

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. PPAR γ is a master regulator of adipocyte differentiation. PPAR γ 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 PPAR γ have been reported to have hypertension. However, the molecular mechanism by which PPAR γ 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 PPAR γ interference, we used transgenic mice containing a dominant negative form of PPAR γ . The dominant negative mutated copies only expressed in the endothelial. Wild-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.

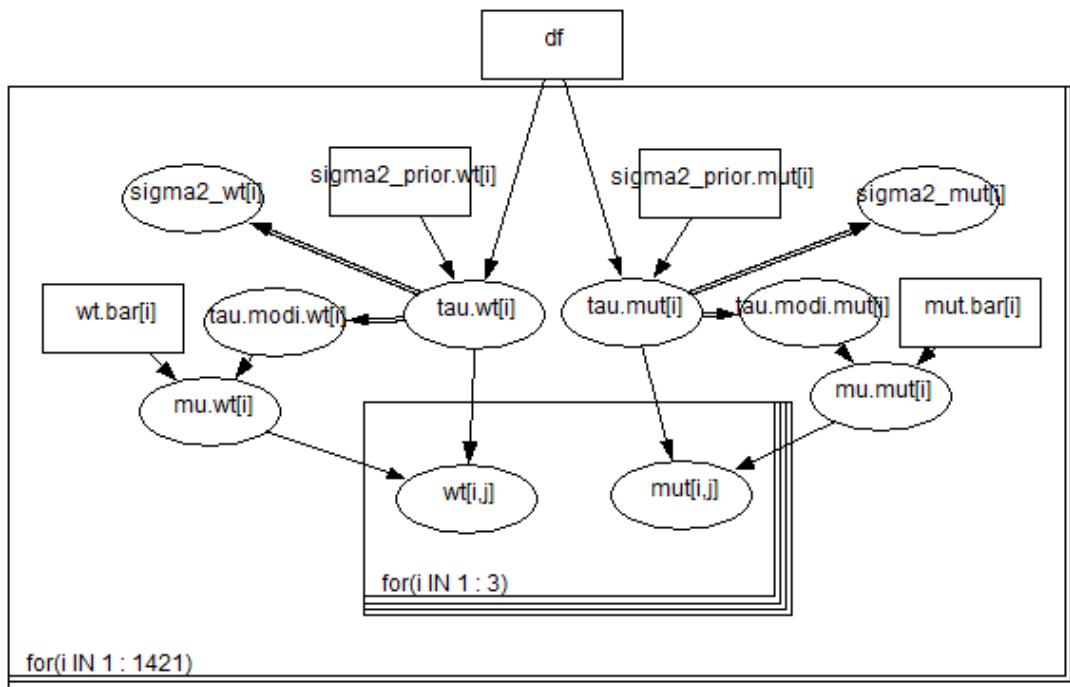
Dataset -- selection of genes (or probe-sets):

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.

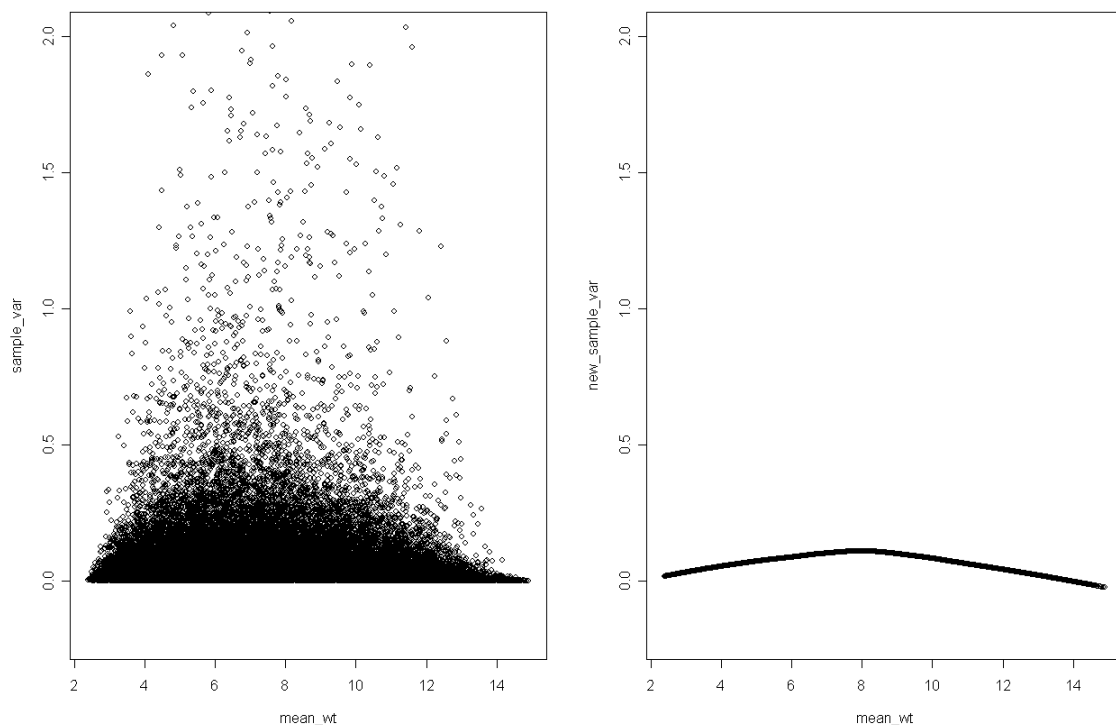
We are interested in estimating the means and the variances of each gene in both control and transgenic groups using a bayesian approach. Then we can conduct the two sample t-test to compare the expression levels of genes between two groups. A three-stage bayesian model is setup as below.



Prior calculation:

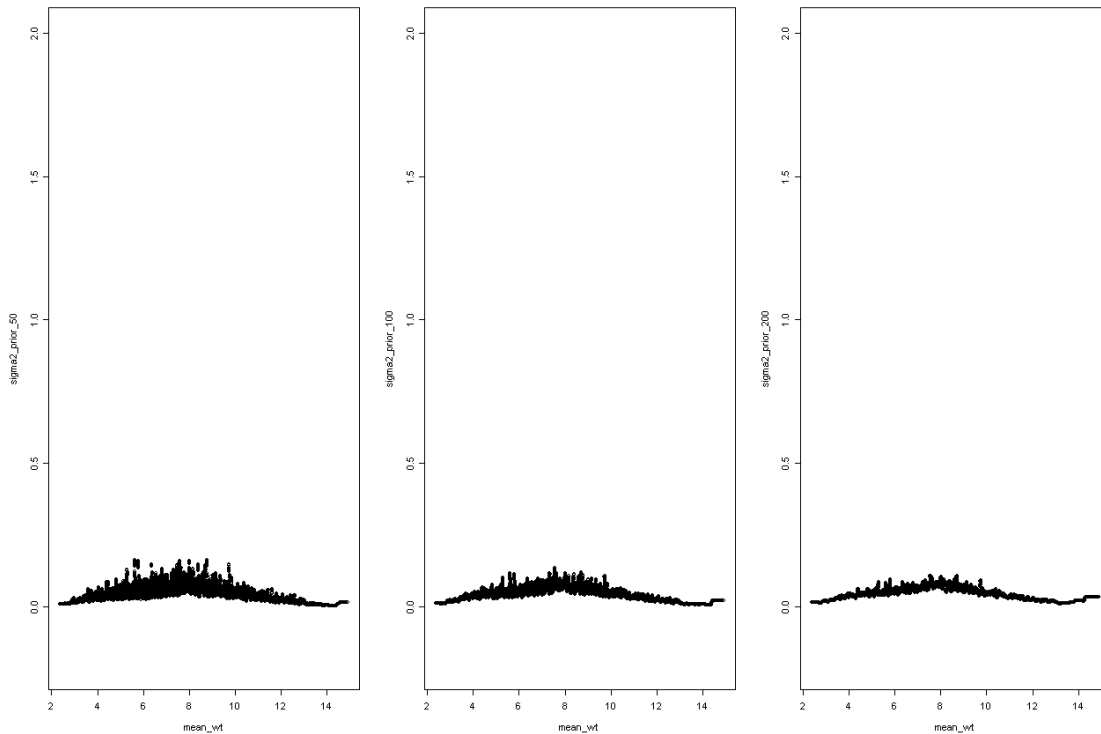
In order to compare the influence of different priors, two different regression models, nominally, non-linear local regression (Loess method) and window-smooth regression, are used to obtain estimates of precisions.

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

50,100,200 respectively)



WINBUGS code:

We have data of 1421 genes and for each gene there are 6 data values (3 for control group, 3 for transgenic group). So we have large vectors/matrices in the data list. Only a few lines of them are listed as the below.

We monitor the difference between the expression of control and transgenic groups directly as well.

```
#WIBUGS CODE
model
{
  for (i in 1:N)
  {
    for (j in 1:3)
    {
      wt[i,j]~dnorm(mu.wt[i],tau.wt[i])
      mut[i,j]~dnorm(mu.mut[i],tau.mut[i])
    }

    mu.wt[i]~dnorm(wt.bar[i],tau.modi.wt[i])
    mu.mut[i]~dnorm(mut.bar[i],tau.modi.mut[i])

    diff[i]~mu.wt[i]-mu.mut[i]

    tau.modi.wt[i]<-tau.wt[i]*.0001
    tau.modi.mut[i]<-tau.mut[i]*.0001

    tau.wt[i]~dgamma(df,sigma2_prior.wt[i])
    tau.mut[i]~dgamma(df,sigma2_prior.mut[i])
  }
}
```

```

        sigma2_wt[i]<-1/tau.wt[i]
        sigma2_mut[i]<-1/tau.mut[i]
    }
}

#data
list(
N=1421,
wt.bar=c( 2.381754,...,2.416665),
mut.bar=c( 10.658698621,...,7.3763444564),
df=302,
sigma2_prior.wt=c( 0.01195393,..., 0.01195393),
sigma2_prior.mut=c( 0.01258357, ...,0.01258357),

wt=structure(
.Data=c(2.435167,2.314355,2.395740,
        2.347672,2.424980,2.477343),
.Dim=c(1421,3)
),

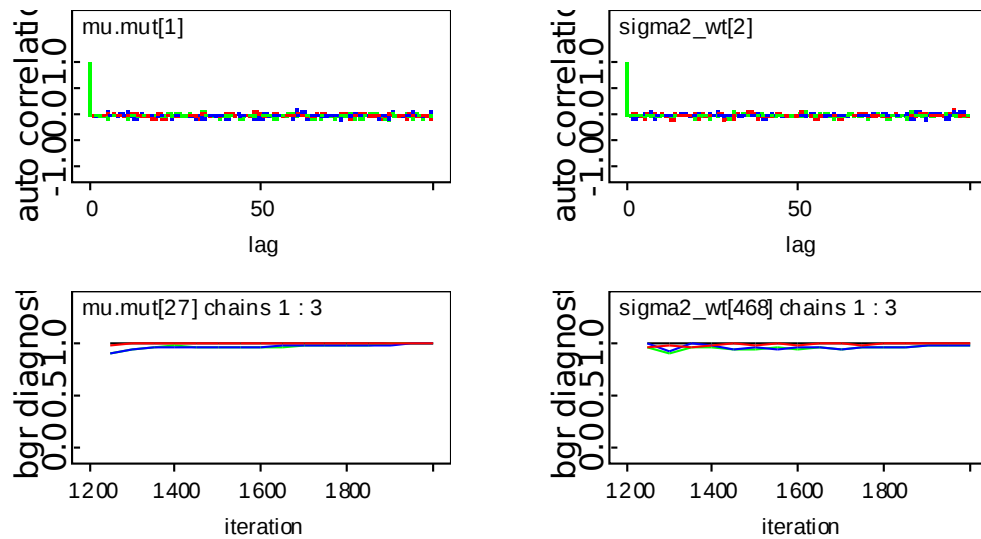
mut=structure(
.Data=c( 2.853000, 2.796998, 2.829541,
        2.687230, 2.765639, 3.027291),
.Dim=c(1421,3)))

```

Convergence assessment:

To better assess convergence of the model, three chains were generated. Initial values for each chain were produced automatically by winbugs.

Initially the model is updated for 2000 iterations. All the parameters have been monitored. The history plots do not show any strong signal against convergence. The autocorrelation series drop significantly from 1 to near 0 after order 2 for all parameters. BGR diagnosis plots indicate that approximately for all parameters R ratios start to converge to around 1.0 after 1200 iterations (though some of the parameters have earlier convergence points). Considering the large number of parameters we have, we decide to discard the first 1200 iterations. The final estimate of parameters including the 95% credit intervals were based upon the last 800 iterations. Here are some graphs of convergence assessment randomly selected as the below.



Winbugs Output:

For example, output of Stat for control group when priors are obtained by local regression are shown as below (only the mean value of the first 10 genes)

	mean	sd	MC_error	val2.5	median	val97.5	start
sample							
mu.mut[1]	10.66	0.01075	2.227E-4	10.64	10.66	10.68	1201 2400
mu.mut[2]	10.07	0.01617	3.02E-4	10.04	10.07	10.11	1201 2400
mu.mut[3]	9.915	0.01462	2.723E-4	9.886	9.914	9.944	1201 2400
mu.mut[4]	11.41	0.01446	2.899E-4	11.39	11.41	11.44	1201 2400
mu.mut[5]	9.365	0.01667	3.559E-4	9.333	9.365	9.398	1201 2400
mu.mut[6]	5.303	0.00984	1.814E-4	5.285	5.303	5.323	1201 2400
mu.mut[7]	9.932	0.01761	3.713E-4	9.898	9.932	9.967	1201 2400
mu.mut[8]	11.87	0.01058	2.105E-4	11.85	11.87	11.89	1201 2400
mu.mut[9]	8.354	0.01609	3.105E-4	8.323	8.354	8.386	1201 2400
mu.mut[10]	9.338	0.02271	4.705E-4	9.294	9.338	9.383	1201 2400

The MC error is quite small compared to mean and this happens to all other parameters as well, which confirms the convergence as well.

Result and discussion:

Final T test:

After estimating mean and variance for each gene by Bayesian framework, we apply the two-sample t test with un-equal variance assumption again, in order to identify genes with robust changes between control and transgenic groups. We also use the cut-off P value=0.05 for the t test.

For the estimations using priors from loess regression, 1175 out of 1421 genes have been selected. Meanwhile, for the estimations using priors from window-smooth method, all genes have been shown significant. Then, we pay attention to the final estimation of variances of expression values after MCMC processing, when the priors from window-smooth method are used. We find that these estimated variances are extremely small, which can help explain why all the 1421 genes are significant.

Difference monitored by winbugs:

As the above winbugs codes show, the difference between control and transgenic groups for each gene has been monitored directly by winbugs. We focus on the 95% credible interval of each $\text{diff}[i]$ and all of them do not include “0” whatever priors have been used for calculation, which indicates that all these 1421 genes significantly change.

Analytical analysis – regularized t-test:

One approach to conduct the test based on the Bayesian method is to construct the t-test using the mean of posterior (MP) estimate. This is called regularized t-test by Baldi. and Long. The MP estimate is given by

$$\hat{\mu} = m \quad \text{and} \quad \hat{\sigma}^2 = \frac{v_0 \sigma_0^2 + (n-1)s^2}{v_0 + n - 2} ,$$

where v_0 and σ_0^2 are the degrees of freedom and scale of the prior Gamma distribution of σ^2 , and s is the sample standard deviation.

The test statistic used by the regularized t-test has the same form as the frequentist t-test, which is

$$t = (m_c - m_t) / \sqrt{\frac{s_c^2}{n_c} + \frac{s_t^2}{n_t}} .$$

It approximately follows a Student distribution, with

$$f = \frac{[(s_c^2/n_c) + (s_t^2/n_t)]^2}{\frac{(s_c^2/n_c)^2}{n_c - 1} + \frac{(s_t^2/n_t)^2}{n_t - 1}}$$

degrees of freedom.

We conducted regularized t-test with our data, using both the Loess priors and the Window-smooth priors. Among all the 1421 genes, the regularized t test are significant for

609 genes when Loess priors were used, and for 959 genes when Window-smooth priors were used.

Discussion:

When we compare the result from final T test using means and variances estimated from MCMC in winbugs and the result of analytical test, fewer genes are significant using Loess priors than Window-smooth priors in both cases, because Window-smooth priors lead to smaller estimation of variances. If we focus the comparison on MCMC and analytic test, fewer genes have been shown significant in analytical test, which might be due to the limited iterations. Based on our convergence assessment, 1200 burnt-in iterations seem to be enough, but more iterations might be needed to approximate the theoretical results.

When we compare the result from final T test using means and variances estimated from MCMC and the result using the $\text{diff}[j]$ monitored directly from MCMC in winbugs, more genes are significant in the latter approach. However, the reason behind this phenomenon is beyond our knowledge.

Appendix

Jinlu Cai:

- Prior calculation,
- Winbugs code writing (model),
- Winbugs code running (convergence assessment);

Jin Gong:

- Data preprocessing,
- Winbugs code writing (data),
- Final T test and analytical analysis.