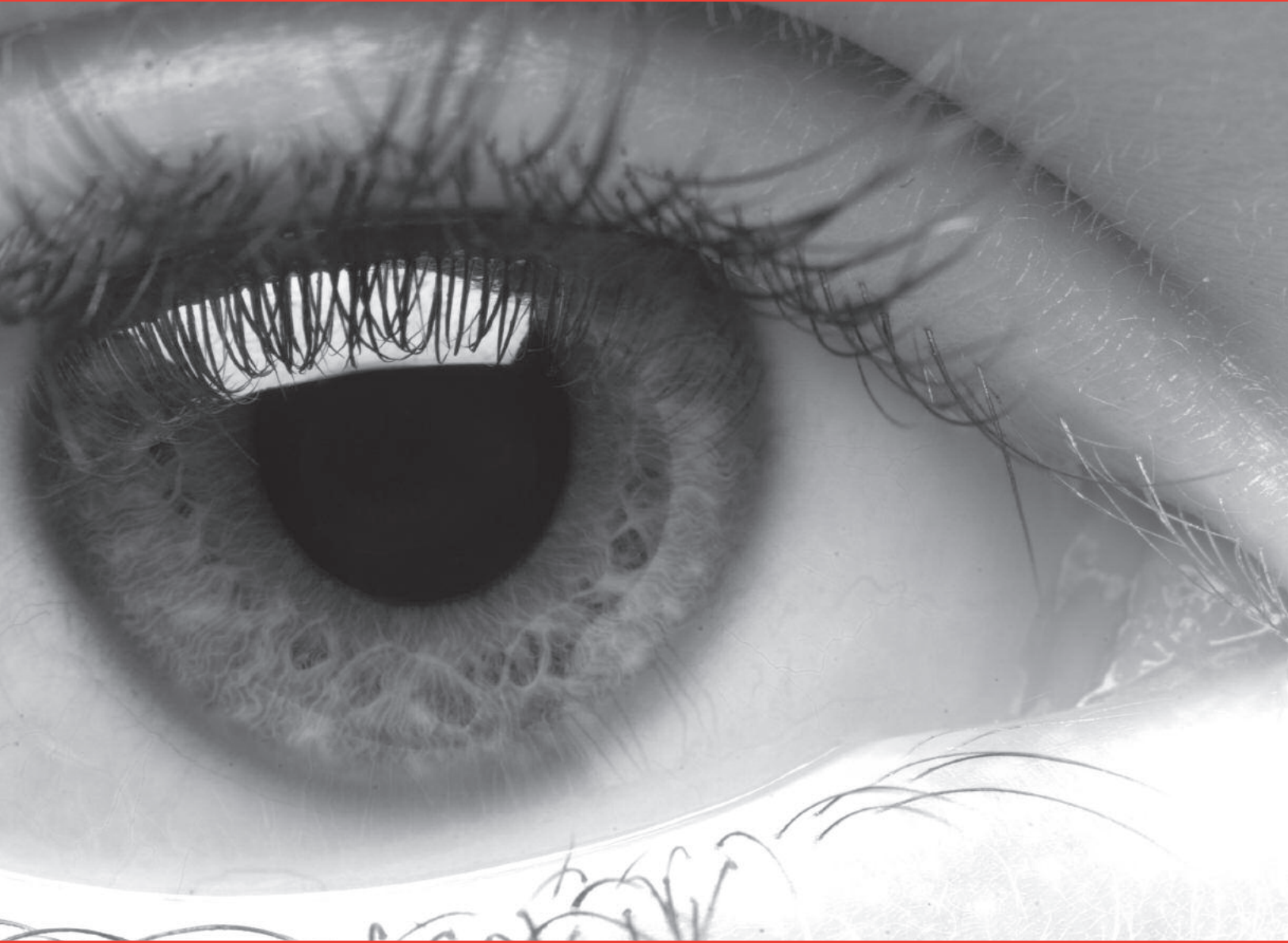


EXPERIMENTAL DESIGN

By Steve Moore



Diagnostics
Sciences
Clinical Services
Clinical Technologies
Pharma Services

www.almacgroup.com

Gene Expression Analysis – Experimental Design

1. Introduction:

Microarray experiments generally focus on three main experimental types:

1. Class Comparison (Figure 1.1)
2. Class Prediction (Figure 1.2)
3. Class Discovery (Figure 1.3)

1.1 Class comparison:

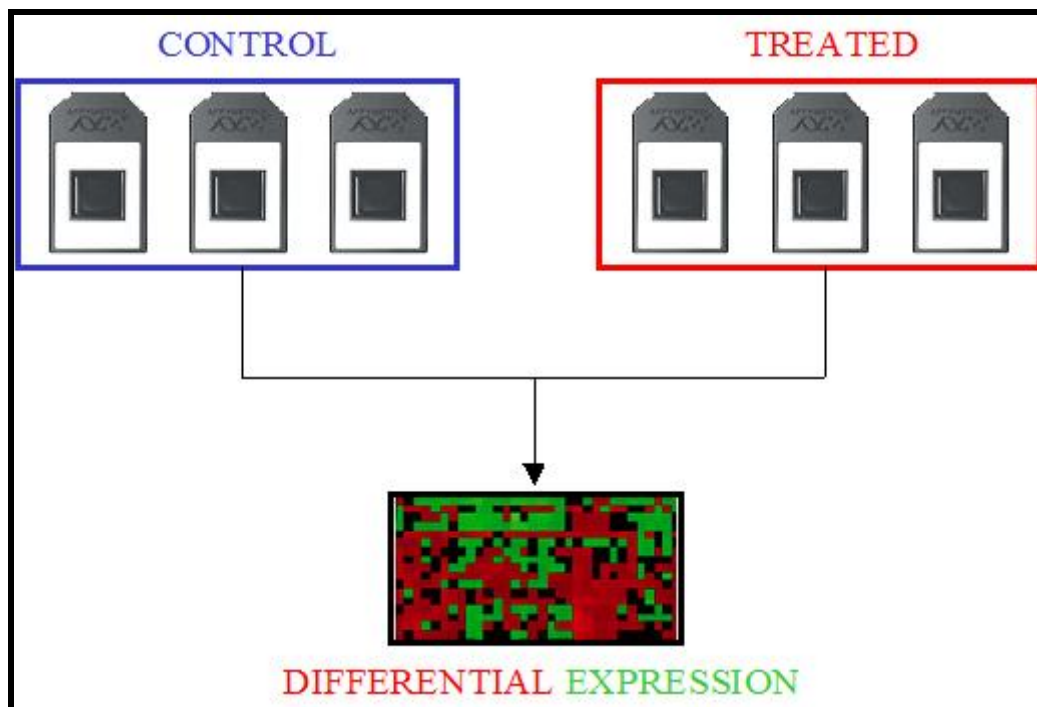


Figure 1.1 An example of a class comparison study with treated and control samples compared to generate a differential expression heatmap.

This experimental type aims to identify genes or transcripts that are differentially expressed between two or more classes. For example, treated vs. untreated, normal vs. tumour or in a multi-conditional experiment a time series of drug treatments.

1.2 Class Prediction:

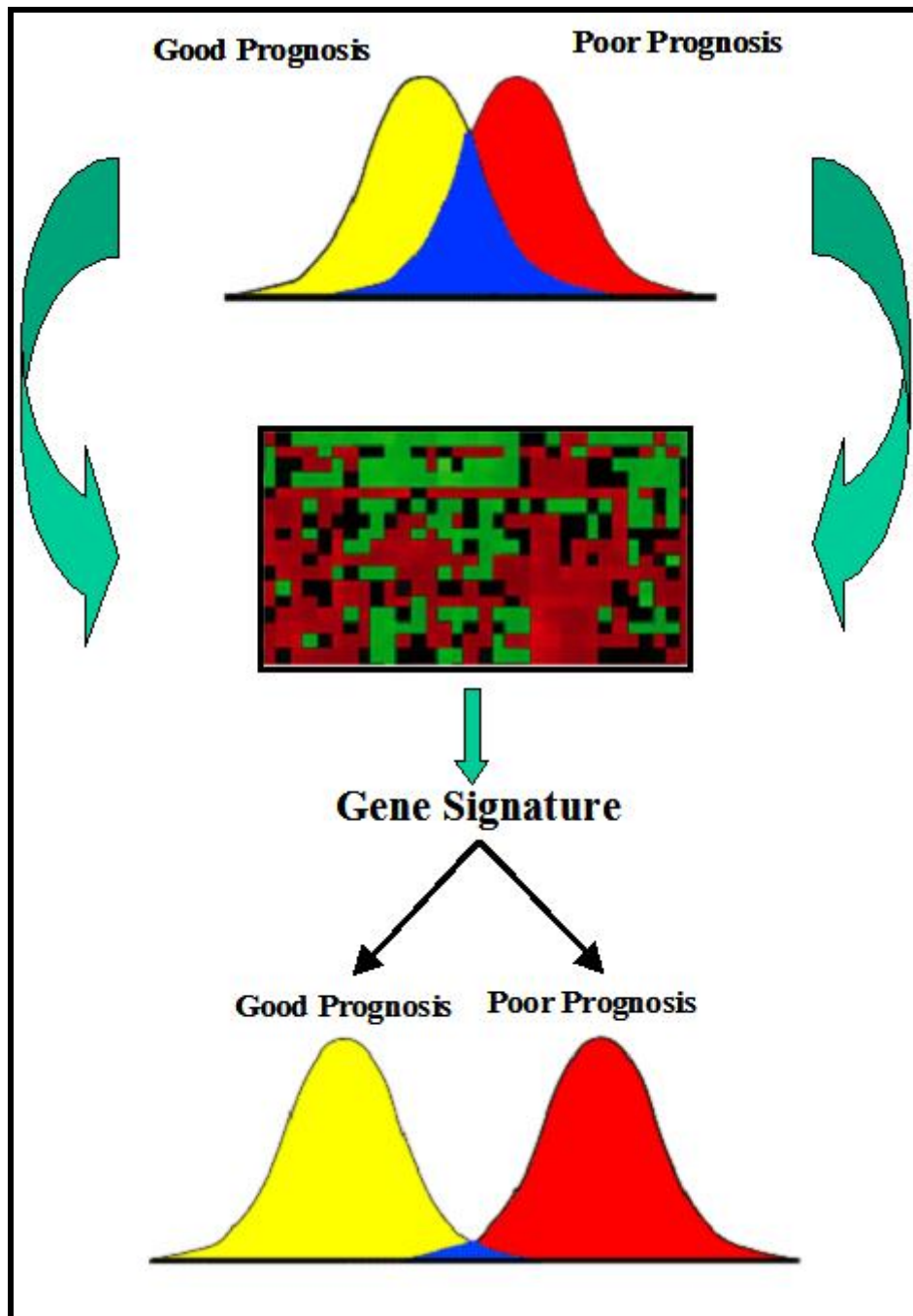


Figure 1.2 An example of a class prediction design. In this case two prognostic outcome sample distributions are assessed in terms of gene expression to generate profiles of the samples within the distributions. The most discriminatory genes are extracted to create a gene signature that will enable the differentiation of the two population distributions for further clinical samples.

In class prediction problems, the aim is to identify a multi-variate predictor between two or more classes that can be used to distinguish the same classes in a blinded fashion. This could be drug treated vs. untreated; normal vs. tumour; good prognosis vs. poor prognosis etc.

1.3 Class Discovery:

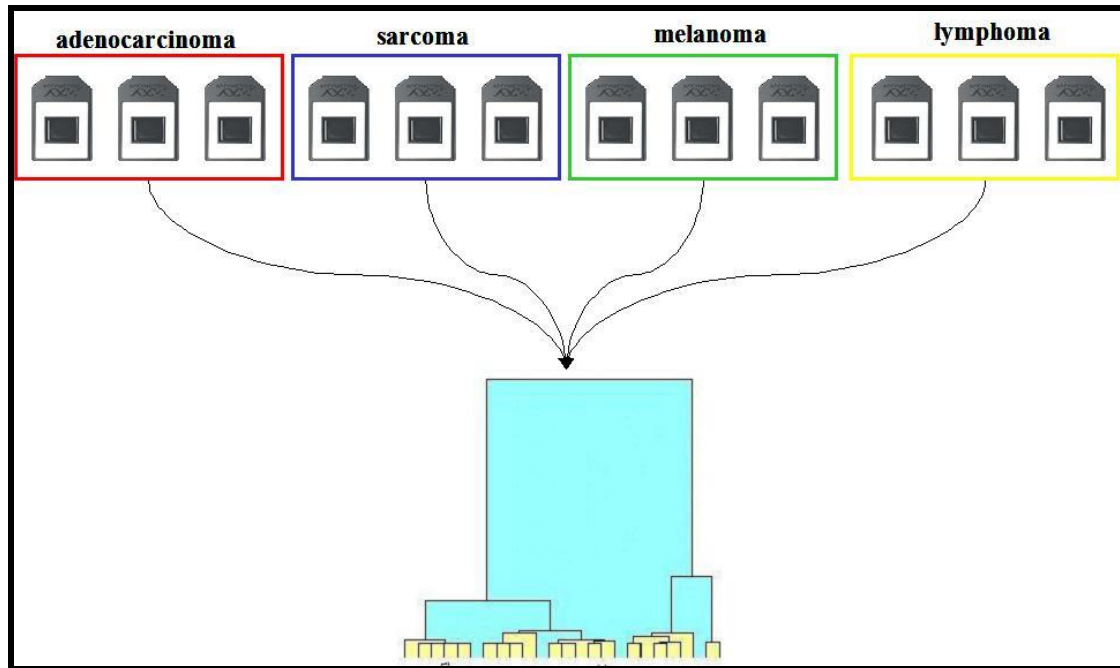


Figure 1.3 This is an example of a class discovery design. The design could be taken as a class comparison design, except that in the above case we have identified the meta-classes, but are more interested in generating expression profiles which can be clustered to isolate possible sub-classes.

This type aims to find sub-classes within known classes using clustering techniques based upon expression data. In this way new taxonomies can be found and identified. This has been employed before as a pre-cursor to class prediction problems.

1.4 Interaction of Designs:

The first two types (class comparison, class prediction), involve a design where the conditions are known and are independent of the experimental results, the third (class discovery), involves the finding or creation of conditions so the final or sub-conditions cannot be known at the outset of the experiment (obviously meta-conditions are known).

The first step of a microarray study will be deciding which of the above 'experimental types' fits your experiment and it is quite likely that more than one or all three could be assessed during one experiment. It is imperative that the experiment therefore is designed properly as requirements for one type may not facilitate another.

For example in class prediction studies, more than 20 samples per group are required by Almac Diagnostics. As such, you cannot design an experiment to assess class comparison and then try to apply it to class prediction as there typically would too few samples, and it also unlikely the samples would cover enough of the population to be effective.

So, in order to carry out a microarray experiment, it is clear that the researcher must understand the objectives of the experiment, have clearly defined questions and be aware of the statistical implications that these questions will invoke.

2. Statistical Considerations:

It is useful to understand general principals of statistical thinking when planning a microarray experiment and the impact this can have on the analysis of the data.

Typically a researcher wants to show the effect of some input (stimulus) on a biological material. This could be a drug treatment on xenograft implants or the transfection of a fully functional gene into a mutant cell line. The obvious aim is to demonstrate that this observation holds for this biological material in general (population) and not just the experimental samples on which the experiment was carried out (samples).

Using a simple example of a drug treatment for colorectal cancer. In an ideal world we would profile all breast cancer sufferers in the world and demonstrate the results of the drug. It would be surprising if you found anyone who thought this was feasible. As such we are required therefore to 'sample' the population. The trick is to make your sample set 'representative of the population'.

Representative of the population means that it include variables found in the population for example male and female subjects, represent the age range, represent the geographical nature of the population etc.

In statistical terms a population is a group of entities on which measurements are being carried out. For example, it could be drug effect on colorectal cancer sufferers, or effect of sunlight on leave lengths of a specific tree type etc. with colorectal cancer sufferers and all leaves of one specific tree type being the population in these examples. Each population is represented by a distribution, which in turn is described by what are known as 'parameters'.

The most well known distribution is the normal distribution. The parameters of this distribution are the mean and standard deviation. As such the normal distribution is not a single distribution but a family of distributions that are described by these parameters (Figure 2.1).

Our sampling procedure will generate a sampling distribution which we hope will approximate as closely as possible to the population distribution, therefore we want the sampling parameters of the mean and standard deviation to approximate the true population mean and standard deviation. This is accomplished by high sample numbers in different experimental groups (in other words high replication).

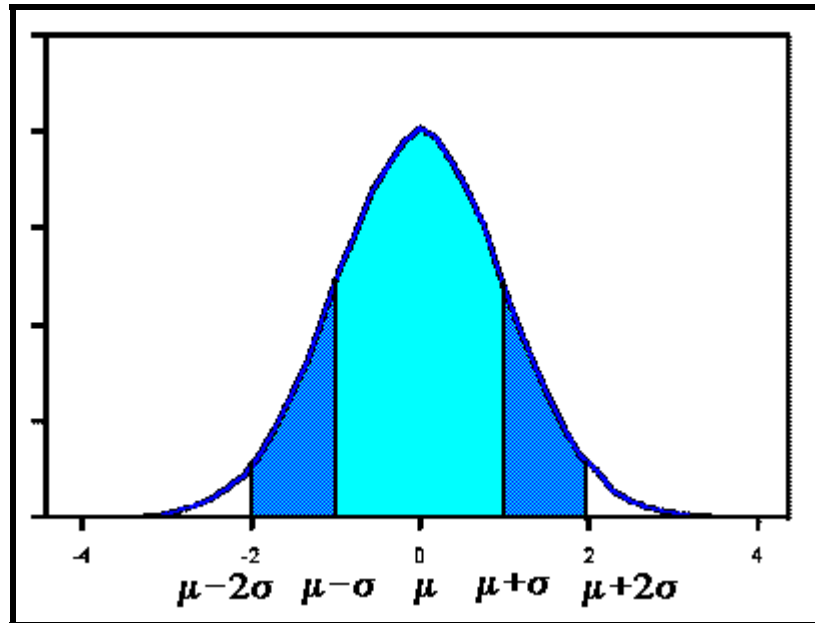


Figure 2.1 A normal distribution. The axis of symmetry for this distribution is described by the mean and the scatter or spread of the distribution described by the standard deviations. These parameters dictate how tall, short, thin or spread the distribution will be. Using these parameters alone, one could sketch the normal distribution in question

A further statistical consideration that researchers must be aware of is that statistical tests such as the student t-test for example, require a good estimation of the standard deviation and the means of the distributions being compared. In addition the test assumes that the distributions are independent of each other.

Good approximation of the standard deviation and means of two distributions can only be accomplished by replicating samples and the independence can only be achieved by generating replicates at different times. For example, creating three biological replicates on the same day, results in a dependency between the samples, whereas creating these replicates one week apart for three weeks will remove this dependency. Dependency can also be introduced with cell lines that have not been suitably passaged and as such still retain some level of dependency. In general conditions can be thought of as independent if a change in one condition makes it neither more nor less probable to occur in another.

3. Experimental Design:

The key to experimental design is good planning. Planning is driven by experimental objectives, with successful experiments based on clear objectives.

Using a simple example of a drug treatment on a cell line, we want our experimental design to show that with input X (drug treatment), outcome Y is observed (changes in certain gene expression). However, we also want to show that in the absence of input X, outcome Y does not occur. This requires the consideration of different principals; internal and external experiment validity.

3.1 Internal Validity:

Firstly, consider the internal validity of the experiment and the causative factors, which can affect this. Internal validity is focused on the fact that it was input X that caused outcome Y and not some other variable. Some of these causative factors include:

- **History:** Events which may occur between replicates. For example changes in operators, environment etc.
- **Maturation:** Collecting samples over time, may involve biological entities getting older and thus causing changes in gene expression.
- **Pre-Testing:** Measures taken from a biological system before a microarray experiment, may affect that system. For example pre-testing is often used to assign samples to experimental groups.
- **Instruments:** Measuring instruments may change over the time of replication or sample collection. For example instruments falling out of calibration, or change of location of equipment etc.
- **Mortality:** Subjects in experimental groups may die. For example, animal studies may suffer from the mortality of one animal in a control or experimental group.

3.2 External Validity:

External validity concerns the causative factors that can jeopardize the generalizability of the observed effect in the experiment upon the population. Some of these factors include:

- **Pre-Testing:** Individuals who have been pre-tested may have a different reaction to a factor (drug, treatment etc.) than a population.
- **Differential Selection:** Small groups of samples may have defining characteristics not shared by the whole population.

- **Experimental Procedures:** These can have an effect on gene expression during an experiment, which wouldn't necessarily be typical of the population.
- **Multiple Treatments:** Treatments administered in linear format can cause changes that wouldn't be observed with the second treatment alone in the population.

3.3 Controlling Internal and External Validity:

Internal and external validity can be controlled using the following techniques:

1. **Pre-Testing:** Although a potential causative factor of internal and external validity, it can control for differential selection by determining presence or knowledge of an experimental variable. It can also be used to identify factors of mortality.
2. **Control Group:** The control group is exposed to all experimental variables except the factor being measured, therefore it can account for the effect of external sources of variation. It can help in the elimination of the effects of history, maturation, instrumentation and interaction effects.
3. **Randomization:** Use of random selection can help in statistical regression, differential selection and interaction of factors. It also aids in generalizability.

In addition to the three methods above, keeping replicated experiments as identical as possible, for example, same researcher, same equipment (where applicable) etc. can help manage the effect of history, experimental procedures and instrumentation effects.

In addition the implementation of a good quality control system can also help in generating reproducible data in the laboratory or between operators, thus controlling experimental procedures, history and instrumentation effects.

3.4 Replication:

Investigators typically ask how many replicates are required for a good microarray experiment. In short, the more replicates the better.

Before considering how many replicates one requires, it is important to know what kind of replication exists and what they measure. The two types of replication are

1. Technical
2. Biological

Technical replication occurs when multiple arrays are run from one RNA sample, these are also sometimes referred to in published text as chip replicates.

Biological replications on the other hand are samples prepared separately of each other from independent biological entities and run on respective arrays. In both cases, the aim of replication is to assess variability either of the process (technical) or the biological system (biological). Given the high reproducibility of the Affymetrix GeneChips™ and the controlled nature of the Almac Diagnostics ISO17025 accredited procedure, technical replicates are not required unless a technical issue is under investigation.

Biological replicates are much more important in getting the correct expression profile from a system. Biological systems typically demonstrate variability (sometimes quite large), even in cell culturing which can introduce variability in terms of experimental manipulation or differences in cell growth and harvesting.

In general variability is greatest in human samples, followed by animal models, where inbreeding of strains can reduce variability somewhat, and cell lines which typically are relatively homogenous and so demonstrate low levels of variability. In terms of replication, it is proportional to the level of variation. In other words more replicates will be required to demonstrate the same change in a human tissue sample as is demonstrated in a cell line sample. This is the fundamental idea behind power analysis which is used to calculate the sample size required for the detection of a gene expression change at a specific power, given the variability of the samples.

In terms of analysing microarray data we are typically concerned with two issues:

1. The number of differentially expressed genes that are detected.
2. The magnitude of change that can be detected between groups of conditions.

In both of the above cases, replication plays a role. Considering the number of differentially expressed genes it is not surprising that the number of differentially expressed genes increases with the number of biological replicates performed.

Table 3.4 is reproduced from a Baylor College of Medicine webpage (<http://www.bcm.edu/mcfweb/?PMID=3101>), where they have reported an experiment carried out by Daniel Amador-Noguez.

In this experiment 24 mouse arrays were run profiling wild-type and Ames dwarf mice. The data was analysed using one-way ANOVA at two p value cut-off values with no multiplicity correction. As you can observe from the table, as replication increases so do the number of differentially expressed genes identified. The increase in replication also decreases the number of false positives or 'ghost' genes that are identified.

Table 3.4 Number of Differentially Expressed Genes vs. Number of Replicates

Number of differentially expressed genes vs. number of replicates (ANOVA)				
# of Replicates	Variance Not Equal		Variance Equal	
	p<0.001	p<0.0001	p<0.001	p<0.0001
2	3	1	32	4
3	74	10	160	27
4	292	61	441	135
5	456	140	618	228
6	791	296	956	420
7	1128	513	1294	628
8	1315	626	1469	727
9	1766	896	1895	990
10	1928	1016	2014	1121

The difference in the magnitude of change detected (Point 2 above), can be measured by a typical hypothesis test such as the Student t-test. The detectable difference is dependent on the means of the two sample distributions and their standard deviations.

This is demonstrated by figures 3.4.1 and 3.4.2. In both Figures the standard deviation is similar between distributions, however the means are seen to be quite different (Figure 3.4.1) or quite similar (Figure 3.4.2). The outcome of this means that in order to detect the same small difference in the distribution means of Figure 3.4.2 as can be detected from the distribution of Figure 3.4.1 then more replication is required.

This is identical in the assessment of gene expression. In both cases the distributions represent that of the expression gene X across two sample groups and as such the to detect the same small difference between the two samples for gene X, more replication is required for the distributions displayed in Figure 3.4.2.

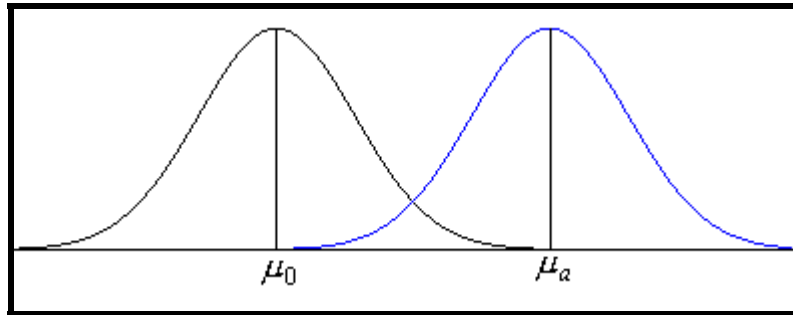


Figure 3.4.1 Normal distributions of two sample sets with associated means and standard deviations. The means of these two distributions can be seen to be quite different, with little overlap between the distributions. These distributions could represent the expression of gene X across two different sample types.

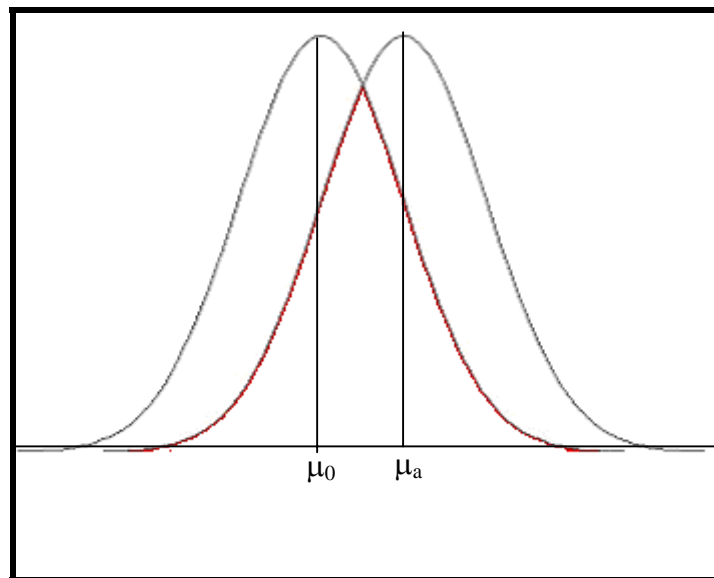


Figure 3.4.2 Normal distributions of two sample sets with associated means and standard deviations. The means of these two distributions can be seen to be quite similar, with a lot of overlap between the distributions. These distributions could represent the expression of gene X across two different sample types.

When considering the Student t-test, it works by using the available degrees of freedom to assess groups of measurements for a difference between the means. If there are three replicates there are, $N - 1$, or 2 degrees of freedom. Now, consider if the replicates increased to 10, there are now 9 degrees of freedom. This means that smaller differences can be detected and when you are faced with highly variable overlapping distributions like Figure 3.4.2 more replication is required to isolate the difference in, for example, gene X that is dysregulated between the experimental conditions.

On a different note, hypothesis tests such as the t-test used the approximation of standard deviation and thus standard error in their equations. If your experiment has two replicates, the approximation of standard deviation is very poor. At a simpler level consider an experiment with two replicates in the control group. When scanned these replicates show that the intensity of gene X is 5000 in one and 50 in the other. How does one decide which replicate is correct?

Finally, as discussed earlier, one aim of microarray experiments is to infer the results of the experiment on to some population. The population parameters as mentioned are approximated from the sampling parameters i.e.

1. The population mean ' μ ' is approximated from the sample mean ' \bar{X} '
2. The population standard deviation ' σ ' is approximated from the sample standard deviation ' s '

Therefore the more replicates there are per group, the more accurate the sample parameters are and hence the better the approximation of the population parameters so your experiment is more representative of the population.

4. Additional variability considerations:

It may not always be possible to get enough RNA to run a GeneChip experiment from a single sample, for example, laser capture microdissection, or you could be working with a population that is extremely heterogeneous and you want to control some of the resultant variability.

One method that researchers sometimes use to control that variability or generate enough material for the GeneChip process is to pool their samples. It should be borne in mind however that the same principals that are applied to individual samples also apply to pooled samples. For example a pooled sample from 5 subjects run on a single chip does not give an estimation of variability. The creation of multiple pooled samples on the same day for example in the same lab, whilst they are biological replicates they still have a dependence structure and as such violate the independence assumption of the t-test.

Pooled samples also remove the estimation of variability between individual samples as they now represent an average variability, therefore if you wish to know the variability between individual samples then pooling is not appropriate.

When pooling samples, the pools should be created in an analogous way to single samples. The pools should be created independently on different days etc. to reduce or eliminate dependency, and pools should be created from individual samples each time, thus if there are 5 rat samples per pool, pool 2 should have a different 5 rats from pool 1. If experiments involving pooled samples are carried out in this way then statistical applications can be performed and there can be inference to the population.

4.1 Classification:

We discussed briefly, the experimental type known as class prediction. This method of analysis aims to generate a multi-variate predictor gene set that will separate two or more classes of samples.

The first consideration for an experiment of this type should be a question. This questions should always be “what do I intend to use this classifier for?”. This will enable the researcher to address the following issues:

1. How large is my population
2. How should I sample the population
3. What test set of samples should I create
4. What independent assessment should I carry out?

These are all important factors in terms of running a classification experiment. The reason for asking what is the classifier to be used for is that, when created, it is capable of classifying the samples that it was intended for.

For example, if a classification model is built upon samples taken from 40 year old men in the UK, to find responders to a drug for lung cancer, then it is quite likely that it will not work for 40 year old women in the United States. Therefore if it is intended to be used to find responders independent of gender, age (there may be a cut-off) and geography, then these variables must be represented in the model building process.

The next consideration is how this signature should be assessed in terms of accuracy, sensitivity and specificity. One method of assessing the usefulness of a classifier is to perform a technique known as cross-validation. This might be leave-one-out cross validation (LOOCV) or k-fold cross validation.

Taking LOOCV as an example, a gene signature is generated from the training set. One sample is then ‘left out’ and the model rebuilt on the remaining samples. The one ‘left out’ sample is then predicted as to whether it is class a, b etc. That sample is then placed back and another removed and the process occurs again. This occurs iteratively until all samples have been classified. This then gives an assessment of accuracy of the classifier. k-fold cross validation works in a similar manner except that a group of k samples are left out rather than one sample.

Cross validation if performed correctly can accurately assess the error inherent in the genetic signature. However, it is still statistically bias as we are using samples that have been involved in the model. The best method of assessing the model and signature is to predict a set of samples that have played no part in the model building process. These are usually referred to as the ‘test set’.

Sometimes being able to represent, for example, all geographical locations in your model is not always feasible. You may have enough samples to generate a classifier from a UK clinical trial but have not included any samples from the United States or Europe. In this case the gene signature must be independently validated. In other

words the signature that was generated from a UK trial should be tested to assess if it works in a US or European trial.

5 Overview:

This document should not be read or interpreted as an exhaustive coverage of experimental design. Experimental design is a field of research in its own right and there are many books, websites and published papers on the subject, many of which are dedicated to microarray studies.

This document has aimed to increase the awareness of our customers understanding of experimental design issues that will result in a better return from their microarray experiments in terms of the volume of data returned and the robustness and confidence in that data.

The take home message is that successful microarray experiments are backed by clear unambiguous questions and a clear design strategy to answer those questions. When considering a microarray experiment in the future, remember the following three steps:

1. Define the objectives

Think about the experiment at hand. Decide on unambiguous questions. If applicable state a hypothesis to be measures versus a null-hypothesis. Clearly record for future reference what is to be tested

2. Devise the Strategy

Work out how to achieve the objectives stated in point 1. For example, size and structure of experiment, how many treatments, are the treatments structured or random, is there pre-testing of samples, how many replicates are required, how are the results to be analysed.

3. Set down the operational strategy

Are the treatments to be random or structured? Who will carry out the sample preparations, how will they do it, when will they do it, how is the microarray experiment to be conducted etc.

Glossary

ANOVA:

Analysis of variance is a collection of models that compare means by splitting the overall observed variance into different parts. One-Way ANOVA is used to assess 3 or more groups by assessing within group variance versus between group variance.

Class Prediction:

Classification is a statistical procedure in which individual items are placed into groups based on quantitative information generated from gene expression data and based on a training set of previously labelled items.

Clustering:

Clustering is the separation or partitioning of samples or genes into groups based on similar measurement traits, for example, measurement magnitude or correlation. The data in each cluster share similar measurement usually defined by some distance metric.

Degrees of Freedom:

Degrees of freedom is the number of independent pieces of information on which a parameter estimate is based. It is a measure of how much precision an estimate of variability has. The degrees of freedom for an estimate equals the number of observations minus the number of additional parameters estimated for that calculation.

Differential expression:

The difference in expression level of a particular gene in two compared samples. Typically displayed as a ratio.

Distribution:

Distributions are probability functions, which can be discrete or continuous for a given population or sample. For discrete variables the distribution gives the probability of a certain point occurring. In continuous distributions this probability is typically given for an interval by measuring the area under the curve.

False Positives:

A positive test result for a sample that inherently does not possess the true nature of a positive test result.

Gene Expression Profiles:

The profile of a sample described in terms of the expression level of all genes / transcripts contained on the specific array on which the sample was measured.

Heterogeneous:

Composed of parts or elements that are dissimilar. Tumour samples are generally regarded heterogeneous as they usually contain different cell types, and even those cells of the same type may exhibit radically different phenotypes.

Homogeneous:

Composed of parts or elements that are of the same type. Cell lines are considered homogeneous as they are all derived from the same cell type.

Hypothesis test:

Hypothesis testing attempts to prove or disprove the null hypothesis, which in turn leads to the acceptance or not of the alternative hypothesis. A typical hypothesis test is the Student t-test.

Laser Capture Microdissection:

This is a method for separating a population of pure cells from a collection of tissue cells. Clear film is applied to the tissue sample on a microscope slide. The cells for isolation are identified and the laser causes fusion of the clear film to the wanted cells, which are then peeled away from the unwanted cells, which remain on the slide.

Leave One Out Cross Validation / k-fold cross validation:

In the case of LOOCV one sample is removed from the data set to behave as the validation data whilst the remainder (N-1) act as the training set for the model. Similarly k-fold cross validation has the data set split into k sets, with each data set dropped once and the remainder (N-k) used as the training set. This validation is performed exactly k times (the folds).

Multiplicity Correction:

When we carry out a hypothesis test, for example, a t-test at a significance level of 0.05 we expect that an extreme result will occur 1 in 20 tests. Now consider that we carry out 6 t-tests together for 6 different genes. Suddenly the probability of getting a chance result changes from 0.05 to $0.05^6 = 0.74$. Therefore the probability that one will give a chance result is now 0.265 and not 0.05 or 1 in 4 and not 1 in 20. Therefore corrections are applied to adjust the p-values of the genes so that the probability of a chance occurrence is still 0.05.

Multi-Variate:

Describes the measurement of more than one variable. In terms of classification, multiple genes are typically part of the classifier, each of them considered a variable therefore it is regarded as a multivariate classifier.

p-value:

The p value is the probability of obtaining a result as extreme as that observed from the test data, assuming null hypothesis is true, so that the result was simply due to chance. The p-value is calculated based on a significance level, say 0.05, which indicates that we would expect a result as extreme as that observed 5% of the time.

Parameter:

In terms of a distribution a statistical parameter that indexes a family of distributions. In terms of normal distribution the parameters are the mean and standard deviation. If these are known then the distribution is known exactly.

Population:

A finite or infinite group of entities subject to statistical study.

Population Inference:

This is inference about a population based on the behaviour of a random sample drawn from it. Typically the population parameters are approximated from the sample parameters.

Prognosis:

Forecasting of the probable course and outcome of disease, specifically the chances of recovery or the likelihood of relapse.

Sample:

A subset of a population

Sampling distribution:

This is the distribution created from samples randomly chosen from the population distribution. As the number of samples chosen increases the more similar the sampling distribution becomes to the population distribution and the closer the distribution parameters become in value. The law of big numbers states that when the number of samples increases above a certain cut-off the sampling parameters can then be used to approximate the population parameters.

Sample Pooling:

This involves creating individual samples from a range of biological entities of the same type or species and then pooling before performing a measurement on the sample.

Standard Error:

The standard error of a sample is the standard deviation adjusted for sample size. As such it is the standard deviation of the sample distribution, and is the expression of the uncertainty in a value.

Statistical Power:

The power of a statistical test is the probability that the test will reject a false null hypothesis i.e. accepts the alternative hypothesis when it is true. As such it reduces the chance of the test making a type II error. Power calculations are typically carried out to determine sample size from an already conducted pilot experiment, which assesses variability of the data.

Stimulus:

An entity that causes functional response in a subject following exposure to that entity.

Student t-test:

A t-test is any statistical test in which the test statistic has a student's t distribution. Student was the pen name of William Sealy Gosset who invented the t-statistic whilst working for Guinness in Dublin.

Taxonomies:

The classification of organisms in an ordered system that indicates natural relationships. This is typically hierarchical.

Time Series:

A sequence of measurements that follow a non-random order. The assumption is that successive values represent consecutive measurements taken at equally spaced time intervals.

Training set:

A training set consists of input samples of known class that can be used to 'train' the classification model to identify further samples of the same classes based on their gene expression profiles

Transfection:

The insertion of genetic material into a cell via the use of a bacterial plasmid. Can be transient (short-lived) or stable (permanent).

Test Set:

A test set is a set of samples, which were independent of a classification model building process, and are now used to assess the validity of the model.

Variable:

A factor that varies or is prone to variation.

Xenograft:

A graft of tissue taken from a donor organism and transplanted into a recipient organism. Usually concerns tumour tissue taken from human and transplanted into mouse or rat model.

References

1. Mei-Ling Ting Lee et al., (2000), PNAS, 97(18), 9834 – 9839
2. Nadon, R & Shoemaker, J., (2002), Trends in Genetics, 18(5), 265 – 271
3. Quackenbush, J., (2002), Nature Genetics Supplement, Vol.32, 496 – 500
4. Sasik, R., et al., (2002), Bioinformatics, 18(12), 1633 – 1640
5. Churchill, G., (2002), Nature Genetics Supplement, Vol.32, 490 – 495
6. Simon , R. & Dobbin, K., (2002), Bioinformatics, 18(11), 1438-45
7. Simon, R & Dobbin, K, (2005), Biostatistics, 6(1),27 - 38

Bibliography

Simon, R., Korn, E., McShane, L., Radmacher, M., (2004), *Design and Analysis of DNA Microarray Investigations (Statistics for Biology & Health)*, 1st ed., New York, Springer-Verlag Inc.

Websites

<http://www.statsoft.com/textbook>

<http://www.wikipedia.com/>

http://epe.lac-bac.gc.ca/100/201/300/cdn_medical_association/cmaj/vol-152/0027.htm#multiple

http://www.brc.dcs.gla.ac.uk/~rb106x/microarray_tips.htm

<http://www.tufts.edu/~gdallal/LHSP.HTM>

<http://www.socialresearchmethods.net/kb/expclass.htm>

<http://www.bcm.edu/mcfweb/?PMID=3101>

<http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage2.htm>