16 Setter, T.L. and Laureles, E.V. (1996) The beneficial effect of reduced elongation growth on submergence tolerance of rice. *J. Exp. Bot.* 47, 1551–1559
17 Fukao, T. *et al.* (2006) A variable cluster of ethylene-responsive-like factors regulates metabolic and developmental acclimation responses to submergence in rice. *Plant Cell* 18, 2021–2034
18 Xu, K. *et al.* (2006) *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* 442, 705–708
19 Alpi, A. and Beevers, H. (1983) Effects of O_2 concentration on rice seedlings. *Plant Physiol.* 71, 30–34
20 Colmer, T.D. (2003) Long-distance transport of gases in plants: a perspective on internal aeration and radial oxygen loss from roots. *Plant Cell Environ.* 26, 17–36
21 Xu, K.N. and Mackill, D.J. (1996) A major locus for submergence tolerance mapped on rice chromosome 9. *Mol. Breed.* 2, 219–224
22 Toojinda, T. *et al.* (2003) Molecular genetics of submergence tolerance in rice: QTL analysis of key traits. *Ann. Bot. (Lond.)* 91, 243–253
23 Loreti, E. *et al.* (2003) Gibberellins are not required for rice germination under anoxia. *Plant Soil* 253, 137–143
24 Loreti, E. *et al.* (2003) Sugar modulation of alpha-amylase genes under anoxia. *Ann. Bot. (Lond.)* 91, 143–148
25 Klok, E.J. *et al.* (2002) Expression profile analysis of the low-oxygen response in *Arabidopsis* root cultures. *Plant Cell* 14, 2481–2494
26 Loreti, E. *et al.* (2005) A genome-wide analysis of the effects of sucrose on gene expression in *Arabidopsis* seedlings under anoxia. *Plant Physiol.* 137, 1130–1138
27 Guglielminetti, L. *et al.* (2001) Carbohydrate-ethanol transition in cereal grains under anoxia. *New Phytol.* 151, 607–612
28 Perata, P. *et al.* (1997) Mobilization of endosperm reserves in cereal seeds under anoxia. *Ann. Bot. (Lond.)* 79, 49–56

1360-1385/$ – see front matter
doi:10.1016/j.tplants.2006.12.005

Understanding sample size: what determines the required number of microarrays for an experiment?

Tommy S. Jørstad[1], Mette Langaas[2] and Atle M. Bones[1]

[1] Department of Biology, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
[2] Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

DNA microarray experiments have become a widely used tool for studying gene expression. An important, but difficult, part of these experiments is deciding on the appropriate number of biological replicates to use. Often, researchers will want a number of replicates that give sufficient power to recognize regulated genes while controlling the false discovery rate (FDR) at an acceptable level. Recent advances in statistical methodology can now help to resolve this issue. Before using such methods it is helpful to understand the reasoning behind them. In this Research Focus article we explain, in an intuitive way, the effect sample size has on the FDR and power, and then briefly survey some recently proposed methods in this field of research and provide an example of use.

Replication in microarray experiments

The results of a DNA microarray experiment are influenced by biological and technical sources of variation. To handle this variation, researchers often replicate the measurements, using different biological cases, and then use statistical tests to identify genes of interest. An essential step in the design of an experiment is, therefore, choosing the number of biological replicates[a] to be used – the sample size. In general, a larger sample size should produce more reliable results. However, the cost of a microarray experiment calls for moderation. Consequently, one should aim to find the smallest sample size that still provides results that are of a 'good enough' quality.

Recently, several statistical approaches have been proposed that could be used to help estimate the optimal sample size. To make the best use of this new methodology it is helpful to first understand its theoretical basis. How does sample size affect the outcome of an experiment? How are quality measures, such as the false discovery rate (FDR) [1,2] and power, used to determine if the results are 'good enough'? Below we examine a much-used setup that compares samples from two conditions. From this example we will try to answer the above questions in an intuitive way. We then discuss some new developments in the field of sample-size estimation.

Comparing two conditions

Assume that we want to compare gene expression in an *Arabidopsis thaliana* wild type with that in a mutant. We make $n = 3$ biological replicates for both groups and run a microarray experiment. After collecting the data we face a challenging task. For each gene we must now decide whether we think it is differentially regulated.

When trying to find regulated genes, statisticians often calculate a *t*-statistic[b]. Based on microarray measurements, one *t*-statistic can be calculated for each gene. The *t*-statistic, in essence, quantifies the evidence of a gene being regulated. The further away from zero a *t*-statistic is, the greater the

Corresponding author: Bones, A.M. (atle.bones@bio.ntnu.no).

[a] In the literature, 'biological replicates' denote replicated measurements using different biological cases, whereas 'technical replicates' use the same biological cases. In this article, we only consider biological replication.

Available online 16 January 2007.

[b] A standard *t*-statistic can roughly be written as d/s, where d is a sample statistic that quantifies group differences in gene expression, and s is the estimated standard deviation of d. We use this statistic here because of its analytically tractable properties. In microarray data analysis, other statistics are now often preferred. However, many of these are closely related to the standard *t*-statistic. This is the case for popular analysis tools such as the regularized *t* test [3], the limma-package [4] and the SAM-package [5].

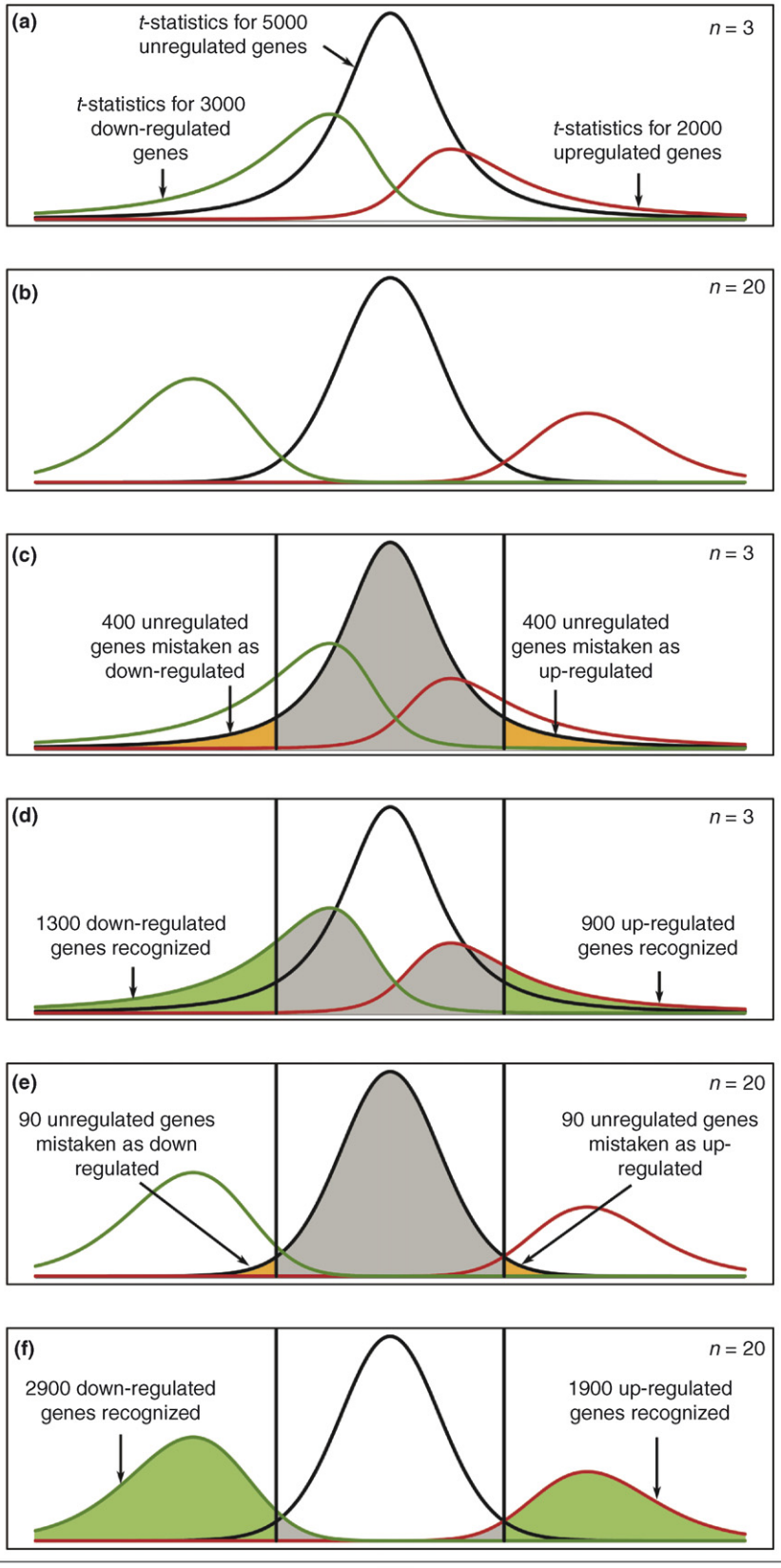

Figure 1. **(a)** The distribution of *t*-statistics for an experiment with 10 000 genes, where three biological replicates were made for each group. 5000 genes are unregulated; their *t*-statistics follow a central *t*-distribution (black). 2000 are up- and 3000 are down-regulated, both with a twofold change; these *t*-statistics follow non-central *t*-distributions, depicted in red and green, respectively. **(b)** The distribution of *t*-statistics for the same experiment as that described in (a), but now using 20 biological replicates for each plant type. **(c)** Distributions from (a) with a cut-off for the *t*-statistics. *t*-statistics for unregulated genes erroneously believed to be regulated at this cut-off are depicted in orange. Unregulated genes recognized as such are depicted in grey. **(d)** Same curves as (c). *t*-statistics for regulated genes recognized as such are depicted in green. Regulated genes not recognized at this cut-off are depicted in grey. **(e,f)** Same curves and colouring as in (c) and (d) but now using 20 biological replicates for each plant type.

Box 1. Power and the false discovery rate

Two important statistical measures are power and false discovery rate (FDR). We illustrate these concepts using the following example. Consider an experiment using mutant and wild-type plants. The only changed phenotype of the mutants is their height: mutants are somewhat taller than the wild type. Also assume we have a mixed set of plants and that we do not know their true state (wild type or mutant). If we wanted to find the mutants in the set, then a good strategy would be choosing the tallest ones. However, there are problems with this strategy. First, because of natural variation, all mutants will not be taller than all wild-type plants. Second, setting a cut-off height is difficult. How tall must a plant be before we are sure that it is a mutant?

In Figure I there is a set of 11 plants: six mutant (m) and five wild type (wt). To find the mutants we set a cut-off height (indicated by a line). The five tallest plants we then believe to be mutants. The outcome of this experimental procedure can, in a sense, be summarized by looking at two key numbers:

(i) The proportion of true mutants that we recognize. In our example this is 4/6 = 0.67.

(ii) The proportion of mistakes among the plants that we believe are mutants. In our example this is 1/5 = 0.2.

Note that if we change the cut-off, the two proportions also change. If we shift the cut-off to the left we recognize more mutants but make more mistakes. If we shift it to the right we make fewer mistakes but recognize fewer true mutants.

The considerations made in the above example are similar to those of a microarray experiment, where, instead of trying to recognize mutant plants based on height, one tries to recognize regulated genes based on test statistics. Because of variability, test statistics for regulated genes are not always larger than those of unregulated genes, and statisticians here also set a cut-off. After doing so, the outcome can be summarized by two numbers, power and FDR:

- Power[l] is the proportion of regulated genes that one recognizes. Researchers will want the power to be close to 1.
- The FDR[m] is the proportion of mistakes among the genes that are believed to be regulated. Researchers will want an FDR close to 0.

Note how power and FDR correspond to the proportions discussed above. As in the mutant plant example, setting a cut-off involves compromise. If the cut-off is changed to improve the FDR, then the power to recognize regulated genes is reduced, and vice versa.

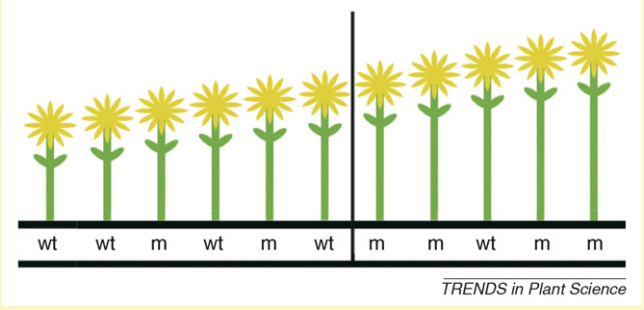

Figure I.

[l] Rigorously, when testing multiple hypotheses, average power (here simply termed power) is usually defined as E(S)/m1, where E(S) is the expected value of S, S = 'number of rejected hypotheses for which the alternative hypothesis holds' and m1 = 'total number of hypotheses for which the alternative hypothesis holds'.

[m] Rigorously, when testing multiple hypotheses, the FDR is usually defined as E(V/R) (i.e. the expected value of the ratio V/R, where V = 'number of rejected hypotheses for which the null hypothesis holds, and R = 'total number of rejected hypotheses'). By definition, FDR = 0 when R = 0.

evidence that the expression level of the gene has changed. Unregulated genes should therefore have *t*-statistics close to zero. The *t*-statistics of up- and down-regulated genes should be positive and negative, respectively.

The *t*-statistics for unregulated genes will never be exactly zero owing to measurement variability. Likewise, the *t*-statistics of, for example, up-regulated genes with a twofold change[c], will also vary. Figure 1a shows an example of the distributions of *t*-statistics one could observe for an experiment with 10 000 genes, where $n = 3$ biological replicates were used for each group. The distribution of *t*-statistics for 5000 unregulated genes is shown in black. It centres on zero but there is variability. The red curve is the distribution of *t*-statistics for 2000 genes up-regulated with a twofold change, and the green curve is for 3000 genes down-regulated with a twofold change.

Now consider a larger sample size such as $n = 20$ biological replicates for each group. How would the distribution of *t*-statistics be affected? For unregulated genes, we expect the distribution to still centre on zero. For regulated genes, we expect the evidence of a change to be greater, and the distribution of *t*-statistics to move away from zero. This is indeed what happens (Figure 1b).

Sample size effects on FDR and power

An introduction to FDR and power is given in Box 1. To see the effect of sample size we continue with the above example of a microarray experiment testing 10 000 genes (5000 unregulated, 2000 up-regulated and 3000 down-regulated). The true state of a gene, regulated or not, is not known to us.

The goal of our microarray experiment is to identify regulated genes. A *t*-statistic measures evidence of a change in the expression level of a gene. A good strategy for finding the regulated genes is, therefore, calculating the *t*-statistic of each gene and then picking those with *t*-statistics that are extreme (i.e. far away from zero).

For a sample size choice of $n = 3$ we could have distributions of *t*-statistics like the ones in Figure 1a. In Figure 1c, we see the same graph but now with two vertical lines. These lines represent the cut-offs that need to be set. All genes with *t*-statistics more extreme than the cut-offs are believed to be regulated, whereas those within the lines are believed to be unregulated. The plot in Figure 1c reveals a problem. Because of measurement variability, the *t*-statistic for unregulated genes will not always be smaller than those of regulated ones. For our chosen cut-off we find that 400 + 400 = 800 unregulated genes will mistakenly be considered regulated. In Figure 1d we see another problem. We are not able to recognize all the regulated genes. Indeed, at this cut-off only 1300 + 900 = 2200 out of the 5000 regulated genes are recognized.

So what is the FDR and power at this cut-off? The total number of genes believed to be regulated is 800 + 2200 = 3000. Of these, 800 are mistakes (unregulated genes) and 2200 are truly regulated. The FDR (i.e. the proportion of mistakes among those believed to be regulated) is, therefore, 800/3000 = 0.27. The power (i.e. the proportion of truly regulated genes recognized) is 2200/5000 = 0.44, which is not satisfactory. 27% of the genes we believe to

[c] A 'twofold change' means the expression level of a gene is twice as high under one condition compared with the other.

be regulated are mistakes, and out of the truly regulated genes we find only 44%. Using different cut-offs will not necessarily help. More extreme cut-offs, will decrease the FDR, but also decrease the power[d]. Less extreme cut-offs give increased power but also increase FDR.

The way to improve both FDR and power is to increase the sample size. In Figure 1e and f we again consider the case of $n = 20$, and use the same cut-offs as before. Now, because the distributions are shifted we can get FDR = 0.04 and power = 0.96[e]. We recognize 96% of the truly regulated genes, and at only a 4% error rate.

New statistical methodology

The FDR has become a much used error measure in the microarray setting. In spite of this, little attention has been given to sample size estimation methods that allow direct control of the FDR. However, recently, progress has been made. Approaches have been suggested that enable researchers to specify the desired FDR and power[f], and that calculate the sample size needed to achieve this goal. Sample size estimation methods that control error measures other than the FDR do exist but these will not be discussed here.

Sample size estimates can be made for two types of experiments. (i) In hypothetical experiments, the researcher can specify all parameters to be used. An example is the one above, where we specified several up- and down-regulated genes, all with a twofold change. For hypothetical experiments exact sample size requirements can be calculated. (ii) In real-life experiments, unlike the hypothetical ones, the true state of the genes is not known. The sample size estimates must then be based on a pilot dataset. Pilot data are data where the distribution of regulated genes is believed to be similar to the experiment at hand, and could, for example, be a small-scale version of the experiment of interest.

For hypothetical experiments that compare two groups, the SAM package[g] [5,6], the OCplus package[h] [7] and the methods of Jianhua Hu *et al.*[i] [8] and Sin-Ho Jung [9][j] can be used to explore sample-size effects. For hypothetical experiments that compare k groups (where $k \geq 2$), the methods of Stan Pounds and Cheng Cheng[k] [10] can be used. Approaches that can be used for sample-size estimation based on real experimental datasets exist but are still being improved upon. Several of the above-mentioned implementations, such as those described in Refs [6,8,10], offer an option to base the estimates on experimental data, but their strategy for doing so, and the input they need, varies. Currently, there is no consensus on which method is better. An example case, where two of the above-mentioned methods are used, is described in Box 2.

Although new and improved sample-size estimation methods will be developed, the theoretical foundation presented here will not change. While designing an experiment one should therefore set a goal in terms of FDR and power (or some other measure), and try estimating the required sample size. However, keep in mind that the goal will depend on the experiment being conducted. For some studies, detecting only the 10% most-regulated genes (i.e. a power as low as 0.1), at an FDR of 0.05, could be sufficient. Other experiments will have more ambitious goals. Regardless of the experimental goal, the sample size issue should be given careful thought.

Box 2. Sample size estimation: an example case

To illustrate the use of sample-size estimation methods, we downloaded an example dataset from The *Arabidopsis* Information Resource (TAIR) (http://www.*Arabidopsis*.org). The chosen dataset (TAIR accession number: ExpressionSet: 1008031444) compares wild type to ARR21-overexpressing seedlings using $n = 3$ Affymetrix slides for each plant type. The data are discussed in Ref. [13]. The example dataset was then treated as pilot data and used to explore the effects of different sample sizes.

Using the sample size estimation methods discussed in Refs [8] and [10] we estimated the per-group sample sizes needed to achieve particular combinations of power[n] and FDR. Table I shows that for moderate power cut-offs the two methods agree. For example, we find that to have power = 0.7 with FDR = 0.05 one needs $n = 5$ slides, whereas power = 0.8 with FDR = 0.05 requires $n = 8$ or $n = 9$ slides per group. The sample size used in the experiment, $n = 3$, is estimated to give power = 0.5 with FDR = 0.05.

For high power the estimates do not fully agree; this is probably related to the difficult task of recognizing regulated genes that have only small changes in expression level. New approaches to recognizing such genes are being developed. However, the estimates do agree that to get high power with low FDR one needs sample sizes considerably larger than those commonly used in studies today.

Table I. Estimated sample size requirements for example data set*

	FDR = 0.10	FDR = 0.05	FDR = 0.01
Power = 0.5	3 / 3	3 / 3	5 / 5
Power = 0.6	3 / 3	3 / 4	7 / 6
Power = 0.7	3 / 4	5 / 5	10 / 9
Power = 0.8	4 / 6	9 / 8	20 / 14
Power = 0.9	13 / 11	30 / 16	75 / 27

*The numbers either side of the solidus indicate sample-size estimates made using the sample-size estimation methods described in Ref. [8] and Ref. [10], respectively.

[n] In the methods of Hu *et al.* [8], one specifies FDR and the expected number of genes taken as regulated, E(R), instead of FDR and power. Each power cut-off was converted to a corresponding E(R) to produce the estimates.

Acknowledgements

We thank Herman Midelfart for helpful discussions on the subject. This work was supported by grants NFR 143250/140 and NFR 151991/S10 from the biotechnology and the functional genomics (FUGE) programs of the Norwegian Research Council.

References

1 Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300

2 Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9440–9445

[d] In microarray data analysis, the cut-off is often set so that the FDR is below 0.05, or sometimes 0.01. The motivation for this is keeping the error rate among the genes that are claimed to be regulated below 5%.

[e] Calculations: 90 + 90 + 1900 + 2900 = 4980 are believed to be regulated. 90 + 90 = 180 are errors and 1900 + 2900 = 4800 are truly regulated. FDR = 180/4980, power = 4800/5000.

[f] Other measures than power can be used together with the FDR. Examples are the false negative rate (FNR) and the number of genes that are labelled as 'regulated'.

[g] Implemented as the package samr for the R environment [11].

[h] Implemented for the R environment, downloadable from Bioconductor [12].

[i] R source code available from the authors.

[j] Implementation available from the author.

[k] R source code available from the authors.

3 Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t* test and statistical inferences of gene changes. *Bioinformatics* 17, 509–519
4 Smyth, G.K. (2005) Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (Gentleman, R. *et al.*, eds), pp. 397–420, Springer
5 Tusher, V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5116–5121
6 Tibshirani, R. (2006) A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics* 7, 106
7 Pawitan, Y. *et al.* (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* 21, 3017–3024
8 Hu, J. *et al.* (2005) Practical FDR-based sample size calculations in microarray experiments. *Bioinformatics* 21, 3264–3272
9 Jung, S.H. (2005) Sample size for FDR-control in microarray data analysis. *Bioinformatics* 21, 3097–3104
10 Pounds, S. and Cheng, C. (2005) Sample size determination for the false discovery rate. *Bioinformatics* 21, 4263–4271
11 R Development Core Team (2006) *R: A Language and Environment for Statistical Computing,* R Foundation for Statistical Computing, (Vienna, Austria 3-900051-07-0 (http://www.R-project.org)
12 Gentleman, R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5, R80
13 Kiba, T. *et al.* (2005) Combinatorial microarray analysis revealing *Arabidopsis* genes implicated in cytokinin responses through the His→Asp phosphorelay circuitry. *Plant Cell Physiol.* 46, 339–355

1360-1385/$ – see front matter
doi:10.1016/j.tplants.2007.01.001