

May, 2015

extTADA for  
one pop

Simulated data  
Real data

extTADA for  
multiple pops

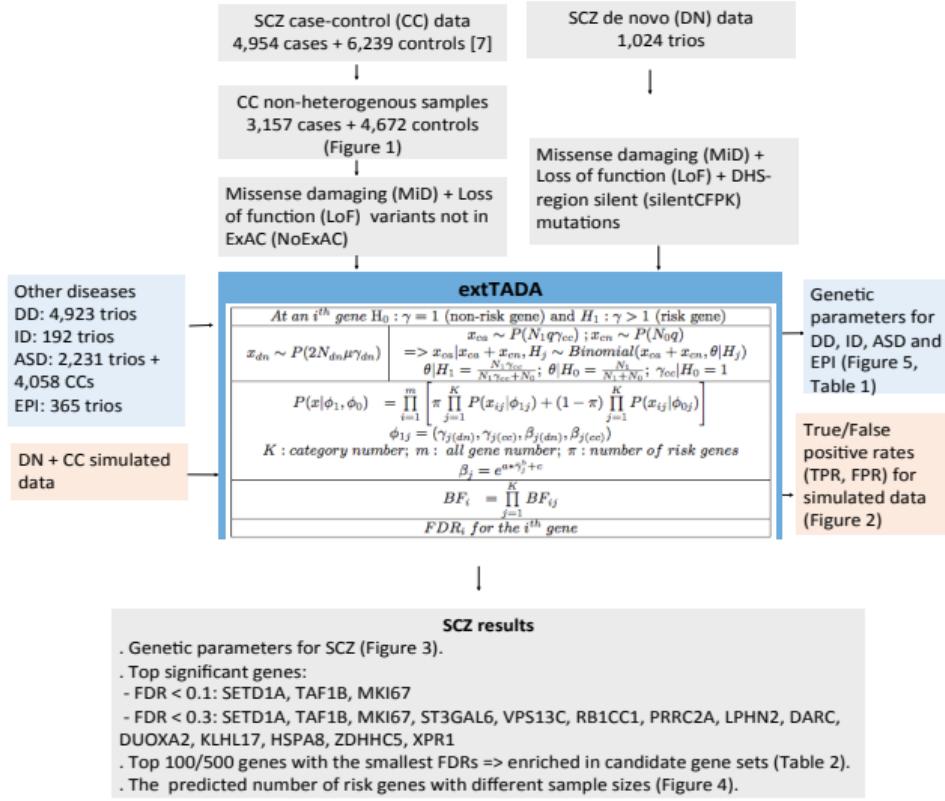
May, 2015

October 26, 2016

# Current results

extTADA for  
one pop  
Simulated data  
Real data

extTADA for  
multiple pops



extTADA for  
one pop

Simulated data  
Real data

extTADA for  
multiple pops

## Going to discuss:

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

# Extended TADA

extTADA for  
one pop  
Simulated data  
Real data  
  
extTADA for  
multiple pops

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).**

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

## Reason for modifying the CC model

extTADA for  
one pop  
Simulated data  
Real data

extTADA for  
multiple pops

- $q \sim \text{Gamma}(\rho, \nu)$  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

extTADA for  
one pop  
Simulated data  
Real data  
extTADA for  
multiple pops

$$P(x_1, x_0 | H_j) = P(x_0 | H_j) P(x_1 | x_0, H_j) \quad (2)$$

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

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

- The second part:

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

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

## Approximate case/control model

extTADA for  
one pop  
Simulated data  
Real data  
extTADA for  
multiple pops

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

- 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.

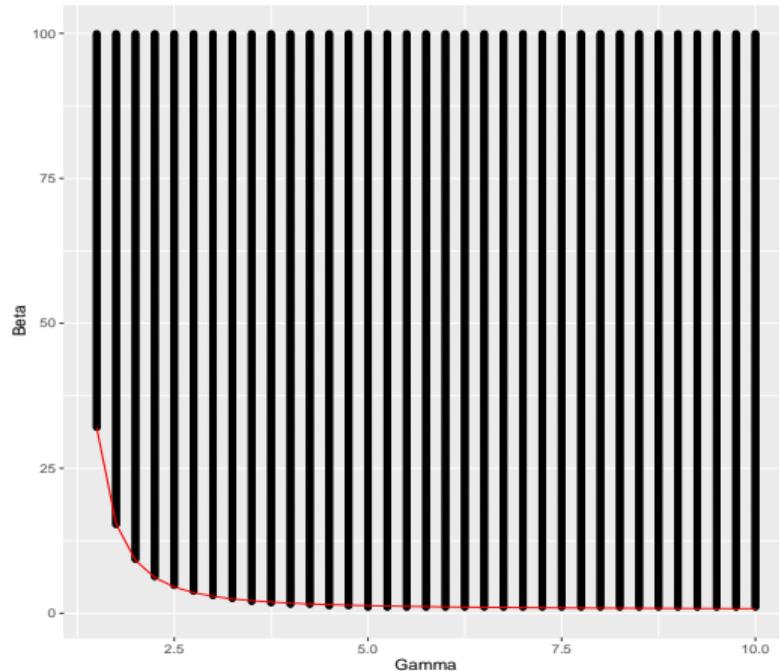
Use original case-control model (TADA 2013) to simulate data, and then estimate using this approximate model => show reliable results.

If we control the proportion of protective variants => nonlinear relationship between  $\beta$  and  $\gamma$ .

extTADA for  
one pop  
Simulated data  
Real data

extTADA for  
multiple pops

$$\beta = e^{a*\gamma^b+c} \quad (6)$$

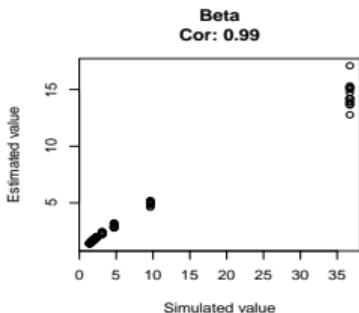
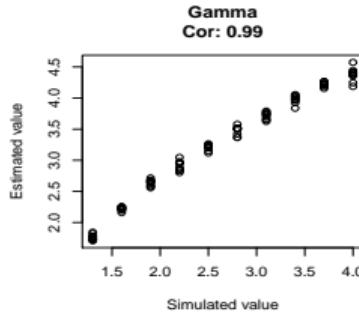
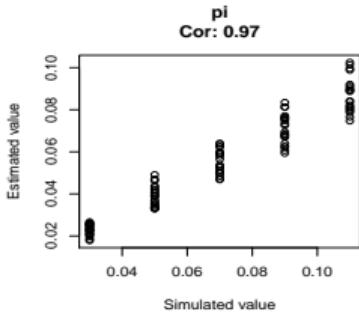


## CC correlation

extTADA for  
one pop

Simulated data  
Real data

extTADA for  
multiple pops



# Calculate true/false positive rates

extTADA for  
one pop

Simulated data  
Real data

extTADA for  
multiple pops

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).
- $\text{TPR} = \frac{\text{the number of overlapping genes of } s\text{Gene and } r\text{Gene}}{\text{the number of } s\text{Gene}}$ .
- $\text{FPR} = \frac{\text{the number of } r\text{Gene not in } s\text{Gene}}{\text{the total gene (1,941) - the number of } s\text{Gene}}$ .

extTADA for  
one pop  
Simulated data  
Real data

extTADA for  
multiple pops

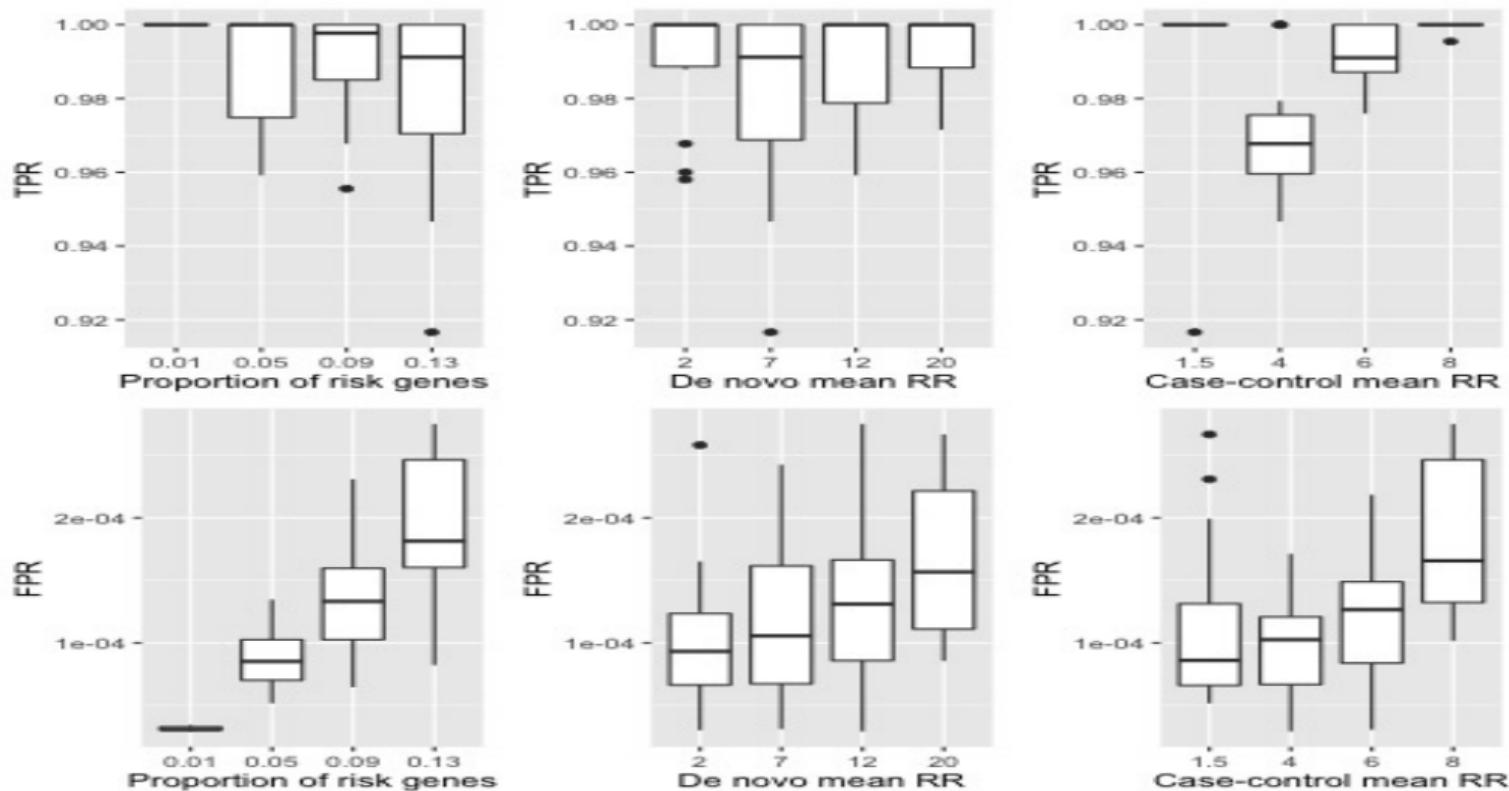


Figure: True positive rates (TPRs) and false positive rate (FPRs) for simulated data.

## CC data: choose a non-heterogeneous pop

extTADA for  
one pop  
Simulated data  
Real data  
extTADA for  
multiple pops

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

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

(7)

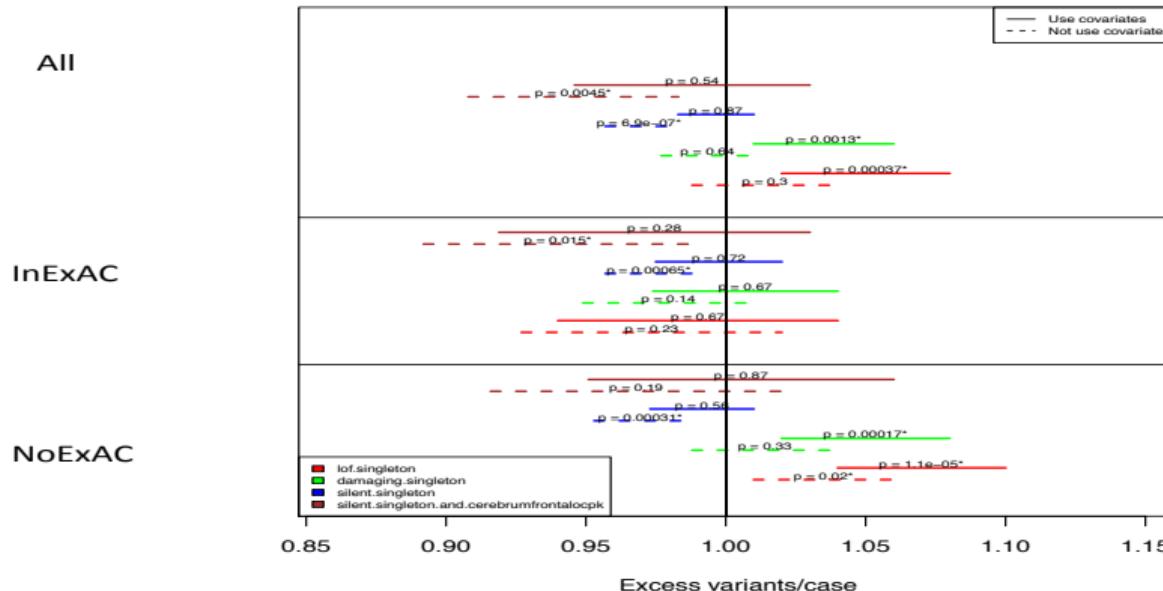
$$\begin{aligned} SCZ &\sim \text{count} \\ \text{count} &\sim SCZ \end{aligned} \tag{8}$$

# SCZ Case control

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

extTADA for  
one pop  
Simulated data  
Real data

extTADA for  
multiple pops



# SCZ Case control

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

extTADA for  
one pop

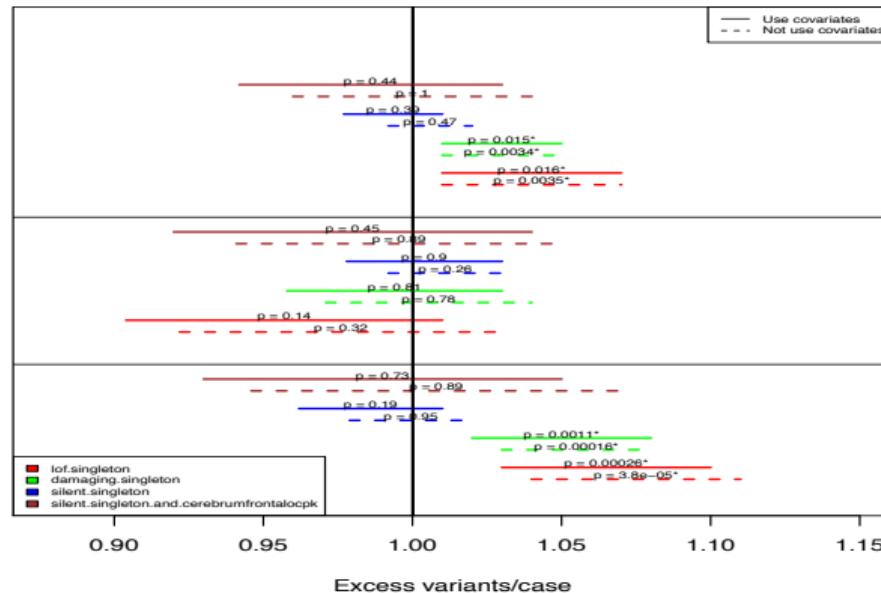
Simulated data  
Real data

extTADA for  
multiple pops

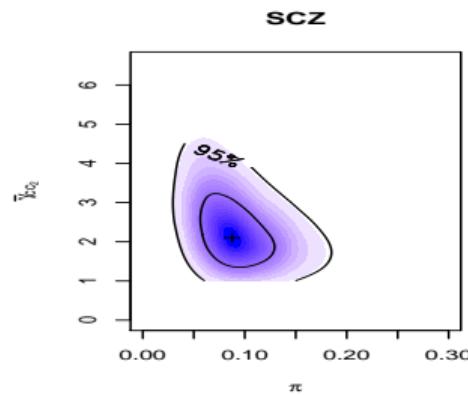
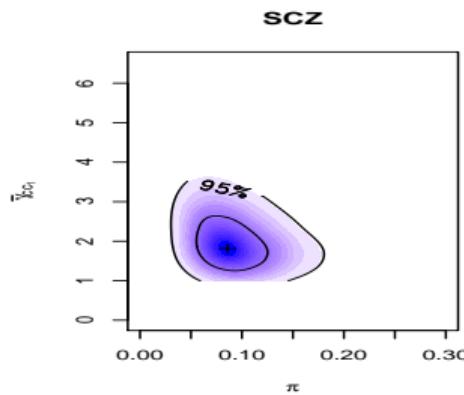
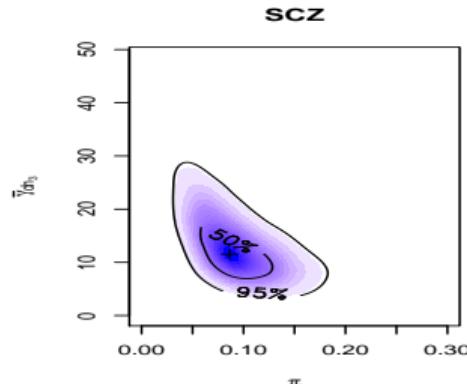
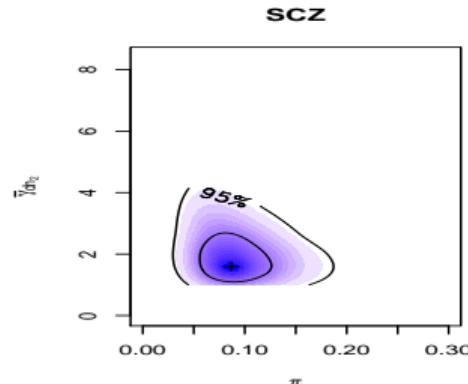
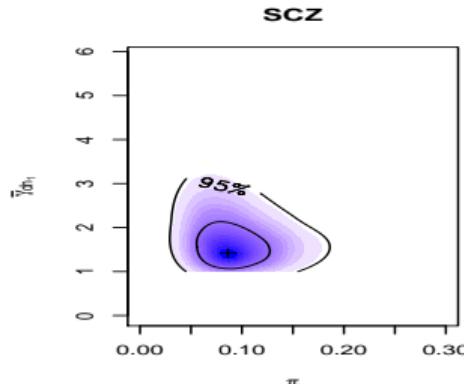
All

InExAC

NoExAC



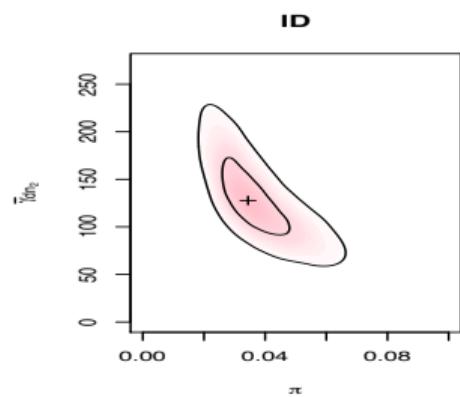
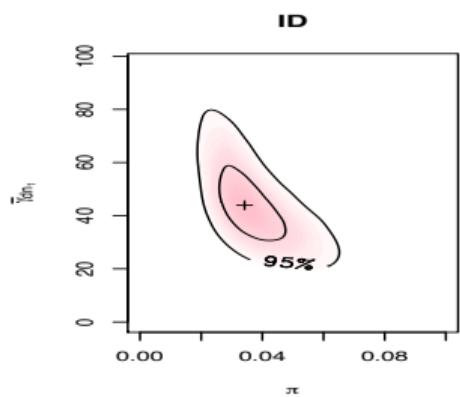
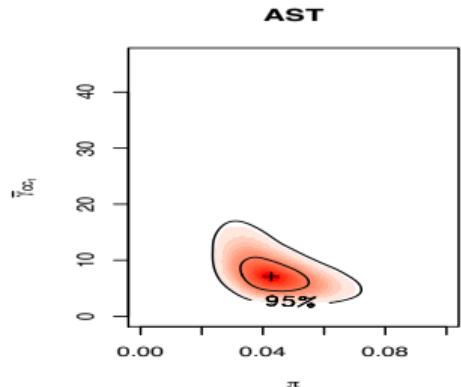
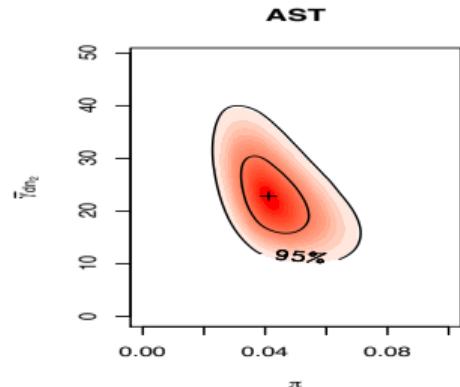
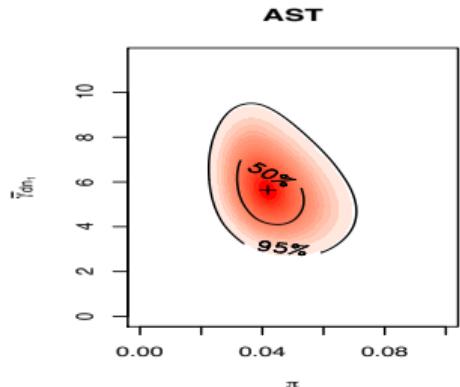
extTADA for  
one pop  
Simulated data  
Real data  
extTADA for  
multiple pops



# Other diseases

extTADA for  
one pop  
Simulated data  
Real data

extTADA for  
multiple pops

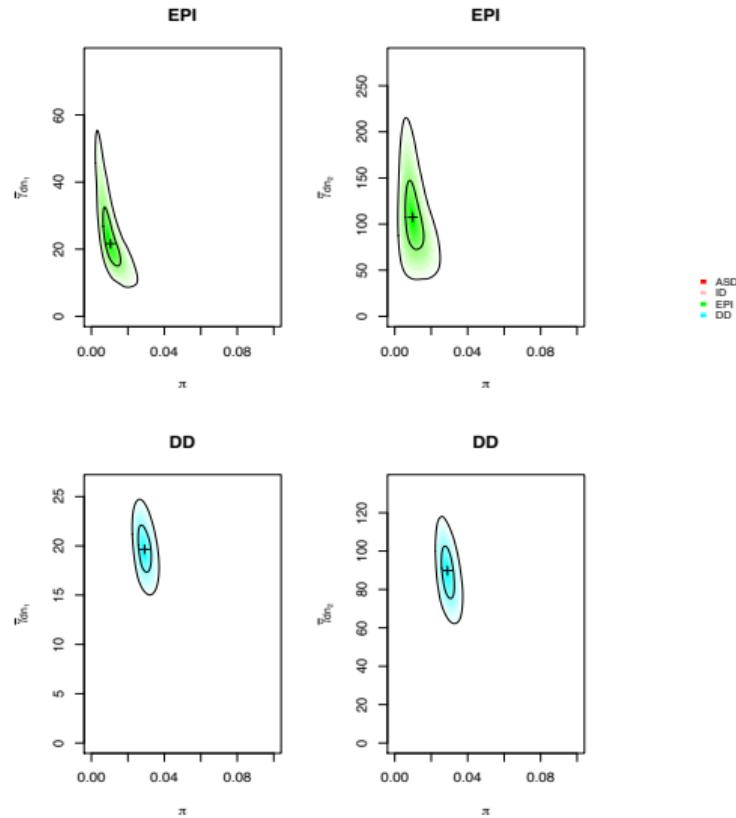


# Other diseases

extTADA for  
one pop

Simulated data  
Real data

extTADA for  
multiple pops



## Other diseases

extTADA for  
one pop  
Simulated data  
Real data

extTADA for  
multiple pops

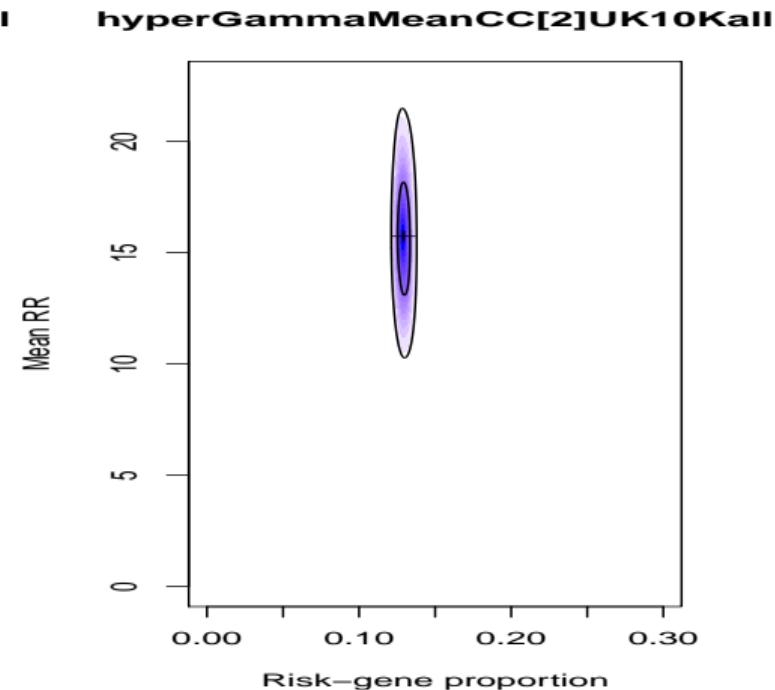
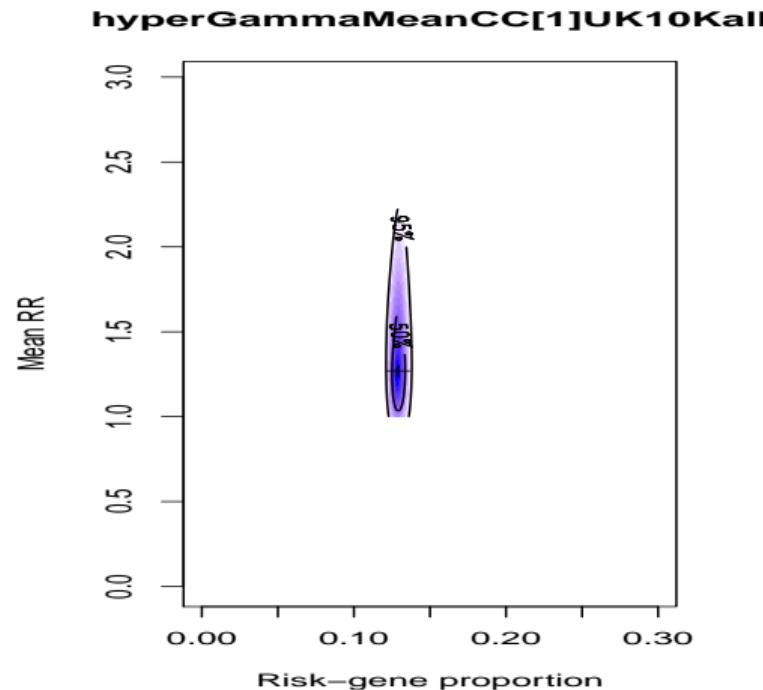
- Intellectual disorder (ID): compared with McRae et al (2016), some new significant genes have been obtained using extTADA.
- Other diseases: need to obtain risk gene sets to compare (?).

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.



extTADA for  
one pop

Simulated data  
Real data

extTADA for  
multiple pops

# Combine the case-control data of UK10K into the model

extTADA for  
one pop  
Simulated data  
Real data

extTADA for  
multiple pops

Divide the data into different populations and combine all Bayes Factors across populations.

Not weigh sample sizes for pops now.

$$BF = \prod_{i=1}^{Nd_{pop}} \left( \prod_{j=1}^{K_1} BF_{dn_{ij}} \right) \left( \prod_{m=1}^{Ncc_{pop}} \prod_{h=1}^{K_2} BF_{cc_{mh}} \right)$$

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

$$P(x|\phi_1, \phi_0) = \prod_{i=1}^m \left[ \pi \left( \prod_{j=1}^{Nd_{pop}} f_{1DN_j} \right) \left( \prod_{h=1}^{Ncc_{pop}} f_{1CC_h} \right) + (1 - \pi) \left( \prod_{j=1}^{Nd_{pop}} f_{0DN_j} \right) \left( \prod_{h=1}^{Ncc_{pop}} f_{0CC_h} \right) \right] \quad (9)$$

extTADA for  
one pop

Simulated data  
Real data

extTADA for  
multiple pops

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.