Advanced Statistical Computing Final Project (2011)

Most human genes have two alleles A and B, one from mother and the other from father. Gene expression can be allele specific (or allele-skewed), that is, one allele can be transcribed (or "expressed") at a higher level than the other allele. Similarly, many other epigenetic signatures are allele-specific. The next-generation sequencing technology provides an opportunity for scientists to measure allelic events by counting sequence reads for each allele at certain genomic loci known as SNPs. Your project data contain allelic counts for a number of different epigenetic signatures at approximately 6000 SNPs. We treat each epigenetic signature as a dataset. Each dataset has two columns. One column is the read count for allele A, measuring the level of expression or epigenetic modification at allele A. The other column is the read count for allele B. Each row is a SNP. Intuitively, the two allele counts at each SNP contain information about the underlying allele-specificity status of the SNP, which is unobserved. As you can see from the data, the counts generally are small due to technology limitations. Because of the small counts, statistical power for detecting allele-specific events is limited if you only look at one dataset. Since many epigenetic signatures are correlated, allele-specificity in one dataset has a strong tendency to (but does not always) co-occur with allele-specificity in other datasets. The purpose of your project is to explore whether jointly modeling all datasets may help you to improve statistical power of detecting allele specificity.

A potential solution is to use a generative probabilistic model to describe the data. Assume that there are S SNPs and D datasets. For each SNP s and dataset d, you have two counts n_{sdA} and n_{sdB}. A SNP s could be potentially allele-specific (a_s =1) with prior probability $p(a_s = 1) = \pi_1$, or not allele-specific with prior probability $p(a_s = 0) = 1 - \pi_1$. Assume that we first use these prior probabilities to generate a_ss for different SNPs independently. Conditional on s is allele specific (i.e., a_s =1), we then generate indicators b_{sd} to indicate whether it is allele-specific in dataset d (b_{sd} =1) or not (b_{sd} = 0), with conditional probability $p(b_{sd} = 1 \mid a_s = 1) = q_d$. We assume that conditional on a_s, b_{sd}s are independent. If a_s =0, then b_{sd} = 0 for all datasets. Finally, conditional on b_{sd}, you generate the observed data n_{sdA} and n_{sdB}. Let $n_{sd} = n_{sdA} + n_{sdB}$. If b_{sd} =0, then $n_{sdA} \mid b_{sd}$ =0, $n_{sd} \sim$ Binomial(n_{sd}, p_{d0}). If b_{sd} =1, then $n_{sdA} \mid b_{sd}$ =1, $n_{sd} \sim$ some other distribution which you will need to specify. Note that p_{d0} is the baseline allele A frequency in dataset d for SNPs without allelic skewing. p_{d0} is close to but not necessarily be 0.5. It may deviates a little from 0.5 due to technology bias. Also, p_{d0} is different for different dataset, but remains a constant for all SNPs within a dataset.

In the model above, π_1, q_d, a_s, b_{sd} are unknown quantities you are interested in. p_{d0} and other parameters for modeling n_{sdA} is also unknown. You need to figure out how to handle them. Your goal is to infer π_1, q_d, a_s, b_{sd} from the observed data. You are asked to do the following:

(1) Develop and implement a model and algorithm to infer π_1, q_d, a_s, b_{sd}.
(2) Use your estimated posterior probability P(a_s | Data) to rank SNPs. Draw a curve $f(n)$ that shows among the top n predictions, how many SNPs in your ranking list are on

chromosome X. chromosome X is special since most SNPs on this chromosome are allele specific due to imprinting. Therefore you can use chrX to benchmark your methodology.

(3) Compare your method with a simple method where you analyze each dataset separately. Does the joint model improve statistical power? To compare different methods, you should plot multiple $f(n)$ curves on the same plot. You should also report the area under the curve (AUC) for each method.

(4) Write a report to describe your model, algorithm and results. The report cannot exceed 6 pages.

If you need some hint, the following paper uses a similar model for a different problem and it may help:

Wu H and Ji HK (2010) JAMIE: joint analysis of multiple ChIP-chip experiments. Bioinformatics. 26:1864-1870