Gene-set analysis

Danielle Posthuma & Kyoko Watanabe

Faculty/Danielle/2019/Wednesday/Pathway.ppt

Understanding biology

A major goal of genetic studies is to gain mechanistic, biological insight into a disease or trait. This will aid in designing treatment and prevention strategies

- Monogenic disorders: one causal SNP, one gene, large effect. Investigating biology and gaining mechanistic insight relatively straightforward

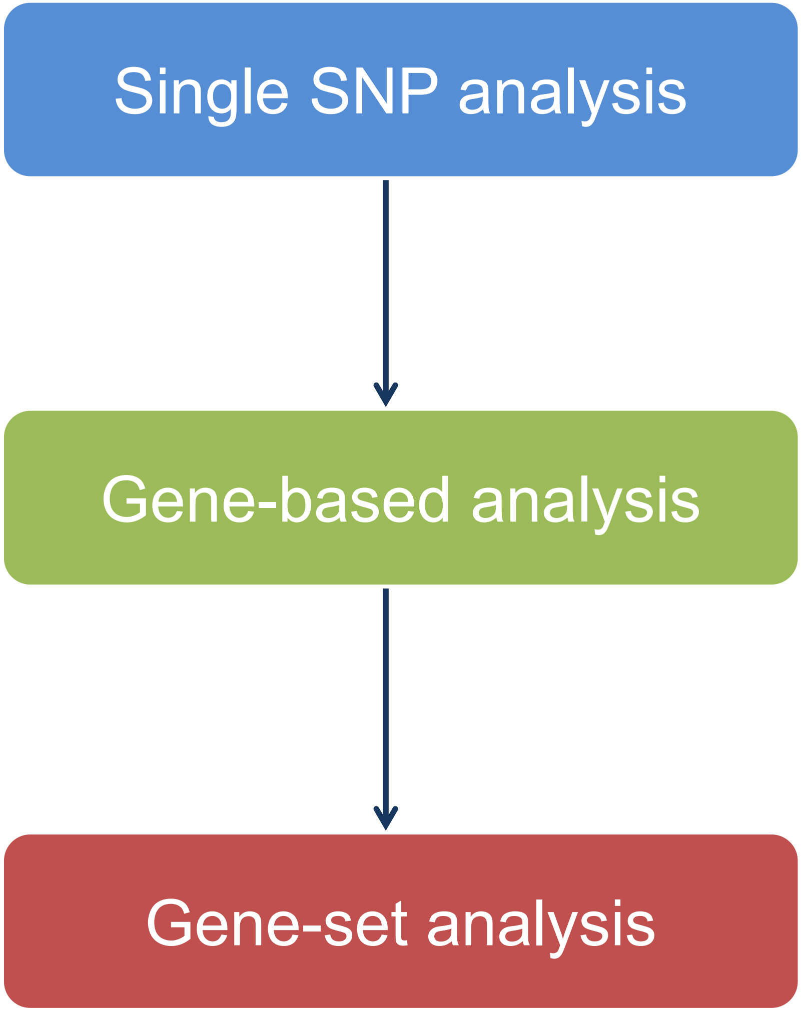
- Polygenic, or 'complex', traits: many SNPs, many genes, small effects. Investigating biology is challenging, gaining mechanistic insight difficult.

Making sense of GWAS results for complex traits

- Annotate SNPs to genes, based on physical location or regulatory relation (also see FUMA talk Friday)

- Conduct gene-based analyses

- Conduct gene-set analyses



Single SNP analysis

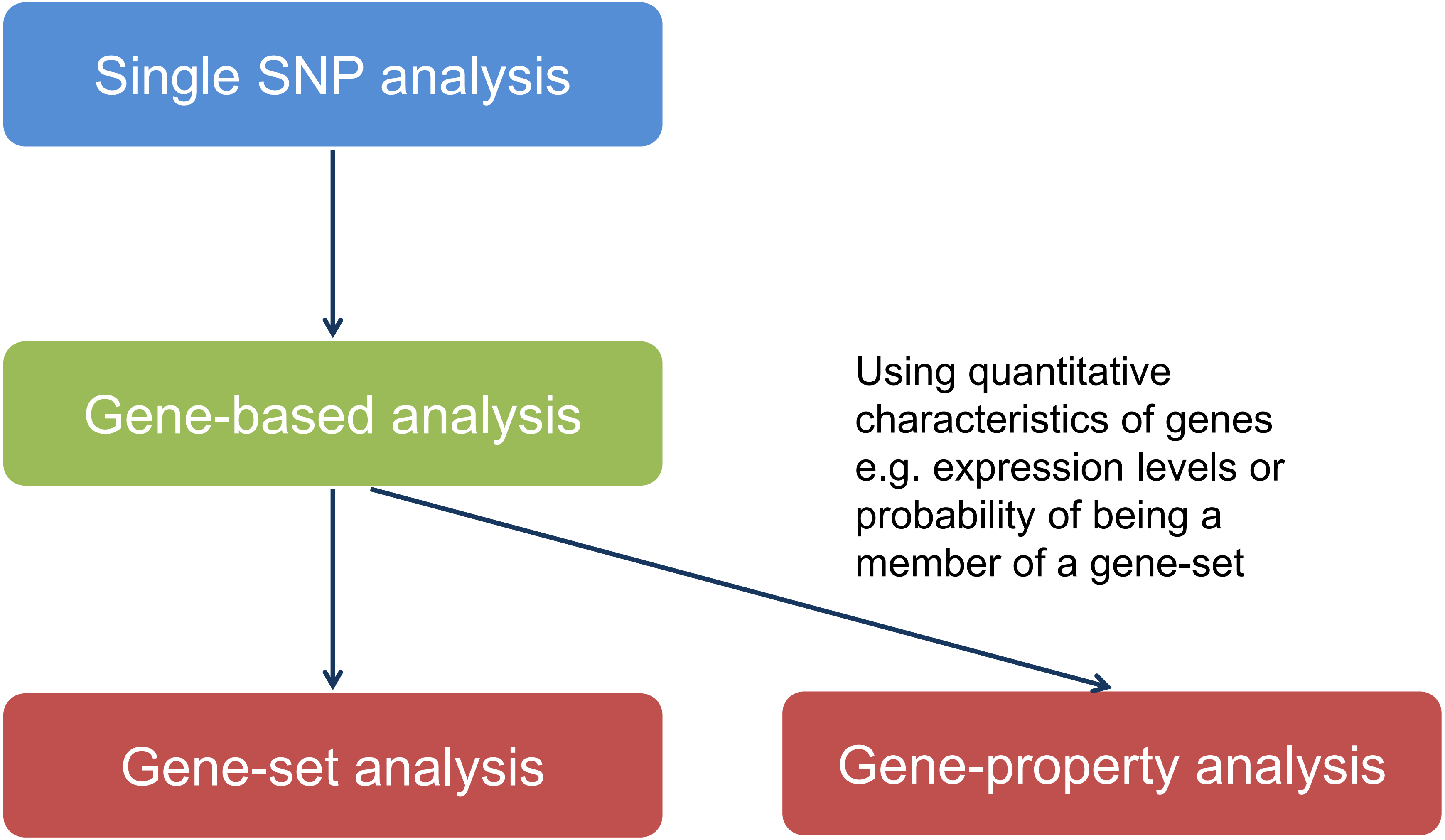
Gene-based analysis

Gene-set analysis

GWAS: SNP by SNP analysis

Joint association effect of all SNPs in a gene, taking into account LD

Gene-set analysis with sets of genes as unit of analysis
- targeted gene-sets/pathways
- all known gene-sets/pathways



Single SNP analysis

Gene-based analysis

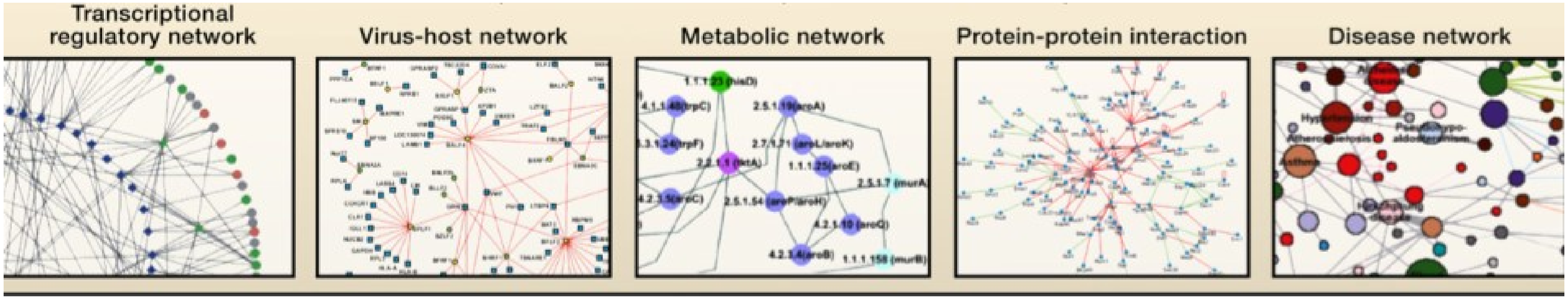
Using quantitative characteristics of genes e.g. expression levels or probability of being a member of a gene-set

Gene-set analysis

Gene-property analysis

Choosing gene-sets

Gene-sets can be based on e.g.
- protein-protein interaction
- co-expression
- shared cellular function
- biological pathway
- *etc*


Transcriptional regulatory network
Virus-host network
Metabolic network
Protein-protein interaction
Disease network

Public databases vs. manual

Information in online databases tends to be

- somewhat biased
  - not all genes included, disease genes tend to be investigated more often
  - genes that are investigated more often will have more interactions
- not always reliable
  - Interactions or functions often not validated, sometimes only predicted.

Tools for gene-set analyses

INRICH, ALIGATOR, MAGENTA, FORGE, SETSCREEN, DAPPLE, DEPICT, MAGMA etc etc

-> do they all provide the same answer..?

Statistical issues in gene-set analyses

- Self-contained vs. competitive tests
- Different statistical algorithms test different alternative hypotheses
- Different statistical algorithms have different sensitivity to LD, ngenes, nSNPs, background h^2

De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

Self-contained vs. competitive tests

Null hypothesis:

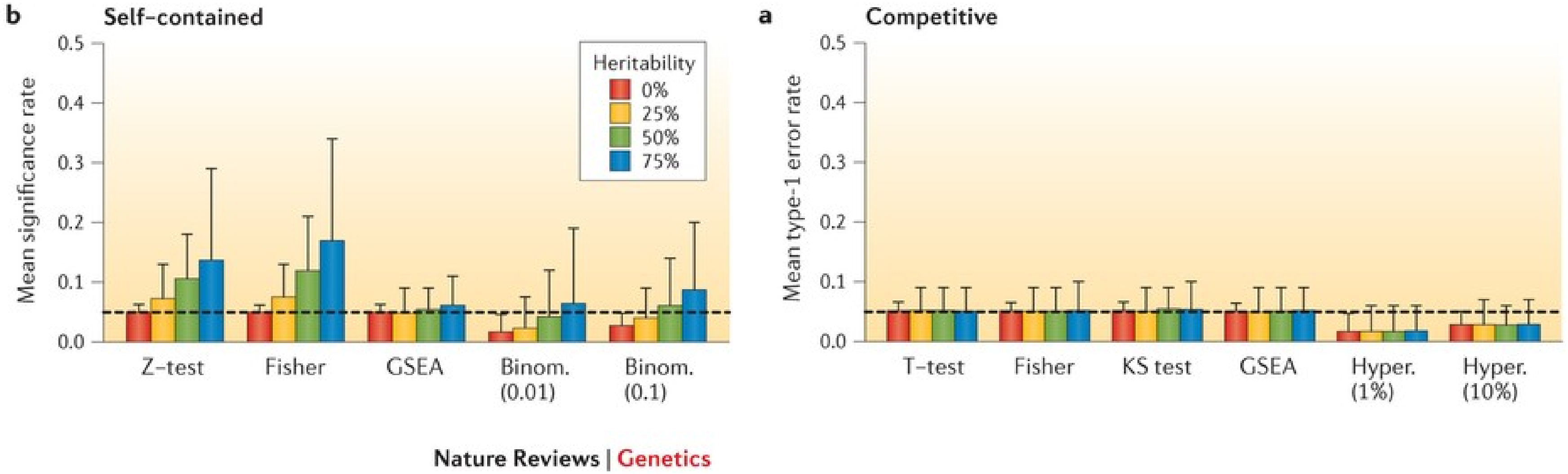
Self-contained:
H0: The genes in the gene-set are not associated with the trait

Competitive:
H0: The genes in the gene-set are not more strongly associated with the trait than the genes not in the gene-set

Why use competitive tests

- Polygenic traits influenced by thousands of SNPs in hundreds of genes
- Very likely that many combinations (i.e. gene-sets) of genes are significantly associated
- Competitive tests define which combinations are biologically most interpretable

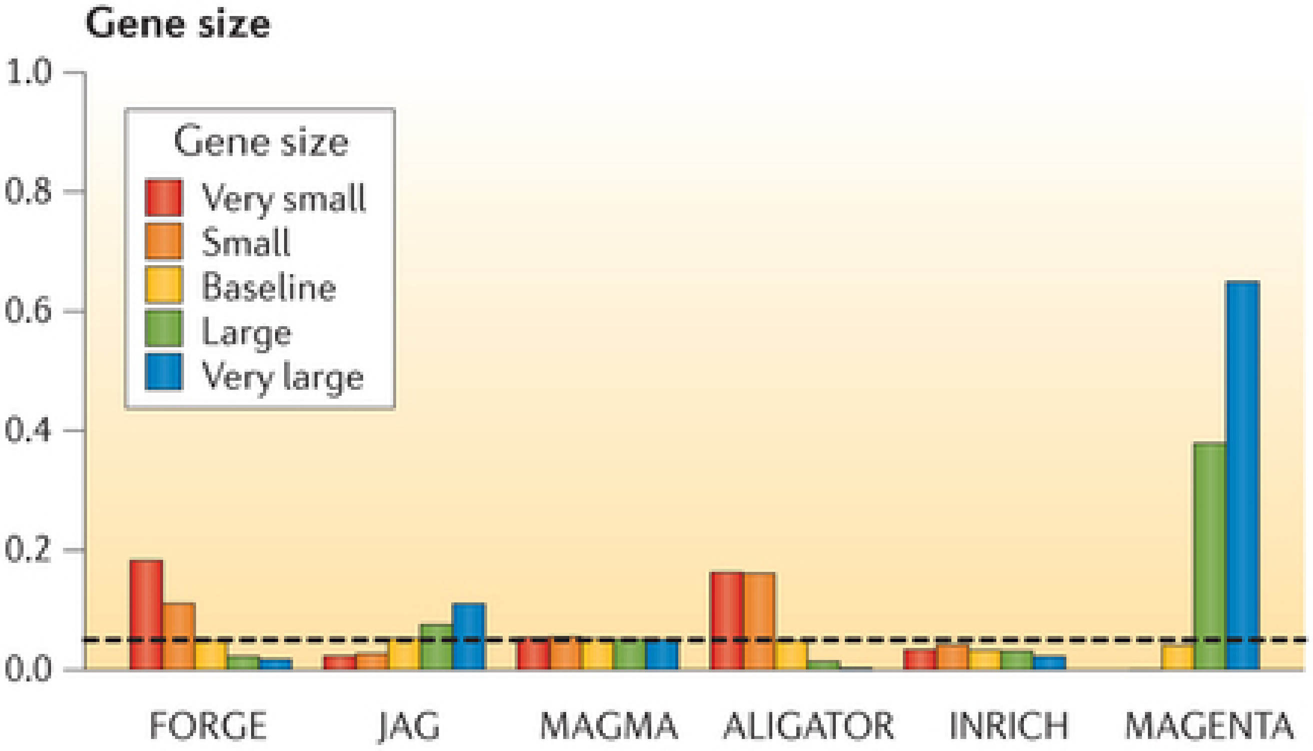
Polygenicity and number of significant gene-sets in self-contained versus competitive testing



For self-contained methods, type I error rates increase with heritability, whereas they are constant for competitive methods.

De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

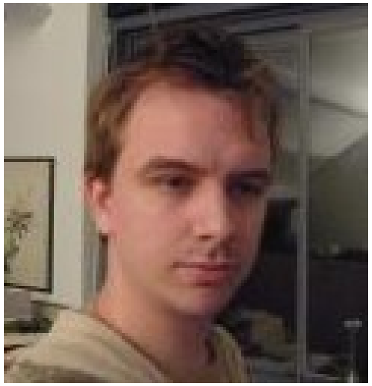
Different tools are differentially affected by gene size



De Leeuw, Neale, Heskes, Posthuma. Nat Rev Genet, 2016

MAGMA

- gene and gene-set analysis
  - Command-line interface



Christiaan de Leeuw

- Input
  - Genotype and phenotype data
    - Or: (full) published GWAS results (plus reference data)
  - Gene definitions
  - Gene sets

de Leeuw CA, Mooij JM, Heskes T, Posthuma D. PLoS Comput Biol. 2015

MAGMA - workflow

- Three main steps
  1. **Annotation**: map SNPs onto genes
  2. **Gene analysis**: compute association of genes with phenotype
  3. **Gene-set analysis**: test gene associations in gene sets
- Generalized gene-set analysis
  4. Continuous ‘sets’
  5. Conditional (joint) analysis
  6. Interaction analysis

1. Annotation

- Map SNPs to a gene based on physical location
  - If located inside the transcription region of the gene
  - Optionally, if located in window around the gene
    - Especially upstream of transcription start site
  - A SNP can be mapped to multiple genes

- Manual annotation of SNPs to genes
  - MAGMA by default annotates SNPs to genes based on distance, but you could create your own annotation manually

2. Gene analysis

4 models available in MAGMA

- Principal component linear regression
  - Performs test on explained phenotypic variance (F-test)
  - Requires raw genotype data
- SNP-wise models: compute SNP associations with phenotype first
  - SNP-wise Mean: performs test on mean SNP association
  - SNP-wise Top: performs test on strongest SNP association
  - SNP-wise Multi: combines SNP-wise Top and Mean

3. Gene-set analysis (GSA)

Gene ID	Association	In gene set
1	1.32	Yes
2	-0.76	Yes
3	0.48	Yes
4	1.12	Yes
5	-0.02	Yes
6	-1.04	No
7	0.86	No
8	-1.27	No
9	0.41	No
10	0.11	No

An analysis of genes:

- Genes are data points in the analysis
- The gene set is a grouping variable
- Genetic association with the phenotype is the outcome variable

gene-set analysis is like a t-test

Testing the mean association of genes in the gene set

3. Gene-set analysis

Gene ID	Association	In gene set
1	1.32	Yes
2	-0.76	Yes
3	0.48	Yes
4	1.12	Yes
5	-0.02	Yes
6	-1.04	No
7	0.86	No
8	-1.27	No
9	0.41	No
10	0.11	No

$\left.\right\}\mu_S$ (genes 1–5)

$\left.\right\}\mu_0$ (genes 6–10)

Competitive analysis:

- Is the mean genetic association of genes in the gene set greater than that of genes outside the gene set?
  - H_0: $\mu_S = \mu_0$

Only competitive analysis allows any inference about the gene set itself

Issues of interpretation in GSA

Statistically significant gene sets are concluded to play a role in the phenotype

Is this a valid conclusion?

Issues of interpretation in GSA

GSA tests for accumulation of genetic association in the set, which may be because:

- **Direct effect:** the set (or biological function) itself is involved
- **Confounding:** the set itself is not involved, but many genes in the set overlap with genes in another set that is involved
- **Interaction:** the set itself is partially involved, with the effect specific to a subset defined by another gene set

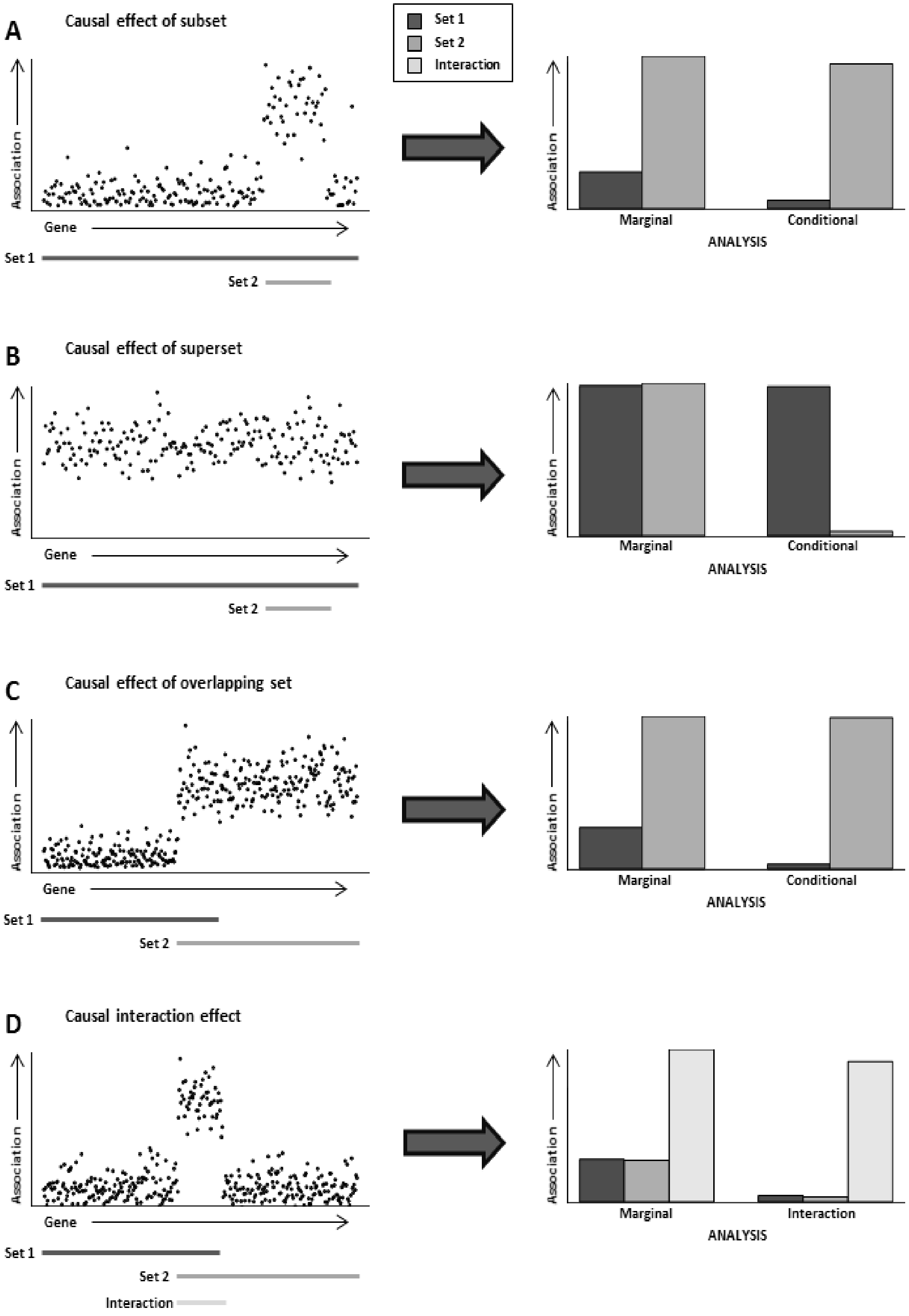
Four general confounding scenarios:

Overlap with actually associated set induces spurious association

- **A:** set1 includes a causal subset
- **B:** set1 is causal and set2 is subset of set1
- **C:** set1 partly overlaps with causal set2

Interaction can be seen as special instance of subset confounding

- **D:** set1 and set2 overlap and the overlapping set of genes is causal



Conditional gene-set analysis

Confounding among gene sets can be tested using a conditional analysis

- In MAGMA: linear regression framework, can add potential confounders as covariates in the analysis to evaluate their influence

When analysing a ‘causal’ set A and an overlapping set B:

- Conditioning set B (on A) will make its association disappear, whereas conditioning set A (on B) will only reduce its association

Confounding remains problematic if ‘causal’ set not available

Interaction gene-set analysis

  - The interaction term is the set AB of genes shared by A and B
  - The interaction can be evaluated by testing AB conditional on A and B

- A gene set interaction arises if the genetic associations are specific to genes that share the same multiple functions

Conclusion

- GSA can identify biologically relevant gene-sets for a trait
- This helps to generate hypotheses that can be tested in functional experiments, with the aim to gain mechanistic insight
- Be aware of statistical issues
- Always check overlap between gene-sets and conduct conditional analyses

- *de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015*
- *de Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties of gene-set analysis. Nat Rev Genet. 2016*
- *de Leeuw CA, Stringer S, Heskes T, Posthuma D. Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure. Nat Comm, 2018*

Setting up for the practical

1. Open terminal

2. Copy practical files into your home directly
   ```
   cp /faculty/danielle/2019/Wednesday/magma_practical.zip ./
   ```

3. Unzip
   ```
   unzip magma_practical.zip
   ```

4. Cd into magma_practical folder
   ```
   cd magma_practical
   ```

5. Open magma_practical.pdf
   Instruction of practical

6. Open magma_commands.txt
   All MAGMA commands used in the practical

7. (**OPTIONAL**) Open followup_scripts.txt
   Some scripts to answer practical questions

Input files

Under your working directory

```
  files
    |-- reactome.sets
    |-- step3a.signif
    |-- tissue_gex.cov
    |-- step6a.partitioned.sets
```

Shared files (**DO NOT COPY TO YOUR WORKING DIR**)

```
 /data/magma/files
    |-- NCBI37.3.gene.loc
    |-- practical.bed/bim/fam
    |-- practical.cov
```

Output files

All output files will be created under `output` folder in your working directory.

Example output files are available at

`/faculty/danielle/2019/Wednesday/magma_practical_example_output.zip`

Practical

Step 1: Gene annotation

Step 2: Gene analysis

Step 3: Gene-set analysis and basic conditional analysis

Step 4: Tissue expression analysis

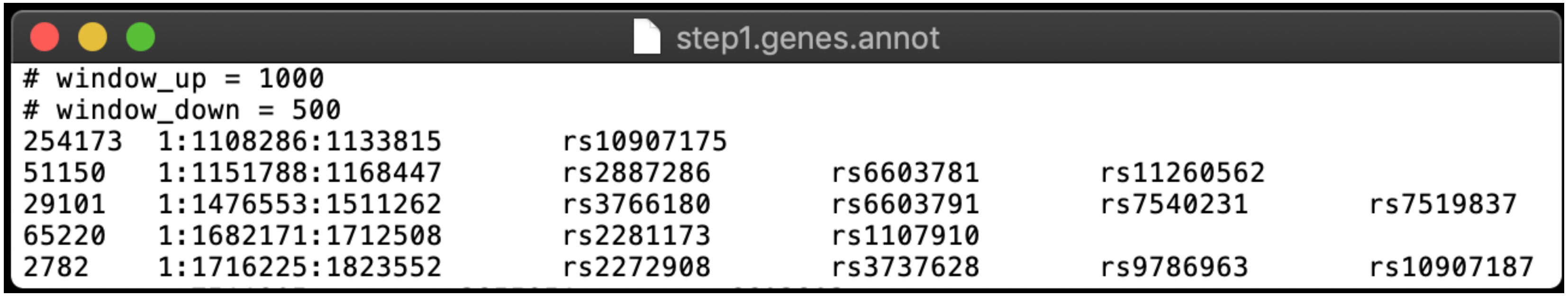
Step 5: Joint analysis of gene-set and tissue expression (**OPTIONAL**)

Step 6: Interaction gene-set analysis (**OPTIONAL**)

Step 1: Gene annotation

Output files

- `step1.genes.annot`
- `step1.log`



```
# window_up = 1000
# window_down = 500
254173  1:1108286:1133815       rs10907175
51150   1:1151788:1168447       rs2887286       rs6603781       rs11260562
29101   1:1476553:1511262       rs3766180       rs6603791       rs7540231       rs7519837
65220   1:1682171:1712508       rs2281173       rs1107910
2782    1:1716225:1823552       rs2272908       rs3737628       rs9786963       rs10907187
```

Number of genes in gene location file

```
### 1.1
wc -l /data/magma/files/NCBI37.3.gene.loc
> 19427
```

Number of genes in genes.annot file

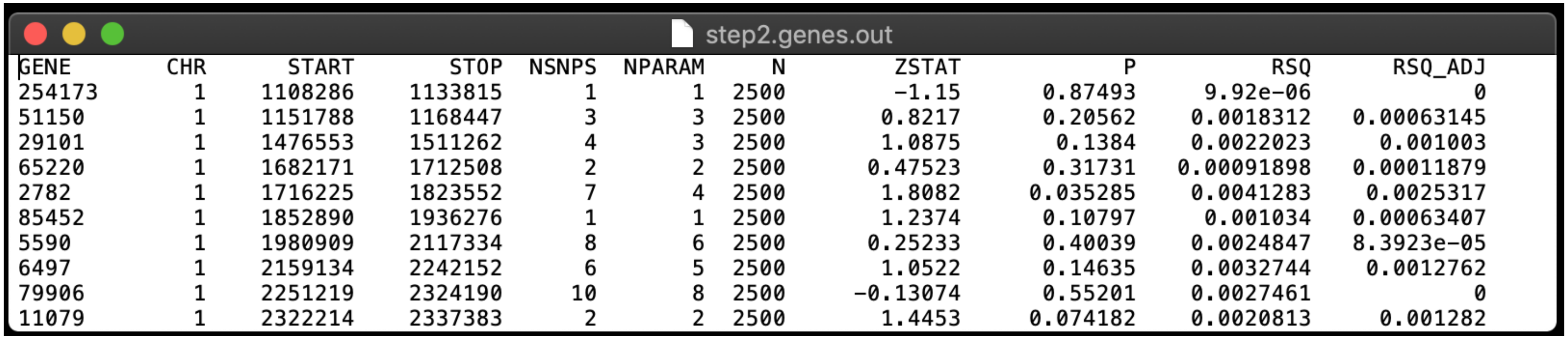
```
### 1.2
grep -v ^# output/step1.genes.annot | wc -l
> 13772
```

5,655 genes were not in genes.annot file because there were not SNPs assigned to those genes within 1kb and 0.5kb window.

Step 2: Gene analysis

Output files

- `step2.genes.out`
- `step2.genes.raw`
- `step2.log`



```
GENE          CHR       START       STOP  NSNPS  NPARAM     N         ZSTAT           P          RSQ      RSQ_ADJ
254173          1     1108286    1133815      1       1  2500         -1.15     0.87493     9.92e-06            0
51150           1     1151788    1168447      3       3  2500        0.8217     0.20562    0.0018312   0.00063145
29101           1     1476553    1511262      4       3  2500        1.0875      0.1384    0.0022023     0.001003
65220           1     1682171    1712508      2       2  2500       0.47523     0.31731   0.00091898   0.00011879
2782            1     1716225    1823552      7       4  2500        1.8082    0.035285    0.0041283    0.0025317
85452           1     1852890    1936276      1       1  2500        1.2374     0.10797     0.001034   0.00063407
5590            1     1980909    2117334      8       6  2500       0.25233     0.40039    0.0024847   8.3923e-05
6497            1     2159134    2242152      6       5  2500        1.0522     0.14635    0.0032744    0.0012762
79906           1     2251219    2324190     10       8  2500      -0.13074     0.55201    0.0027461            0
11079           1     2322214    2337383      2       2  2500        1.4453    0.074182    0.0020813     0.001282
```

Number of significant genes after Bonferroni correction

```
### 2.1
awk '($9<0.05/13772)' output/step2.genes.out | wc -l
> 2
```

Number of genes with P<0.05

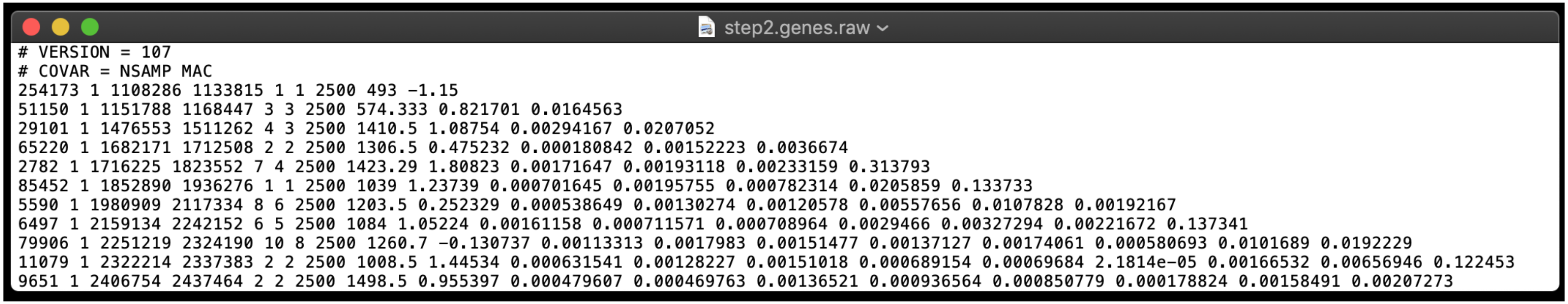
```
### 2.2
awk '($9<0.05)' output/step2.genes.out | wc -l
> 857
```

This is 6.2% of all the tested genes. We would expect 5% if there is no genetic signal.

Step 2: Gene analysis

Output files

- `step2.genes.out`
- `step2.genes.raw`
- `step2.log`



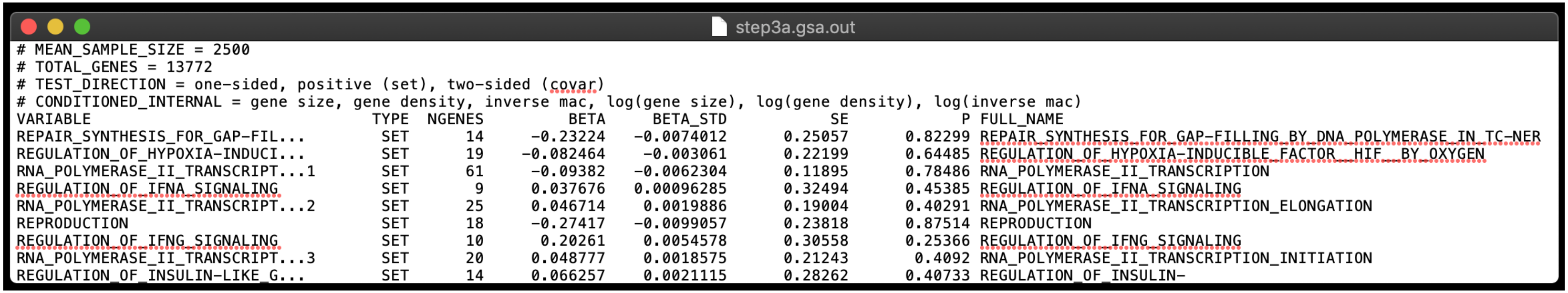
```
# VERSION = 107
# COVAR = NSAMP MAC
254173 1 1108286 1133815 1 1 2500 493 -1.15
51150 1 1151788 1168447 3 3 2500 574.333 0.821701 0.0164563
29101 1 1476553 1511262 4 3 2500 1410.5 1.08754 0.00294167 0.0207052
65220 1 1682171 1712508 2 2 2500 1306.5 0.475232 0.000180842 0.00152223 0.0036674
2782 1 1716225 1823552 7 4 2500 1423.29 1.80823 0.00171647 0.00193118 0.00233159 0.313793
85452 1 1852890 1936276 1 1 2500 1039 1.23739 0.000701645 0.00195755 0.000782314 0.0205859 0.133733
5590 1 1980909 2117334 8 6 2500 1203.5 0.252329 0.000538649 0.00130274 0.00120578 0.00557656 0.0107828 0.00192167
6497 1 2159134 2242152 6 5 2500 1084 1.05224 0.00161158 0.000711571 0.000708964 0.0029466 0.00327294 0.00221672 0.137341
79906 1 2251219 2324190 10 8 2500 1260.7 -0.130737 0.00113313 0.0017983 0.00151477 0.00137127 0.00174061 0.000580693 0.0101689 0.0192229
11079 1 2322214 2337383 2 2 2500 1008.5 1.44534 0.000631541 0.00128227 0.00151018 0.000689154 0.00069684 2.1814e-05 0.00166532 0.00656946 0.122453
9651 1 2406754 2437464 2 2 2500 1498.5 0.955397 0.000479607 0.000469763 0.00136521 0.000936564 0.000850779 0.000178824 0.00158491 0.00207273
```

`genes.raw` file contains gene Z-score and pair-wise correlation (LD) between tested genes. This file is input of gene-set analysis but in practice, you don’t need to read this file to obtain results of gene based analysis.

Step 3a: Basic competitive gene set analysis

Output files

- `step3a.gsa.out`
- `step3a.gsa.genes.out`
- `step3a.gsa.sets.genes.out`
- `step3a.log`



```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13772
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene density), log(inverse mac)
VARIABLE                                TYPE   NGENES          BETA      BETA_STD            SE             P FULL_NAME
REPAIR_SYNTHESIS_FOR_GAP-FIL...          SET       14      -0.23224    -0.0074012       0.25057       0.82299 REPAIR_SYNTHESIS_FOR_GAP-FILLING_BY_DNA_POLYMERASE_IN_TC-NER
REGULATION_OF_HYPOXIA-INDUCI...          SET       19     -0.082464     -0.003061       0.22199       0.64485 REGULATION_OF_HYPOXIA-INDUCIBLE_FACTOR__HIF__BY_OXYGEN
RNA_POLYMERASE_II_TRANSCRIPT...1         SET       61      -0.09382    -0.0062304       0.11895       0.78486 RNA_POLYMERASE_II_TRANSCRIPTION
REGULATION_OF_IFNA_SIGNALING             SET        9      0.037676    0.00096285       0.32494       0.45385 REGULATION_OF_IFNA_SIGNALING
RNA_POLYMERASE_II_TRANSCRIPT...2         SET       25      0.046714     0.0019886       0.19004       0.40291 RNA_POLYMERASE_II_TRANSCRIPTION_ELONGATION
REPRODUCTION                             SET       18      -0.27417    -0.0099057       0.23818       0.87514 REPRODUCTION
REGULATION_OF_IFNG_SIGNALING             SET       10       0.20261     0.0054578       0.30558       0.25366 REGULATION_OF_IFNG_SIGNALING
RNA_POLYMERASE_II_TRANSCRIPT...3         SET       20      0.048777     0.0018575       0.21243        0.4092 RNA_POLYMERASE_II_TRANSCRIPTION_INITIATION
REGULATION_OF_INSULIN-LIKE_G...          SET       14      0.066257     0.0021115       0.28262       0.40733 REGULATION_OF_INSULIN-
```

Number of significant gene sets after Bonferroni correction

```
### 3.1
grep -v ^# output/step3a.gsa.out | awk '($7<0.05/1013)’ | wc -l
> 10
```

Check significant gene sets

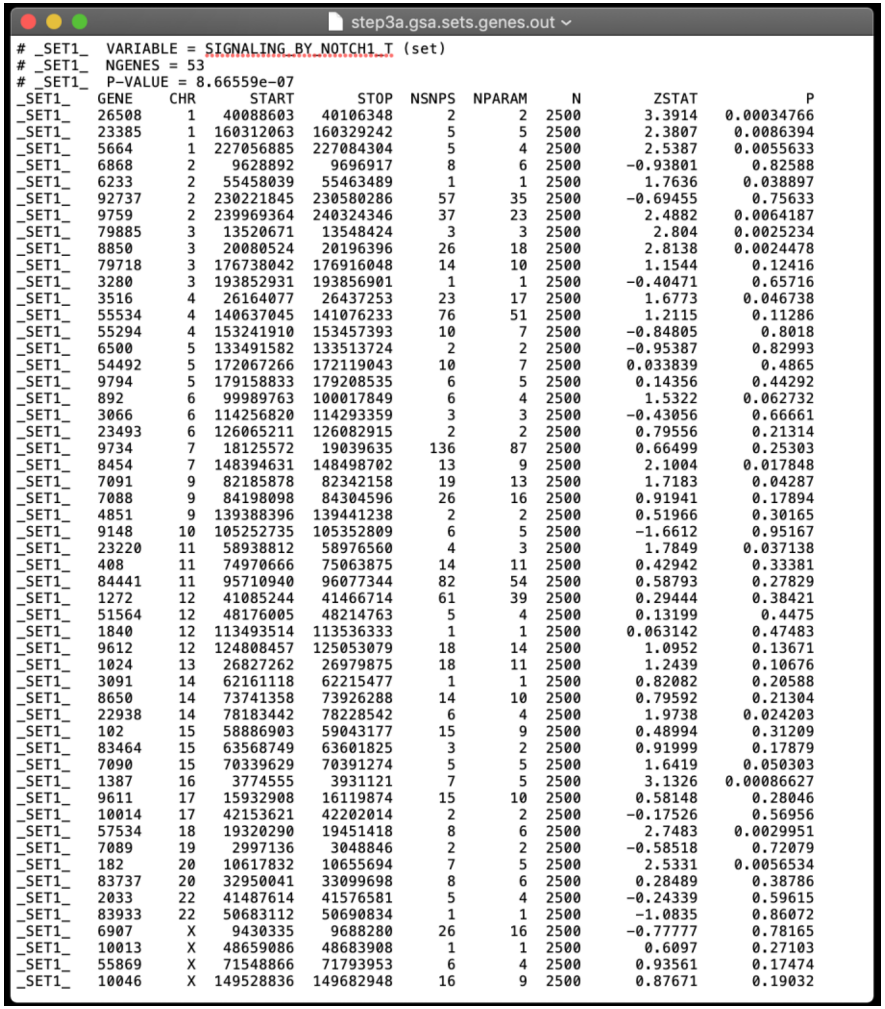
```
grep -v ^# output/step3a.gsa.out | awk '(NR==1 || $7<0.05/1013)'
| sort -k 7g
```

Significant gene set in the competitive analysis means that mean association of genes in the gene set is higher than mean association of genes outside of the gene set.

Step 3a: Basic competitive gene set analysis

Output files

- `step3a.gsa.out`
- `step3a.gsa.genes.out`
- `step3a.gsa.sets.genes.out`
- `step3a.log`



```
# _SET1_  VARIABLE = SIGNALING_BY_NOTCH1_T (set)
# _SET1_  NGENES = 53
# _SET1_  P-VALUE = 8.66559e-07
_SET1_   GENE    CHR      START       STOP  NSNPS  NPARAM     N        ZSTAT           P
_SET1_   26508     1   40088603   40106348      2       2  2500       3.3914  0.00034766
_SET1_   23385     1  160312063  160329242      5       5  2500       2.3807   0.0086394
_SET1_   5664      1  227056885  227084304      5       4  2500       2.5387   0.0055633
_SET1_   6868      2    9628892    9696917      8       6  2500     -0.93801     0.82588
_SET1_   6233      2   55458039   55463489      1       1  2500       1.7636    0.038897
_SET1_   92737     2  230221845  230580286     57      35  2500     -0.69455     0.75633
_SET1_   9759      2  239969364  240324346     37      23  2500       2.4882   0.0064187
_SET1_   79885     3   13520671   13548424      3       3  2500        2.804   0.0025234
_SET1_   8850      3   20080524   20196396     26      18  2500       2.8138   0.0024478
_SET1_   79718     3  176738042  176916048     14      10  2500       1.1544     0.12416
_SET1_   3280      3  193852931  193856901      1       1  2500     -0.40471     0.65716
_SET1_   3516      4   26164077   26437253     23      17  2500       1.6773    0.046738
_SET1_   55534     4  140637045  141076233     76      51  2500       1.2115     0.11286
_SET1_   55294     4  153241910  153457393     10       7  2500     -0.84805      0.8018
_SET1_   6500      5  133491582  133513724      2       2  2500     -0.95387     0.82993
_SET1_   54492     5  172067266  172119043     10       7  2500     0.033839      0.4865
_SET1_   9794      5  179158833  179208535      6       5  2500      0.14356     0.44292
_SET1_   892       6   99989763  100017849      6       4  2500       1.5322    0.062732
_SET1_   3066      6  114256820  114293359      3       3  2500     -0.43056     0.66661
_SET1_   23493     6  126065211  126082915      2       2  2500      0.79556     0.21314
_SET1_   9734      7   18125572   19039635    136      87  2500      0.66499     0.25303
_SET1_   8454      7  148394631  148498702     13       9  2500       2.1004    0.017848
_SET1_   7091      9   82185878   82342158     19      13  2500       1.7183     0.04287
_SET1_   7088      9   84198098   84304596     26      16  2500      0.91941     0.17894
_SET1_   4851      9  139388396  139441238      2       2  2500      0.51966     0.30165
_SET1_   9148     10  105252735  105352809      6       5  2500      -1.6612     0.95167
_SET1_   23220    11   58938812   58976560      4       3  2500       1.7849    0.037138
_SET1_   408      11   74970666   75063875     14      11  2500      0.42942     0.33381
_SET1_   84441    11   95710940   96077344     82      54  2500      0.58793     0.27829
_SET1_   1272     12   41085244   41466714     61      39  2500      0.29444     0.38421
_SET1_   51564    12   48176005   48214763      5       4  2500      0.13199      0.4475
_SET1_   1840     12  113493514  113536333      1       1  2500     0.063142     0.47483
_SET1_   9612     12  124808457  125053079     18      14  2500       1.0952     0.13671
_SET1_   1024     13   26827262   26979875     18      11  2500       1.2439     0.10676
_SET1_   3091     14   62161118   62215477      1       1  2500      0.82082     0.20588
_SET1_   8650     14   73741358   73926288     14      10  2500      0.79592     0.21304
_SET1_   22938    14   78183442   78228542      6       4  2500       1.9738    0.024203
_SET1_   102      15   58886903   59043177     15       9  2500      0.48994     0.31209
_SET1_   83464    15   63568749   63601825      3       2  2500      0.91999     0.17879
_SET1_   7090     15   70339629   70391274      5       5  2500       1.6419    0.050303
_SET1_   1387     16    3774555    3931121      7       5  2500       3.1326  0.00086627
_SET1_   9611     17   15932908   16119874     15      10  2500      0.58148     0.28046
_SET1_   10014    17   42153621   42202014      2       2  2500     -0.17526     0.56956
_SET1_   57534    18   19320290   19451418      8       6  2500       2.7483   0.0029951
_SET1_   7089     19    2997136    3048846      2       2  2500     -0.58518     0.72079
_SET1_   182      20   10617832   10655694      7       5  2500       2.5331   0.0056534
_SET1_   83737    20   32950041   33099698      8       6  2500      0.28489     0.38786
_SET1_   2033     22   41487614   41576581      5       4  2500     -0.24339     0.59615
_SET1_   83933    22   50683112   50690834      1       1  2500      -1.0835     0.86072
_SET1_   6907      X    9430335    9688280     26      16  2500     -0.77777     0.78165
_SET1_   10013     X   48659086   48683908      1       1  2500       0.6097     0.27103
_SET1_   55869     X   71548866   71793953      6       4  2500      0.93561     0.17474
_SET1_   10046     X  149528836  149682948     16       9  2500      0.87671     0.19032
```

Number of significant genes in the gene set SIGNALING_BY_NOTCH_T

```
### 3.2
grep ^_SET1_
output/step3a.gsa.sets.genes.out
| awk '($10<0.05/13772)' | wc -l
> 0
```

Number of genes with P<0.05 in the gene set SIGNALING_BY_NOTCH_T

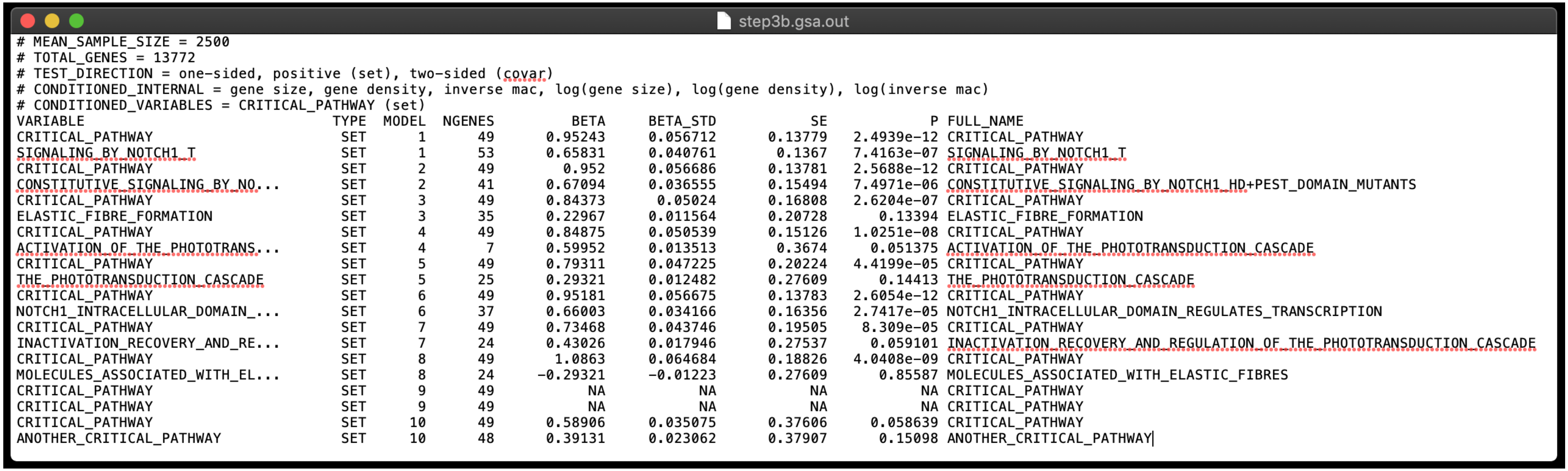
```
grep ^_SET1_
output/step3a.gsa.sets.genes.out
| awk '($10<0.05)' | wc -l
> 15
```

The gene set does not have significant gene but 28.3% (15/53) of genes have P<0.05 which is much higher than 6.2% across the genome.

Step 3b: Basic conditional gene set analysis

Output files

- step3b.gsa.out
- step3b.gsa.genes.out
- step3b.gwa.sets.genes.out
- step3b.log

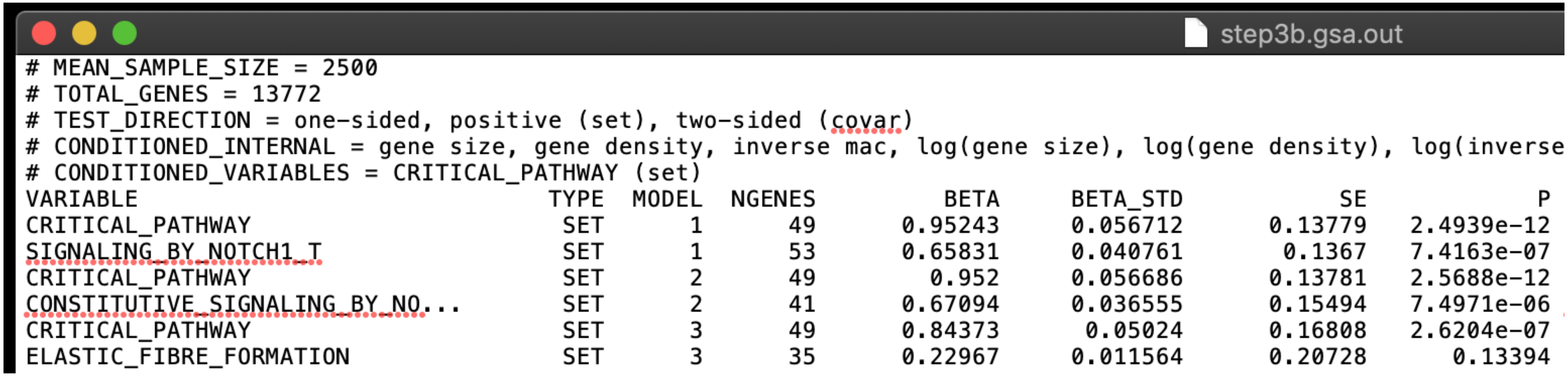


```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13772
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene density), log(inverse mac)
# CONDITIONED_VARIABLES = CRITICAL_PATHWAY (set)
VARIABLE                             TYPE MODEL NGENES         BETA     BETA_STD           SE            P FULL_NAME
CRITICAL_PATHWAY                      SET     1     49      0.95243     0.056712      0.13779   2.4939e-12 CRITICAL_PATHWAY
SIGNALING_BY_NOTCH1_T                 SET     1     53      0.65831     0.040761       0.1367   7.4163e-07 SIGNALING_BY_NOTCH1_T
CRITICAL_PATHWAY                      SET     2     49        0.952     0.056686      0.13781   2.5688e-12 CRITICAL_PATHWAY
CONSTITUTIVE_SIGNALING_BY_NO...       SET     2     41      0.67094     0.036555      0.15494   7.4971e-06 CONSTITUTIVE_SIGNALING_BY_NOTCH1_HD+PEST_DOMAIN_MUTANTS
CRITICAL_PATHWAY                      SET     3     49      0.84373      0.05024      0.16808   2.6204e-07 CRITICAL_PATHWAY
ELASTIC_FIBRE_FORMATION               SET     3     35      0.22967     0.011564      0.20728      0.13394 ELASTIC_FIBRE_FORMATION
CRITICAL_PATHWAY                      SET     4     49      0.84875     0.050539      0.15126   1.0251e-08 CRITICAL_PATHWAY
ACTIVATION_OF_THE_PHOTOTRANS...       SET     4      7      0.59952     0.013513       0.3674     0.051375 ACTIVATION_OF_THE_PHOTOTRANSDUCTION_CASCADE
CRITICAL_PATHWAY                      SET     5     49      0.79311     0.047225      0.20224   4.4199e-05 CRITICAL_PATHWAY
THE_PHOTOTRANSDUCTION_CASCADE         SET     5     25      0.29321     0.012482      0.27609      0.14413 THE_PHOTOTRANSDUCTION_CASCADE
CRITICAL_PATHWAY                      SET     6     49      0.95181     0.056675      0.13783   2.6054e-12 CRITICAL_PATHWAY
NOTCH1_INTRACELLULAR_DOMAIN_...       SET     6     37      0.66003     0.034166      0.16356   2.7417e-05 NOTCH1_INTRACELLULAR_DOMAIN_REGULATES_TRANSCRIPTION
CRITICAL_PATHWAY                      SET     7     49      0.73468     0.043746      0.19505    8.309e-05 CRITICAL_PATHWAY
INACTIVATION_RECOVERY_AND_RE...       SET     7     24      0.43026     0.017946      0.27537     0.059101 INACTIVATION_RECOVERY_AND_REGULATION_OF_THE_PHOTOTRANSDUCTION_CASCADE
CRITICAL_PATHWAY                      SET     8     49       1.0863     0.064684      0.18826   4.0408e-09 CRITICAL_PATHWAY
MOLECULES_ASSOCIATED_WITH_EL...       SET     8     24     -0.29321     -0.01223      0.27609      0.85587 MOLECULES_ASSOCIATED_WITH_ELASTIC_FIBRES
CRITICAL_PATHWAY                      SET     9     49           NA           NA           NA           NA CRITICAL_PATHWAY
CRITICAL_PATHWAY                      SET     9     49           NA           NA           NA           NA CRITICAL_PATHWAY
CRITICAL_PATHWAY                      SET    10     49      0.58906     0.035075      0.37606     0.058639 CRITICAL_PATHWAY
ANOTHER_CRITICAL_PATHWAY              SET    10     48      0.39131     0.023062      0.37907      0.15098 ANOTHER_CRITICAL_PATHWAY
```

Step 3b: Basic conditional gene set analysis

Output files

- `step3b.gsa.out`
- `step3b.gsa.genes.out`
- `step3b.gwa.sets.genes.out`
- `step3b.log`



```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13772
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene density), log(inverse
# CONDITIONED_VARIABLES = CRITICAL_PATHWAY (set)
VARIABLE                             TYPE  MODEL  NGENES         BETA     BETA_STD           SE            P
CRITICAL_PATHWAY                      SET      1      49      0.95243     0.056712      0.13779   2.4939e-12
SIGNALING_BY_NOTCH1_T                 SET      1      53      0.65831     0.040761       0.1367   7.4163e-07
CRITICAL_PATHWAY                      SET      2      49        0.952     0.056686      0.13781   2.5688e-12
CONSTITUTIVE_SIGNALING_BY_NO...       SET      2      41      0.67094     0.036555      0.15494   7.4971e-06
CRITICAL_PATHWAY                      SET      3      49      0.84373      0.05024      0.16808   2.6204e-07
ELASTIC_FIBRE_FORMATION               SET      3      35      0.22967     0.011564      0.20728      0.13394
```

Gene sets that are no longer significant by conditioning CRITICAL_PATHWAY

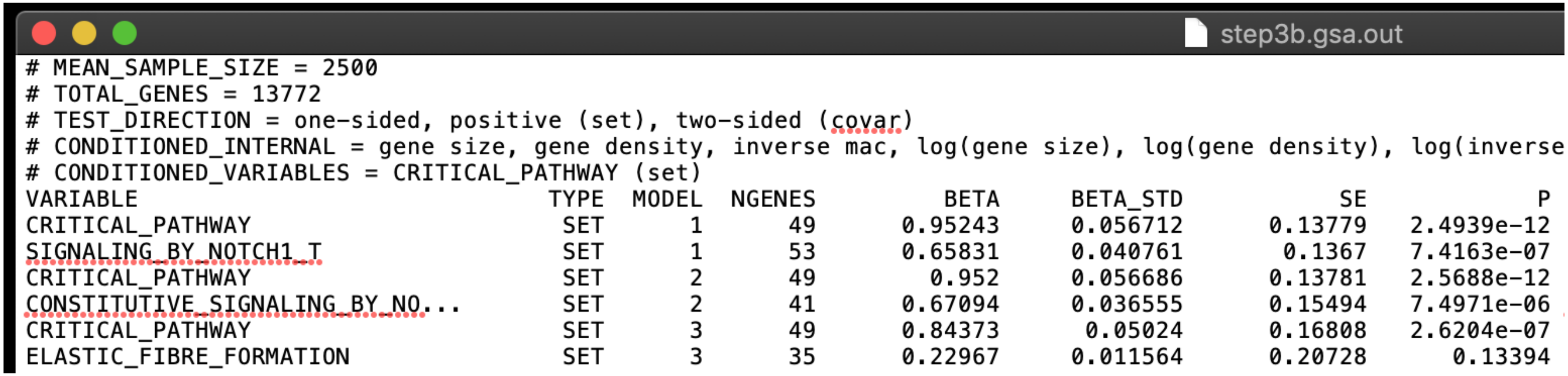
```
### 3.4
grep -v ^# output/step3b.gsa.out | grep -v ^CRITICAL_PATHWAY | awk
'(NR==1 || $8>=0.05/1013)’
```

Association of those gene sets are confounding of the CRITICAL_PATHWAY.

Step 3b: Basic conditional gene set analysis

Output files

- `step3b.gsa.out`
- `step3b.gsa.genes.out`
- `step3b.gwa.sets.genes.out`
- `step3b.log`



```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13772
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene density), log(inverse
# CONDITIONED_VARIABLES = CRITICAL_PATHWAY (set)
VARIABLE                              TYPE  MODEL  NGENES         BETA     BETA_STD           SE           P
CRITICAL_PATHWAY                       SET      1      49      0.95243     0.056712      0.13779  2.4939e-12
SIGNALING_BY_NOTCH1_T                  SET      1      53      0.65831     0.040761       0.1367  7.4163e-07
CRITICAL_PATHWAY                       SET      2      49        0.952     0.056686      0.13781  2.5688e-12
CONSTITUTIVE_SIGNALING_BY_NO...        SET      2      41      0.67094     0.036555      0.15494  7.4971e-06
CRITICAL_PATHWAY                       SET      3      49      0.84373      0.05024      0.16808  2.6204e-07
ELASTIC_FIBRE_FORMATION                SET      3      35      0.22967     0.011564      0.20728     0.13394
```

Find models where CRITICAL_PATHWAY is no longer significant

```
### 3.5
grep -v ^# output/step3b.gsa.out | grep ^CRITICAL_PATHWAY | awk
'($8>=0.05)’
```

Extract results of these models

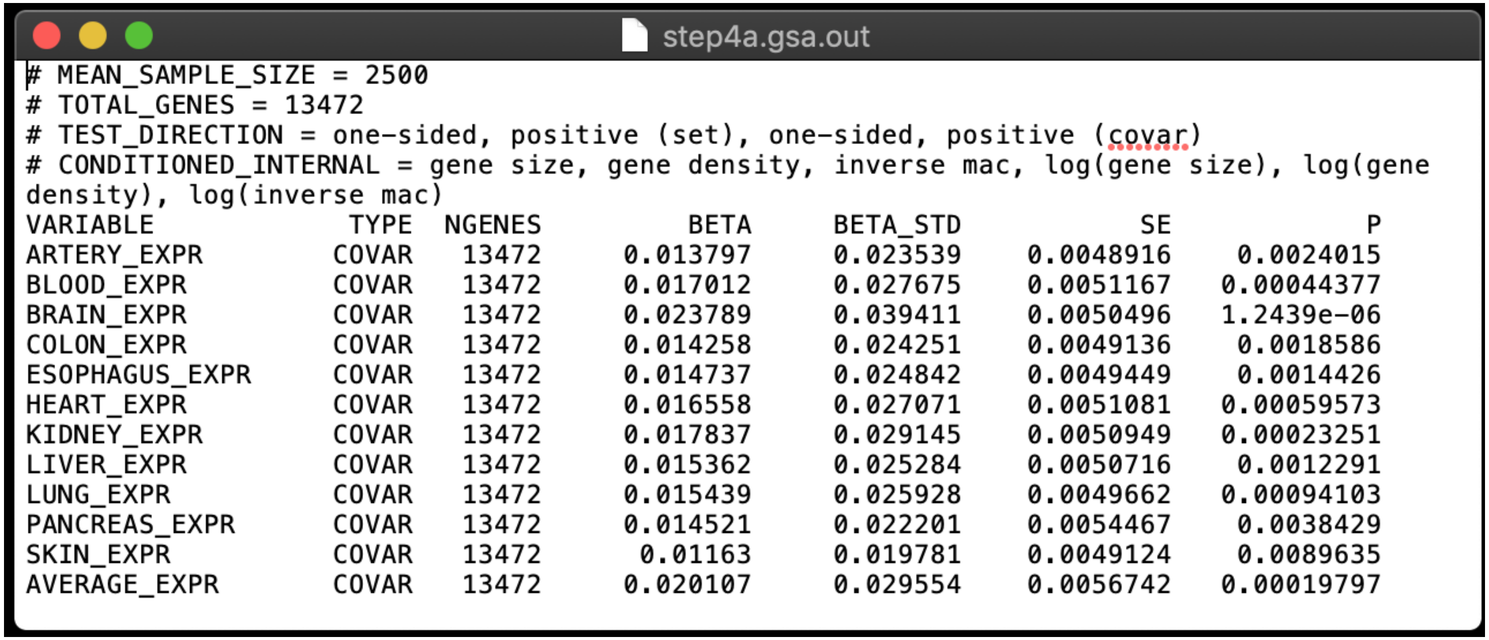
```
### 3.6
grep -v ^# output/step3b.gsa.out | awk '(NR==1 || $3==10)’
```

Association of both CRITICAL_PATHWAY and ANOTHR_CRITICAL_PATHWAY completely disappear, mainly due to large overlap of genes. Both pathways are contributing into the same signal but the model cannot distinguish which is the true signal.

Step 4a: Tissue expression analysis

Output files

- `step4a.gsa.out`
- `step4a.log`



```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13472
# TEST_DIRECTION = one-sided, positive (set), one-sided, positive (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene
density), log(inverse mac)
VARIABLE             TYPE  NGENES        BETA    BETA_STD         SE          P
ARTERY_EXPR         COVAR   13472    0.013797    0.023539  0.0048916  0.0024015
BLOOD_EXPR          COVAR   13472    0.017012    0.027675  0.0051167 0.00044377
BRAIN_EXPR          COVAR   13472    0.023789    0.039411  0.0050496 1.2439e-06
COLON_EXPR          COVAR   13472    0.014258    0.024251  0.0049136  0.0018586
ESOPHAGUS_EXPR      COVAR   13472    0.014737    0.024842  0.0049449  0.0014426
HEART_EXPR          COVAR   13472    0.016558    0.027071  0.0051081 0.00059573
KIDNEY_EXPR         COVAR   13472    0.017837    0.029145  0.0050949 0.00023251
LIVER_EXPR          COVAR   13472    0.015362    0.025284  0.0050716  0.0012291
LUNG_EXPR           COVAR   13472    0.015439    0.025928  0.0049662 0.00094103
PANCREAS_EXPR       COVAR   13472    0.014521    0.022201  0.0054467  0.0038429
SKIN_EXPR           COVAR   13472     0.01163    0.019781  0.0049124  0.0089635
AVERAGE_EXPR        COVAR   13472    0.020107    0.029554  0.0056742 0.00019797
```

Number of tissues significant after Bonferroni correction

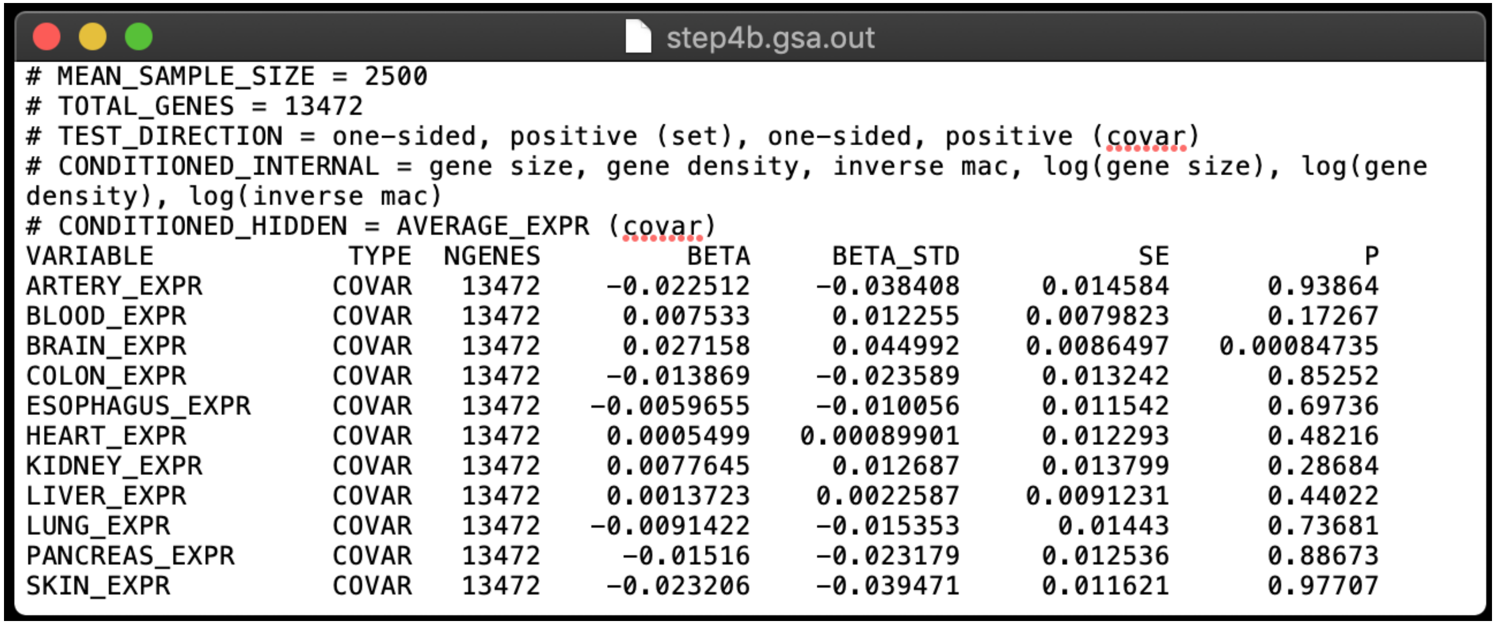
```
### 4.1
grep -v ^# output/step4a.gsa.out | awk '($7<0.05/12)' | wc -l
> 11
```

11 out of 12 tissues showed significant (positive) association including AVERAGE_EXPR. This represent the trait is associated with general gene expression level but does not tell tissue specificity.

Step 4b: Conditional tissue expression analysis

Output files

- `step4b.gsa.out`
- `step4b.log`



```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13472
# TEST_DIRECTION = one-sided, positive (set), one-sided, positive (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene
density), log(inverse mac)
# CONDITIONED_HIDDEN = AVERAGE_EXPR (covar)
VARIABLE            TYPE  NGENES        BETA     BETA_STD          SE            P
ARTERY_EXPR        COVAR   13472   -0.022512    -0.038408    0.014584      0.93864
BLOOD_EXPR         COVAR   13472    0.007533     0.012255   0.0079823      0.17267
BRAIN_EXPR         COVAR   13472    0.027158     0.044992   0.0086497   0.00084735
COLON_EXPR         COVAR   13472   -0.013869    -0.023589    0.013242      0.85252
ESOPHAGUS_EXPR     COVAR   13472  -0.0059655    -0.010056    0.011542      0.69736
HEART_EXPR         COVAR   13472   0.0005499   0.00089901    0.012293      0.48216
KIDNEY_EXPR        COVAR   13472   0.0077645     0.012687    0.013799      0.28684
LIVER_EXPR         COVAR   13472   0.0013723    0.0022587   0.0091231      0.44022
LUNG_EXPR          COVAR   13472  -0.0091422    -0.015353     0.01443      0.73681
PANCREAS_EXPR      COVAR   13472    -0.01516    -0.023179    0.012536      0.88673
SKIN_EXPR          COVAR   13472   -0.023206    -0.039471    0.011621      0.97707
```

Tissue still significant after conditioning on average expression

```
### 4.3
grep -v ^# output/step4b.gsa.out | awk '($7<0.05/11)' | wc -l
> 1
```

Check the tissue type

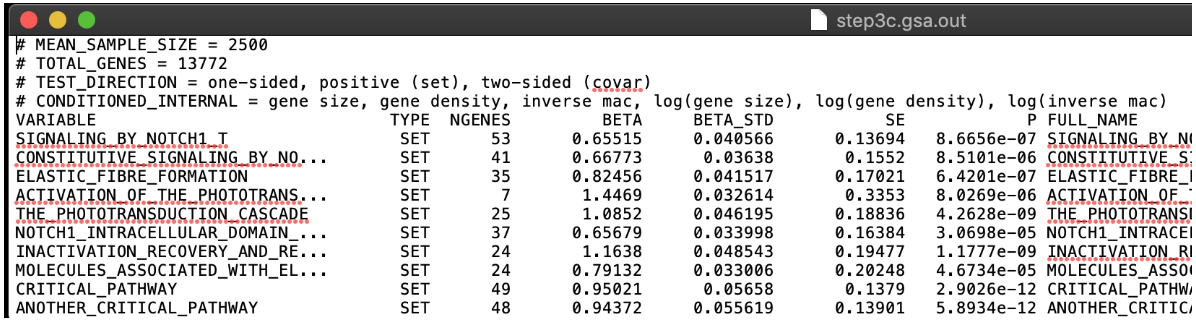
```
### 4.4
grep -v ^# output/step4b.gsa.out | awk ‘(NR==1 || $7<0.05/11)’
```

Only brain remain significant, association of other tissues completely disappeared. This result suggest association of the trait with brain specific gene expression.

Step 5: Tissue expression analysis

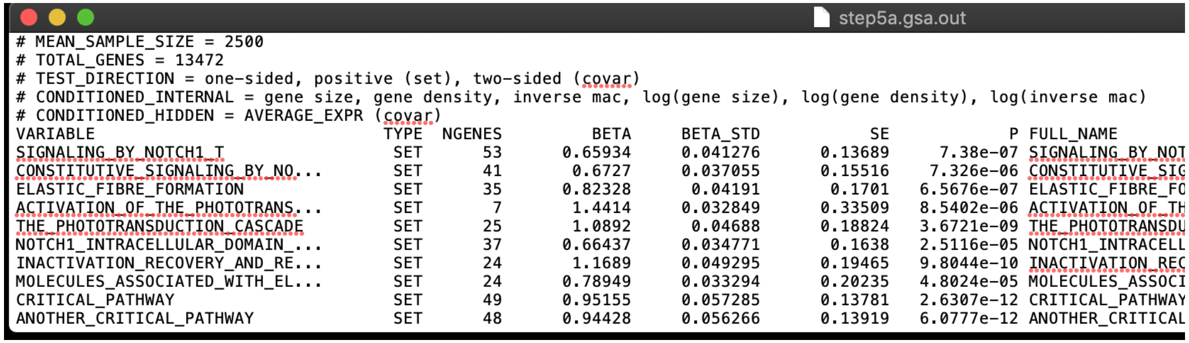
Output files

- `step5a.gsa.out`
- `step5a.gsa.genes.out`
- `step5a.gsa.sets.genes.out`
- `step5a.log`



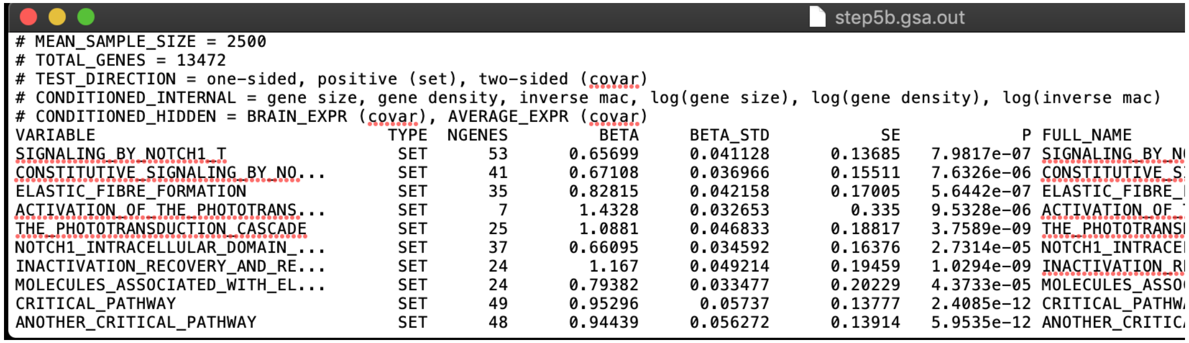
```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13772
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene density), log(inverse mac)
VARIABLE                              TYPE  NGENES        BETA    BETA_STD          SE           P FULL_NAME
SIGNALING_BY_NOTCH1_T                  SET      53     0.65515    0.040566     0.13694  8.6656e-07 SIGNALING_BY_N
CONSTITUTIVE_SIGNALING_BY_NO...        SET      41     0.66773     0.03638      0.1552  8.5101e-06 CONSTITUTIVE_S
ELASTIC_FIBRE_FORMATION                SET      35     0.82456    0.041517     0.17021  6.4201e-07 ELASTIC_FIBRE_
ACTIVATION_OF_THE_PHOTOTRANS...        SET       7      1.4469    0.032614      0.3353  8.0269e-06 ACTIVATION_OF_
THE_PHOTOTRANSDUCTION_CASCADE          SET      25      1.0852    0.046195     0.18836  4.2628e-09 THE_PHOTOTRANS
NOTCH1_INTRACELLULAR_DOMAIN_...        SET      37     0.65679    0.033998     0.16384  3.0698e-05 NOTCH1_INTRACE
INACTIVATION_RECOVERY_AND_RE...        SET      24      1.1638    0.048543     0.19477  1.1777e-09 INACTIVATION_R
MOLECULES_ASSOCIATED_WITH_EL...        SET      24     0.79132    0.033006     0.20248  4.6734e-05 MOLECULES_ASSO
CRITICAL_PATHWAY                       SET      49     0.95021     0.05658      0.1379  2.9026e-12 CRITICAL_PATHW
ANOTHER_CRITICAL_PATHWAY               SET      48     0.94372    0.055619     0.13901  5.8934e-12 ANOTHER_CRITIC
```

Step 3c
No covariate



```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13472
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene density), log(inverse mac)
# CONDITIONED_HIDDEN = AVERAGE_EXPR (covar)
VARIABLE                             TYPE  NGENES        BETA    BETA_STD          SE           P FULL_NAME
SIGNALING_BY_NOTCH1_T                 SET      53     0.65934    0.041276     0.13689    7.38e-07 SIGNALING_BY_NOT
CONSTITUTIVE_SIGNALING_BY_NO...       SET      41      0.6727    0.037055     0.15516   7.326e-06 CONSTITUTIVE_SIG
ELASTIC_FIBRE_FORMATION               SET      35     0.82328     0.04191      0.1701  6.5676e-07 ELASTIC_FIBRE_FO
ACTIVATION_OF_THE_PHOTOTRANS...       SET       7      1.4414    0.032849     0.33509  8.5402e-06 ACTIVATION_OF_TH
THE_PHOTOTRANSDUCTION_CASCADE         SET      25      1.0892     0.04688     0.18824  3.6721e-09 THE_PHOTOTRANSDU
NOTCH1_INTRACELLULAR_DOMAIN_...       SET      37     0.66437    0.034771      0.1638  2.5116e-05 NOTCH1_INTRACELL
INACTIVATION_RECOVERY_AND_RE...       SET      24      1.1689    0.049295     0.19465  9.8044e-10 INACTIVATION_REC
MOLECULES_ASSOCIATED_WITH_EL...       SET      24     0.78949    0.033294     0.20235  4.8024e-05 MOLECULES_ASSOCI
CRITICAL_PATHWAY                      SET      49     0.95155    0.057285     0.13781  2.6307e-12 CRITICAL_PATHWAY
ANOTHER_CRITICAL_PATHWAY              SET      48     0.94428    0.056266     0.13919  6.0777e-12 ANOTHER_CRITICAL
```

Step 5a
Conditioning average expression



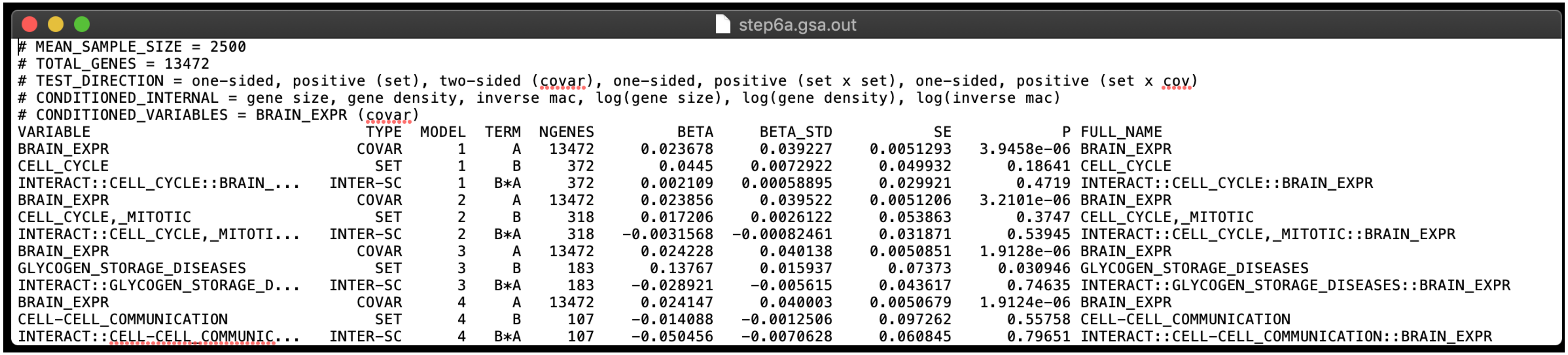
```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13472
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene density), log(inverse mac)
# CONDITIONED_HIDDEN = BRAIN_EXPR (covar), AVERAGE_EXPR (covar)
VARIABLE                              TYPE  NGENES        BETA    BETA_STD          SE           P FULL_NAME
SIGNALING_BY_NOTCH1_T                  SET      53     0.65699    0.041128     0.13685  7.9817e-07 SIGNALING_BY_N
CONSTITUTIVE_SIGNALING_BY_NO...        SET      41     0.67108    0.036966     0.15511  7.6326e-06 CONSTITUTIVE_S
ELASTIC_FIBRE_FORMATION                SET      35     0.82815    0.042158     0.17005  5.6442e-07 ELASTIC_FIBRE_
ACTIVATION_OF_THE_PHOTOTRANS...        SET       7      1.4328    0.032653       0.335  9.5328e-06 ACTIVATION_OF_
THE_PHOTOTRANSDUCTION_CASCADE          SET      25      1.0881    0.046833     0.18817  3.7589e-09 THE_PHOTOTRANS
NOTCH1_INTRACELLULAR_DOMAIN_...        SET      37     0.66095    0.034592     0.16376  2.7314e-05 NOTCH1_INTRACE
INACTIVATION_RECOVERY_AND_RE...        SET      24       1.167    0.049214     0.19459  1.0294e-09 INACTIVATION_R
MOLECULES_ASSOCIATED_WITH_EL...        SET      24     0.79382    0.033477     0.20229  4.3733e-05 MOLECULES_ASSO
CRITICAL_PATHWAY                       SET      49     0.95296     0.05737     0.13777  2.4085e-12 CRITICAL_PATHW
ANOTHER_CRITICAL_PATHWAY               SET      48     0.94439    0.056272     0.13914  5.9535e-12 ANOTHER_CRITIC
```

Step 5b
Conditioning average and brain expression

Step 6a: Interaction analysis

Output files

- `step6a.gsa.out`
- `step6a.gsa.genes.out`
- `step6a.gsa.inter.genes.out`
- `step6a.log`

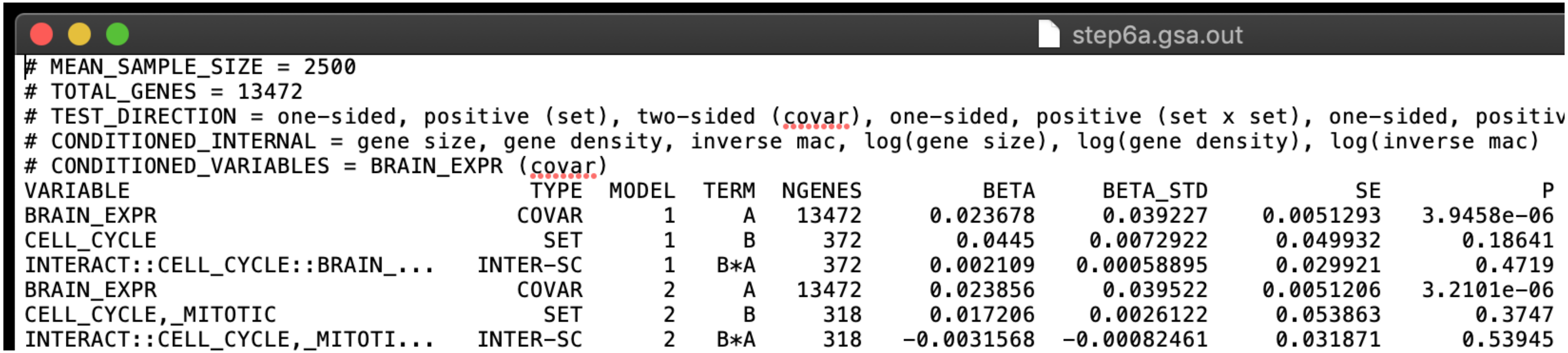


```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13472
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar), one-sided, positive (set x set), one-sided, positive (set x cov)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene density), log(inverse mac)
# CONDITIONED_VARIABLES = BRAIN_EXPR (covar)
VARIABLE                              TYPE  MODEL  TERM  NGENES         BETA      BETA_STD           SE            P FULL_NAME
BRAIN_EXPR                           COVAR      1     A   13472     0.023678      0.039227    0.0051293   3.9458e-06 BRAIN_EXPR
CELL_CYCLE                             SET      1     B     372       0.0445     0.0072922     0.049932      0.18641 CELL_CYCLE
INTERACT::CELL_CYCLE::BRAIN_...   INTER-SC      1   B*A     372     0.002109    0.00058895     0.029921       0.4719 INTERACT::CELL_CYCLE::BRAIN_EXPR
BRAIN_EXPR                           COVAR      2     A   13472     0.023856      0.039522    0.0051206   3.2101e-06 BRAIN_EXPR
CELL_CYCLE,_MITOTIC                    SET      2     B     318     0.017206     0.0026122     0.053863       0.3747 CELL_CYCLE,_MITOTIC
INTERACT::CELL_CYCLE,_MITOTI...   INTER-SC      2   B*A     318   -0.0031568   -0.00082461     0.031871      0.53945 INTERACT::CELL_CYCLE,_MITOTIC::BRAIN_EXPR
BRAIN_EXPR                           COVAR      3     A   13472     0.024228      0.040138    0.0050851   1.9128e-06 BRAIN_EXPR
GLYCOGEN_STORAGE_DISEASES              SET      3     B     183      0.13767      0.015937      0.07373     0.030946 GLYCOGEN_STORAGE_DISEASES
INTERACT::GLYCOGEN_STORAGE_D...   INTER-SC      3   B*A     183    -0.028921     -0.005615     0.043617      0.74635 INTERACT::GLYCOGEN_STORAGE_DISEASES::BRAIN_EXPR
BRAIN_EXPR                           COVAR      4     A   13472     0.024147      0.040003    0.0050679   1.9124e-06 BRAIN_EXPR
CELL-CELL_COMMUNICATION                SET      4     B     107    -0.014088    -0.0012506     0.097262      0.55758 CELL-CELL_COMMUNICATION
INTERACT::CELL-CELL_COMMUNIC...   INTER-SC      4   B*A     107    -0.050456    -0.0070628     0.060845      0.79651 INTERACT::CELL-CELL_COMMUNICATION::BRAIN_EXPR
```

Step 6a: Interaction analysis

Output files

- `step6a.gsa.out`
- `step6a.gsa.genes.out`
- `step6a.gsa.inter.genes.out`
- `step6a.log`



```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13472
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar), one-sided, positive (set x set), one-sided, positiv
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene density), log(inverse mac)
# CONDITIONED_VARIABLES = BRAIN_EXPR (covar)
VARIABLE                                TYPE  MODEL  TERM  NGENES        BETA      BETA_STD          SE           P
BRAIN_EXPR                             COVAR      1     A   13472    0.023678      0.039227   0.0051293  3.9458e-06
CELL_CYCLE                               SET      1     B     372      0.0445     0.0072922    0.049932     0.18641
INTERACT::CELL_CYCLE::BRAIN_...     INTER-SC      1   B*A     372    0.002109    0.00058895    0.029921      0.4719
BRAIN_EXPR                             COVAR      2     A   13472    0.023856      0.039522   0.0051206  3.2101e-06
CELL_CYCLE,_MITOTIC                      SET      2     B     318    0.017206     0.0026122    0.053863      0.3747
INTERACT::CELL_CYCLE,_MITOTI...     INTER-SC      2   B*A     318  -0.0031568   -0.00082461    0.031871     0.53945
```

Number of significant interaction terms

```
### 6.1
grep -v ^# output/step6a.gsa.out | awk '($2=="INTER-SC" &&
$9<0.05/74)' | wc -l
> 1
```

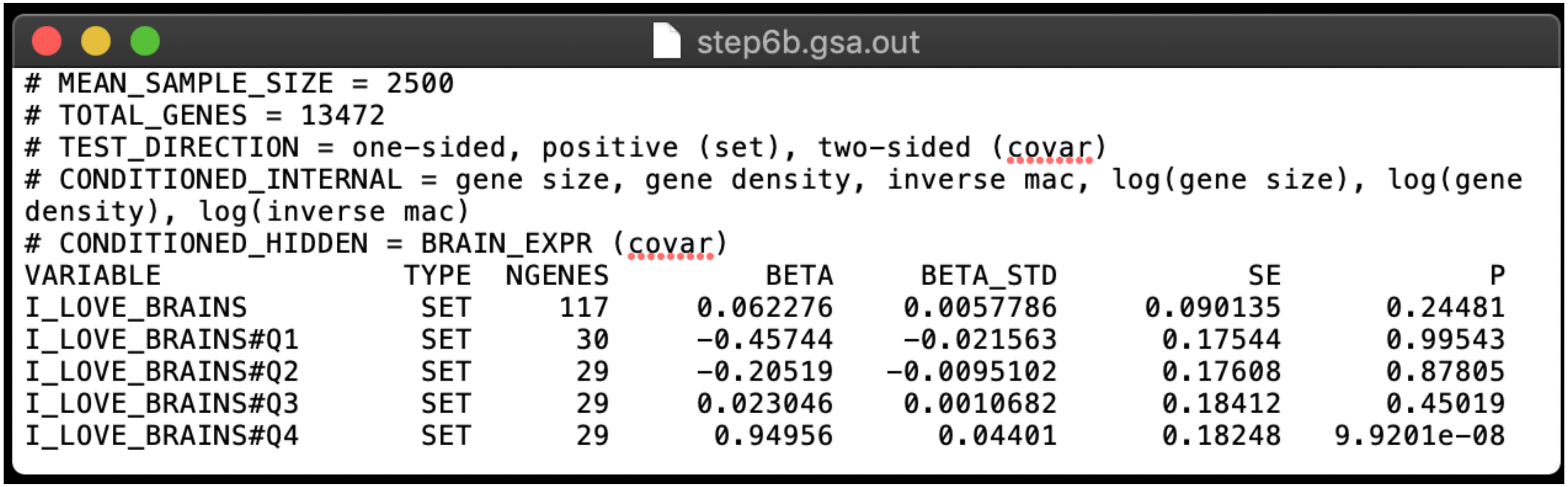
Check significant interaction term

```
grep -v ^# output/step6a.gsa.out | awk '(NR==1 || ($2=="INTER-SC"
&& $9<0.05/74))' | sort -k 9g
```

Step 6b: Interaction analysis (follow up)

Output files

- `step6b.gsa.out`
- `step6b.gsa.genes.out`
- `step6b.gsa.sets.genes.out`
- `step6b.log`



```
# MEAN_SAMPLE_SIZE = 2500
# TOTAL_GENES = 13472
# TEST_DIRECTION = one-sided, positive (set), two-sided (covar)
# CONDITIONED_INTERNAL = gene size, gene density, inverse mac, log(gene size), log(gene
density), log(inverse mac)
# CONDITIONED_HIDDEN = BRAIN_EXPR (covar)
VARIABLE                TYPE  NGENES        BETA      BETA_STD            SE           P
I_LOVE_BRAINS            SET     117    0.062276     0.0057786      0.090135     0.24481
I_LOVE_BRAINS#Q1         SET      30    -0.45744     -0.021563       0.17544     0.99543
I_LOVE_BRAINS#Q2         SET      29    -0.20519    -0.0095102       0.17608     0.87805
I_LOVE_BRAINS#Q3         SET      29    0.023046     0.0010682       0.18412     0.45019
I_LOVE_BRAINS#Q4         SET      29     0.94956       0.04401       0.18248  9.9201e-08
```