Application of hierarchical bayesian models to PPAR related microarray data (part 1)

Jinlu Cai, Jin Gong

Introduction

Background:

Peroxisome proliferator-activated receptors (PPARs) are transcription factors. PPARγ is a master regulator of adipocyte differentiation. PPARγ is also expressed in endothelial and has been shown to have an important role in the regulation of vascular function. Furthermore, patients with dominant negative mutations of PPARγ have been reported to have hypertension. However, the molecular mechanism by which PPARγ exerts its effect in the genome-wide transcriptional regulatory network of its target genes remains to be elucidated.

Experimental design:

To assess the response to PPARγ interference, we used transgenic mice containing a dominant negative form of PPARγ. The dominant negative mutated copies only expressed in the endothelial. Wile-type mice from the same strain were used as the control. For the microarray hybridizations, Affymetrix GeneChip Mouse Genome 430 2.0 array was used for the experiments and 3 biological replicates from each group were used. So, we have 3 controls and 3 transgenic groups in total and each group includes 45101 genes (or probe-sets).

Data analysis

Overview:

The gene level analysis requires determining whether observed differences between control and transgenic groups in expression are significant or not. Using the observed data directly for 2-sample T test is lack of robustness, due to the low replication. We propose to apply a bayesian framework to better estimate the difference between control and transgenic groups. Hierarchical Bayesian model will be set up and MCMC will be carried out via winbugs.

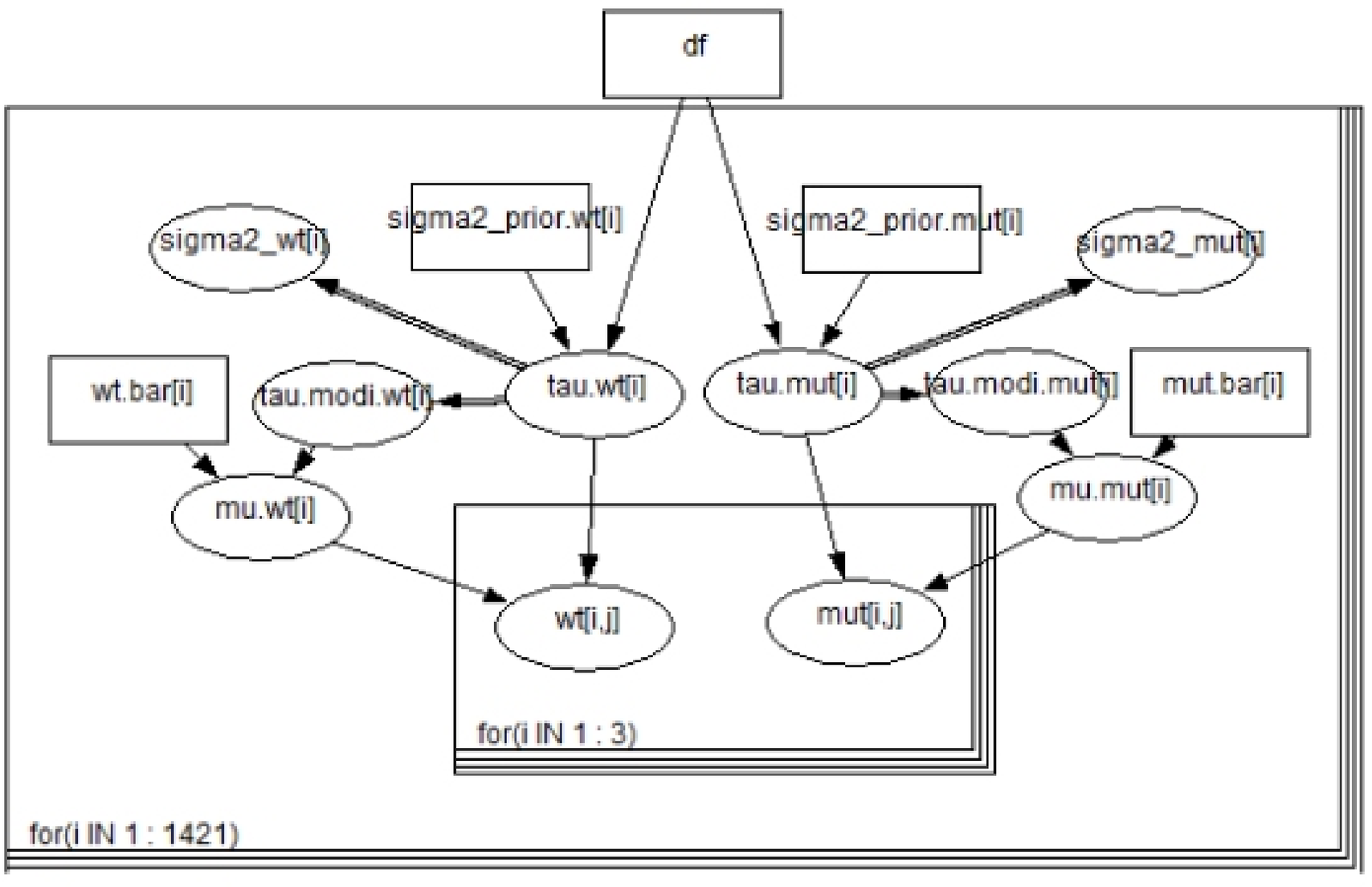
Dataset -- selection of genes (or probe-sets):

As there are 45101 genes on the microarray platform, it is time-consuming and not realistic for us to process all the genes by MCMC in winbugs. Therefore, we apply a filtering scheme to select genes to fit our Bayesian model. We have selected genes with at least 1.5 fold change (both up- and down- regulated), and as well as significant at P value=0.05 from two-sample t test, in which un-equal variance is assumed. We result in 1421 genes and they are 374 up-regulated and 1047 down-regulated respectively.

Model setup:

Due to the small size of samples (N=3 for each gene), frequentist method tends to underestimate the variance, which in turn would lead to a higher type I error. A Bayesian approach would capture background information (from priors) and integrated it with current samples to generate more robust estimates. Thus it could address the problem of gene comparison with small sample size better than the frequentist method.

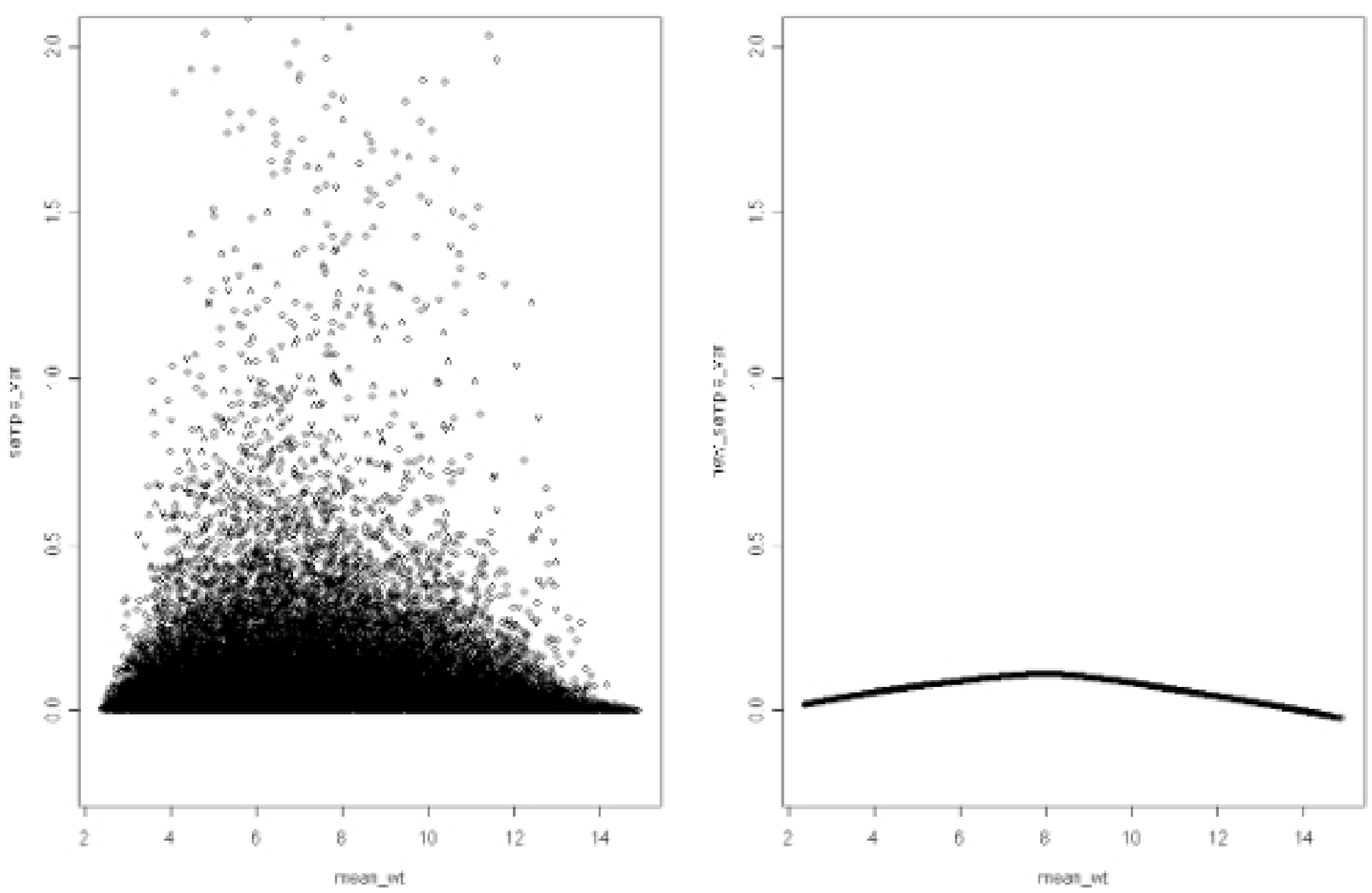
We are interested in estimating the means and the variances of each gene in both control and transgenic groups using a bayesian approach. Then we can conduct the two sample t-test to compare the expression levels of genes between two groups. A three-stage bayesian model is setup as below.



Prior calculation:

In order to compare the influence of different priors, two different regression models, nominally, non-linear local regression (Loess method) and window-smooth regression, are used to obtain estimates of precisions.

For both methods, genes are ranked according to their expression levels first (all 45101 genes are included for prior calculation), separately for control and transgenic groups. Loess local regression is performed using R, in which sample mean serves as the predictor and sample variance is the response variable. Here we assume the degree of freedom for the prior of variance as 2, which is a quite conservative estimation, as we are not able to know how many data points have been taken by Loess local method for estimation. Take the control group for example, the scatter plots before and after regression are shown as the below.



For the window-smooth regression, the prior of variance of one gene is calculated based on 100 neighboring genes with similar expression level. Technically, in the ranked list, the above 50 and below 50 genes of a specific one are included for the calculation. Therefore, we have the degree of freedom equal to 303-1=302. In fact, the size of the window (default = 100) can be adjusted. We do not have a good argument for selection of the window size, therefore, we follow a previous study and fix on 100. Plus, we have tried window size equal to 50 and 200 as well, shown as the below. (From left to right, the window size is