Genome Informatics: Assignment 1

University of Cambridge

Henrik Åhl

November 4, 2016

Preface

This is an assignment report in connection to the *Functional Genomics* module in the Computational Biology course at the University of Cambridge, Michaelmas term 2016. All related code is as of November 4, 2016 available on https://github.com/supersubscript/compbio/tree/master/src/fg_assignments/assignment_1/, or available per request by contacting hpa22@cam.ac.uk. Likewise, the corresponding assignment can be found on https://github.com/supersubscript/compbio/tree/master/general/fg_assignment_1.pdf.

1 Introduction

2 Problems

Part A

Describe the principles guiding the design of a microarray experiment

Microarray experiments are by construct in need of rigourous underlying design in order to enable and improve subsequent analysis.

The core principles of microarray sequences can be summarised, as by Fisher [1], as *randomisation*, *replication* and *blocking*.

Randomisation is the principle of assigning samples to groups at random, e.g. by constructing a set of labels and then accordingly assign these to the relevant compounds of interest, forming arrays with a randomised setup of samples. This serves to counter uncontrolled factors which might affect the outcome of the experiment, should statistically dampen the effects of these.

Replication is the process of repeating the data acquisition, in principle from scratch, in order to account for the variability in the experiment outcome.

Blocking signifies the notion behind the distribution of samples such that comparisons, for example between different microarrays, can be performed adequately.

Give a description of Illumina microarray platform including the advantages compared to other commercial platforms

Illumina BeadArray bases itself on fiber optic bundles or silica slides in which microwells have been arranged. Into these wells, silica beads are placed an coated with $\sim$ 100000 copies of specific oligonucleotides, which act to capture certain genes or sequences. The captures consist of a 23 base pair long address, tied to a 50 base pair sequence specific probe.

In the process of capture, a decoding procedure is undergone to determine which beads occupy which well. The sample of interest is then applied, whereafter strands who have bound are measer using a fluorescent label.

In particular, Illumina microarrays are advantageous as they are highly reproducible, and allow for fast multiplex processing. Microarrays in general have also traditionally been highly used, which means that the corresponding analysis is well-developed, and the workflow comparably easy.https://www.ncbi.nlm.nih.gov/probe/docs/techbeadarray/ http://www.illumina.com/techniques/m

SPREAD OUT THING? (SEE SLIDES)

Describe at least two different normalisation methods for single- channel microarrays. What could help in choosing the most appropriate method?

Quantile Normalisation is a way of relativising data points so that data between arrays get the same distribution. This works by assigning data points within arrays with labels according

to their rank, sorting the data in columns within each array separately, and subsequently normalising each resulting row by the row mean. This allows These methods rely on the assumption that changes between samples are due to technical variance, and not a product of the data itself. In transforming the distributions, the identification of patterns is simplified [2, 3].

LOESS is a regression model working with a nearest-neighbour approach. LOESS works by separating the data set into subgroups to which simple models are fitted in order to build up a function describing the deterministic part of the data. A set of low-degree polynomials are fitted using a weighted least-squares approach at each value in the variable range(s) of the function, effectively creating a higher-order polynomial fitted to the data, which is used instead of the singular data points [4].

Explain why a probe filtering step is important and give an example of a meaningful filtering criterion

Part B

References

[1] Sir Ronald Aylmer Fisher et al. "The design of experiments". In: (1960).

[2] Dhammika Amaratunga and Javier Cabrera. *Exploration and analysis of DNA microarray and protein array data*. Vol. 446. John Wiley & Sons, 2004.

[3] Stephanie C. Hicks and Rafael A. Irizarry. "When to use Quantile Normalization?" In: *bioRxiv* (2014). DOI: 10.1101/012203. eprint: http://biorxiv.org/content/early/2014/12/04/012203.full.pdf. URL: http://biorxiv.org/content/early/2014/12/04/012203.

[4] William S. Cleveland and Susan J. Devlin. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting". In: *Journal of the American Statistical Association* 83.403 (1988), pp. 596–610. DOI: 10.1080/01621459.1988.10478639. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1988.10478639. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478639.

A Appendix 1

B Appendix 2