Using Stepwise Regression to Find Independent eQTLs

Tony Jiang
Montgomery Lab Meeting
Stanford University School of Medicine
8/13/14

Introduction

- Matrix eQTL reports many eQTLs, but not all of them are independent (i.e. two eQTLs may not explain much more variance in gene expression than just one of them)
- We can use stepwise regression methods to find independent eQTLs
- Final model: $GE = \beta_0 + \beta_1 \times SNP_1 + \beta_2 \times SNP_2 + \ldots + \beta_n \times SNP_n$

Methods

- Current model: $GE = \beta_0$

SNP_i	p-value of SNP_i in $GE = \beta_0 + \beta_1 \times SNP_i$
rs1	0.03
rs2	0.045
rs3	$5.4*10^{-54}$
rs4	0.01

Methods

- Current model: $GE = \beta_0 + \beta_1 \times rs3$

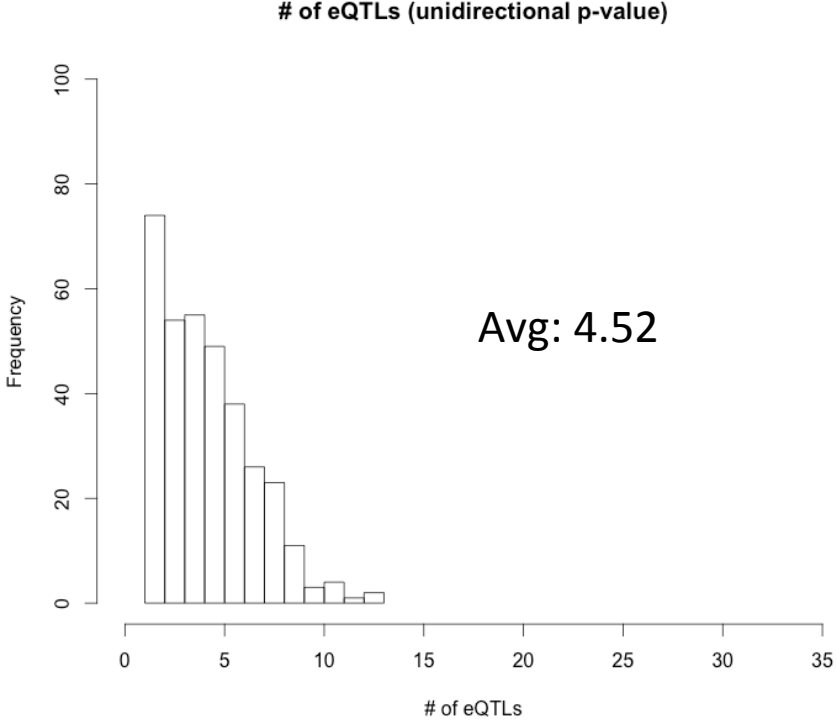
SNP_i	p-value of SNP_i in $GE = \beta_0 + \beta_1 \times rs3 + \beta_2 \times SNP_i$
rs1	0.4
rs2	0.08
rs3	
rs4	0.1

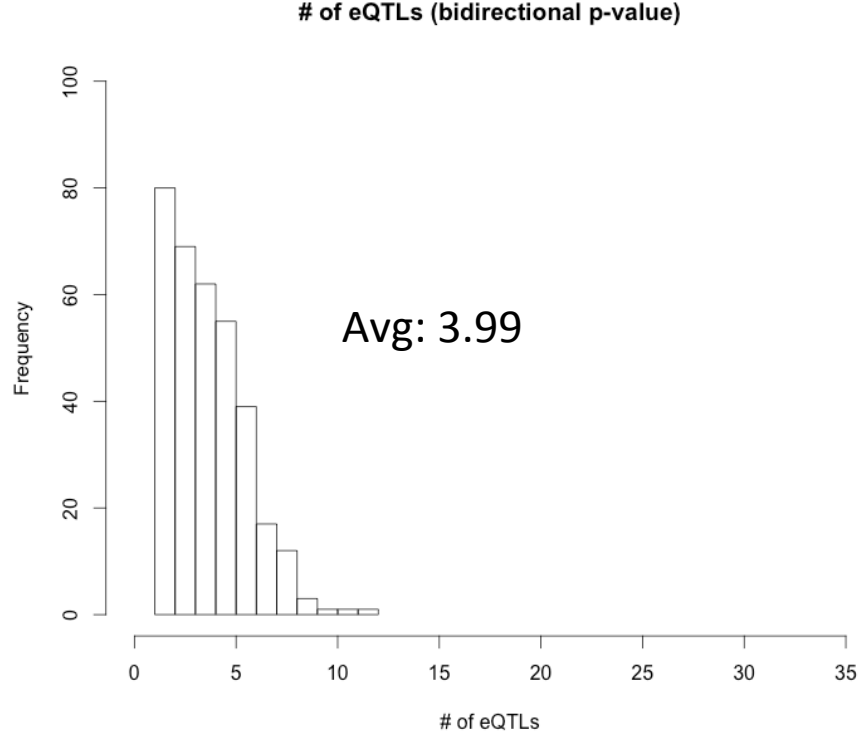
Final model: $GE = \beta_0 + \beta_1 \times rs3$

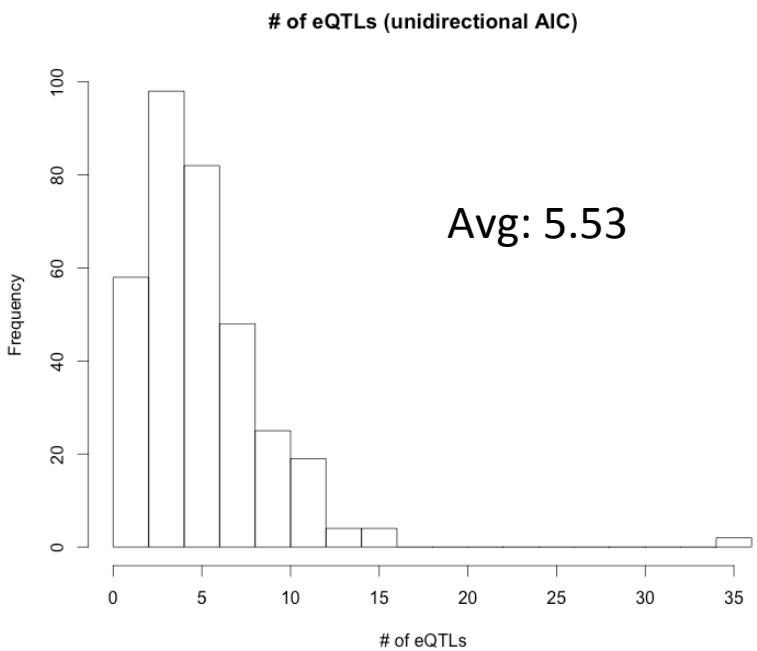
Methods

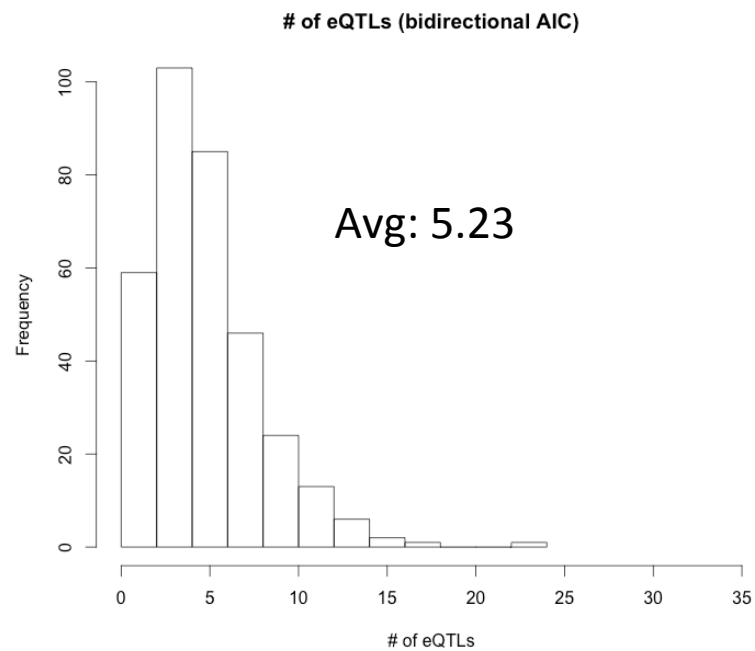
- We don't have to select by p-value; we can also select using AIC (Akaike information criterion)
- We can also remove SNPs from the model if they become insignificant upon adding another SNP, or if the AIC would decrease from removing it
- Gives a total of 4 sub-methods of performing stepwise regression

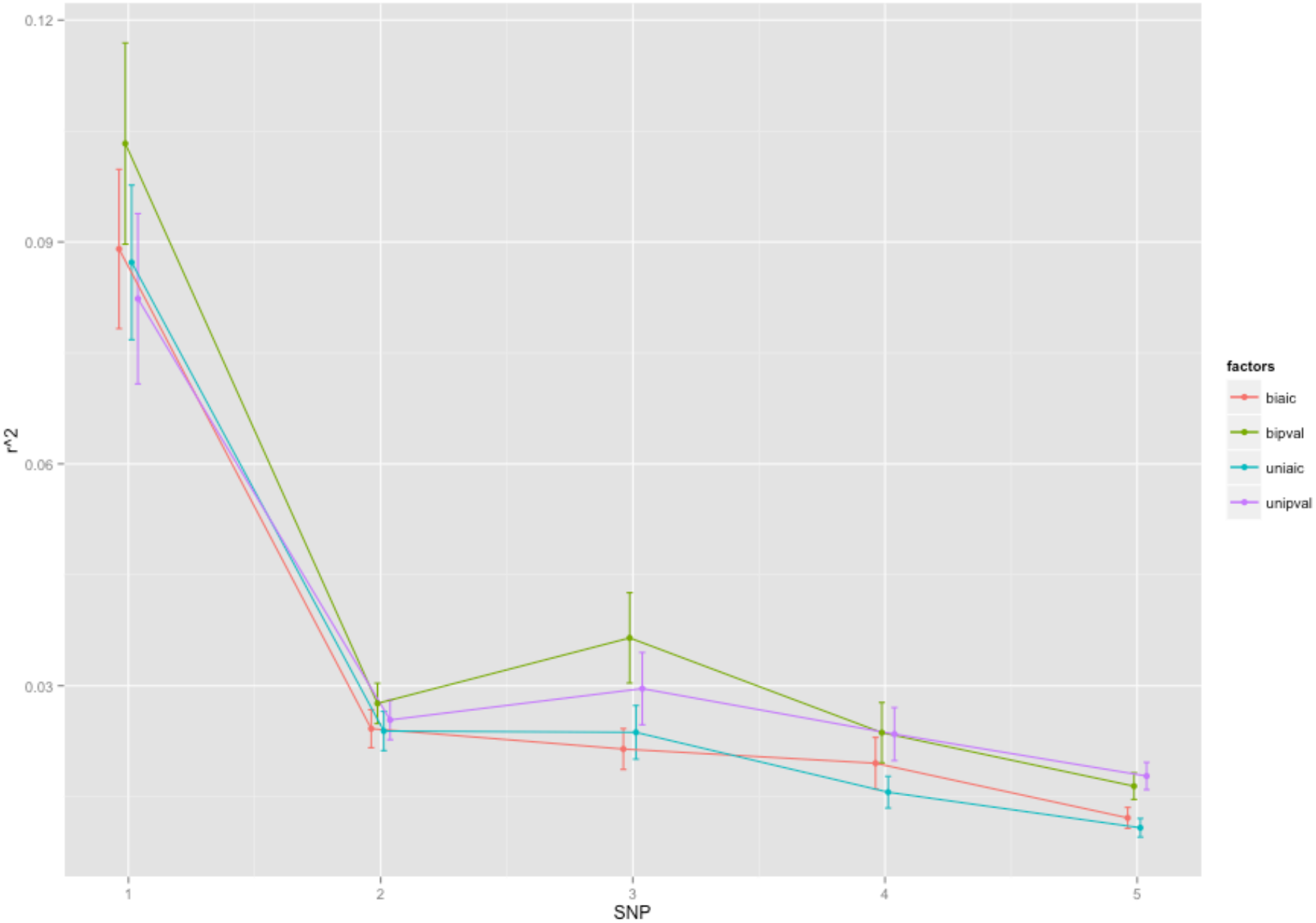
Results


# of eQTLs (unidirectional p-value)
Frequency
0
20
40
60
80
100
0
5
10
15
20
25
30
35
# of eQTLs
Avg: 4.52

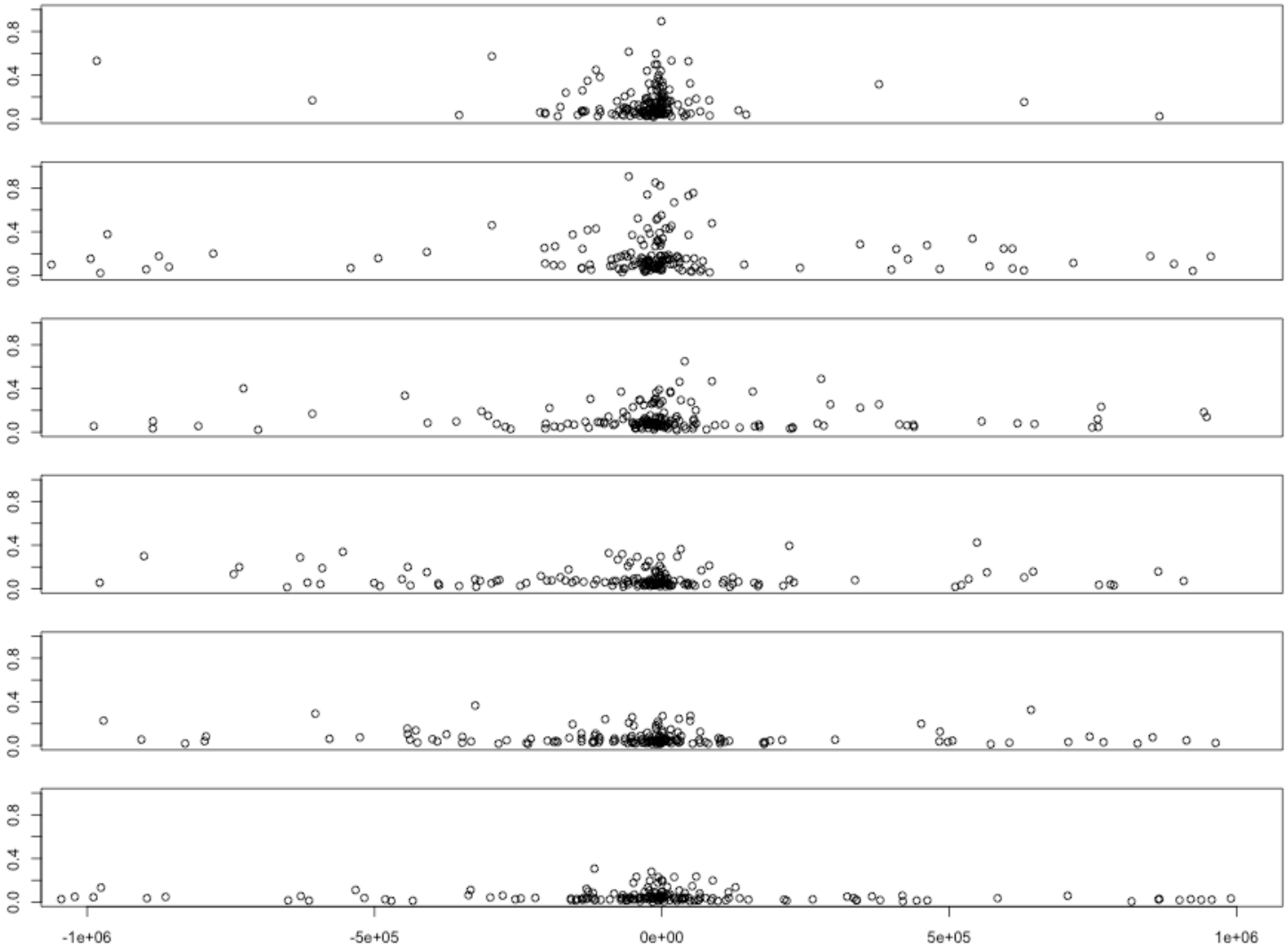

# of eQTLs (bidirectional p-value)
Frequency
0
20
40
60
80
100
0
5
10
15
20
25
30
35
# of eQTLs
Avg: 3.99


# of eQTLs (unidirectional AIC)
Frequency
0
20
40
60
80
100
0
5
10
15
20
25
30
35
# of eQTLs
Avg: 5.53


# of eQTLs (bidirectional AIC)
Frequency
0
20
40
60
80
100
0
5
10
15
20
25
30
35
# of eQTLs
Avg: 5.23


0.12
0.09
0.06
0.03
r^2
1
2
3
4
5
SNP
factors
biaic
bipval
uniaic
unipval

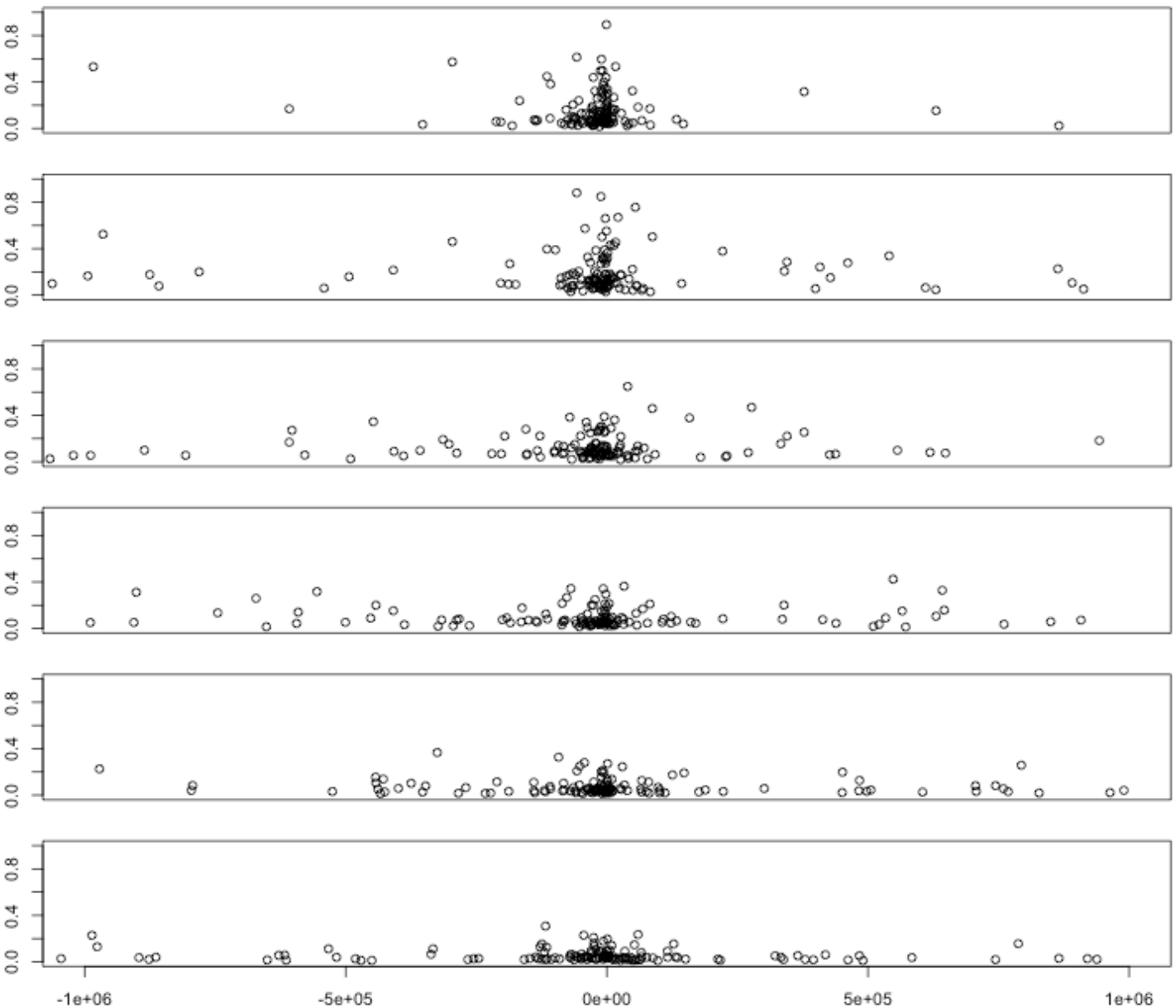
Unidirectional FSR (p-value)



Standard deviations:
145763.8
325652.1
286088.5
297085.3
319457.2
341613.1

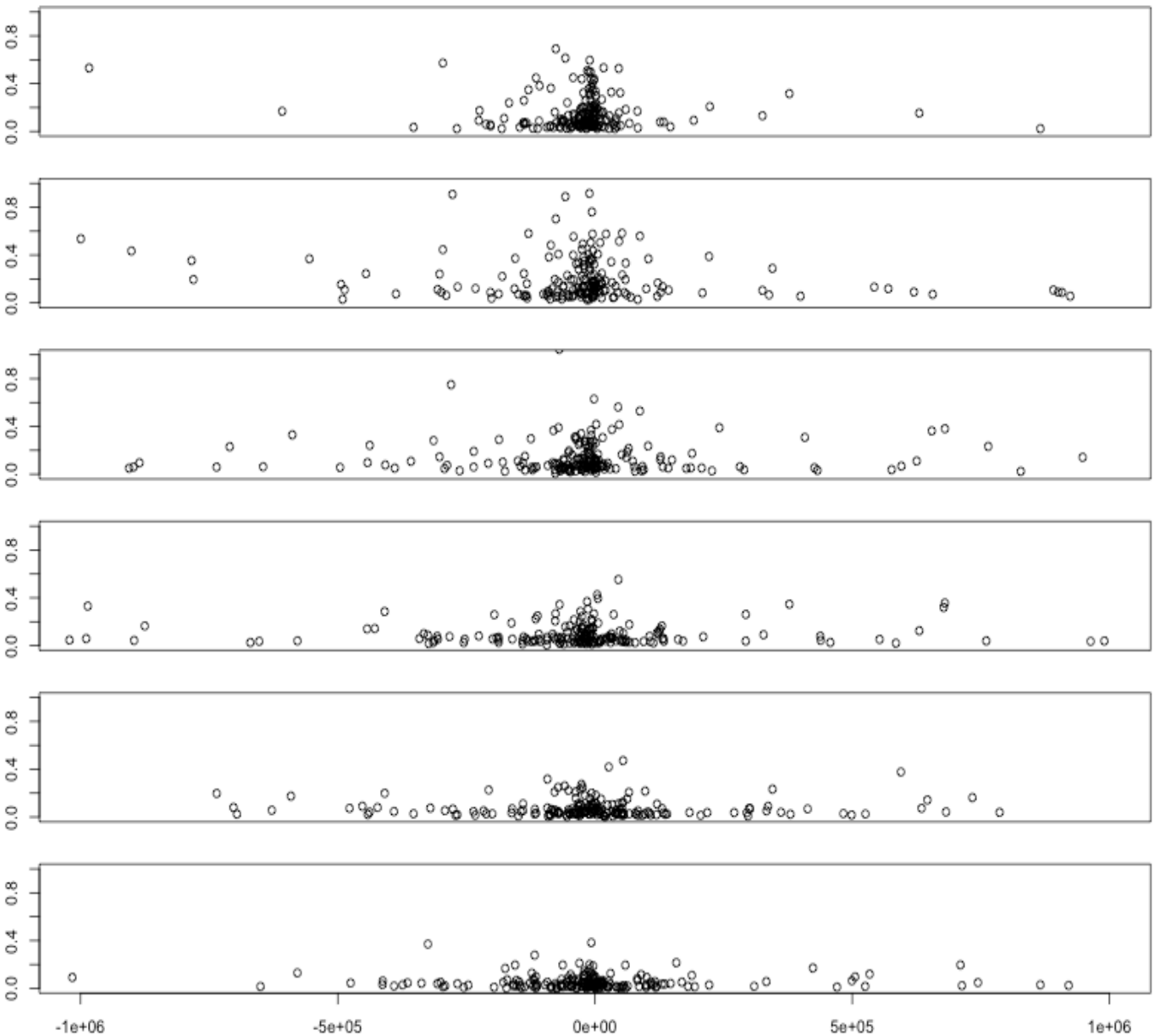
To reproduce these plots: Run plotsix.R

Bidirectional FSR (p-value)



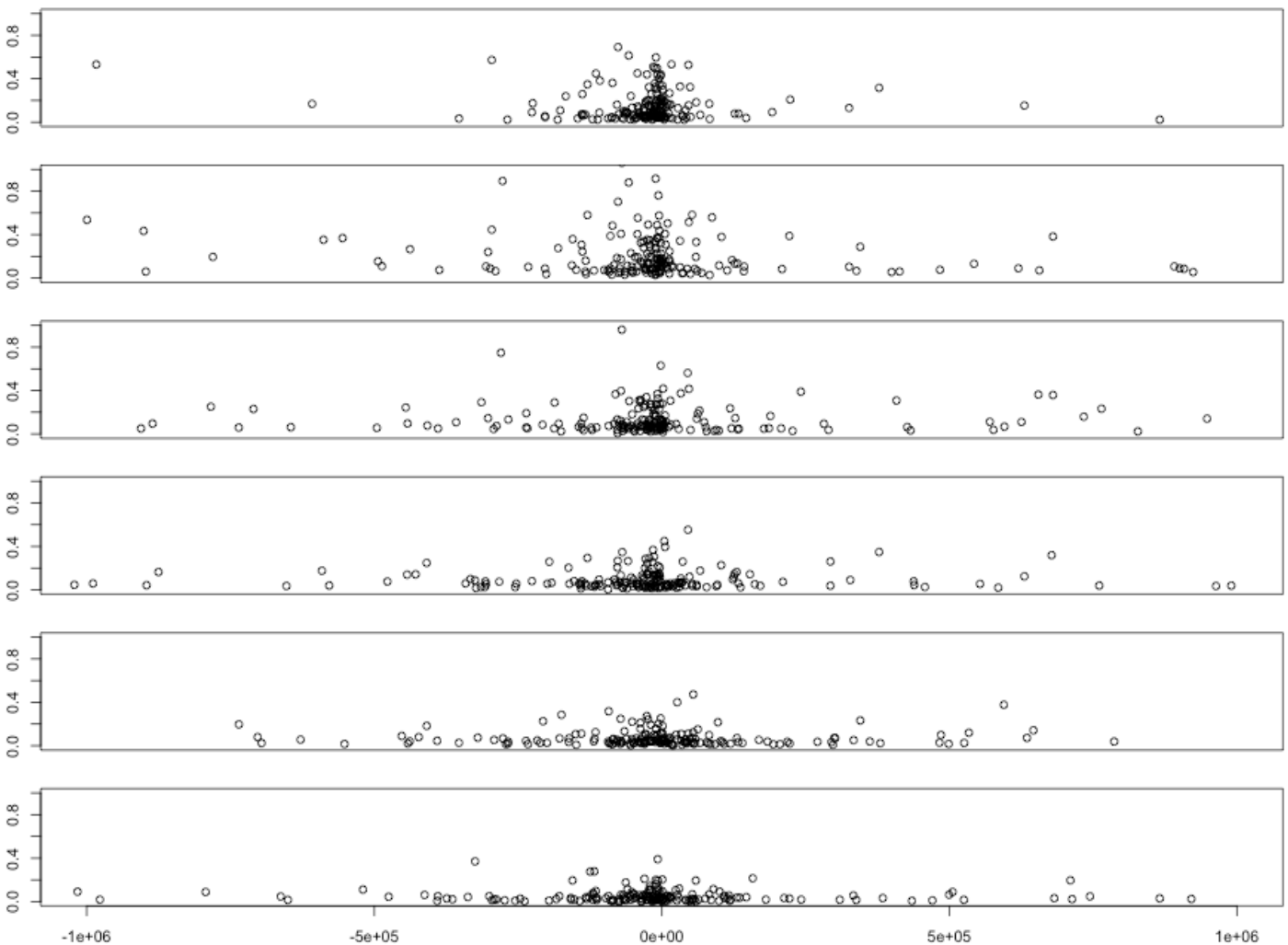
Standard deviations:
157576.5
288903.5
279522.3
315391.9
331471.7
337548.4

Unidirectional FSR (AIC)



Standard deviations:
143157.4
258628.3
251230.1
272980.8
232340.1
227413.1

Bidirectional FSR (AIC)



Standard deviations:
144710.1
271350.3
260022.4
262815.4
225550.9
257592.3

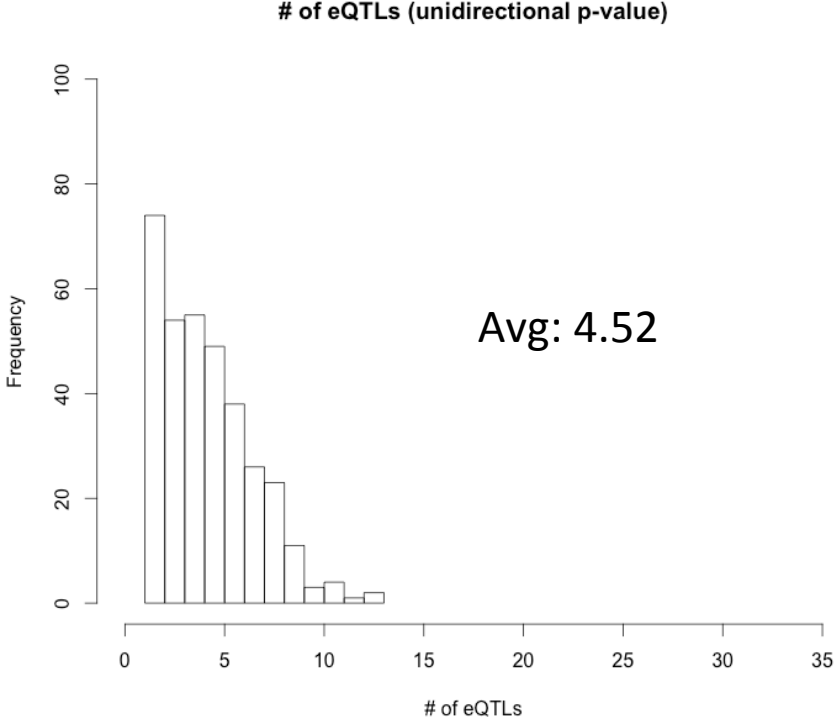
Collinearity / Linear Dependence

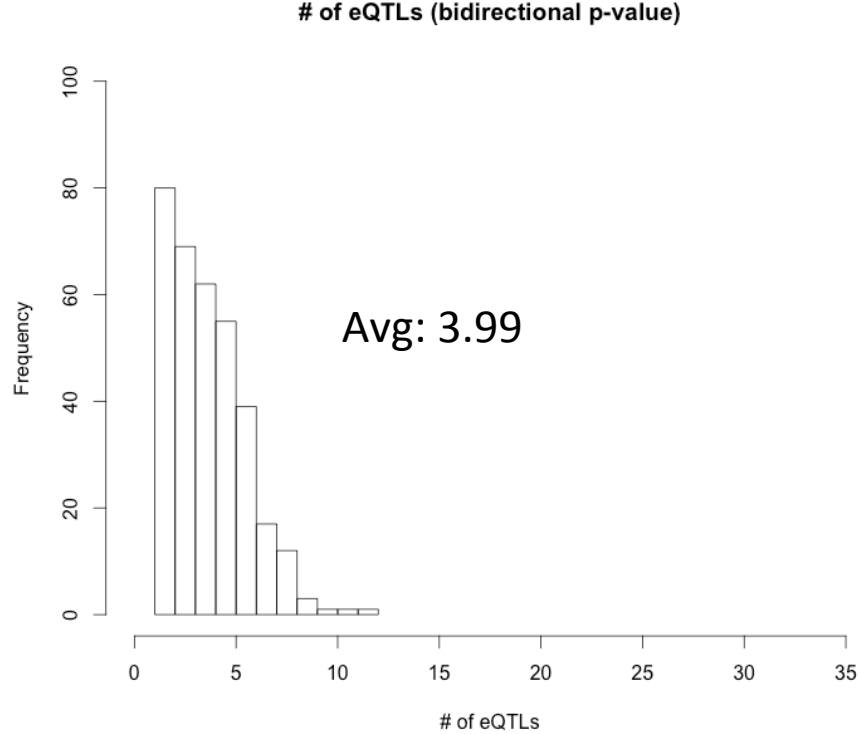
The Problem

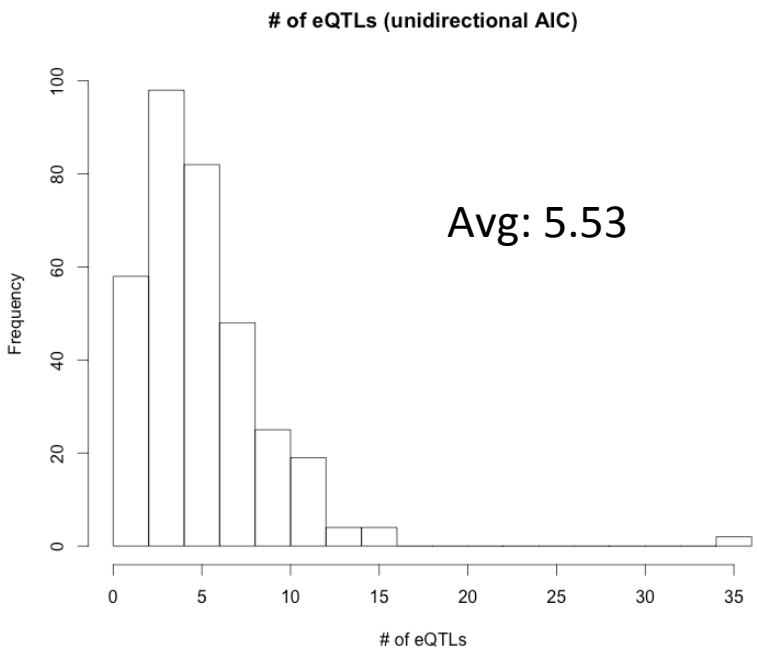
- If 2 SNPs are perfectly collinear, or multiple SNPs are linearly dependent, it's not clear how to include them into the model.
- My scripts get rid of these collinear/linearly dependent SNPs, but they do not make an informed decision about which SNPs to keep in the model.

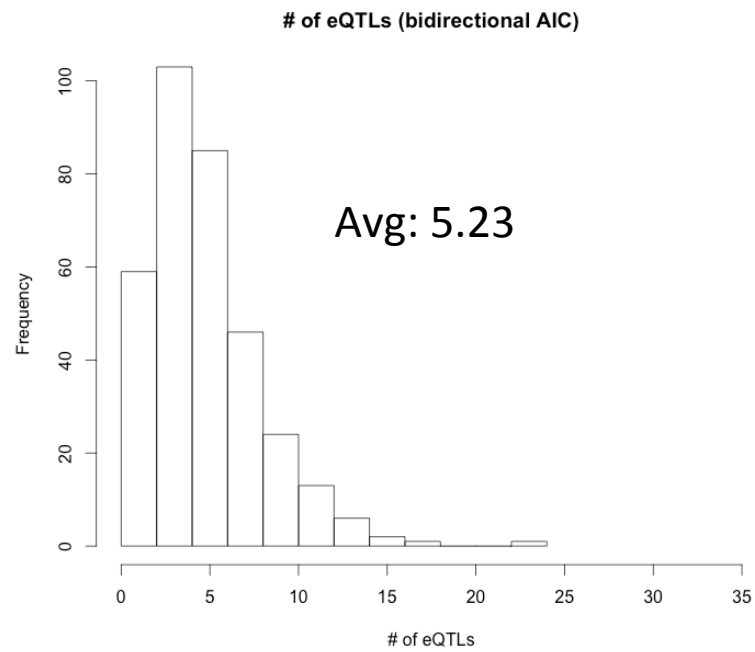
Possible Solutions

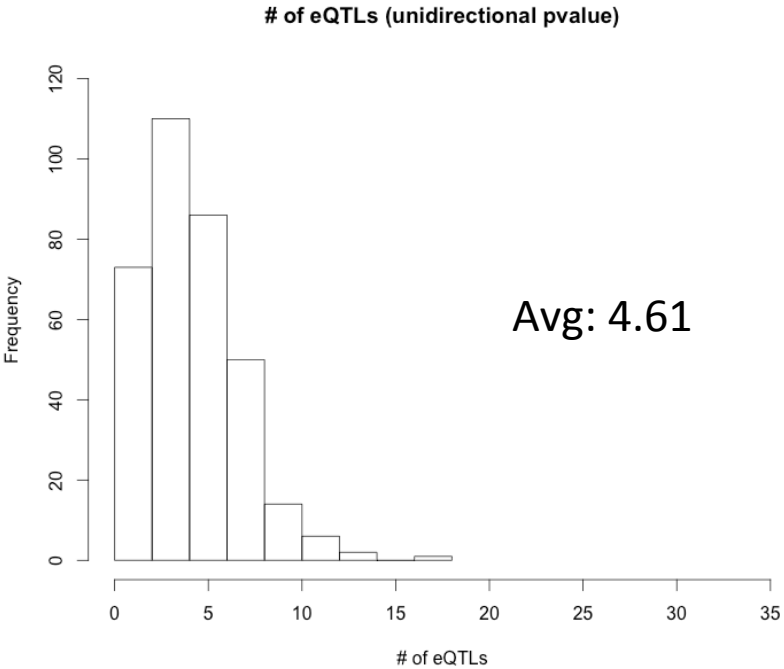
- Keep the best SNP reported by Matrix eQTL
- Assign weights to SNPs based on biological factors (e.g. distance from TSS, conservation score, DNASE I HS, etc…)

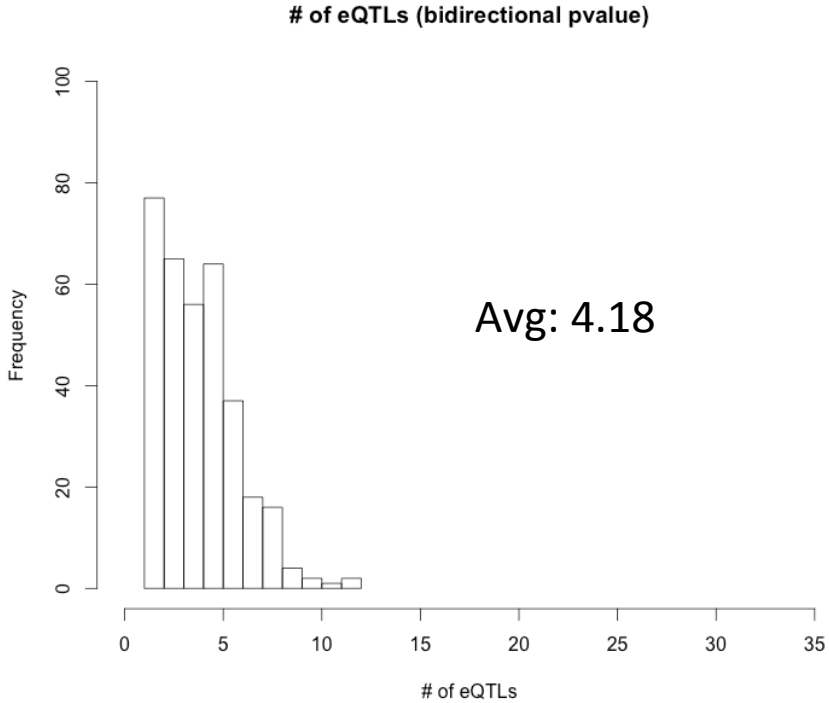

# of eQTLs (unidirectional p-value)
Frequency
0
20
40
60
80
100
0
5
10
15
20
25
30
35
# of eQTLs
Avg: 4.52

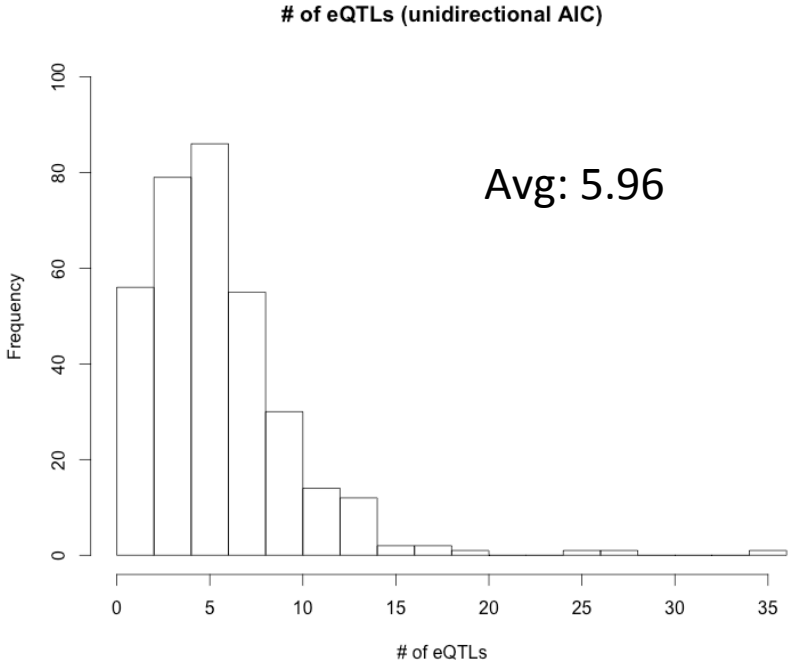

# of eQTLs (bidirectional p-value)
Frequency
0
20
40
60
80
100
0
5
10
15
20
25
30
35
# of eQTLs
Avg: 3.99

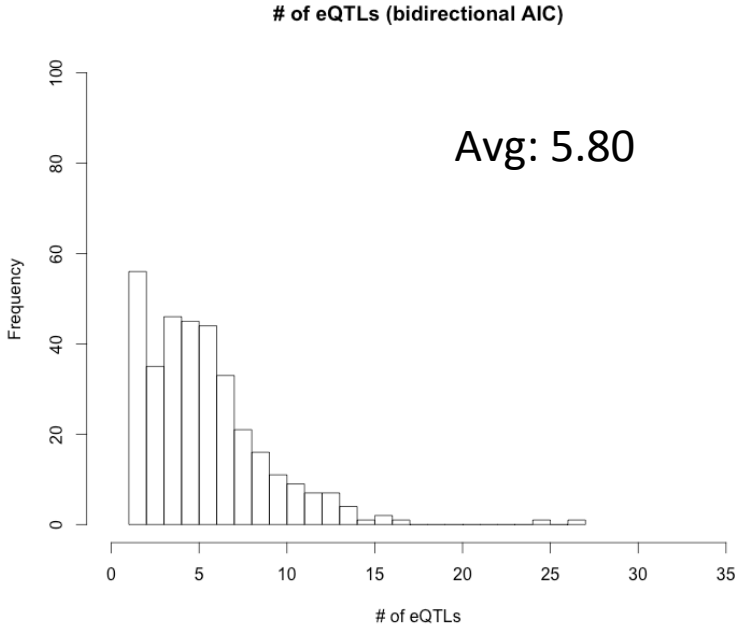

# of eQTLs (unidirectional AIC)
Frequency
0
20
40
60
80
100
0
5
10
15
20
25
30
35
# of eQTLs
Avg: 5.53

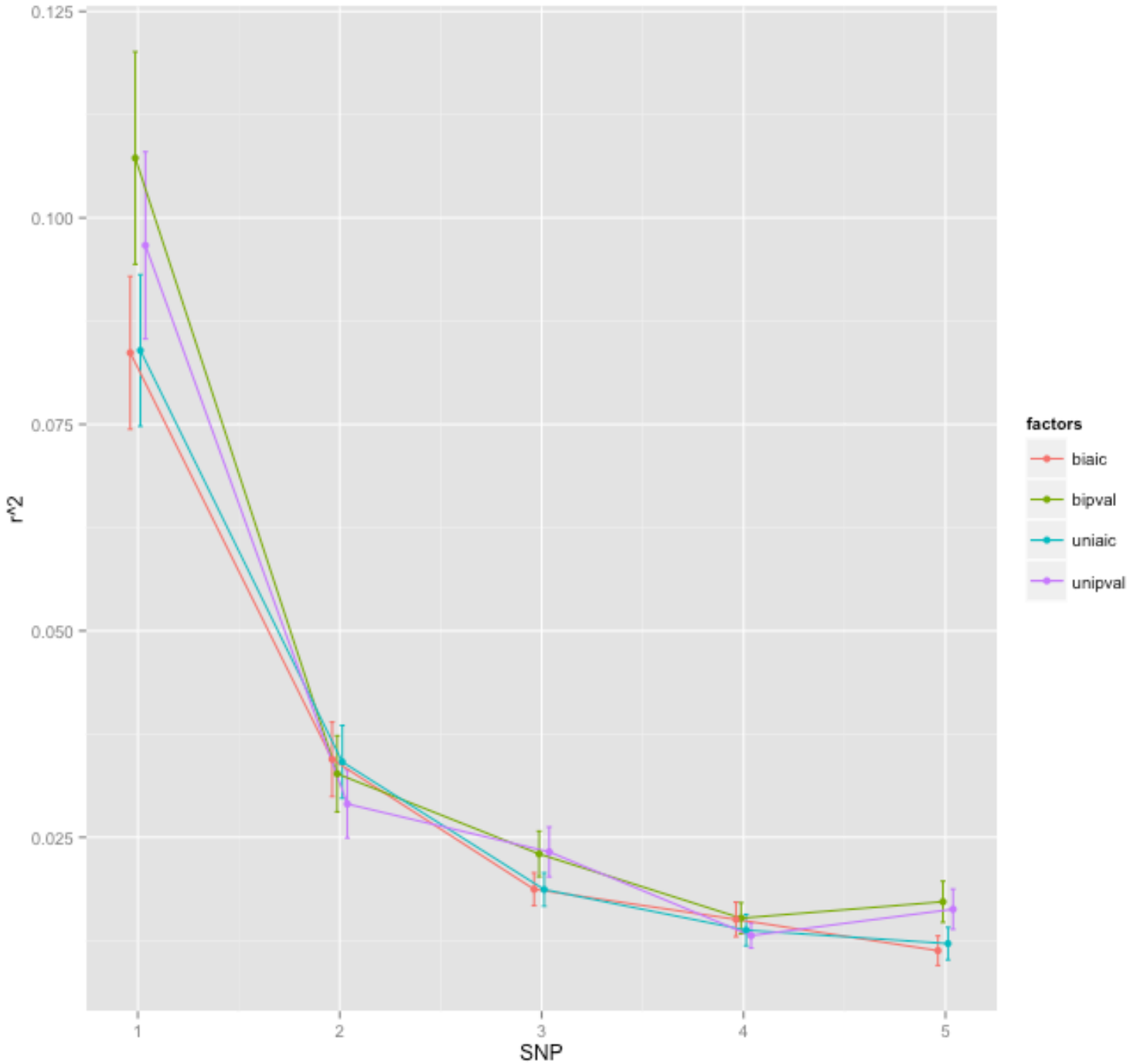

# of eQTLs (bidirectional AIC)
Frequency
0
20
40
60
80
100
0
5
10
15
20
25
30
35
# of eQTLs
Avg: 5.23


# of eQTLs (unidirectional pvalue)
Frequency
# of eQTLs
Avg: 4.61


# of eQTLs (bidirectional pvalue)
Frequency
# of eQTLs
Avg: 4.18


# of eQTLs (unidirectional AIC)
Frequency
# of eQTLs
Avg: 5.96


# of eQTLs (bidirectional AIC)
Frequency
# of eQTLs
Avg: 5.80


0.125
0.100
0.075
0.050
0.025
r^2
1
2
3
4
5
SNP
factors
biaic
bipval
uniaic
unipval

Cross-validation

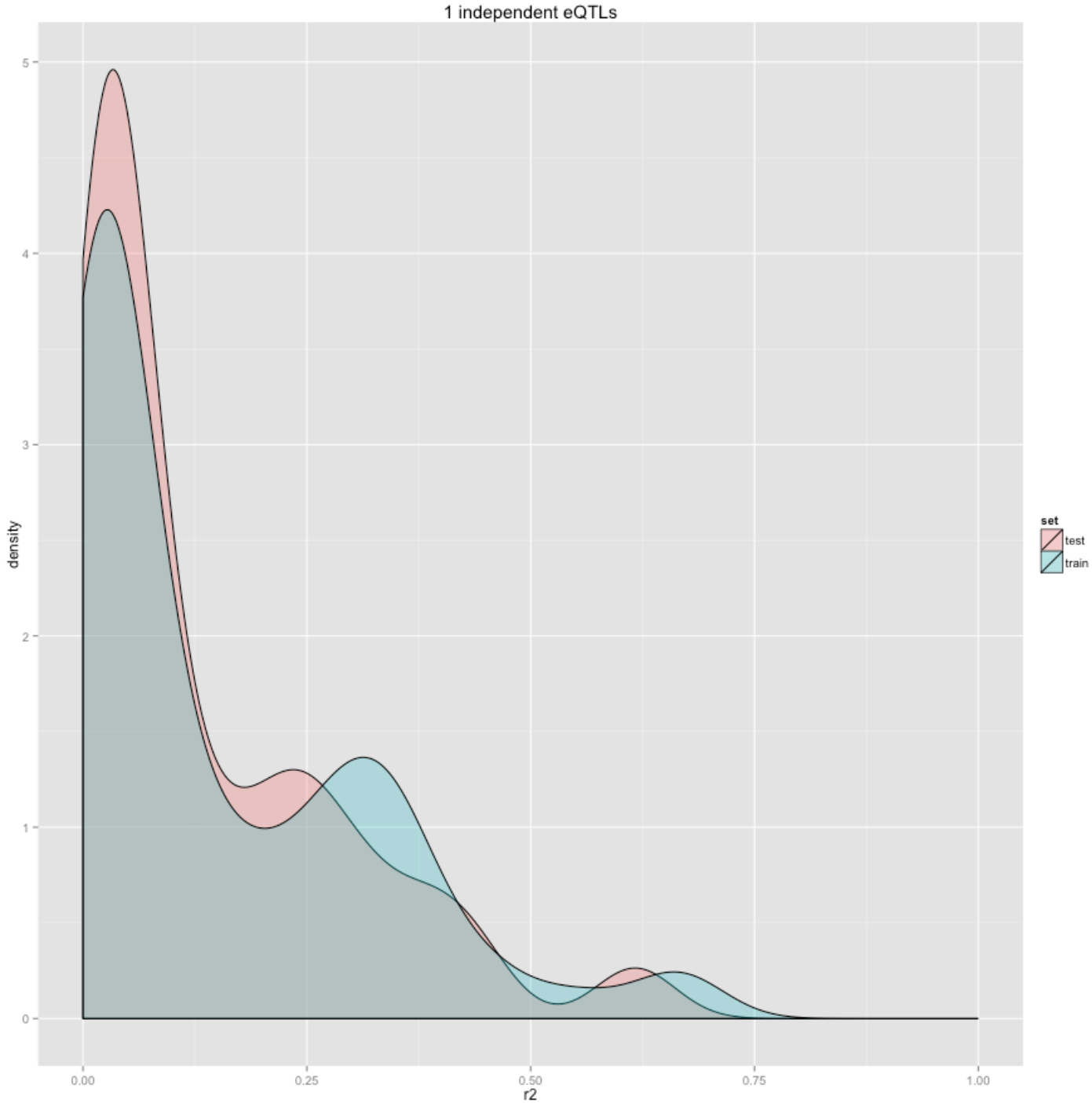
- Average r^2 across all genes for the four methods:

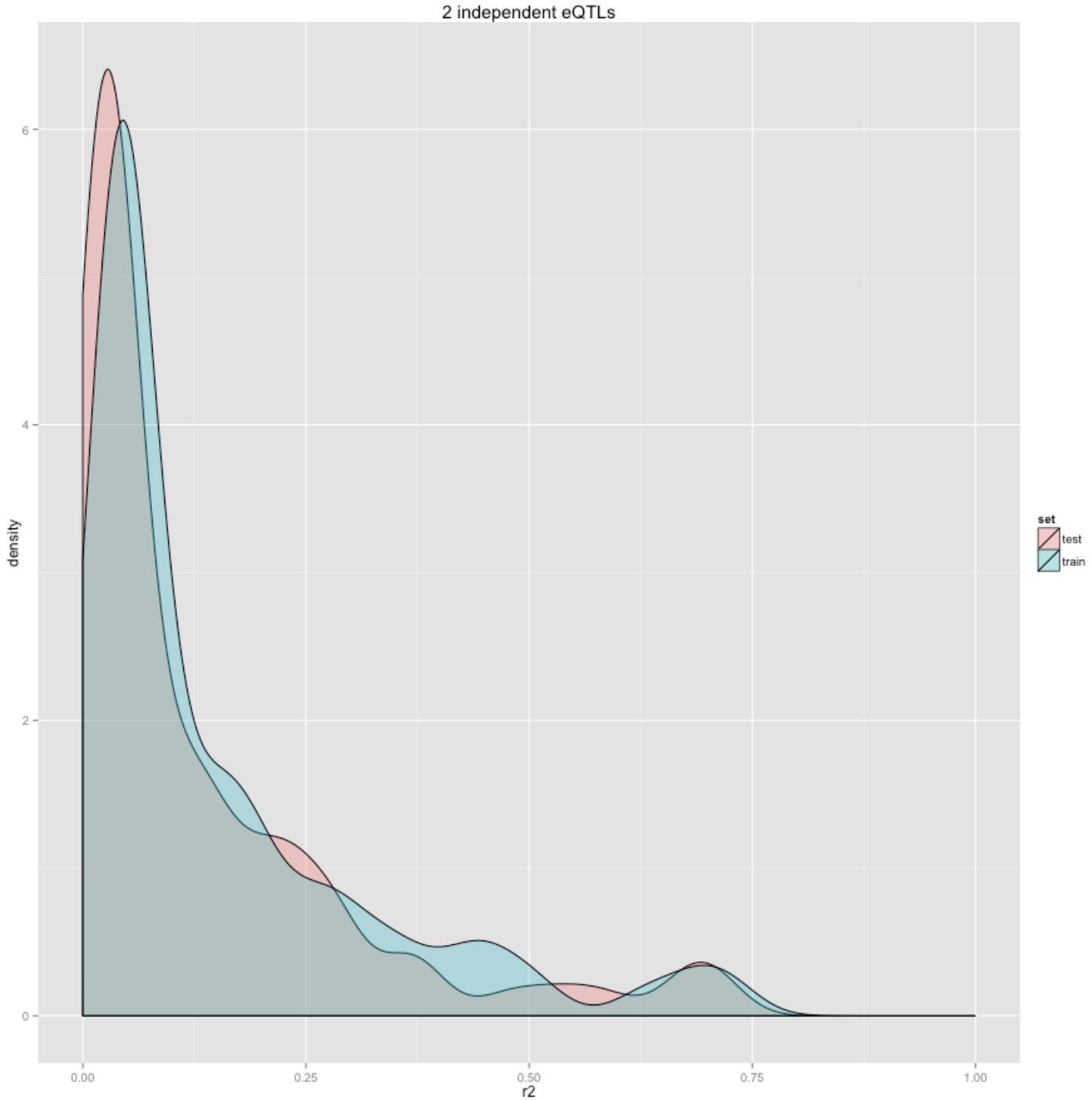
Method	Avg r^2 for training set	Avg r^2 for testing set
p-value unidirectional	0.174	0.115
p-value bidirectional	0.172	0.114
AIC unidirectional	0.228	0.065
AIC bidirectional	0.227	0.064

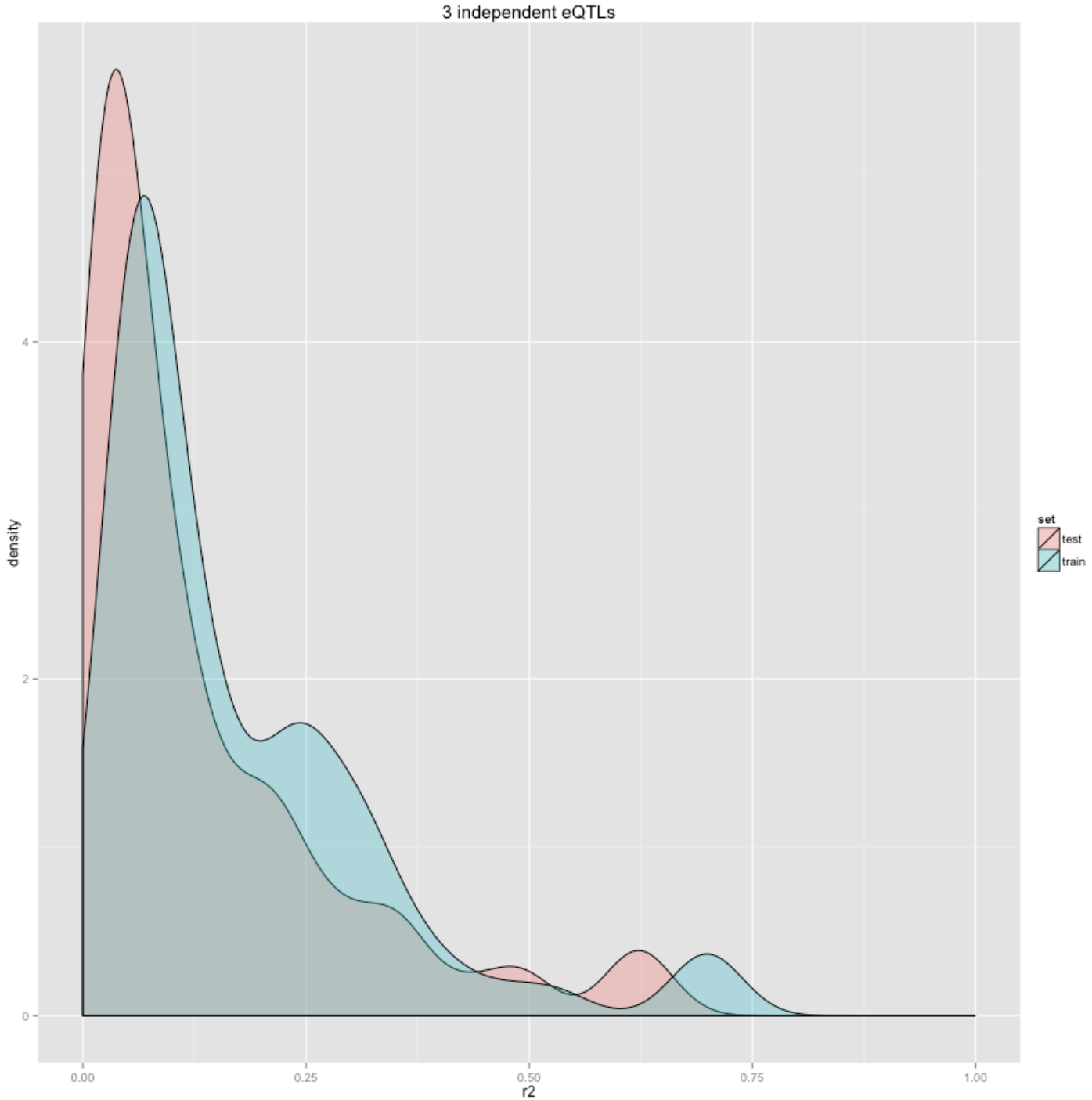
Cross-validation

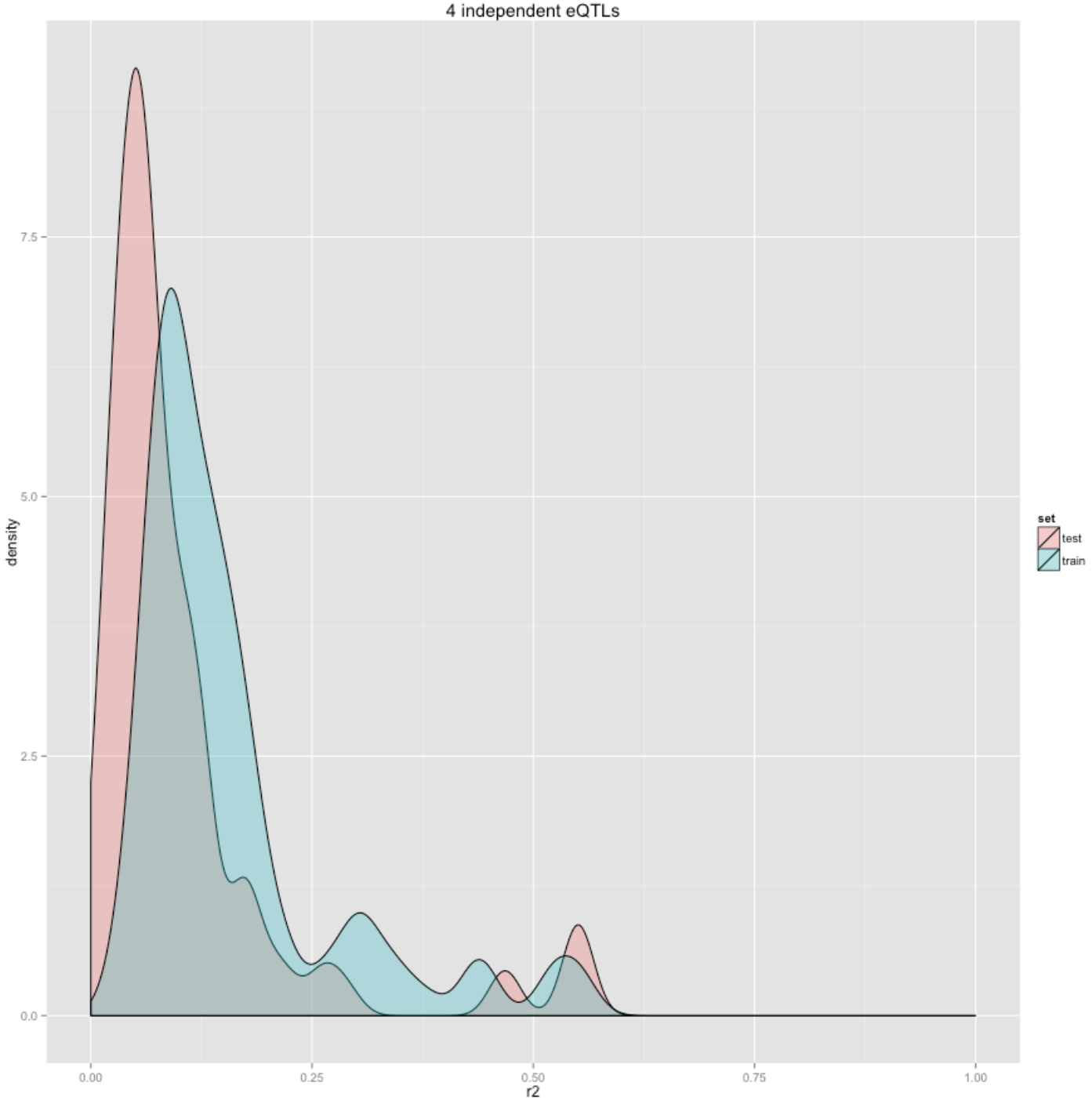
- Average r^2 across all genes for the four methods when all four methods are forced to take the same # of SNPs in the model:

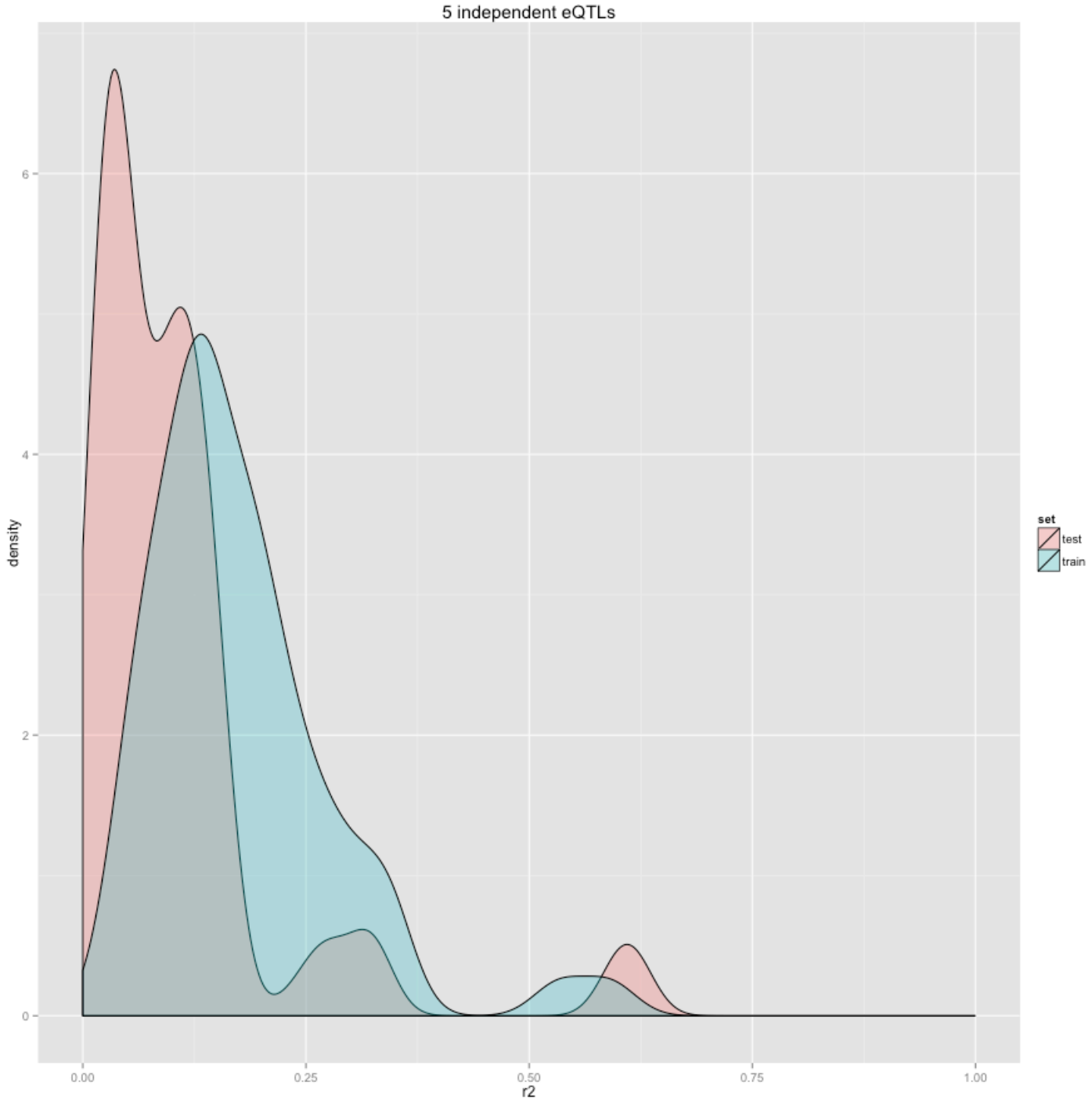
Method	Avg r^2 for training set	Avg r^2 for testing set
p-value unidirectional	0.171	0.118
p-value bidirectional	0.172	0.118
AIC unidirectional	0.156	0.114
AIC bidirectional	0.157	0.114

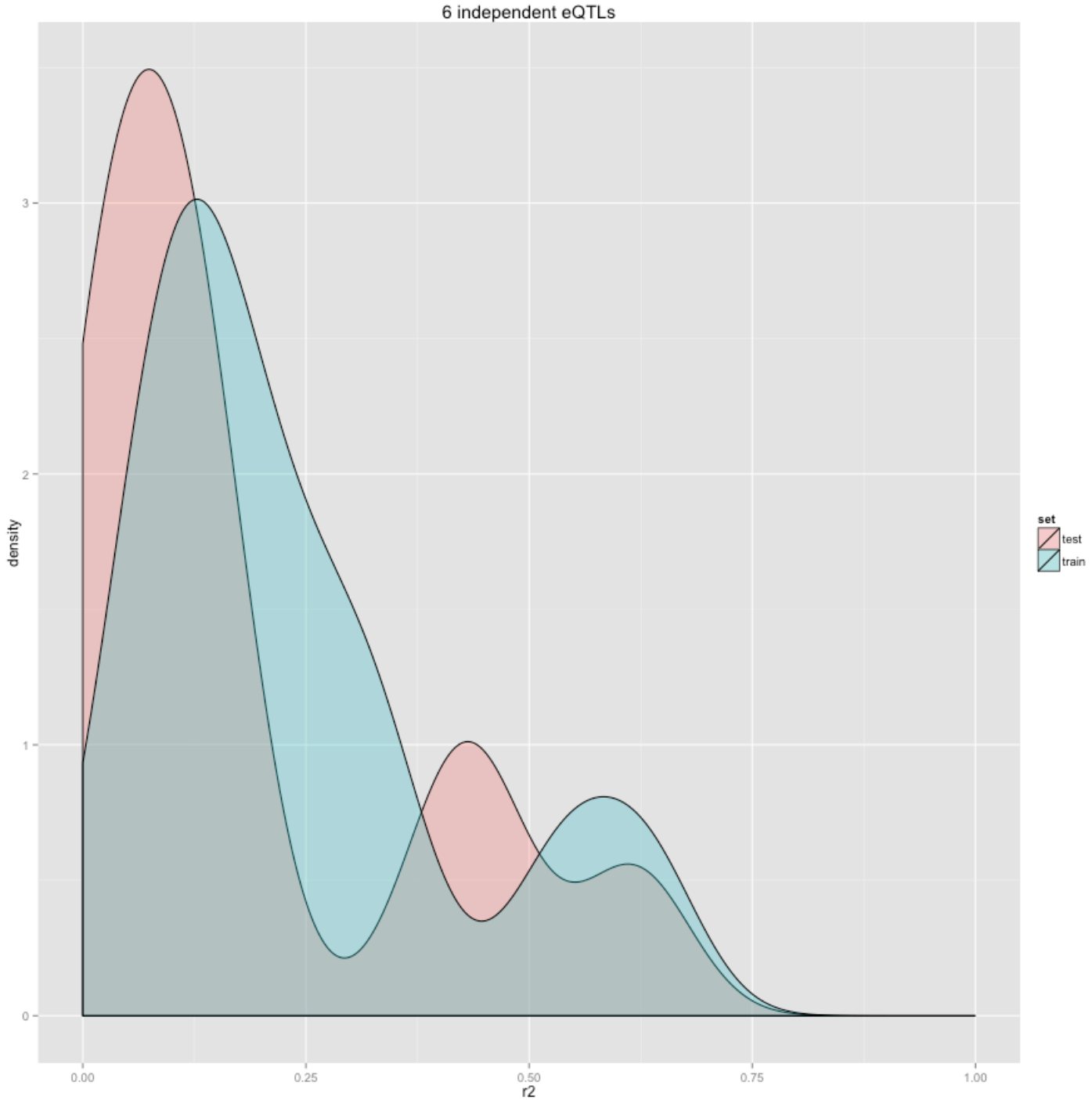

1 independent eQTLs
density
r2
0.00
0.25
0.50
0.75
1.00
0
1
2
3
4
5
set
test
train

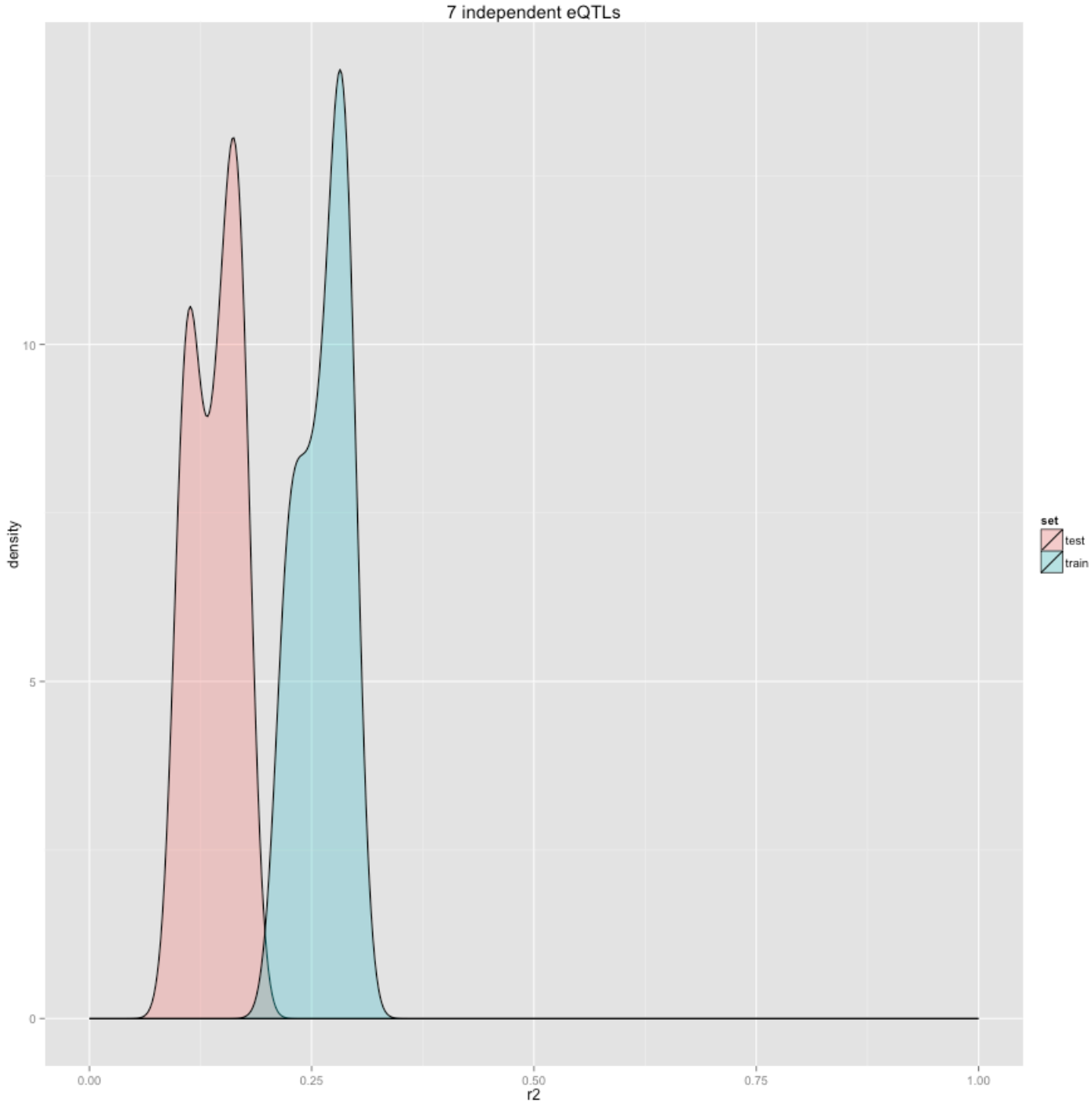

2 independent eQTLs
density
6
4
2
0
0.00
0.25
0.50
0.75
1.00
r2
set
test
train


3 independent eQTLs
density
4
2
0
0.00
0.25
0.50
0.75
1.00
r2
set
test
train

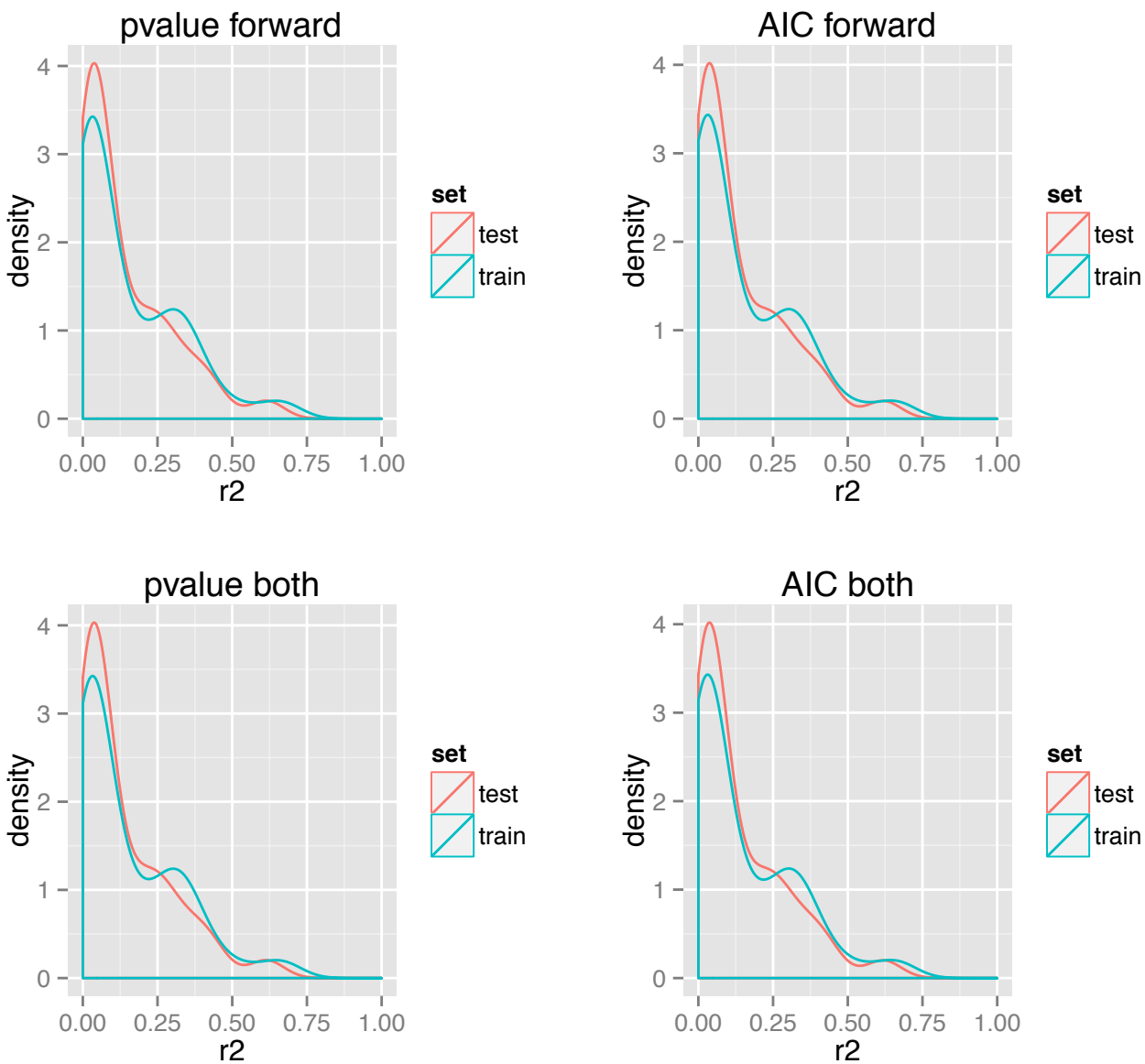

4 independent eQTLs
density
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
r2
set
test
train


5 independent eQTLs
density
6
4
2
0
0.00
0.25
0.50
0.75
1.00
r2
set
test
train

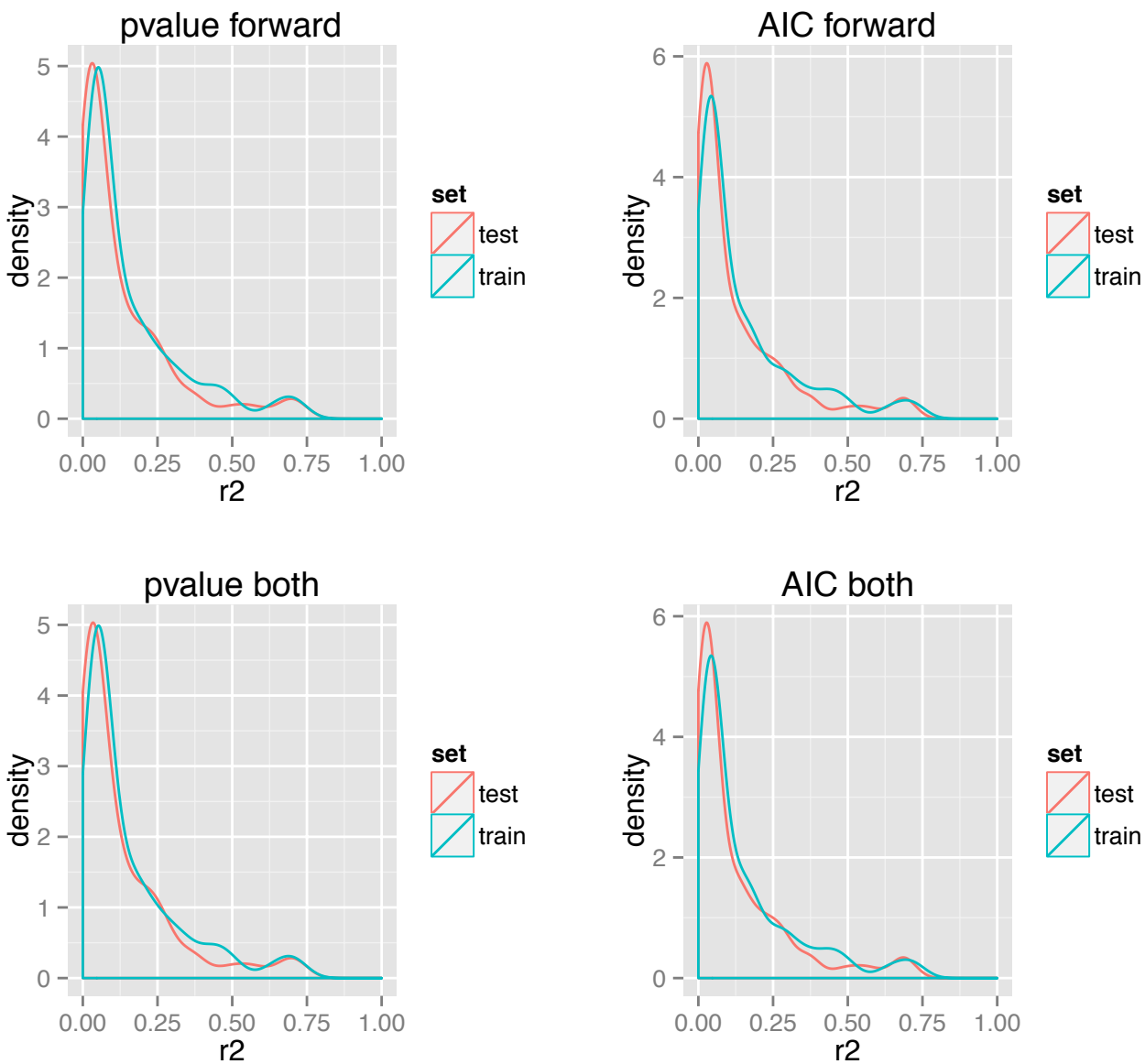

6 independent eQTLs
density
3
2
1
0
0.00
0.25
0.50
0.75
1.00
r2
set
test
train


7 independent eQTLs
density
10
5
0
0.00
0.25
0.50
0.75
1.00
r2
set
test
train

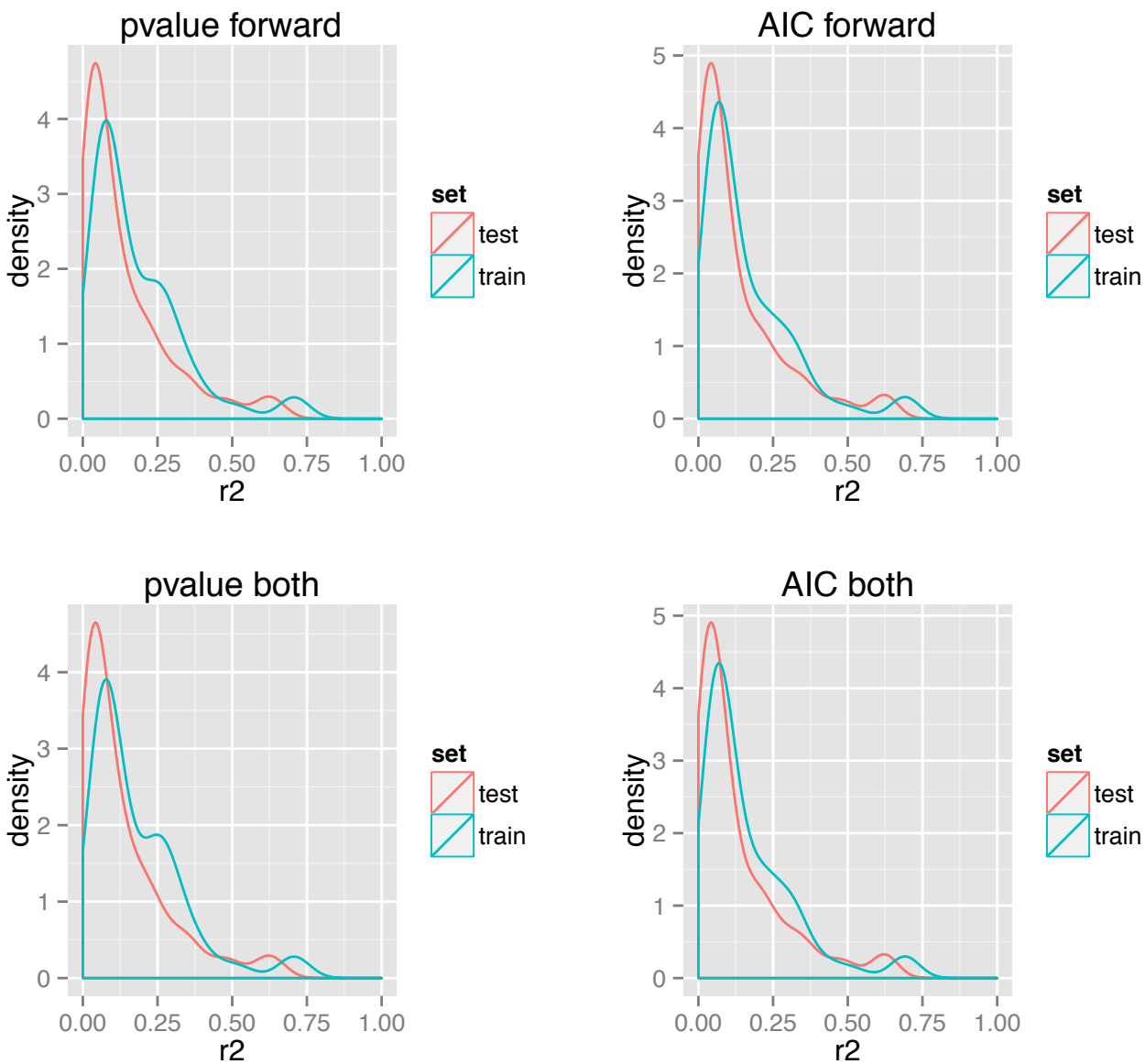
1 independent eQTLs



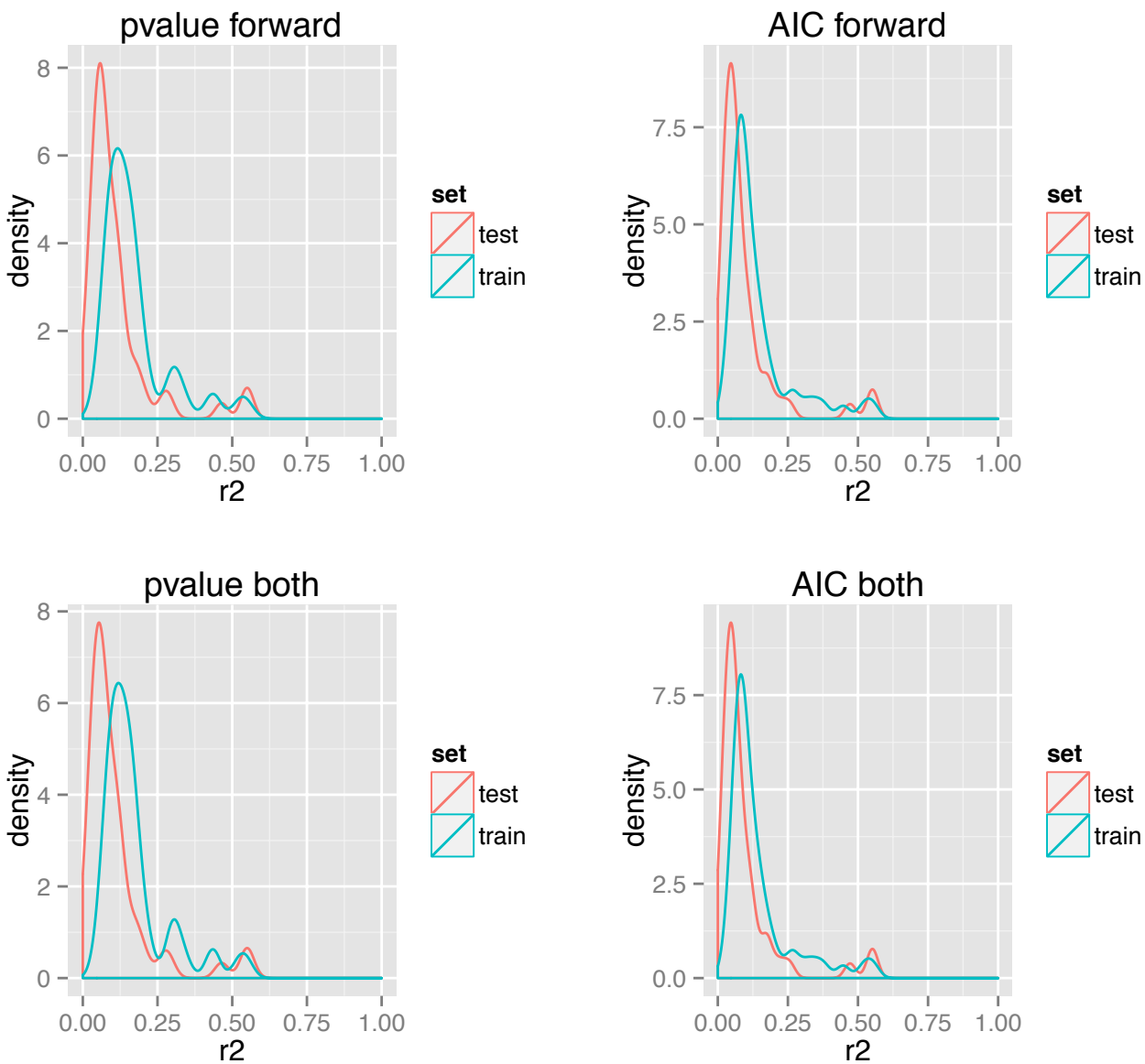
2 independent eQTLs



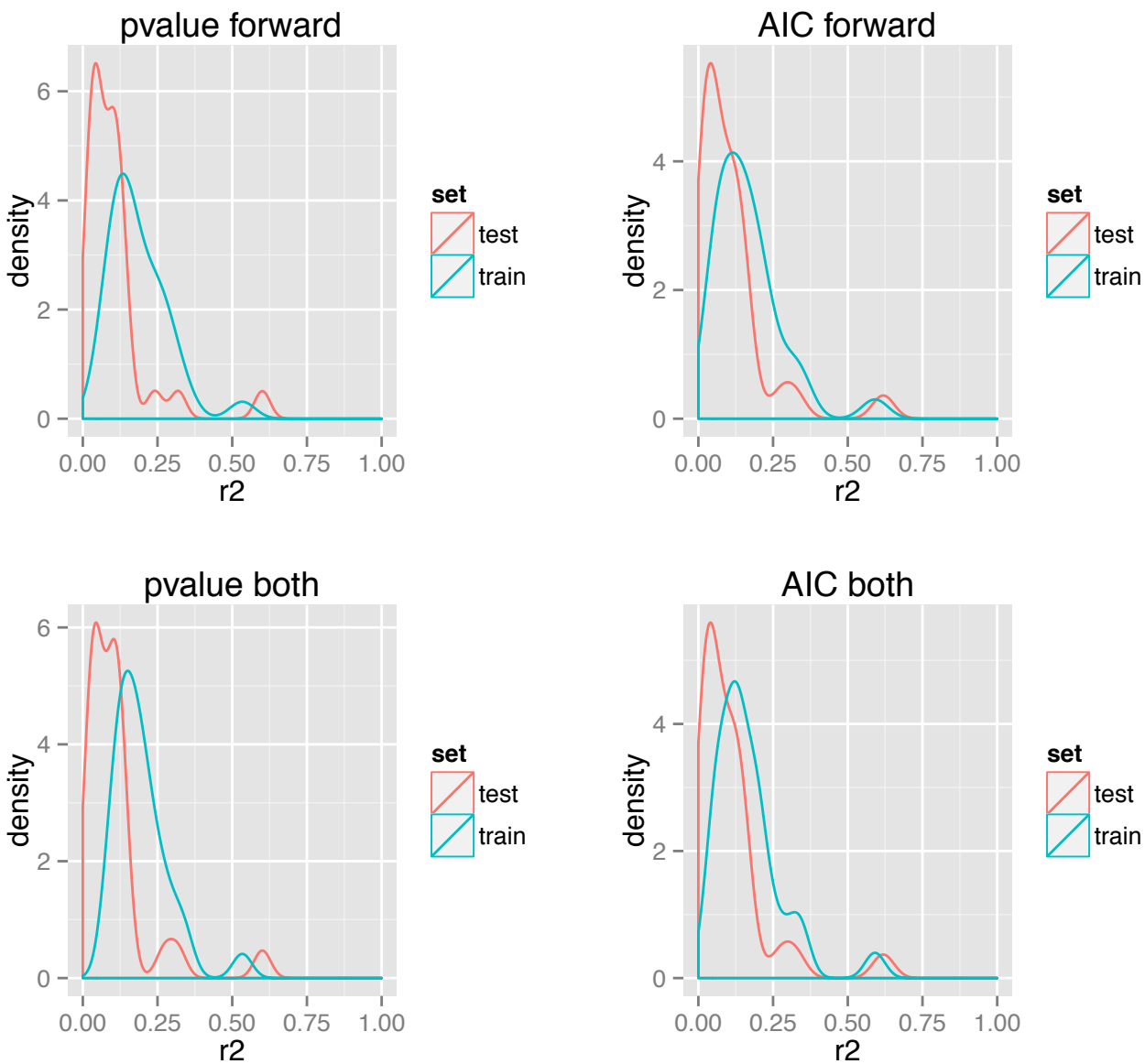
3 independent eQTLs



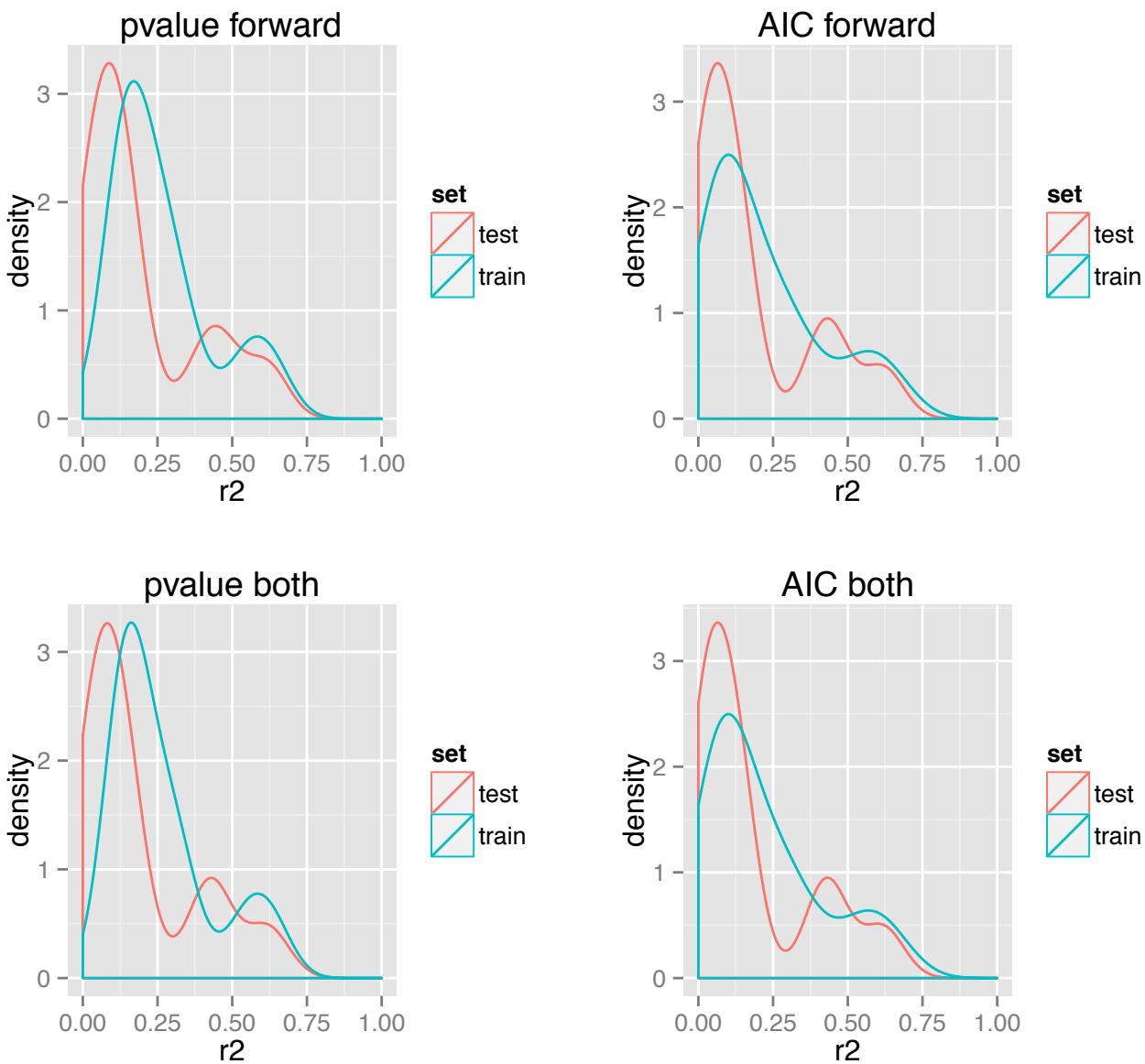
4 independent eQTLs



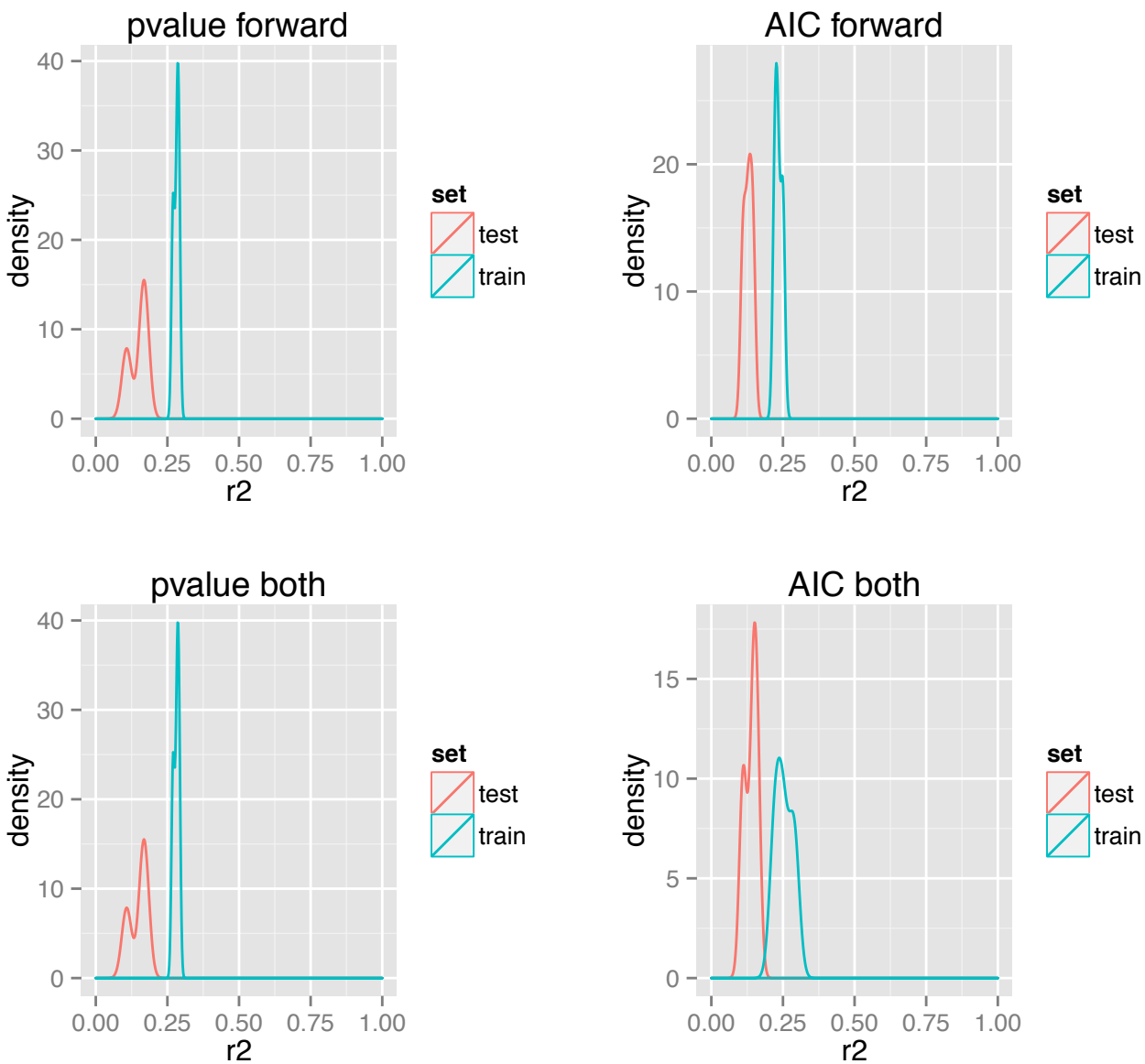
5 independent eQTLs



6 independent eQTLs



7 independent eQTLs



Future Directions/Questions

- Is effect size or p-value the best criterion by which to order SNPs?
- Improve the collinear/linearly dependent SNP problem
- The SNPs studied here were cis-eQTLs; do their corresponding trans-eQTLs influence the same transcriptional network?

Future Directions/Questions

- LASSO???