parametric approaches with the Wilcoxon's test and the SAM test, which do not rely on assumptions that the data are drawn from a given probability distribution.

Besides, variances estimated on a set of genes are thought to lead to an undesirable amount of false-positives. Attributing a common variance to all the genes is clearly not a solution, even when sample sizes are small. Several proposals make the assumption that genes with the same expression level have approximatively the same variance [22,23]. However this is not realistic and also leads to false-positives [24]. We find VarMixt more subtle: it makes the assumption that classes of genes can be identified based on similar response to the various sources of variability (mixture model); the variance of each homogeneous class is then accurately estimated from a large set of observations; the individual gene variance is then replaced by its "class" variance.

Comparison process

Gene list analysis. An intuitive first step to compare the tests is to investigate the consistency between gene lists resulting from the application of each test on real data. Here we apply this approach to five publicly available data sets (Table 1) to assess the overlap between gene lists and to identify similar behaviors among the variance modeling strategies.

Table 1. Data sets used for the gene list analysis.

Data-set	Groups	Sample size	Publication
Lymphoid tumors	Disease staging	37	Lamant et al. 2007 [26]
Liver tumors	TP53 mutation	65	Boyault et al. 2007 [27]
Head and neck tumors	Gender	81	Rickman et al. 2008 [28]
Leukemia	Gender	104	Soulier et al. 2006 [29]
Breast tumors	ESR1 expression	500	Bertheau et al. 2007 [30]

The five data sets come from the *Cartes d'Identité des Tumeurs* (CIT, http://cit.ligue-cancer.net) program and are publicly available. All the microarrays are Affymetrix U133A microarrays with 22,283 genes.
doi:10.1371/journal.pone.0012336.t001

In addition to the eight tests, we define a "control" test that draws for each gene a *p*-value from a Uniform distribution between 0 and 1. Then, we applied the tests to the five data-sets to identify gene differentially expressed by setting a *p*-value threshold of 0.05.

Gene list similarities between tests are analyzed and visualized using a Hierarchical Clustering (binary metric and the Ward's aggregation algorithm, R package *stats*) and Principal Component Analysis (R package *ade4* [25]). For more details please refer to Methods S1 and Table S1.

Simulation study. The purpose of simulations is to estimate power and false-positive rate on a large range of simulated data sets, in order to compare the tests under simple and sometimes extreme situations. We define a reference model (denoted M_1), frequently adopted in the literature and that matches the assumptions of the *t*-test. Under M_1, gene expressions for the groups 1 and 2 are drawn from Gaussian distributions of same variance ($\sigma = 1$):

$$\begin{cases} Y_{g1r} & \sim & \mathcal{N}(\mu_{g1}, \sigma^2) \\ Y_{g2r} & \sim & \mathcal{N}(\mu_{g2}, \sigma^2) \end{cases}$$

Under H_0: $\{\mu_{g1} = \mu_{g2}\}$ while under H_1: $\{\mu_{g2} = \mu_{g1} + \delta\}$, with $\delta = 0.5$.

Then, we propose three extensions of M_1 (denoted M_2, M_3 and M_4) designed to be less to the *t*-test advantage. M_2 is quite similar but expression levels are now drawn from a Uniform distribution of same parameters. M_3 applies a mixture model on variances and corresponds to the VarMixt hypothesis; genes are then divided into three classes of variance. Under M_4, 10% of the genes are simulated with small variances ($\sigma^2 = 0.05$) since they can lead to an increase of false-positive rate when the *t*-test is applied.

For each model we simulate 10,000 independent genes under H_0 to assess the false-positive rate attached to each test, and 10,000 under H_1 to compute their respective power. False-positive rate and power are both assessed at a *p*-value threshold of 0.05. Sample size ranges from 5 to 100 samples per group. The simulated data matrix is given Figure 1.

Spike-in data set. The Human Genome U133 data set is used to test and validate microarray analysis methods (http://www.affymetrix.com). The data set consists in 14 hybridizations of 42 spiked transcripts in a complex human background at concentrations ranging from 0.125 pM to 512 pM. Each group includes three replicates. We perform the 13 pairwise comparisons for which "spike-in" genes have a true fold-change of two [5].

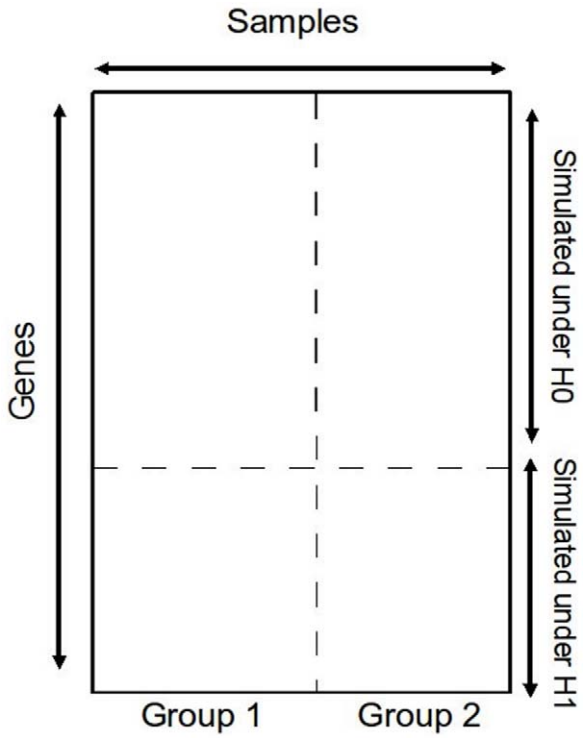

Figure 1. Data matrix resulting from simulations. Rows refer to genes simulated under H_0 and H_1, columns refer to samples of both groups to compare.
doi:10.1371/journal.pone.0012336.g001

The whole data set contains 22,300 genes. The 42 spike-in genes are designed to be differentially expressed (under H_1) and used for power estimation. To be able to compute the false-positive rate, the 22,258 remaining genes are forced to be under H_0 by permutation of the group labels. False-positive rate and power are both assessed at a *p*-value threshold of 0.05.

Re-sampling approach. The main idea is to assess the ability of a test to select from small subsets of samples ($n = 5$ and $n = 10$), genes determined as differentially expressed from the full data set. The strategy can be summarized in four steps:

Step 1: From the 500 samples data set (Table 1) split into two groups to compare, we define a set of differentially expressed genes (*p*-value $\leq 10^{-4}$ with the Welch's *t*-test). This set is considered in Step 3 as the "truth" to estimate power.

Step 2: n samples are drawn from each group and the eight tests are performed on this subset of the initial data. We apply the Benjamini and Hochberg correction at a 0.1 FDR level [31].

Step 3: From Step 2 we estimate power as the proportion of genes defined as differentially expressed at Step 1 and detected at Step 2.

Step 4: Steps 2 and 3 are iterated 1,000 times. Finally power is averaged over the 1,000 iterations.

Results

Gene list analysis

Figure 2 represents PCAs and dendrograms resulting from gene list analysis. The cumulative inertia explained by the two first axes of PCA is about 80%. Both representations underline the same tendencies.

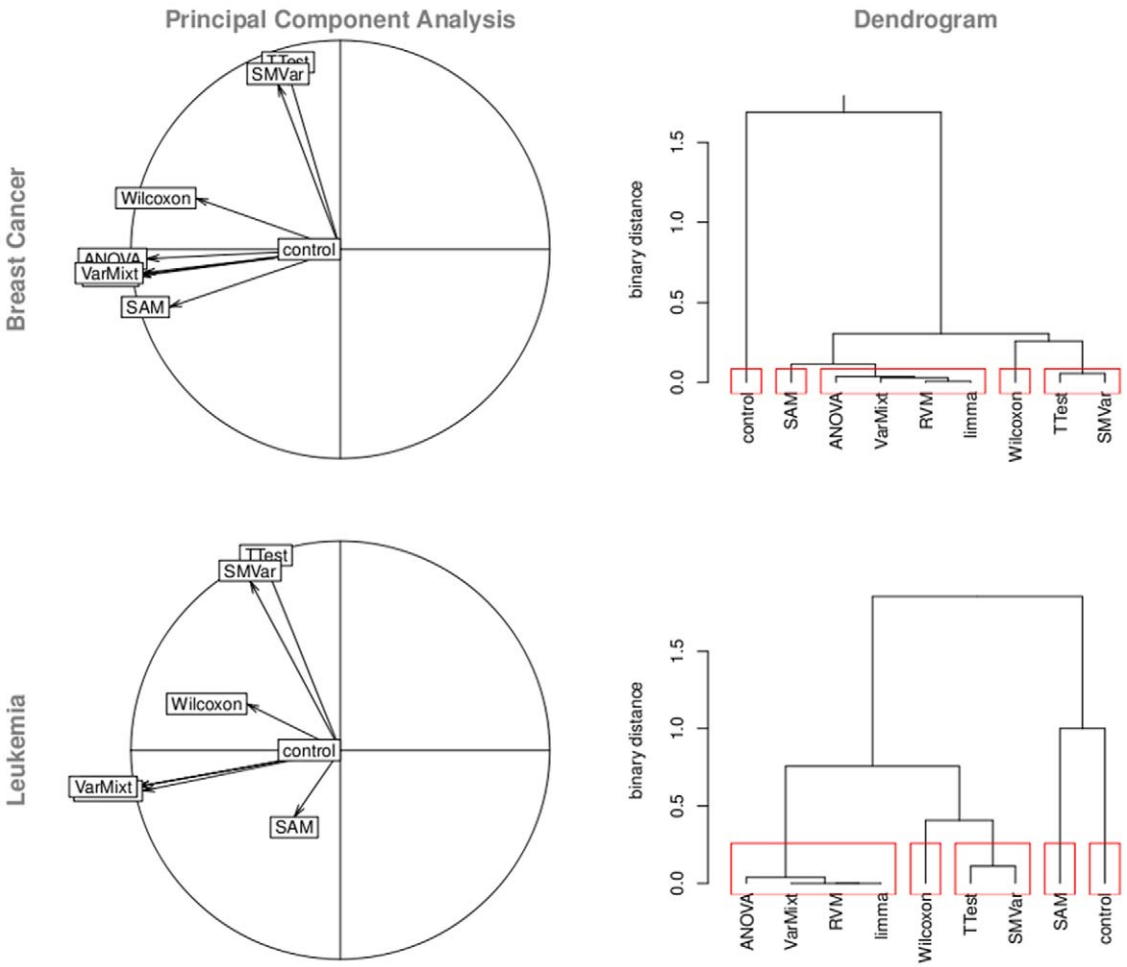

Figure 2. Gene list analysis. PCAs and dendrograms are generated based on the gene lists resulting from the application of the eight tests of interest and the control-test. Here we show results for two data sets comparing ESR1 expression in breast cancer and gender in leukemia. Both outline five clusters of tests.
doi:10.1371/journal.pone.0012336.g002

As expected, gene lists resulting from the control-test are clearly independent from the other ones, since it selects genes (differentially expressed or not) uniformly. Then, the eight tests show various behaviors. Six tests clusterize in two distinct groups: {*t*-test; SMVar} and {VarMixt; limma; RVM; ANOVA}. The proportion of common genes selected by two tests of the same cluster is about 90%. On the other hand, Wilcoxon and SAM do not clearly fall in one of the two main groups: Wilcoxon tends to consistently lie between them, whereas SAM does not present a reproducible behavior.

To summarize, homoscedastic (VarMixt, limma, RVM and ANOVA), heteroscedastic (*t*-test and SMVar) variance modeling strategies are well discriminated by a similarity analysis of gene lists. It outlines the interesting property that similar modeling strategies in theory imply similar results in practice.

Simulation study

First, we evaluate power according to sample size under the simulation model M_1 (Figure 3). On Figure 3-A, we notice little difference between the tests (less than 0.08), particularly for large samples as expected. Wilcoxon is not as good as the other tests in most cases. SAM and ANOVA show equivalent performance to the *t*-test. VarMixt, RVM and limma tend to provide an increase in power, and SMVar slightly outperforms all the tests (Figures 3-A and B).

As we know, these preliminary results are valid only if all the tests meet the theoretical 5% false-positive rate when applying a *p*-value threshold of 0.05. Table 2 gives the observed false-positive rate for each test under small and large sample sizes and sheds light on the fact that some tests clearly deviate from the 5% level and return biased *p*-values. Observed deviations are more accentuated for small sample sizes compared to large ones. SMVar and RVM inflate the expected number of false-positives whereas Wilcoxon and the *t*-test tend to be conservative; ANOVA, SAM, limma and VarMixt show no deviation.

Regarding these observations, the tests inefficient to control the false-positive rate at the expected 5% level have to be adjusted by a time consuming Monte-Carlo procedure. Figures 3-C and D present power results at adjusted and hence valid false-positive rates. Differences are clearly reduced compared to Figures 3-A and B which confirms that part of the difference in power observed is due to actual difference in false-positive rate, particularly concerning SMVar. After adjustment VarMixt, RVM and limma tend to be the best tests although they provide an insignificant gain compared to the *t*-test; Wilcoxon remains the less powerful. ANOVA has performance comparable to the *t*-test which is

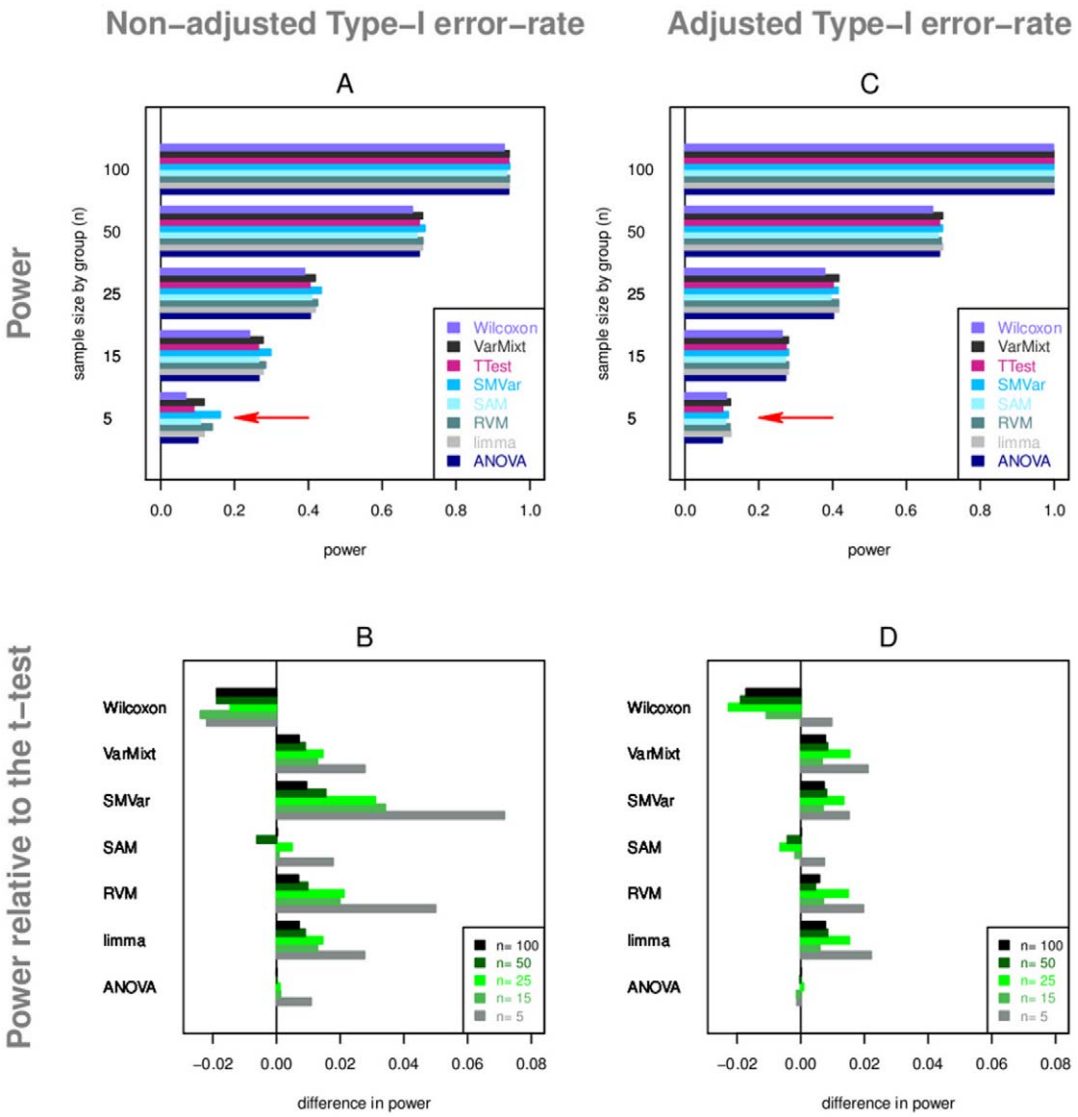

Figure 3. Power study from simulations (Gaussian model, M1). Power values are calculated at the 5% level and displayed according to the sample size. Figures A and C represent power values. Red arrows highlight the effect of false-positive rate adjustment on power values. Figures B and D represent power values relative to t-test. Figures A and B concern power values calculated at the actual false-positive rate. Figures C and D concern power values calculated at the adjusted false-positive rate.
doi:10.1371/journal.pone.0012336.g003

interesting: under the same variance between the two groups, tests that make the corresponding homoscedastic assumption (ANOVA) do not show improved power compared to heteroscedastic ones (Welch *t*-test).

Surprisingly, model M_2 leads to the same conclusions (data not shown). Here expression values follow a Uniform distribution instead of a Gaussian one, which does not match the assumption of parametric approaches. Compared to model M_1, we were expecting to note a more striking increase in power for Wilcoxon, which is not observed. This result confirms that *t*-test and assimilated approaches are quite robust to the Gaussian assumption. Indeed the Central Limit Theorem implies that even if expression values are not Gaussian, the *t*-statistic resulting from the comparison of two groups is likely to be. It should be noticed that the structural model of SMVar is not able to provide results for the uniform model.

Finally models M_3 and M_4 also lead to the same conclusions, with an overall loss of power (data not shown).

Spike-in data set

Spike-in data confirm observations and conclusions made on the simulations. SMVar and RVM inflate the expected number of false-positives whereas Wilcoxon and the *t*-test tend to be conservative. Power values adjusted to a valid false-positive rate present more significant differences than in simulations (Figure 4): with an average decrease of almost 0.6, Wilcoxon is the less powerful and similar to the ''control'' test; ANOVA shows equivalent performance than the *t*-test; VarMixt, RVM, SMVar and limma provide a significant increase in power with an average gain of 0.25. With performance comparable to the best tests, SAM has a different behavior than in simulations.

Table 2. False-positive rate study from simulations.

	M1		M2		M3		M4	
Sample size	**n = 5**	**n = 100**	**n = 5**	**n = 100**	**n = 5**	**n = 100**	**n = 5**	**n = 100**
t-test▼	3.8 – 4.6	4.5 – 5.4	4.0 – 4.8	4.6 – 5.5	3.8 – 4.6	4.7 – 5.6	3.9 – 4.7	4.4 – 5.3
ANOVA	4.5 – 5.2	4.5 – 5.4	4.7 – 5.6	4.6 – 5.5	4.5 – 5.4	4.7 – 5.6	4.5 – 5.3	4.4 – 5.3
Wilcoxon▼	2.8 – 3.5	4.6 – 5.5	2.6 – 3.3	4.5 – 5.4	2.8 – 3.5	4.7 – 5.6	2.7 – 3.4	4.5 – 5.4
SAM	4.6 – 5.5	4.5 – 5.3	4.2 – 5.1	4.5 – 5.4	4.7 – 5.6	4.7 – 5.6	4.3 – 5.2	4.4 – 5.3
RVM▲	5.7 – 6.7	4.5 – 5.4	5.6 – 6.5	4.5 – 5.4	5.4 – 6.3	4.7 – 5.6	5.3 – 6.2	4.7 – 5.5
limma	4.6 – 5.5	4.6 – 5.5	4.2 – 5.1	4.5 – 5.4	4.7 – 5.6	4.7 – 5.6	4.4 – 5.3	4.3 – 5.1
SMVar▲	7.0 – 8.1	4.7 – 5.6	–	–	5.9 – 6.8	4.8 – 5.7	4.6 – 5.5	4.5 – 5.3
VarMixt	4.7 – 5.5	4.6 – 5.5	4.3 – 5.2	4.6 – 5.5	4.8 – 5.6	4.6 – 5.5	4.5 – 5.4	4.5 – 5.3

For small and large samples, this table presents the 95% confidence-interval of false-positive rate obtained by applying a threshold of 0.05 to the *p*-values. Up triangles ▲ (resp. down triangles ▼) indicate an increase (resp. a decrease) of the false-positive rate compared to the expected level of 5%. Two triangles inform of a deviation in both small and large sample sizes.

doi:10.1371/journal.pone.0012336.t002

Re-sampling approach

This approach corroborates tendencies obtained with simulations and spike-in data (Figure 5): limma, VarMixt and RVM perform much better than other tests in identifying differentially expressed genes, while SMVar is somewhat less efficient than the three top-tests. ANOVA and the *t*-test still show equivalent performance, although ANOVA presents here a slight but significant improvement.

Wilcoxon and SAM were never able to detect genes determined as differentially expressed. Indeed the calibration performed can not reach *p*-value lower than 10^{-3} for small sample sizes. After the Benjamini-Hochberg correction at a 0.1 FDR level (corresponding here to a 10^{-6} *p*-value threshold), they do not detect any gene as differentially expressed.

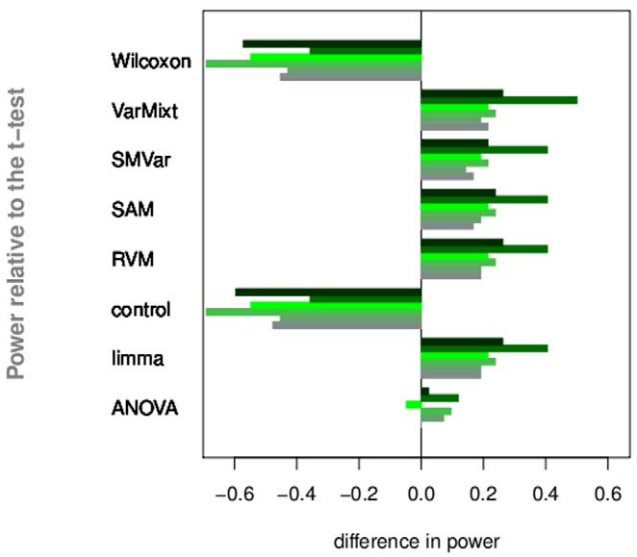

Figure 4. Spike-in data set. Power values are calculated at the 5% level and displayed according to six of the 13 pairwise comparisons.

doi:10.1371/journal.pone.0012336.g004

Practical comparison

Concerning time of execution and ease of use, the *t*-test and ANOVA are the most efficient as they rely on standard statistical considerations and have benefited of improved implementations. On real high-throughput data, both take few seconds to treat tens of thousands of genes. In terms of time of execution, limma appears as efficient as the *t*-test and ANOVA, which is a noteworthy point. SMVar, RVM and SAM run in longer but still reasonable time (up to 8 minutes in our case). Varmixt turns out to be the slowest approach (up to 80 minutes) as it relies on a time consuming EM algorithm.

Discussion

Given the current tendency to apply the *t*-test to gene expression data and the wealth of available alternatives, finding the most appropriate approach to handle differential analysis is critical.

To address this problematic and provide some answers, we develop a comparison process of eight tests for differential expression. It is based on gene list analysis, simulations, spike-in data and re-sampling, with the intention to benefit from the specificity and advantages of each strategy.

Gene list analysis do not properly compare test performance and hence lead to limited conclusions. However it is an appropriate preliminary approach that focuses on similarities between test results. An analysis of the consistency between gene lists outlines general tendencies that can help in interpreting differential analysis results. In our case, we observed comparable results between tests based on similar variance modeling strategies.

The three other approaches (simulations, spike-in data and re-sampling) propose a direct comparison of power values. Simulations represent a convenient statistical framework as genes under H_0 and H_1 are known in advance. In addition different hypotheses on data structure can be specified under different simulation models. Here, the three further models (M_2, M_3 and M_4) lead actually to the same conclusions than the reference Gaussian one (M_1). If simulations do not allow to observe significant differences in power between the tests, they still reveal reproducible tendencies. In addition, simulations turn out to be the gold standard to check possible deviations from the expected false-positive rate. However it is unclear whether simulated data sets can sufficiently and realistically reflect the noise inherent in real microarray data [32].

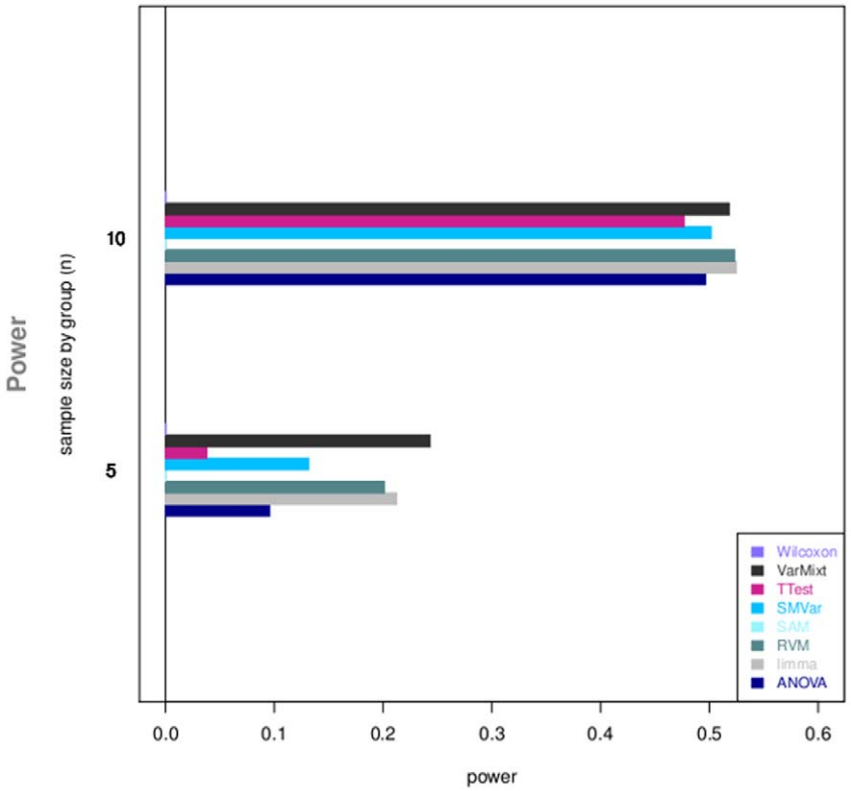

Figure 5. Re-sampling approach. Power values are calculated at a 0.1 FDR level and displayed according to the sample size.
doi:10.1371/journal.pone.0012336.g005

More empirical alternatives include the use of spike-in data and re-sampling. Spike-in genes can represent gene expression better than simulations. In our case it confirms conclusions from simulations with more significant differences in power. Regarding the Affymetrix data set we used, a criticism of this approach could be that the small number of actual spike-in genes does not allow a very accurate power estimation. Moreover variation across technical replicates is likely to be lower than that typically observed across true biological replicates, and many biological effects of interest may be smaller than two-fold [12].

In this context, a re-sampling approach takes advantage of the complexity found in real data. Differentially expressed genes are not known but determined from a large data set (500 samples in our case); power is then evaluated on a subset of the data. Results are comparable to those obtained with simulations and spike-in data. However this approach can be considered as limited in that it assumes that gene lists generated on the full dataset are correct; besides it is fastidious to implement and extremely time consuming.

By applying four distinct comparison strategies with specific advantages and drawbacks: **(i)** we ensure to offset the limitations of each strategy and **(ii)** we provide robust conclusions on test performance.

We applied this comparison process to eight tests representative of different variance modeling strategies. Results are summarized in Table 3. A first important result concerns the control of the false-positive rate, which is often disregarded in the literature. Under H_0, distribution of p-values is supposed to be uniform and the false-positive rate resulting from a p-value threshold of 0.05 should be controlled at 5%. Deviation from this major assumption may indicate biased p-values. In both simulations and spike-in data, some tests deviate from the expected false-positive rate, which partly explains some differences in power (namely SMVar, RVM and Wilcoxon). For the purpose of our study, we performed a Monte-Carlo based adjustment of the false-positive rate to formulate comparable conclusions across all the tests. However in practice this adjustment remains fastidious to implement. In consequence, we strongly advocate to avoid using these tests until a proper corrected version is made available.

Overall, Wilcoxon and SAM show weak performance. One of our simulation model (M_2) clearly outlines the robustness of parametric tests to the Gaussian assumption. Concerning SAM, our results do not allow to formulate clear conclusions and reflect existing doubts about its efficacy [18,33].

Compared to the *t*-test, limma and VarMixt consistently show real improvement, in particular on small sample sizes. Limma has often been discussed in the biostatistical field and its good performance has been reported [12,18,24]. Surprisingly VarMixt does not appear as weak as similar methods evaluated by Kooperberg et al. [24]. Presumably it benefits from a more realistic mixture model on variances, less likely to generate false-positives.

If limma and VarMixt are equivalent regarding both power and false-positive rate, in practice limma presents several further advantages in terms of execution time. In addition, limma can be generalized to more than two groups which makes it relevant to many broader situations.

To conclude, we have developed a comprehensive process to compare statistical tests dedicated to differential analysis. This approach can be used as the basis to evaluate performance of methods developed in the near future. In addition, to answer our initial question "Should we abandon the *t*-test", limma provides a

Table 3. Summary table.

	False-positive rate		Power		In practice	
	Small samples	**Large samples**	**Small samples**	**Large samples**	**Ease of use**	**Execution time**
t-test	+	+++	+	+++	+++	+++
ANOVA	+++	+++	+	+++	+++	+++
Wilcoxon	+	+	+	++	+++	++
SAM	+++	+++	+	++	++	++
RVM	+	++	+++	+++	++	+
limma	+++	+++	+++	+++	++	+++
VarMixt	+++	+++	+++	+++	+	+
SMVar	+	+	++	+++	++	+++

This table summarizes the results of our study in terms of false-positive rate, power and practical criteria. The number of "+" indicates the performance, from weak (+), to very good one (+++).

doi:10.1371/journal.pone.0012336.t003

substantial improvement compared to the *t*-test, particularly for small samples. However the *t*-test remains easy to apply through a wide-range of genomic analysis tools whereas limma can appear more difficult to implement at a first sight. To promote its application we make available on demand a simplified R version of limma dedicated to the analysis of two groups of samples.

Supporting Information

Methods S1 A detailed description of (i) the eight tests included in the study and (ii) the gene list analysis process.
Found at: doi:10.1371/journal.pone.0012336.s001 (0.09 MB PDF)

Table S1 Example of binary matrix. For a given test, the genes identified as differentially expressed (''1'') and not differentially expressed (''0'') at a given p-value threshold are reported in the binary matrix.
Found at: doi:10.1371/journal.pone.0012336.s002 (0.01 MB PDF)

Acknowledgments

We thank Emilie Thomas, Laure Vescovo, Anne-Sophie Valin, Fabien Petel, Renaud Schiappa, Matthieu Bouaziz and Antoine Canu for helpful discussions. We also thank Jacqueline Metral, Jacqueline Godet and Christophe Ambroise for their support.

Author Contributions

Conceived and designed the experiments: MJ GN MG. Performed the experiments: MJ. Analyzed the data: MJ MG. Wrote the paper: MJ MG. Implemented tools in R: MJ, AdR. Significantly contributed to the paper: AdR, LM, CP, GN.

References

1. Kerr M, Martin M, Churchill G (2000) Analysis of variance for gene expression microarray data. Journal of Computational Biology 7: 819–837.
2. Tusher V, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America 98: 5116–5121.
3. Wright G, Simon R (2003) A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics 19: 2448–2455.
4. Smyth G (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 3: doi: 10.2202/1544–6115.1027.
5. Delmar P, Robin S, Daudin J (2005) Varmixt: efficient variance modelling for the differential analysis of replicated gene expression data. Bioinformatics 21: 502–508.
6. Jaffrezic F, Marot G, Degrelle S, Hue I, Foulley J (2007) A structural mixed model for variances in differential gene expression studies. Genetics Research 89: 19–25.
7. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron J, et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America 100: 8418–8423.
8. Van 't Veer L, Dai J, Van de Vijver M, He Y, Hart A, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.
9. Allison D, Cui X, Page G, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics 7: 55–65.
10. Zhou L, Rocke D (2005) An expression index for Affymetrix GeneChips based on the generalized logarithm. Bioinformatics 21: 3983–3989.
11. Simon R, Radmacher M, Dobbin K, McShane L (2003) Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute 95: 14–18.
12. Murie C, Woody O, Lee A, Nadon R (2009) Comparison of small n statistical tests of differential expression applied to microarrays. BMC Bioinformatics 10: 45.
13. Dudoit S, Popper Shaffer J, Boldrick J (2002) Multiple hypothesis testing in microarray experiments. UC Berkeley Division of Biostatistics Working Paper Series. Available: http://www.bepress.com/ucbbiostat/paper110.
14. Marot G, Foulley J, Mayer C, Jaffrezic F (2009) Moderated effect size and P-value combinations for microarray meta-analyses. Bioinformatics 25: 2692–2699.
15. Mary-Huard T, Picard F, Robin S Introduction to statistical methods for microarray data analysis: 56126.
16. Pirooznia M, Yang J, Yang M, Deng Y (2008) A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics 9: S13.
17. Yauk C, Berndt M (2007) Review of the literature examining the correlation among dna microarray technologies. Environmental and Molecular Mutagenesis 48: 380–394.
18. Jeffery I, Higgins D, Culhane A (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 7: 359.
19. Sullivan Pepe M, Longton G, Anderson G, Schummer M (2003) Selecting differentially expressed genes from microarray experiments. Biometrics 59: 133–142.
20. McCall M, Irizarry R (2008) Consolidated strategy for the analysis of microarray spike-in data. Nucleic Acids Research Advance 36: e108.
21. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100: 9440–9445.
22. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, et al. (2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. Bioinformatics 19: 1945–1951.

23. Huang X, Pan W (2002) Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. Functional and Integrative Genomics 2.
24. Kooperberg C, Aragaki A, Strand A, Olson J (2005) Significance testing for small microarray experiments. Statistics in medicine 24: 2281–2298.
25. Chessel D, Dufour A, Thioulouse J (2004) The ade4 package - I : One-table methods. R News 4: 5–10.
26. Lamant L, de Reynies A, Duplantier M, Rickman D, Sabourdy F, et al. (2007) Gene-expression profiling of systemic anaplastic large-cell lymphoma reveals differences based on ALK status and two distinct morphologic ALK+ subtypes. Blood 109: 2156–2164.
27. Boyault S, Rickman D, de Reynies A, Balabaud C, Rebouissou S, et al. (2007) Transcriptome classification of hcc is related to gene alterations and to new therapeutic targets. Hepatology 45.
28. Rickman D, Millon R, De Reynies A, Thomas E, Wasylyk C, et al. (2008) Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays. Oncogene 27: 6607–6622.
29. Soulier J, Clappier E, Cayuela J, Regnault A, Garcia-Peydro M, et al. (2005) HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). Blood 106: 274–286.
30. Bertheau P, Turpin E, Rickman D, Espie M, de Reynies A, et al. (2007) Exquisite sensitivity of *TP53* mutant and basal breast cancers to a dose-dense epirubicin-cyclophosphamide regimen. PLoS Med 4: e90.
31. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289–300.
32. Wu B (2005) Differential gene expression detection using penalized linear regression models: the improved SAM statistics. Bioinformatics 21: 1565–1571.
33. Zhang S (2007) A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. BMC Bioinformatics 8: 230.