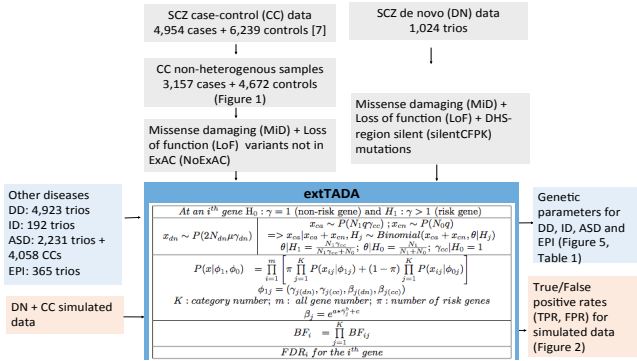
May, 2015

October 26, 2016

Current results



SCZ results

. Genetic parameters for SCZ (Figure 3).
. Top significant genes:
- FDR < 0.1: SETD1A, TAF1B, MKI67
- FDR < 0.3: SETD1A, TAF1B, MKI67, ST3GAL6, VPS13C, RB1CC1, PRRC2A, LPHN2, DARC, DUOXA2, KLHL17, HSPA8, ZDHHC5, XPR1

. Top 100/500 genes with the smallest FDRs => enriched in candidate gene sets (Table 2).
. The predicted number of risk genes with different sample sizes (Figure 4).

Going to discuss:

- Evaluate extTADA using simulation data.
- extTADA for multiple populations.
- Multiple traits with extTADA??

Extended TADA

Extending TADA method to **automatically analyse**: only DN, only CC, DN+CC.

Steps

- De novo mutations: the same as original TADA (Or using binomial distribution).
- **Inherited/Case-control**:
  1. Only use a non heterogeneous population (obtain by using LM/GLM).
  2. Use an approximate model in the estimation process.
- **Estimate all parameters using a MCMC method from Equation 1 (known risk genes are not necessary).**

$$P(x|\phi_1,\phi_0) = \prod_{i=1}^{m}\left[\pi\prod_{j=1}^{K_1} f_{1_{DN_j}}\prod_{h=1}^{K_2} f_{1_{CC_h}} + (1-\pi)\prod_{j=1}^{K_1} f_{0_{DN_j}}\prod_{h=1}^{K_2} f_{0_{CC_h}}\right] \quad (1)$$

Reason for modifiying the CC model

- $q \sim Gamma(\rho, \nu)$ values do not affect much final results => use the same calculation as TADA 2014.
- Original case-control model => not easy to estimate parameters => Two ways to obtain parameters for case-control parameters:
  - Change the order of integrals in the original case-control model.
  - Use an approximate model.

Change the order of integrals to rely only on relative risks

$$P(x_1, x_0|H_j) = P(x_0|H_j)P(x_1|x_0, H_j) \tag{2}$$

- The first part $P(x_0|H_j)$ was the same as TADA 2014:

$$P(x_0|H_j) = \int P(x_0|q, H_j)P(q|\rho, \nu, H_j)dq = NegBin(x_0|\rho, \frac{N_0}{\nu + N_0}), j = 0, 1 \tag{3}$$

- The second part:

$$\begin{aligned} P(x_1|H_j, x_0) &= \int P(x_1|q, \gamma)P(q|H_j, x_0)P(\gamma|H_j)dqd\gamma \\ &= \int [P(x_1|q, \gamma)P(q|H_j, x_0)dq]\, P(\gamma|H_j)d\gamma \\ &= \int NegBin(x_1|\rho + x_0, \tfrac{N_0+\nu}{N_1\gamma+N_0+\nu})P(\gamma|H_j)d\gamma \end{aligned} \tag{4}$$

Use simulated data => it can converge to simulated values, but sometimes it is not good as expected.

Approximate case/control model

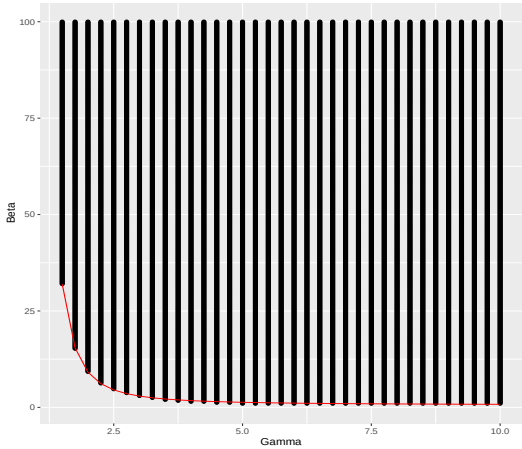
$$\begin{aligned} P(x_1, x_0|H_j) &= P(x_1, x_1 + x_0|H_j) \\ &= P(x_1|x_1 + x_0, H_j)P(x_1 + x_0|H_j) \end{aligned} \tag{5}$$

- The first part: $P(x_1|x_1 + x_0, H_j)$ Because of $x_1 \sim Pois(N_1 q\gamma)$ and $x_0 \sim Pois(N_0 q)$, we assumed that x_1 and x_0 were **independent**, we had:
  $x_1|x_1 + x_0, H_j \sim Binomial(x_1 + x_0, \theta|H_j)$
  with $\theta|H_1 = \frac{N_1\gamma}{N_1\gamma+N_0}$ and $\theta|H_0 = \frac{N_1}{N_1+N_0}$
  The marginal likelihood was
  $P(x_1|x_1 + x_0, H_j) = \int P(x_1|x_1 + x_0, \gamma, H_j)P(\gamma|x_1 + x_0, H_j)d\gamma$
- The second part $P(x_1 + x_0|H_j)$ was not used in the estimation process.

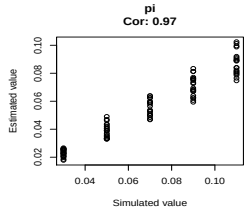
Use original case-control model (TADA 2013) to simulate data, and then estimate using this approximate model => show reliable results.

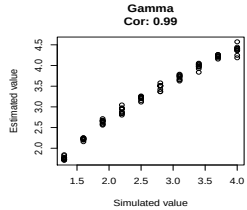
If we control the proportion of protective variants => nonlinear relationship between β and γ.

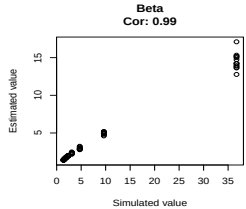
$$\beta = e^{a*\gamma^b+c} \tag{6}$$



CC correlation







Calculate true/false positive rates

Simulate DN+CC data:

- Set a threshold FDR (0.1 in this study).
- For each set of simulated parameters:
  - Calculate BFs, and then count the number of risk genes with the FDR threshold (sGene).
  - Use extTADA to estimate parameters => calculate BFs => count the number of risk genes with the FDR threshold (rGene).

- TPR $= \frac{\textit{the number of overlapping genes of sGene and rGene}}{\textit{the number of sGene}}$.
- FPR $= \frac{\textit{the number of rGene not in sGene}}{\textit{the total gene } (1{,}941) - \textit{the number of sGene}}$.

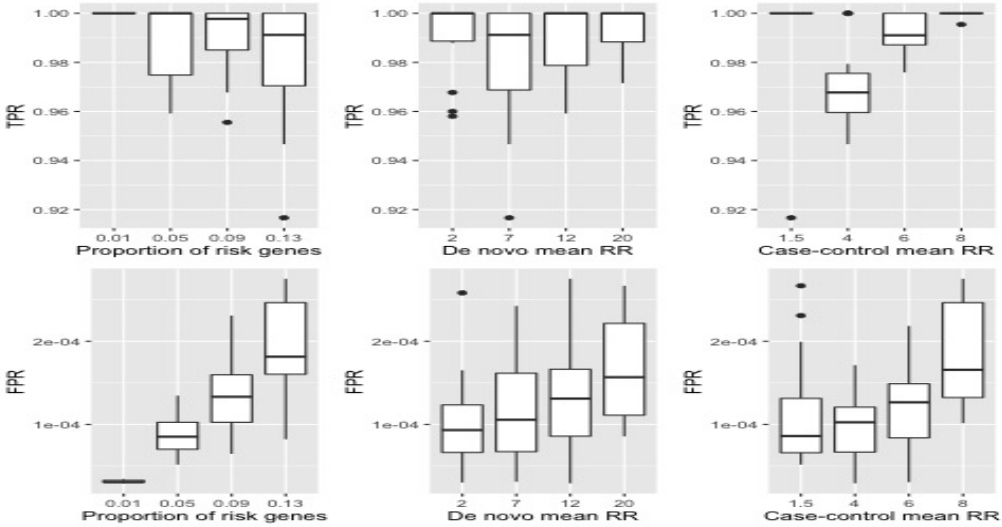


Figure: True postive rates (TPRs) and false postive rate (FPRs) for simulated data.

CC data: choose a non-heterogeneous pop

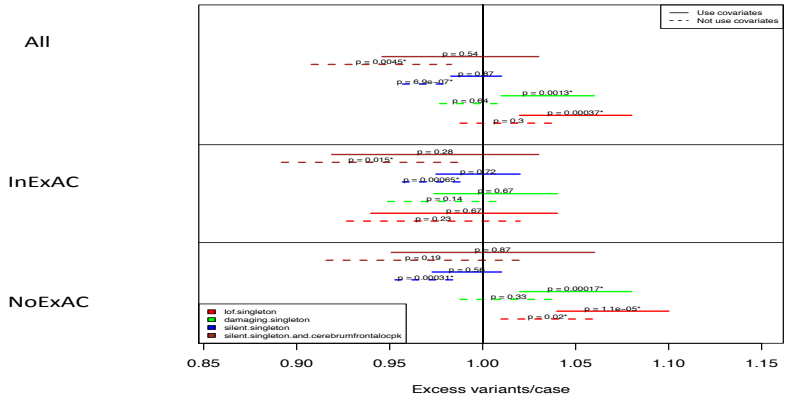
Ran multiple clustering processes, and chose a population whose results of Equation 7 and 8 were not much different.

$$\begin{aligned} &logit(P(SCZ = 1)) \sim count + countAll + sex + birth + kit + PC1 + ... + PC20 \\ &count \sim SCZ + countAll + sex + birth + kit + PC1 + ... + PC20 \end{aligned} \tag{7}$$

$$\begin{aligned} SCZ &\sim count \\ count &\sim SCZ \end{aligned} \tag{8}$$

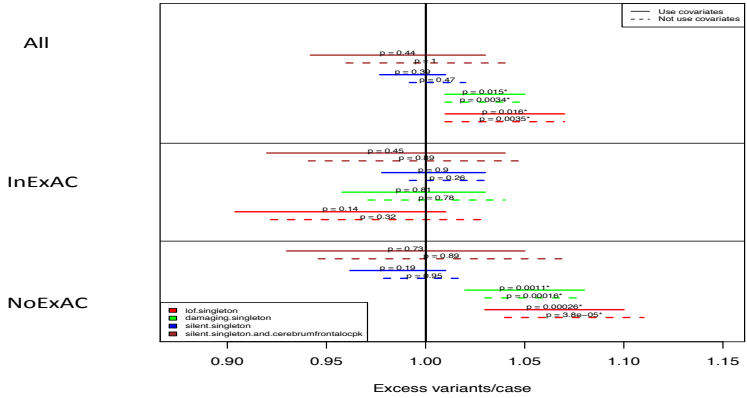
SCZ Case control

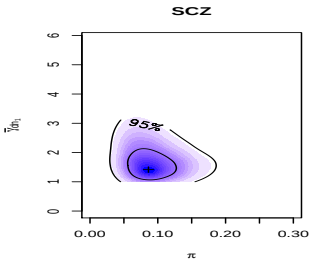
Whole samples: the results are very different if we use covariates or not use covariates.

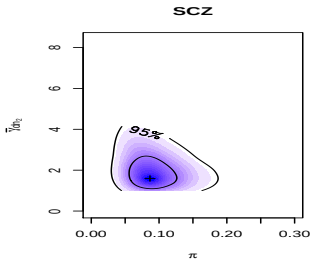


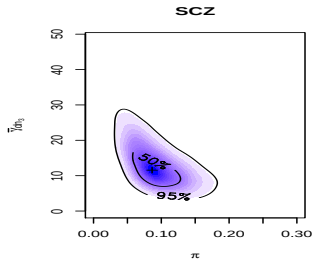
SCZ Case control

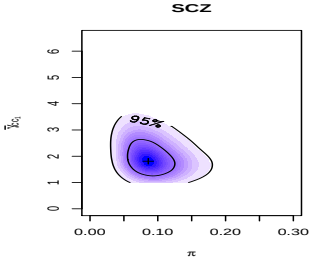
After the clustering process => only non-heteregeneous pop: the results are similar.

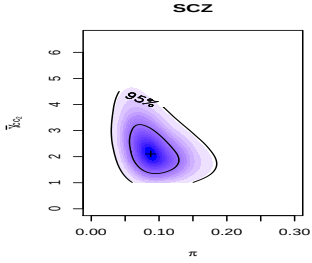



SCZ
95%
π

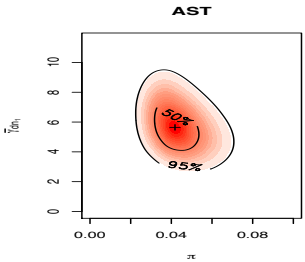

SCZ
95%
π

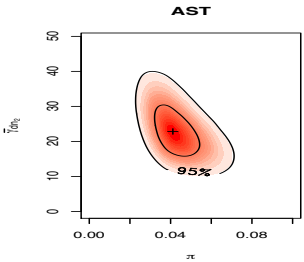

SCZ
50%
95%
π

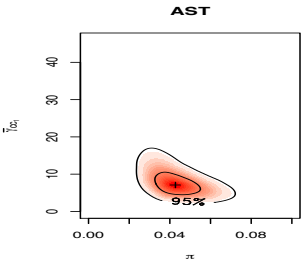

SCZ
95%
π

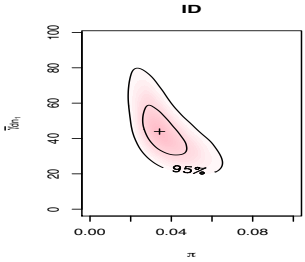

SCZ
95%
π

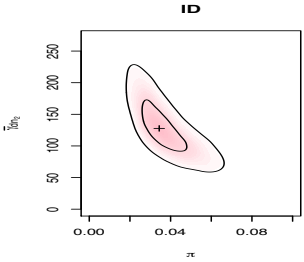
Other diseases



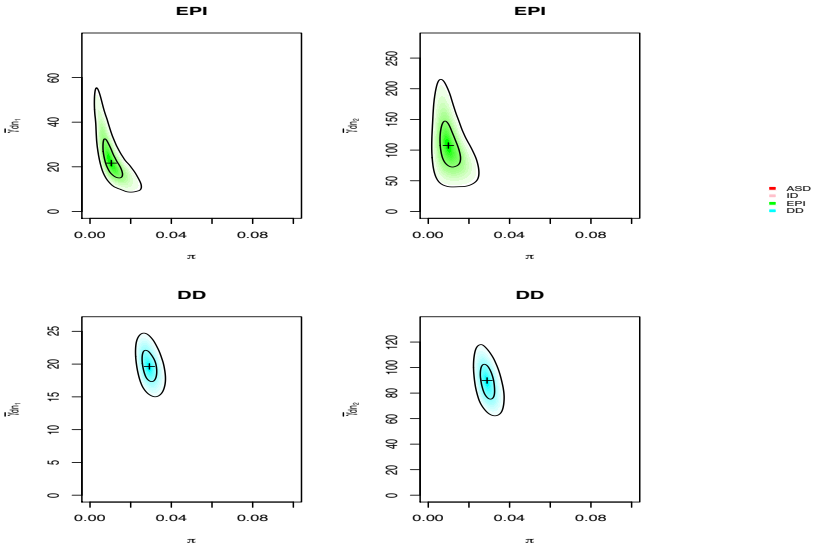








Other diseases



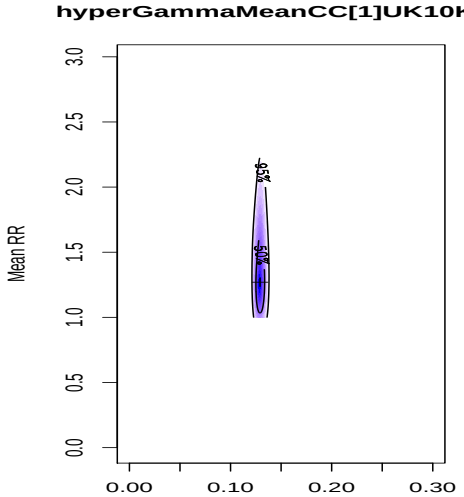
Other diseases

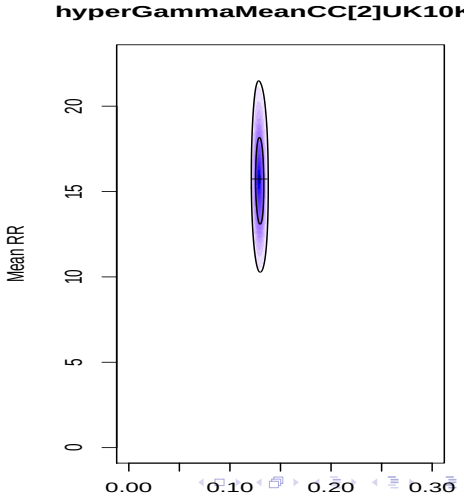
- Intellectual disorder (ID): compared with McRae et al (2016), some new significant genes have been obtained using extTADA.
- Other diseases: need to obtain risk gene sets to compare (?).

`Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. McRae et al. (2016). http://biorxiv.org/content/early/2016/04/20/0490`

Only case-control data from UK10K

New data last week (UK + Fin + Sweden2): used extTADA for only case-control data.





Combine the case-control data of UK10K into the model

Divide the data into different populations and combine all Bayes Factors across populations.
Not weigh sample sizes for pops now.

$$BF = \prod_{i=1}^{Ndn_{pop}} (\prod BF_{dn_{ij}})(\prod_{m=1}^{Ncc_{pop}} \prod BF_{cc_{mn}})$$

Going to use the same internal model, but more flexible to input for multiple populations:

$$P(x|\phi_1, \phi_0) = \prod_{i=1}^{m} \left[\pi(\prod^{Ndn_{pop}} \prod_{j=1}^{K_1} f1_{DN_j})(\prod^{Ncc_{pop}} \prod_{h=1}^{K_2} f1_{CC_h}) + (1-\pi)(\prod^{Ndn_{pop}} \prod_{j=1}^{K_1} f0_{DN_j})(\prod^{Ncc_{pop}} \prod_{h=1}^{K_2} f0_{CC_h}) \right] \quad (9)$$

Is that useful to extend TADA to multiple traits?

$$\pi_{00} H_{10} H_{20} + \pi_{01} H_{10} H_{21} + \pi_{10} H_{11} H_{20} + \pi_{11} H_{11} H_{21}$$

H_{1j} is the j^{th} hypothesis for the first trait.
H_{2j} is the j^{th} hypothesis for the second trait.