Dec, 2015

February 9, 2016

Meeting

- Review TADA (Transmission And De novo Association) model.
- Test the likelihood values of TADA on grids of parameters on **D0**.
- Estimate parameters using box-constrained optimization and MCMC on **D1** and **D2**:

D0. De Rubeis, Silvia, et al. "Synaptic, transcriptional and chromatin genes disrupted in autism." Nature 515.7526 (2014): 209-215.
D1. He, Xin, et al. "Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes." PLoS Genet 9.8 (2013): e1003671.
D2. Only the FMRP gene set from the data set D0.
From Darnell: 842 genes.

TADA

Let π be the fraction of risk genes in all genes.

Main parameters		Variants (LoF and mis3)
Mutation rate (μ)		De novo mutation
Relative risk (γ)	=>	Transmitted variations
Population frequency (q)		Variants in case-control studies

For each i^{th} gene, TADA uses a Bayesian approach to test the hypothesis
$H_0 : \gamma_i = 1$ against the alternative $H_1 : \gamma_i \neq 1$
=> A fraction π of risk genes (per total genes) follows the H_1 model.

The model incorporates information across genes, assumming that:

- Relative risk γ:
  $\gamma \sim Gamma(\bar{\gamma} * \beta, \beta)$.
- Population frequency of variants q:
  - Risk genes: $q_1 \sim Gamma(\rho_1, \nu_1)$
  - Normal genes: $q_0 \sim Gamma(\rho_0, \nu_0)$

=> Need to estimate $\boxed{\bar{\gamma}, \beta, \rho_1, \nu_1, \rho_0, \nu_0, \pi}$.

With each type of data/variants (x), at each gene:

$$P(x|paramterers) = \pi P(x|H_1) + (1 - \pi)P(x|H_0)$$

Two types of mutations are tested: Loss-of-function (LoF) and probably damaging (Mis3).

Test two models:

External product

$$\boxed{P(x|paramterers) = \prod_{i=1}^{m} \left[\pi P(x_{i_{LoF}}|H_1) + (1-\pi)P(x_{i_{LoF}}|H_0)\right] \left[\pi P(x_{i_{mis3}}|H_1) + (1-\pi)P(x_{i_{mis3}}|H_0)\right]}$$

Internal product

$$\boxed{P(x|paramterers) = \prod_{i=1}^{m} \left[\pi P(x_{i_{LoF}}|H_1)P(x_{i_{mis3}}|H_1) + (1-\pi)P(x_{i_{LoF}}|H_0)P(x_{i_{mis3}}|H_0)\right]}$$

Bayes factor

$$\boxed{BF = \frac{P(x|H_1)}{P(x|H_0)} = \frac{P(x_{LoF}|H_1)P(x_{mis3}|H_1)}{P(x_{LoF}|H_0)P(x_{mis3}|H_0)}}$$

Model for case-control data

Relative riks:
$\gamma|H_1 \sim Gamma(\bar{\gamma} * \beta, \beta)$
Frequency of variants:
$q|H_1 \sim Gamma(\rho, \nu)$
$q|H_0 \sim Gamma(\rho_0, \nu_0)$
The model:
$P(x|H_0) = \int p(x|q, \gamma = 1)p(q|H_0)dq$
$P(x|H_1) = \int p(x|q, \gamma)p(q|H_1)p(\gamma|H_1)dqd\gamma$

Model for case-control data (cont)

For simplicity, let $\boxed{q|H_1 = q|H_0 = q \;\sim Gamma(\rho, \nu)}$,
we will calculate the conditional distribution of variants in cases on total variants of cases and controls.
$X_{case}\ Pois(\lambda_1)$; $X_{control}\ Pois(\lambda_0)$
With $\lambda_1 = 2N_{case} * q * \gamma; \lambda_0 = 2N_{control} * q$
$X = X_{case} + X_{control}$
$X \sim Pois(\lambda_1 + \lambda_0)$
Question:
1) should let $\boxed{q = \epsilon * q_0}$.
2) Relationship between q and μ.

Model for case-control data (cont2)

At the i^{th} gene,

$$P(X_{case} = k|X = n) = \frac{P(X_{case}=k,X=n)}{P(X=n)} = \frac{P(X_{case}=k,X_{contro}=n-k)}{P(X=n)}$$

$$= \frac{\frac{e^{-\lambda_1}\lambda_1^k}{k!}\frac{e^{-\lambda_0}\lambda_0^{n-k}}{(n-k)!}}{\frac{e^{-(\lambda_1+\lambda_0)}(\lambda_1+\lambda_0)^n}{n!}}$$

$$= \frac{n!}{(n-k)!k!}\frac{\lambda_1^k\lambda_0^{n-k}}{(\lambda_1+\lambda_0)^n} = C_n^k p^k(1-p)^{n-k}$$

With

$$\boxed{p = \frac{\lambda_1}{\lambda_1+\lambda_0} = \frac{2N_{case}q\gamma}{2N_{case}q\gamma+2N_{control}q} = \frac{N_{case}\gamma}{N_{case}\gamma+N_{control}}}$$

- First way (**Poisson distribution)**
  $X_{dn}|H_1 \sim Pois(2N_{dn}\mu\gamma)$
  $X_{dn}|H_0 \sim Pois(2N_{dn}\mu)$
- **Second way (Binomial distribution)**
  $X_{dn}|H_1 \sim Binomial(2N_{dn}, \mu\gamma)$
  $X_{dn}|H_0 \sim Binomial(2N_{dn}, \mu)$

With
Relative risk:
$\gamma \sim Gamma(\bar{\gamma} * \beta_{dn}, \beta_{dn})$

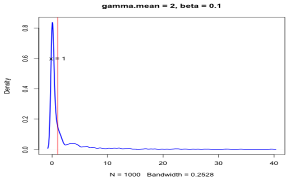
Can we simultaneously estimate all parameters based data + prior information from publications? => it will be easier to incorporate other information.

Some simple steps last month:

- => Re-write likelihood functions with two types of model.
- => Check there are overlapping intervals π between different variants.
- => Test whether we can constrain parameterers to estimate simultaneously.

Some issues we have had:

- Internal OR external models.
- Constrain relative risks (γ).
    - Some relative risk parameters imply substantial proportions of protective genes.

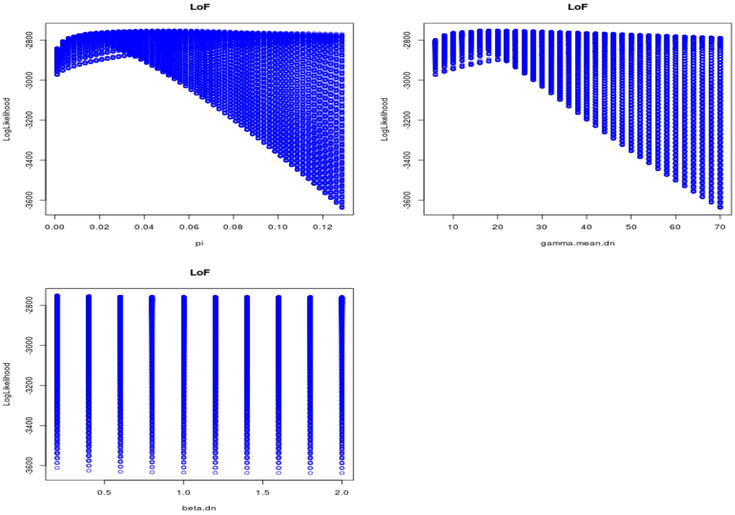


- Improve the calculation of mutation rates for each genes/annotations.
- Use adjusted counts as data.
- Improve algorithm (eliminate numerical integration).

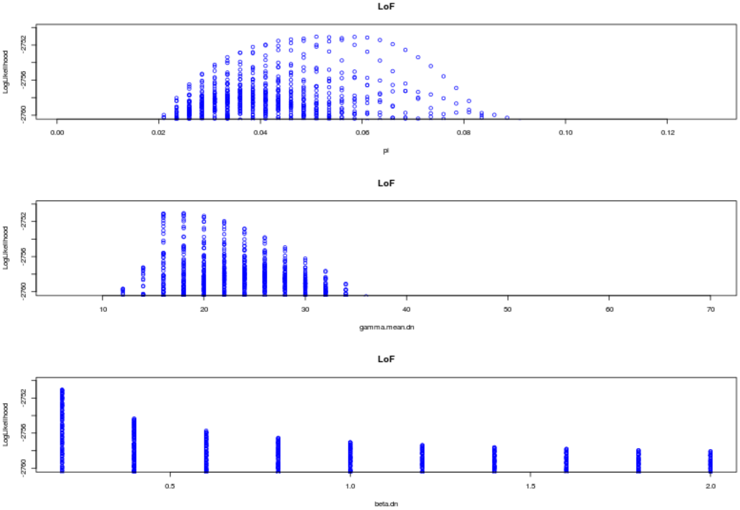
Grid on each data type

- Denovo LoF.
- Denovo Mis3.
- Case-control LoF.

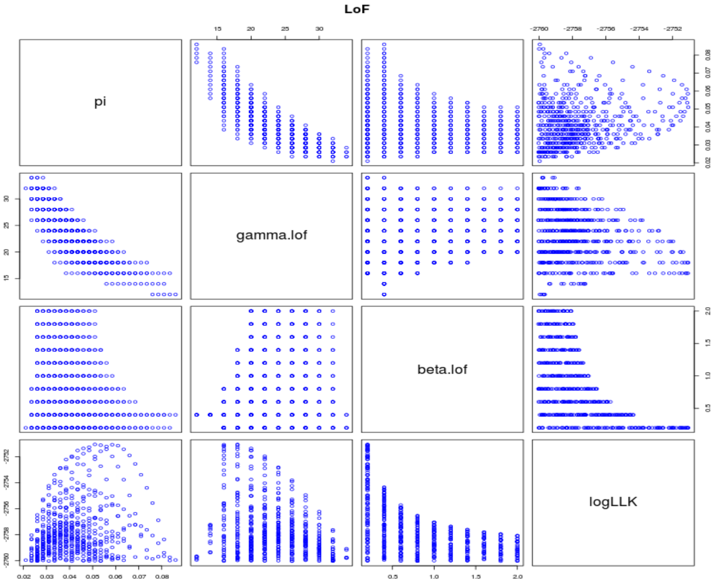
LoF de novo


LoF
LogLikelihood
-2800
-3000
-3200
-3400
-3600
0.00
0.02
0.04
0.06
0.08
0.10
0.12
pi
LoF
LogLikelihood
-2800
-3000
-3200
-3400
-3600
10
20
30
40
50
60
70
gamma.mean.dn
LoF
LogLikelihood
-2800
-3000
-3200
-3400
-3600
0.5
1.0
1.5
2.0
beta.dn

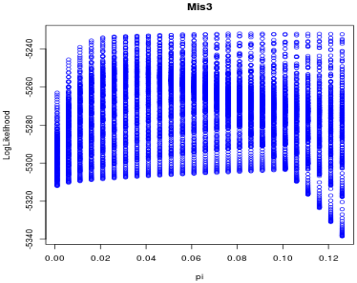
LoF de novo: top log LLK

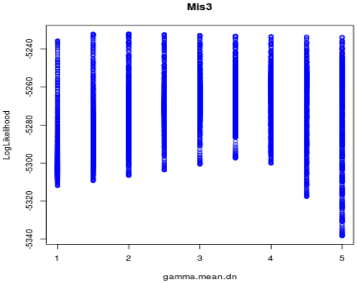


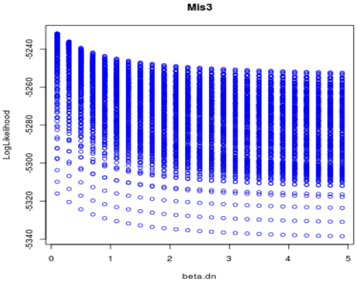
LoF de novo: check correlations between hyperparameters



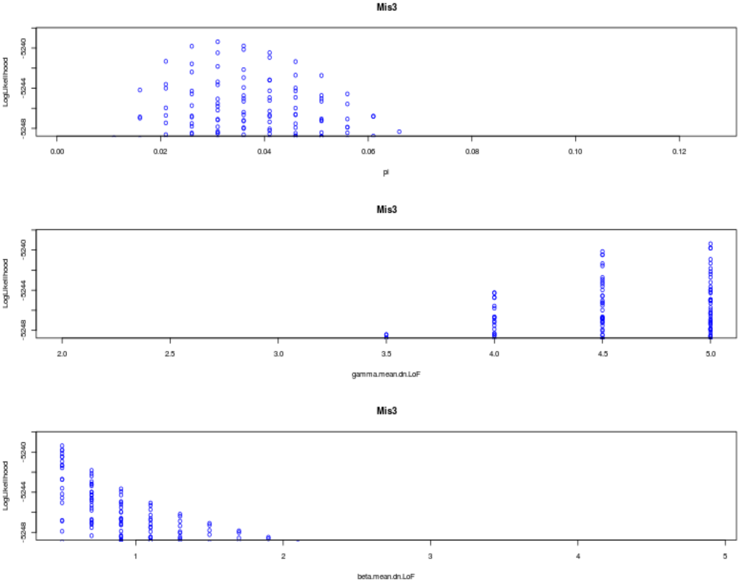
Mis3 de novo


Mis3
LogLikelihood
-5240
-5260
-5280
-5300
-5320
-5340
0.00
0.02
0.04
0.06
0.08
0.10
0.12
pi

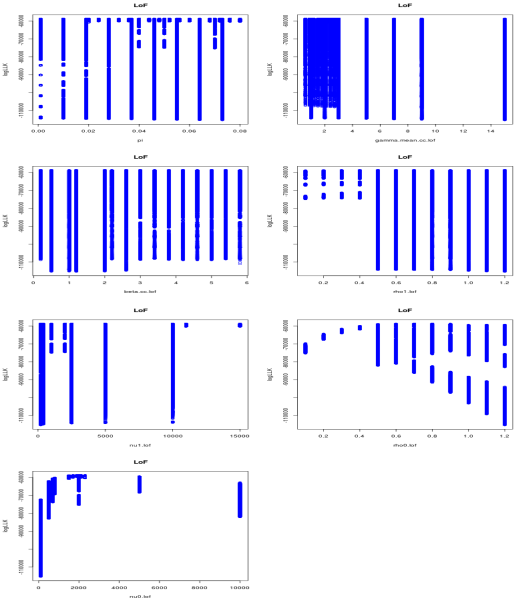

Mis3
LogLikelihood
-5240
-5260
-5280
-5300
-5320
-5340
1
2
3
4
5
gamma.mean.dn


Mis3
LogLikelihood
-5240
-5260
-5280
-5300
-5320
-5340
0
1
2
3
4
5
beta.dn

Mis3 de novo: top LLK
less than 0.6% protective variants.



LoF case-control



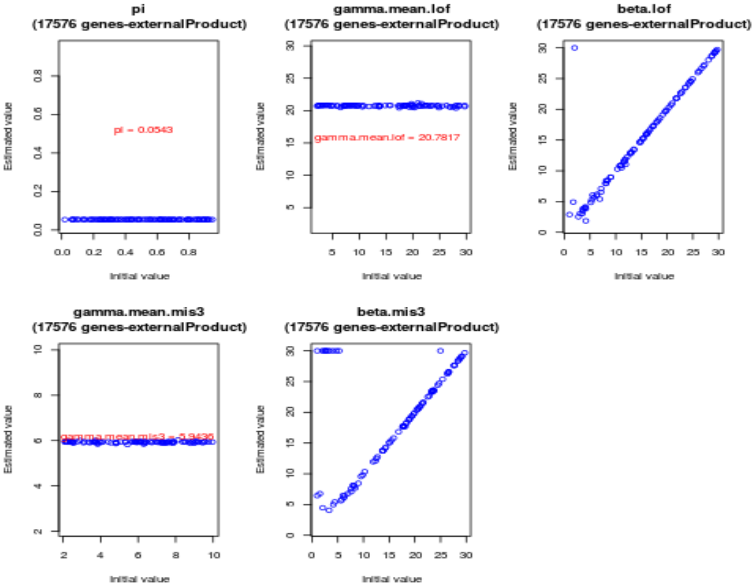
Estimate parameters

Use intervals of hyperparameters to set uniform priors for hyperparameters.

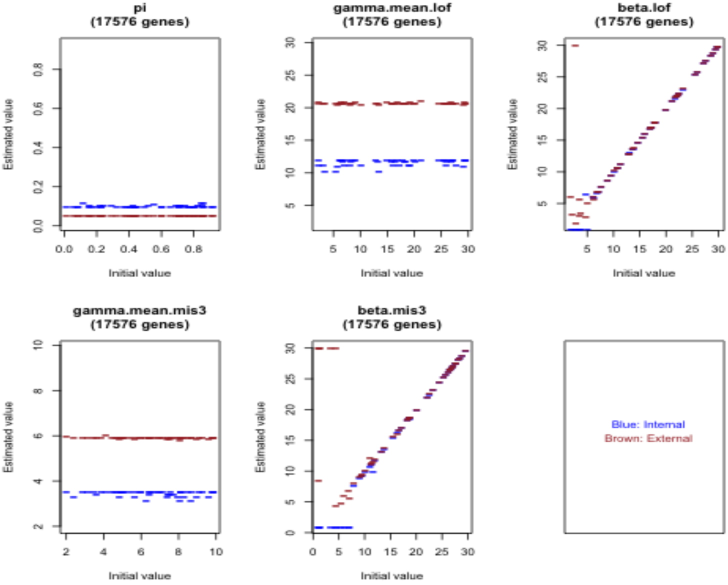
Constrained optimization

1. A set of random initial values was used $=>$ they can converge to approximately optimal values.
2. Some different algorithms (built in R) are used.

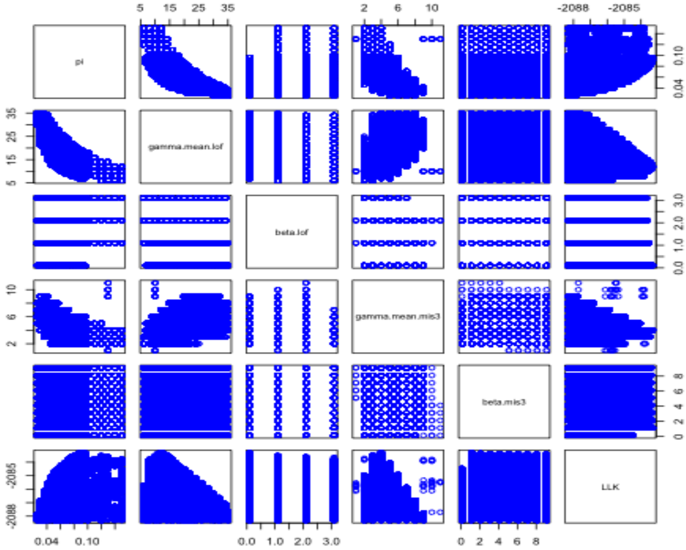
Test (external) for de novo data: LoF + mis3 for D1!



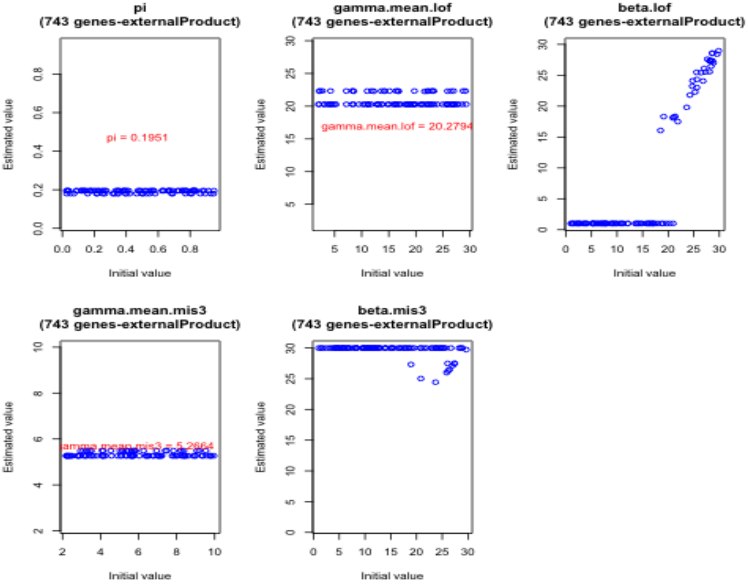
Test (internal/external) for de novo data: LoF + mis3 for D1!



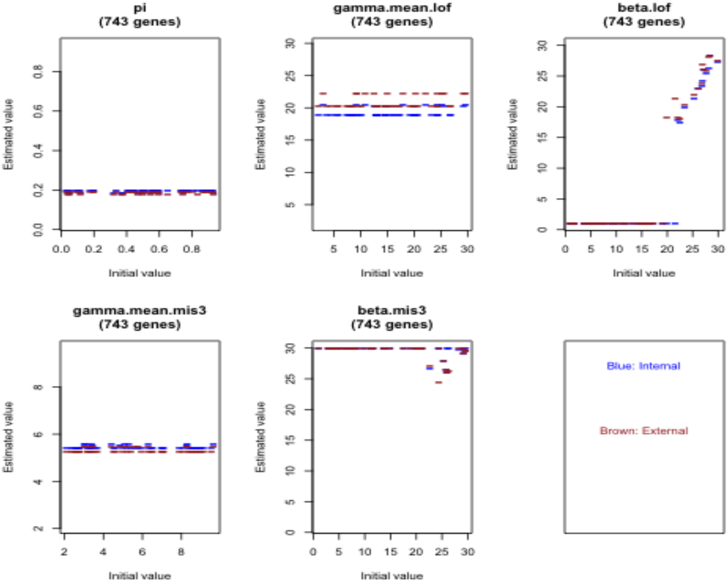
Grid top LLK (internal) for de novo data: LoF + mis3 for D1!



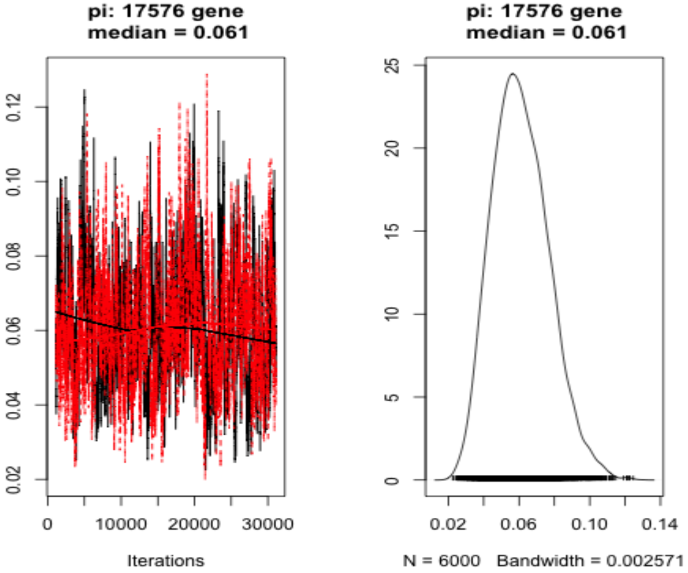
Test (external) for de novo data: LoF + mis3 for D2 (the FMRP gene set)!



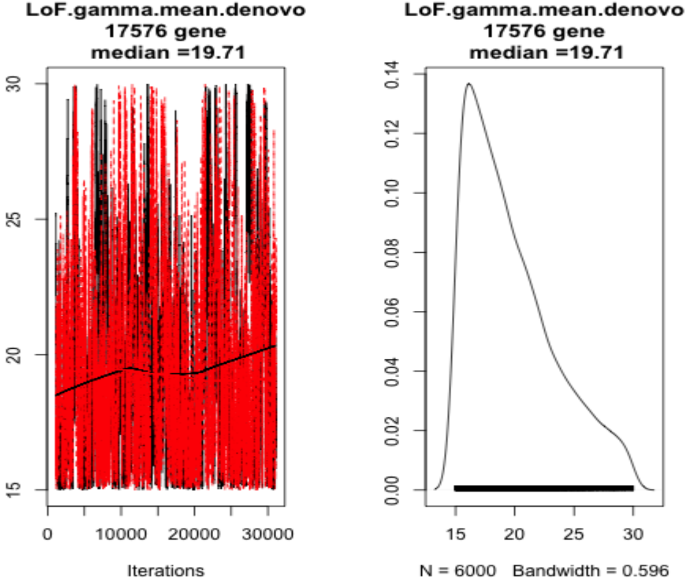
Test (internal/external) for de novo data: LoF + mis3 for D2 (the FMRP gene set)!



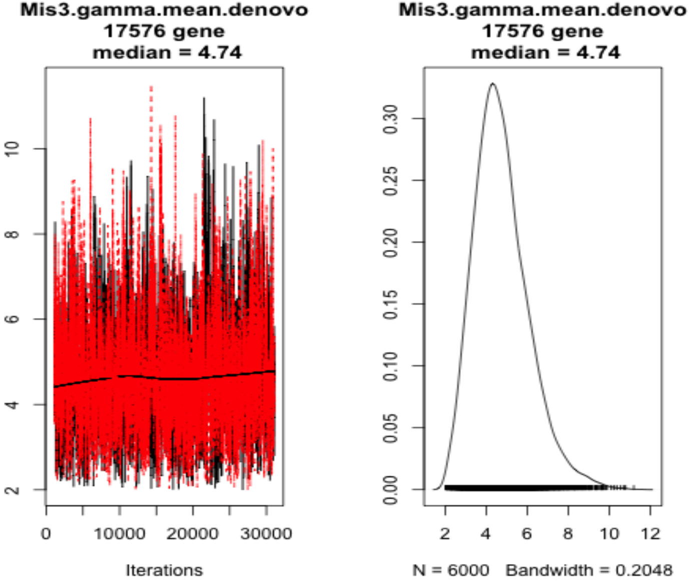
MCMC for D1: external product.



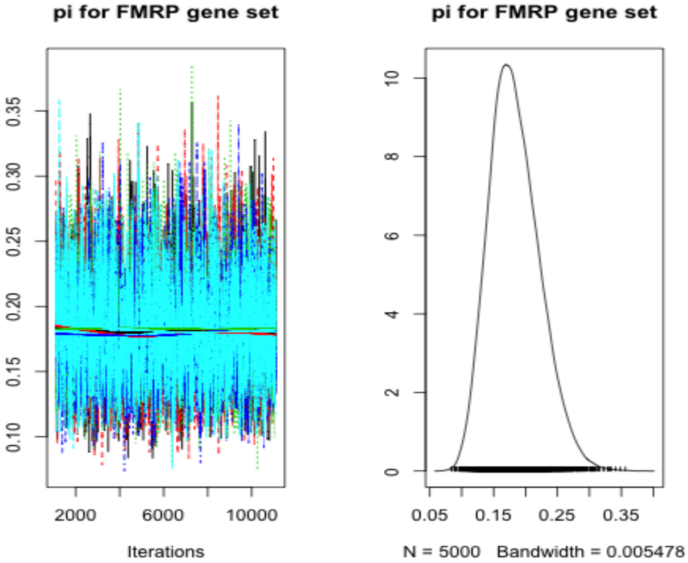
MCMC for D1: external product.



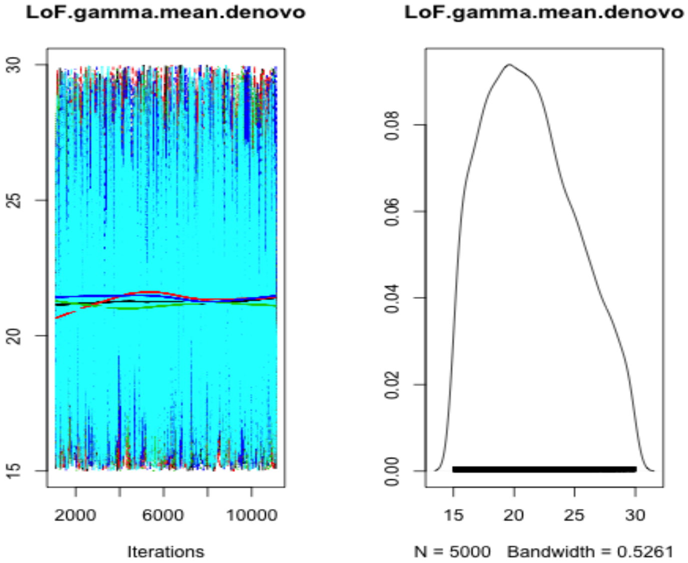
MCMC for D1: external product.



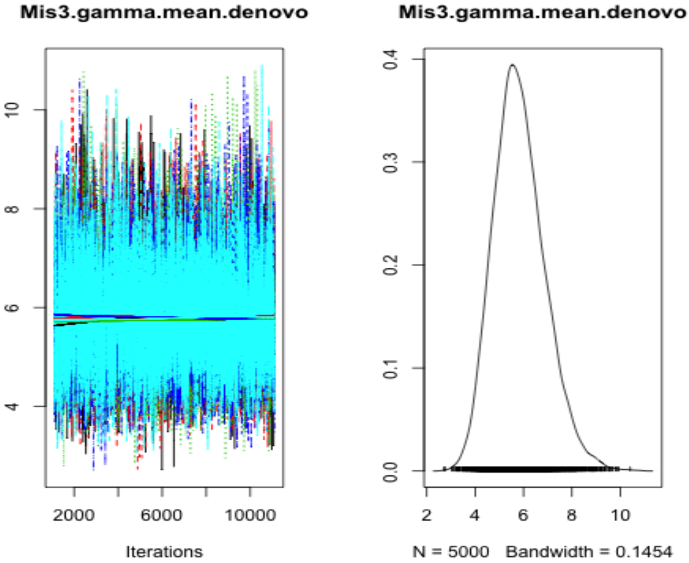
MCMC for **D2** **(the FMRP gene set,)**



MCMC for **D2** (**the FMRP gene set,**)



MCMC for **D2** (the FMRP gene set,)


Mis3.gamma.mean.denovo
Iterations
Mis3.gamma.mean.denovo
N = 5000 Bandwidth = 0.1454