Exercices in linear models and experimental design

Alex Sánchez

Introduction

These exercises describe experimental situations that have been studied using some omics technology like microarrays.

Read each statement, identify the experimental design and write down the linear model -that is the design matrix- that describes it and build the contrast matrix needed to answer the questions posed by the researchers.

Case studies

1. Effect of treatment with SHAM1 on T-ALL cells expression

The dataset for the exercise is available at the entry Series GSE18198 of the in Gene Expression Omnibus. It consists in the analysis of expression profiles of human T-ALL cell lines treated with DMSO or SAHM1.

In short, NOTCH proteins regulate signaling pathways involved in cellular differentiation, proliferation and death. Overactive Notch signaling as been observed in numerous cancers and has been extensively studied in the context of T-cell acute lymphoblastic leukemia (T-ALL) where more than 50% of pateints harbour mutant NOTCH1. Small molecule modulators of these proteins would be important for understanding the role of NOTCH proteins in malignant and normal biological processes. In this stuy, researchers were interested in measuring the global changes in gene expression upon treatment of the human T-ALL cell lines HPB-ALL and KOPT-K1 with either vehicle alone (DMSO) or SAHM1, an alpha-helical hydrocarbon stapled peptide derived from the MAML1 co-activator protein.

Therefore, they designed an experiment that consists in triplicate cultures of KOPT-K1 or HPB-ALL cells treated with either DMSO alone or SAHM1 (20 uM) for 24 hours. Total RNA was extracted and hybridized to Affymetrix human U133 plus 2.0 microarrays (three arrays per treatment per cell line for a total of 12 arrays).

1. Describe -and name- the experimental design. Identify the experimental factors and their levels.
2. Write down the design matrix associated with this experimental design.
3. Build the contrast matrix that can be used to answer the following questions:
    1. Compare the effect of SHAM1 in KOPT-K1 cell line: KOPT-K1 treated with SHAM1 vs KOPT-K1 treated with DMSO (the vehicle)

2. The effect of SHAM1 in HPB-ALL cell line: HPB-ALL treated with SHAM1 vs HPB-ALL treated with DMSO.
3. The interaction: the differences between the two previous effects.

2. Comparison between three types of breast cancer

This case study is based on a paer published in http://www.ncbi.nlm.nih.gov/pubmed/15897907 (http://www.ncbi.nlm.nih.gov/pubmed/15897907) whose data are available in GEO as series GSE1561 series on the following link http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1561 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1561)

The researchers investigated three types of breast cancer tumors: apocrine (APO), basal (BAS) and luminal (LUMI). The classification is based on the resistance of tumors to estrogen and androgen receptors.

- Tumors classified as “APO” are negative for estrogen receptor (ER-) and positive for the androgen receptor (AR +).
- Those classified as “LUMI” are ER + and AR + and
- Those classified as “BAS” are ER- and AR.

The assignment of each sample to an experimental group can be obtained from this link: http://www.ncbi.nlm.nih.gov/geo/gds/profileGraph.cgi?gds=1329 (http://www.ncbi.nlm.nih.gov/geo/gds/profileGraph.cgi?gds=1329)

Obviously this is an observational study but its analysis can be done using a linear model approach as well.

1. Identify the experimental factors and their levels.
2. Write down the design matrix associated with this study design.
3. Build the contrast matrix needed to compare each tumor type with the oher two, that is:
    1. “APO” vs “LUMI”
    2. “APO” vs “BAS”
    3. “LUMI” vas “BAS”

3. Comparing healthy vs endometriosic patients in a RNA-seq study.

This has been directly translated from the description that a researcher has made to a core facility to request a study

We would like to have 40 samples of RNA-High-Seq analyzed.

We have 6 patients with endometriosis and 5 healthy donors. In each group we have 4 different cell populations call them A, B, C, D. We would like to compare each population between healthy and affected. Besides this we would like to compare populations C and D in healthy and the same comparison in affected.

1. Identify the experimental factors and their levels.
2. Write down the design matrix associated with this study design.
3. Build the contrast matrix needed to do the comparisons required.

4. Effect of the thermogenic gene program during adipocyte differentiation

The data for this study had been uploaded into the Gene Expression Omnibus (GEO). The dataset selected is identified with the accession number: GSE100924.

The study that generated the data investigated the function of gene ZBTB7B (http://www.genecards.org/cgi-bin/carddisp.pl?gene=ZBTB7B (http://www.genecards.org/cgi-bin/carddisp.pl?gene=ZBTB7B)). This gene activates the thermogenic gene program during brown and beige adipocyte differentiation regulating brown fat gene expression at ambient room temperature and following cold exposure. The experiment compared 10 weeks old mice with the gene deactivated ("KO" or knockout) or not ("WT" or Wild type) at two different temperatures, ambient room temperature (RT, C) or following cold exposure (COLD, C) for 4 hrs. The sample size of the experiment is 12 samples, three replicates of each group. The microarrays used for this experiment were of type Mouse Gene 2.1 from Affymetrix, now Thermofisher, one of the most popular vendors of microarray technology.

In this example we want to check the effect of knocking out a gene ("KO vs WT") separately for cold and RT temperature. Also we want to test if knocking out the gene affects the (distinct) way the temperature influences adipocyte differentiation.

1. Identify the experimental factors and their levels.
2. Write down the design matrix associated with this study design.
3. Build the contrast matrix needed to do the comparisons required.

5. Molecular basis of age-associated cytokine dysregulation in LPS-stimulated macrophages

This study was published in the Journal of Leukocyte Biology (2006;79:1314-1327).

The goal of the experiment which generated the data was to understand the molecular basis of processes regulated by a molecule (cytokine) in aged mouse.

To do this a microarray analysis was performed on RNA from resting and lipopolysaccharide (LPS) stimulated mice using the Affymetrix Mouse Genome 430 2.0 gene chip.

The database that contained the study has been deprectaed but information on all the characteristics of the dataset and the experiment can be found in the following link: ftp://caftpd.nci.nih.gov/pub/caARRAY/experiments/caArray_bonda-00136/Readme.txt (ftp://caftpd.nci.nih.gov/pub/caARRAY/experiments/caArray_bonda-00136/Readme.txt)

1. Identify the experimental factors and their levels.
2. Write down the design matrix associated with this study design.
3. What do you think may be the "reasonable questions" in this study? Build the contrast matrix needed to do the comparisons required.

A "raw" approach to using linear models for expression analysis

One of the best known applications of the linear model in bioinformatics is its use to detect differentially expressed genes. There is a well-known methodology called `limma` which basically consists of estimating a linear model for each of the rows of a data matrix and using it to make the desired comparisons.

In our case, we will suppose that we have an expression matrix made up of the expressions of 10 genes measured in 15 individuals. These genes have been selected because they are considered markers of a certain pathology. Specifically, the treatments are expected to reduce the expression of each of the genes. Each individual has been assigned to one of three possible groups:

- CTL (Control), no treatment applied
- T1 (Classic) ,, the traditional treatment has been applied
- T2 (Innovative), a new treatment has been applied

Our goal will be to build a linear model, *for each gene* relating the expression to the treatment and use it to answer, gene by gene, the following questions:

- Does the classical treatment determine less expression of the gene than the control?
- Does the Innovative Treatment determine less gene expression than the control?
- Is there a difference between the expression associated with the classic and the innovative treatment?

To answer the question, start with a gene.

The procedure to follow will be:

1. Write the linear model that describes the relationship between the expression of a gene and the treatment that the individual has received.
2. Describe the contrasts you would use to answer the three questions.
3. Estimate the model using the ordinary least squares estimator.
4. Make the requested comparisons and provides an estimate of the treatment effect, a confidence interval for that effect, and a significance p-value of the calculated difference.
5. Considering the basic regression diagnoses, do you think that the conditions of application of the general linear model are met?

All of the above -except for question 5- must be repeated for each gene. Answer one of the following two questions

1. Write in R the code necessary to iterate the previous procedure through the matrix.
2. If you don’t know how to do it, repeat the procedure manually for the first three genes of the expression matrix.