Quantitative Understanding in Biology Module I: Statistics
Lab: The HELP DNA Methylation Assay

Background

Cytosine methylation is an important epigenetic modification commonly found in eukaryotes. DNA methylation is known to play an important role in the regulation of gene expression, and perturbations in genome-wide DNA methylation patterns are associated with cancer (Herman and Baylin, Gene Silencing in Cancer in Association with Promoter Hypermethylation; *New England Journal of Medicine*, 2003).

The HELP assay interrogates cytosine methylation status on a genomic scale (Khulan et. al., Comparative isoschizomer profiling of cytosine methylation: The HELP assay; *Genome Research*, 2006). It makes use of two restriction enzymes (HpaII and MspI), which both cleave DNA at CCGG sites. However, while MspI will cleave DNA at all such sites, HpaII will only cleave sites where the cytosine in the CpG is not methylated. In the assay, DNA fragments that are produced by each of these enzymes are separately amplified by PCR and labeled with different fluorescent dyes. The particular PCR process used in the HELP assay only works well for DNA fragments with a size of 200bp – 2000bp; such fragments are known as HTFs (HpaII Tiny Fragments).

The key idea behind the HELP assay is that by comparing the quantity of HTFs derived from MspI and HpaII treatment, one can say something about the methylation state of sites in the genome. In simple terms, if, for a particular fragment, you see a strong MspI signal but no HpaII signal, you might conclude that the site was not methylated in the genome. If you see both a strong MspI and HpaII signal, you might conclude that the site was not methylated.

To compare the amounts of MspI and HpaII fragments, a mixture of the two labeled fragments are run over a custom designed microarray. For each of a large set of possible fragments (determined bioinformatically), a set of probes has been placed on the array. Each of the probes (`PROBE_ID` in the pair files) in a probe set (`SEQ_ID` in the pair files) is designed to be unique in the genome, and is expected to only specifically bind to its target fragment. To assist (or complicate?) analysis, the chip is also designed with a number of random probes to help quantify non-specific binding. Each array is then scanned twice (once for each fluorescent dye), and a set of pair files giving spot intensities (PM in the pair files) for each probe is generated.

The Project

In the project, we will compare the DNA methylation profiles from two samples obtained as part of a larger Leukemia study (these data are unpublished). The raw data for your analysis consists of five files. The first two files, `105958_532.pair.txt` and `105958_635.pair.txt`, hold HELP microarray

results for MspI and HpaII channels (respectively) for the first biological sample. The second two files, `105961_532.pair.txt` and `105961_635.pair.txt`, hold similar results for the second biological sample. The last file, `2006-10-26_HG17_HELP_Promoter.ndf.txt`, details the design of the custom microarray used in this study. Note that these files are provided exactly as received by Nimblegen (the company that ran the microarray experiments in this study).

The goal of the project is to provide a list of which loci on the genome (probe sets) show differential methylation across the two samples. There are other possible areas of exploration, and you should feel free to pursue them as well if you want (this is optional). For example, you might investigate whether the arrays show systematic spatial variations or other artifacts (you are given the x, y coordinates of every probe on the array) and what impact they might have on the results.

Hints and Guidelines

- This not a toy example or text-book exercise. Real-world data are not perfect, and there is no single correct answer. You will need to make decisions and assumptions in your analysis, and a significant portion of your evaluation will be based on the logic and justification for these assumptions.
- Analysis of HELP data is still very much an open field. It is very possible that you'll come up with something new and significant (appropriate attribution will be given to any new findings).
- There is a lot of data to handle here. Use the right tools and methods. Do not attempt to perform repetitive tasks manually; instead, educate yourself on how to automate them.
    - If you haven't worked through a good R tutorial yet, you might want to do so before you start working on this.
    - You will probably want to learn R commands such as: `read.delim`, `data.frame`, `subset`, and several others. You will also want to learn how to write small functions in R, and run loops, etc.
    - Keep in mind that these data represent only two of over 100 samples that were tested in the overall study, so this is actually pretty small as far as full studies go.
    - Parts of the analysis may be computationally intensive and time consuming. You may want to develop your analysis on a subset of the data.
- You may want to test/validate your analysis against other data sets. For example, HELP data for technical replicates (samples taken from the same tissue, for which you therefore expect no differences) can be found on the GEO web site here (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10894).
- As part of your submission, include **scripts** that will reproduce all of your results (reported quanties, lists, and figures). Assuming that the reader has the original data files available, you should include everything needed to rerun you analyses (ideally with a single command on my part).
- You are free (and **strongly** encouraged) to work in groups. You may submit your results in pairs.

