Causal and statistical inference

What is missing, what is desired?

Counterfactuals?

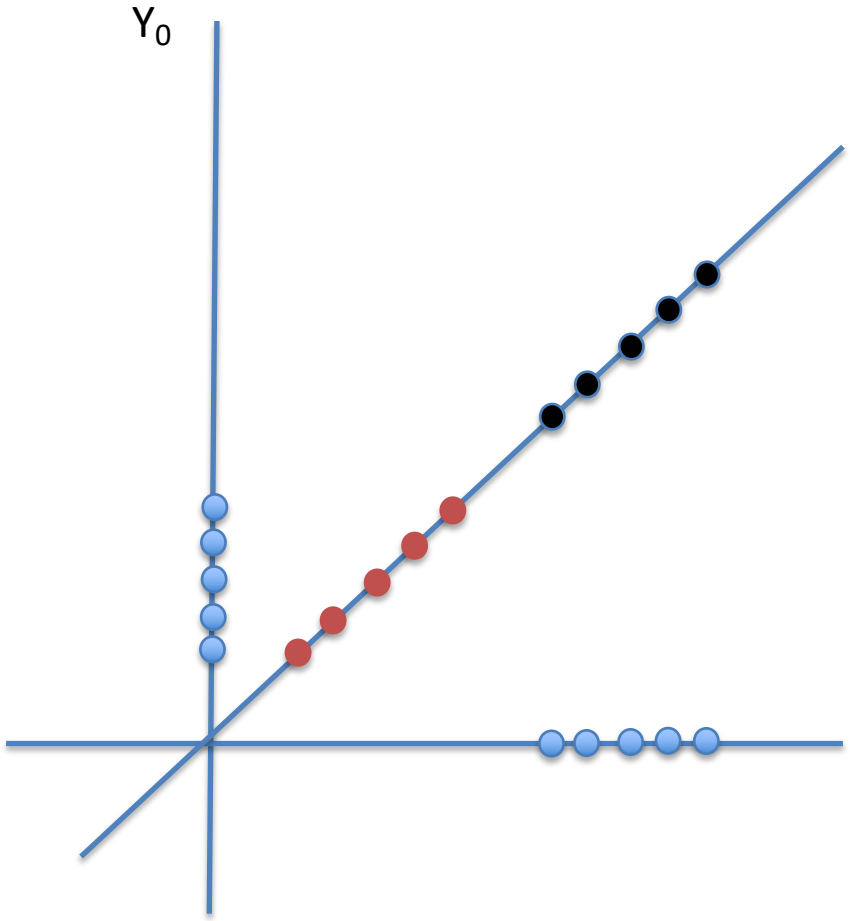
Missingness	No	Yes
No	Complete information	Finite sample causal inference (Fisher)
Yes	Standard statistical inference “to the population”	Sample & assignment selection (Neyman, Rubin)

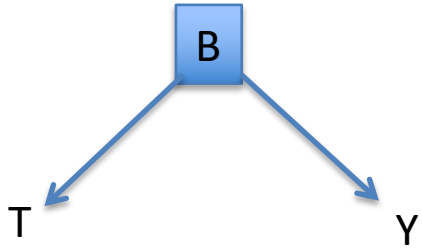
Units?

$$Y_{OBS} = Y_0 + (Y_1 - Y_0)E$$

$\{X : E \perp\!\!\!\perp (Y_0, Y_1) \mid X\}$

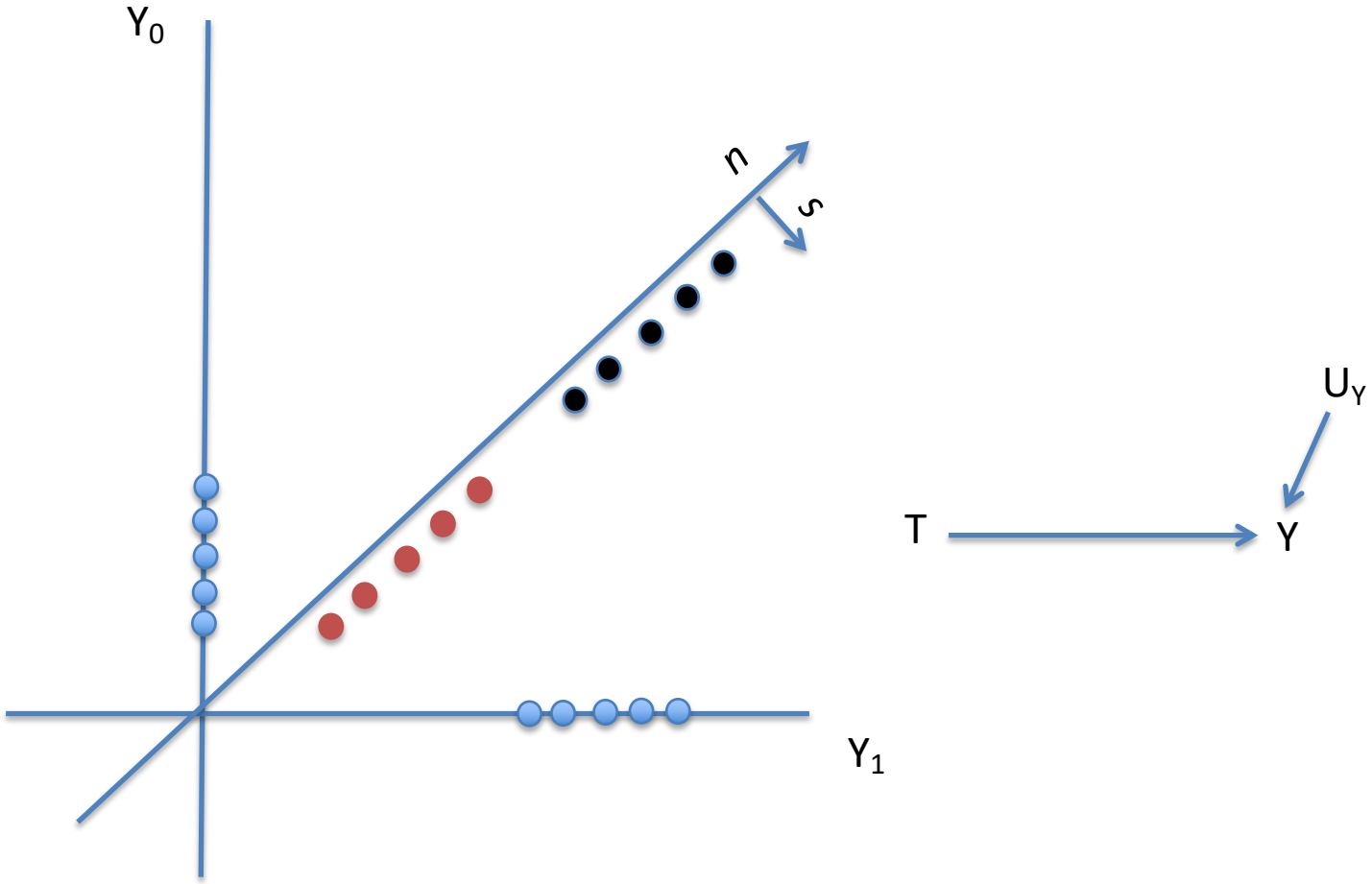
$\{X : E \perp\!\!\!\perp (Y_0, Y_1) \mid e(X)\}$





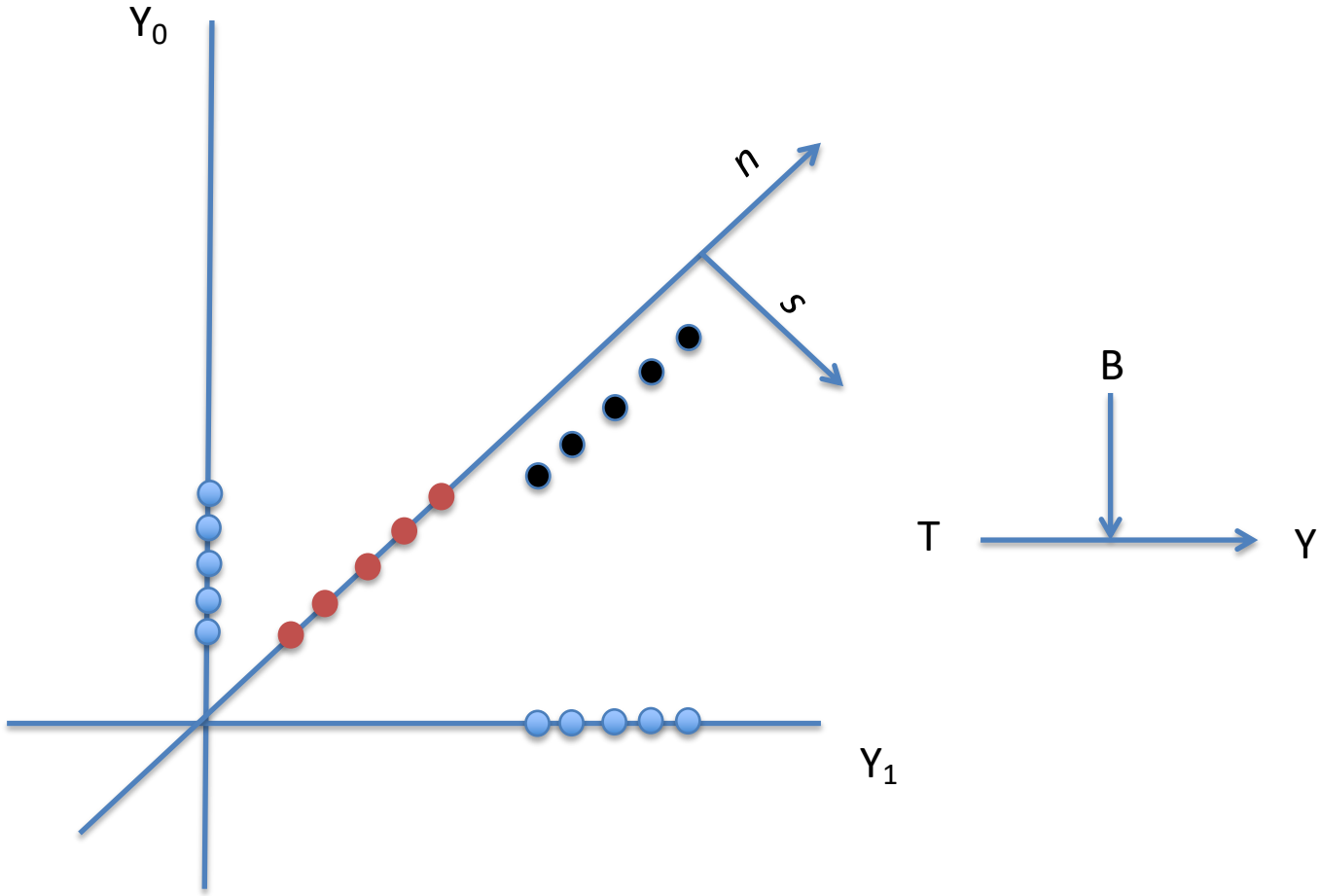
blocking can (e.g. regression, matching,…)

- ***reduce bias***
- *reduce noise*
- *explain heterogeneity.*



blocking can (e.g. regression, matching,…)

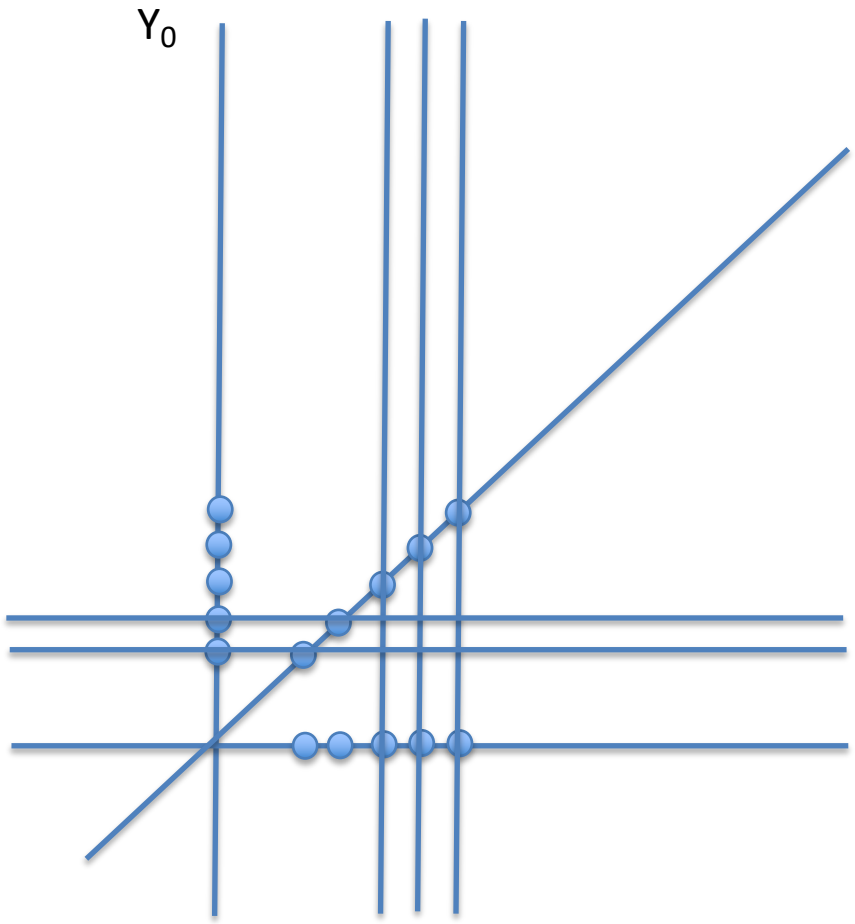
- *reduce bias*
- ***reduce noise***
- *explain heterogeneity.*

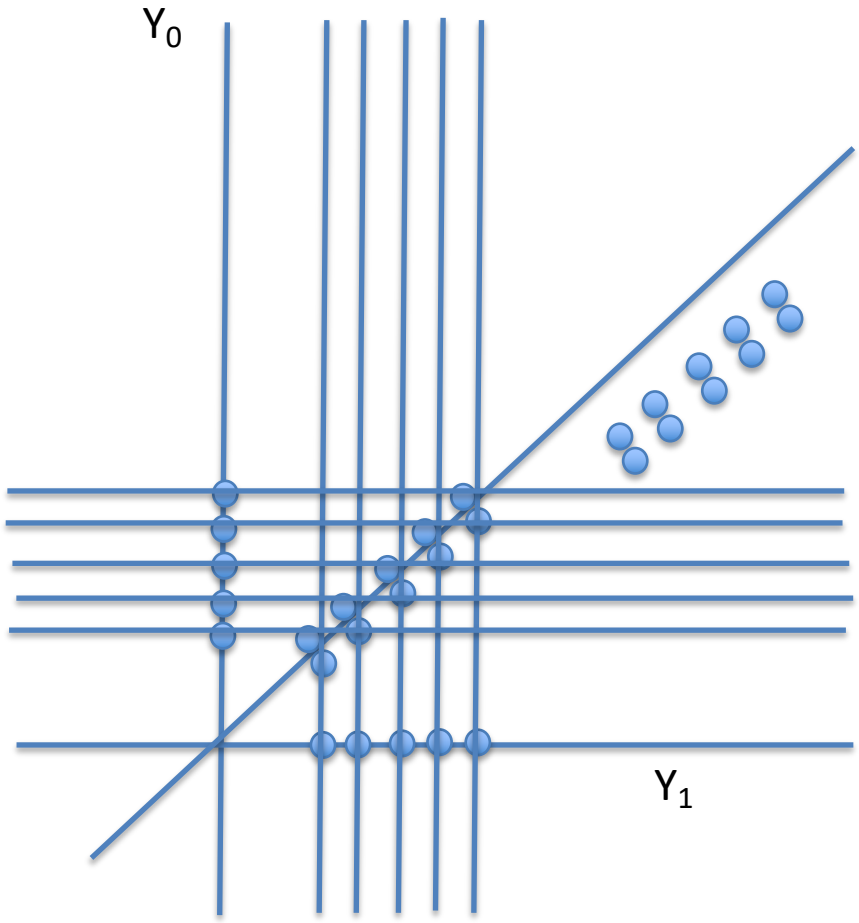


blocking can (e.g. regression, matching,…)

- *reduce bias*
- *reduce noise*
- ***explain heterogeneity***

Bias (sampling & assignment)



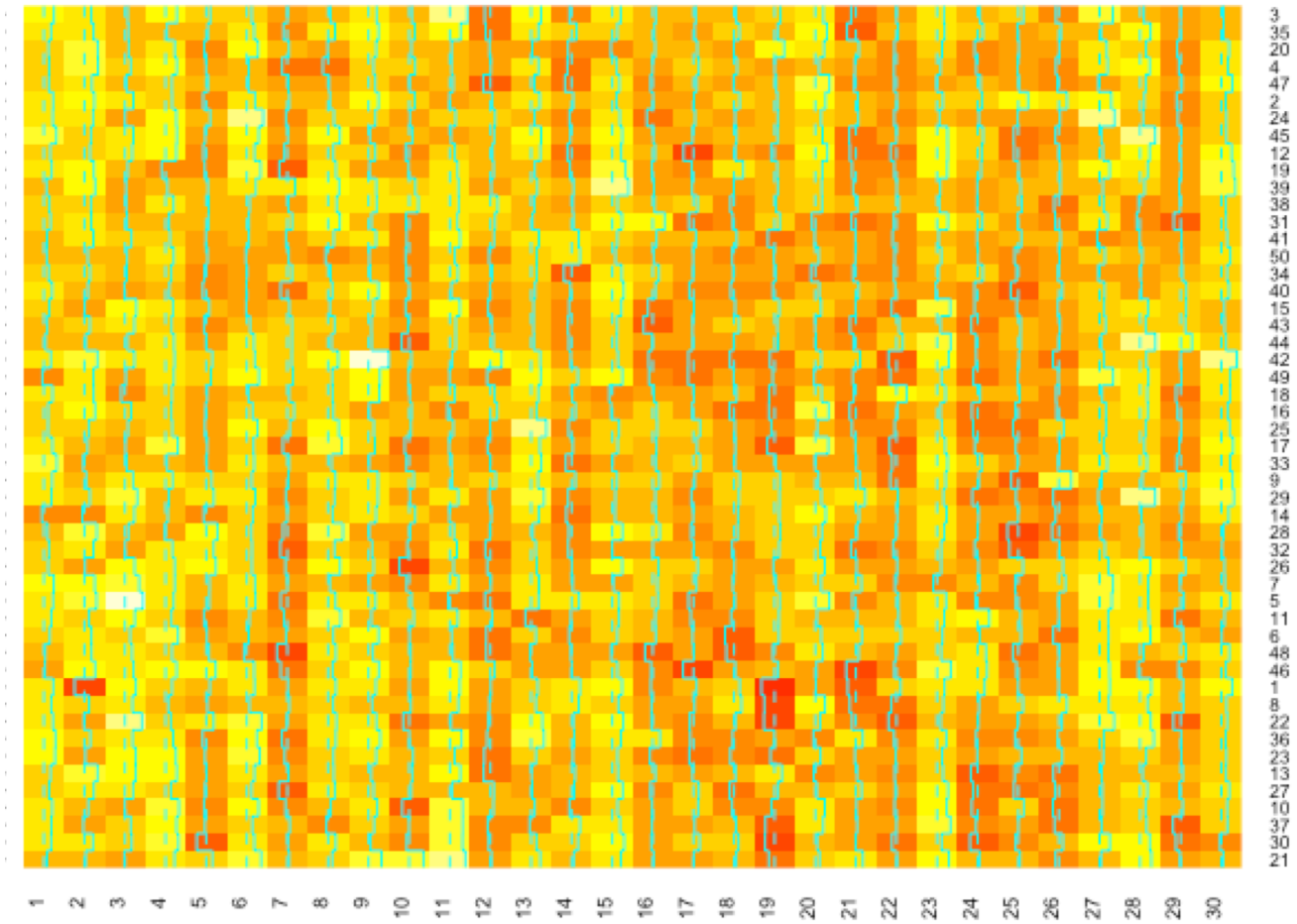


Fisher but Simpson, Neyman but Berkson
Actually, this should be a 2 x 2 table of graphs (with/without each bias).

Correlation and causation in social genomics

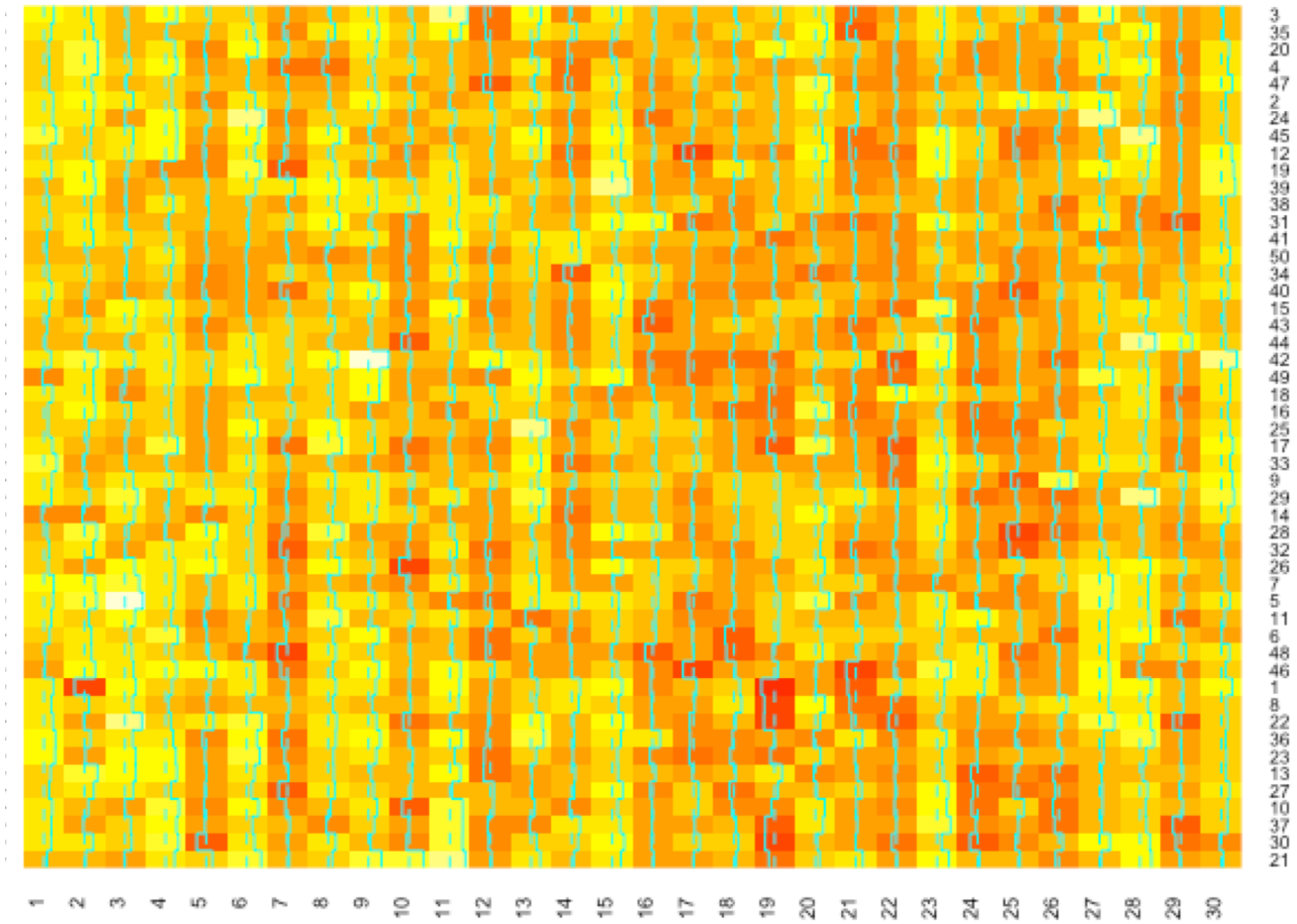
Justin Chumbley

Gene expression



Dogma of statistical inference: y ~ P

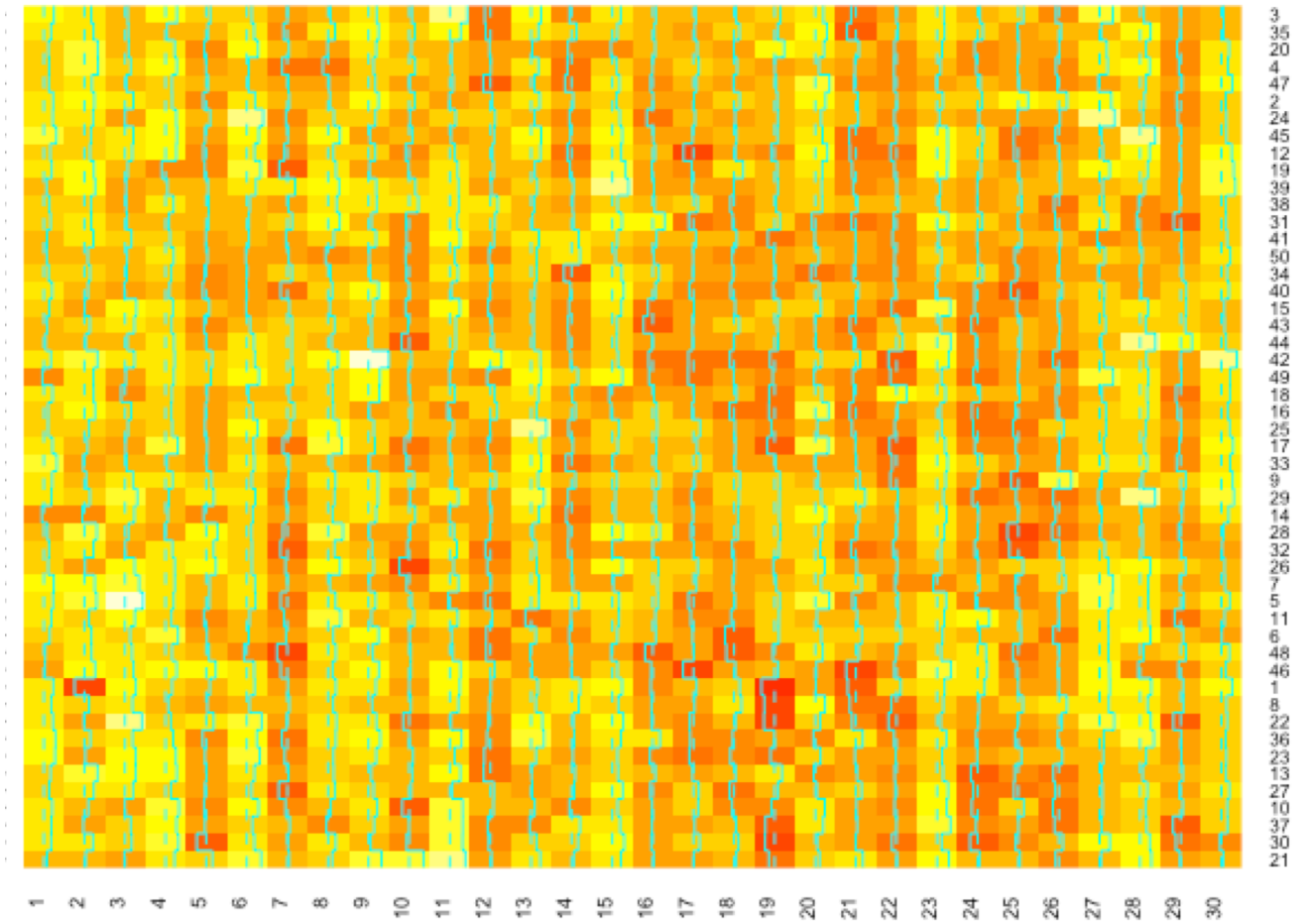
Gene expression



Dogma of statistical inference: (y,t) ~ P

Exposure

Gene expression



Dogma of statistical inference: (y,t) ~ P

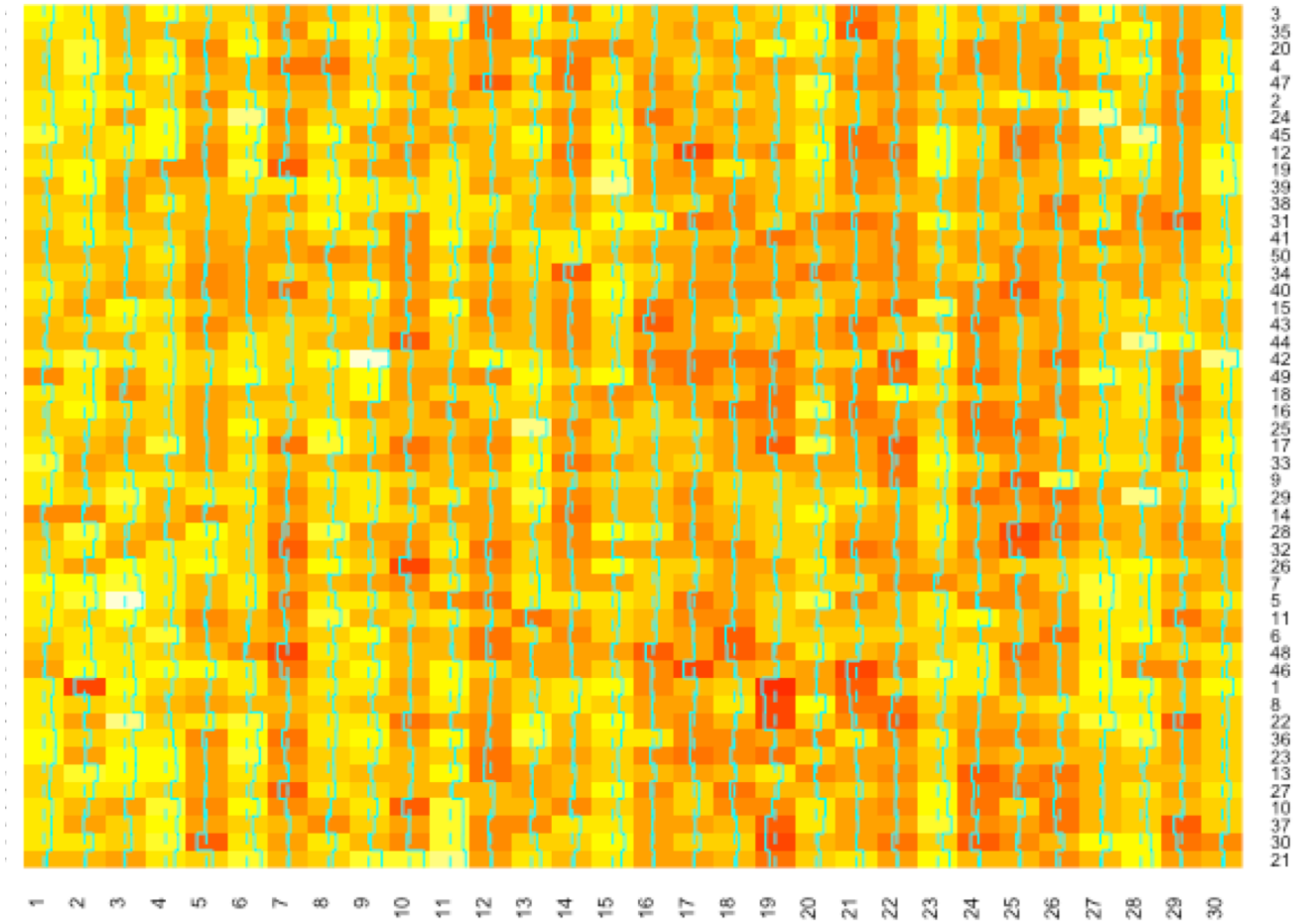
P(T)

Exposure

P(Y, T)

Gene expression

P(Y)



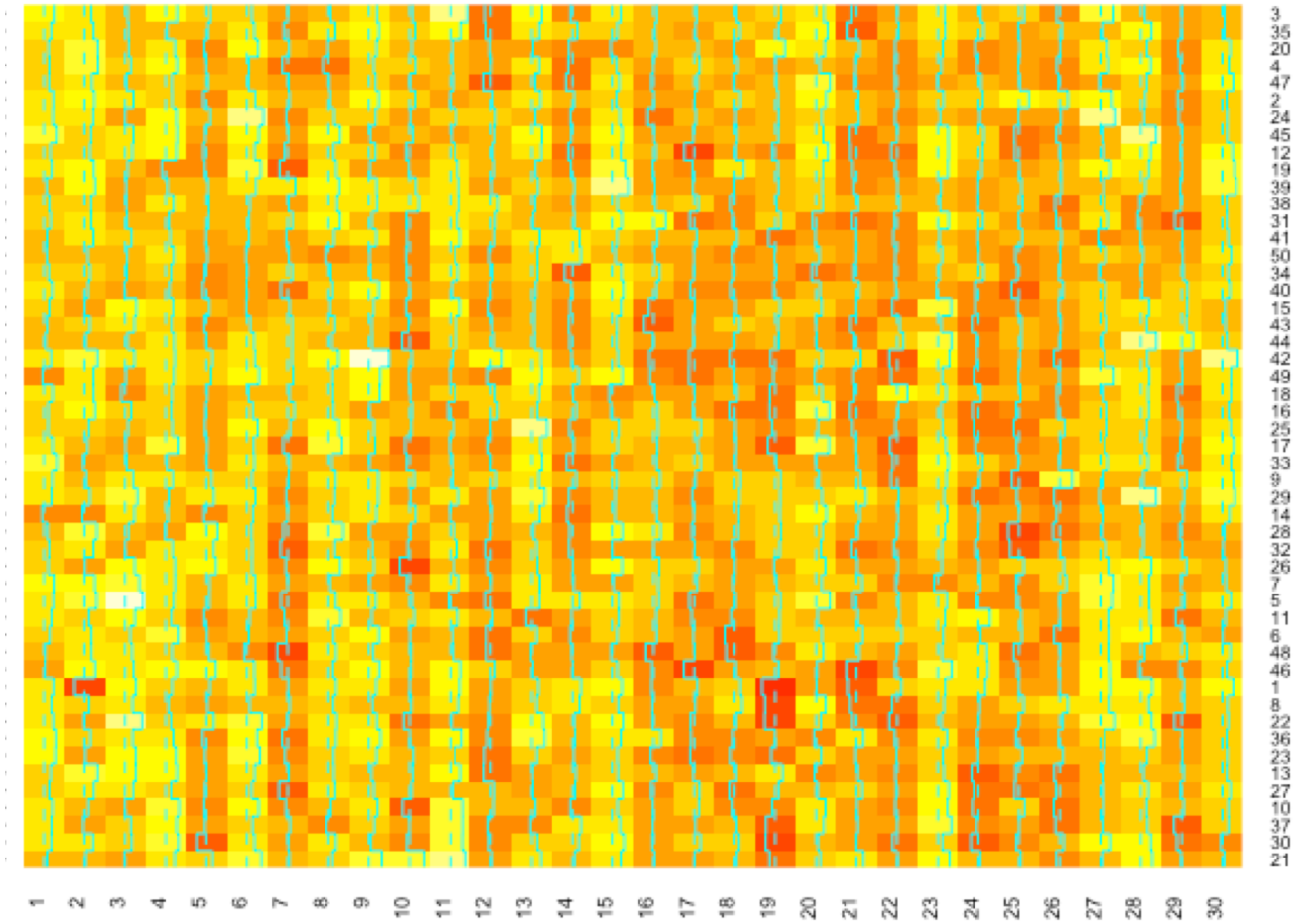
Dogma of statistical inference: (y,t) ~ P

P(T) Exposure

P(Y, T) = P(Y)P(T)

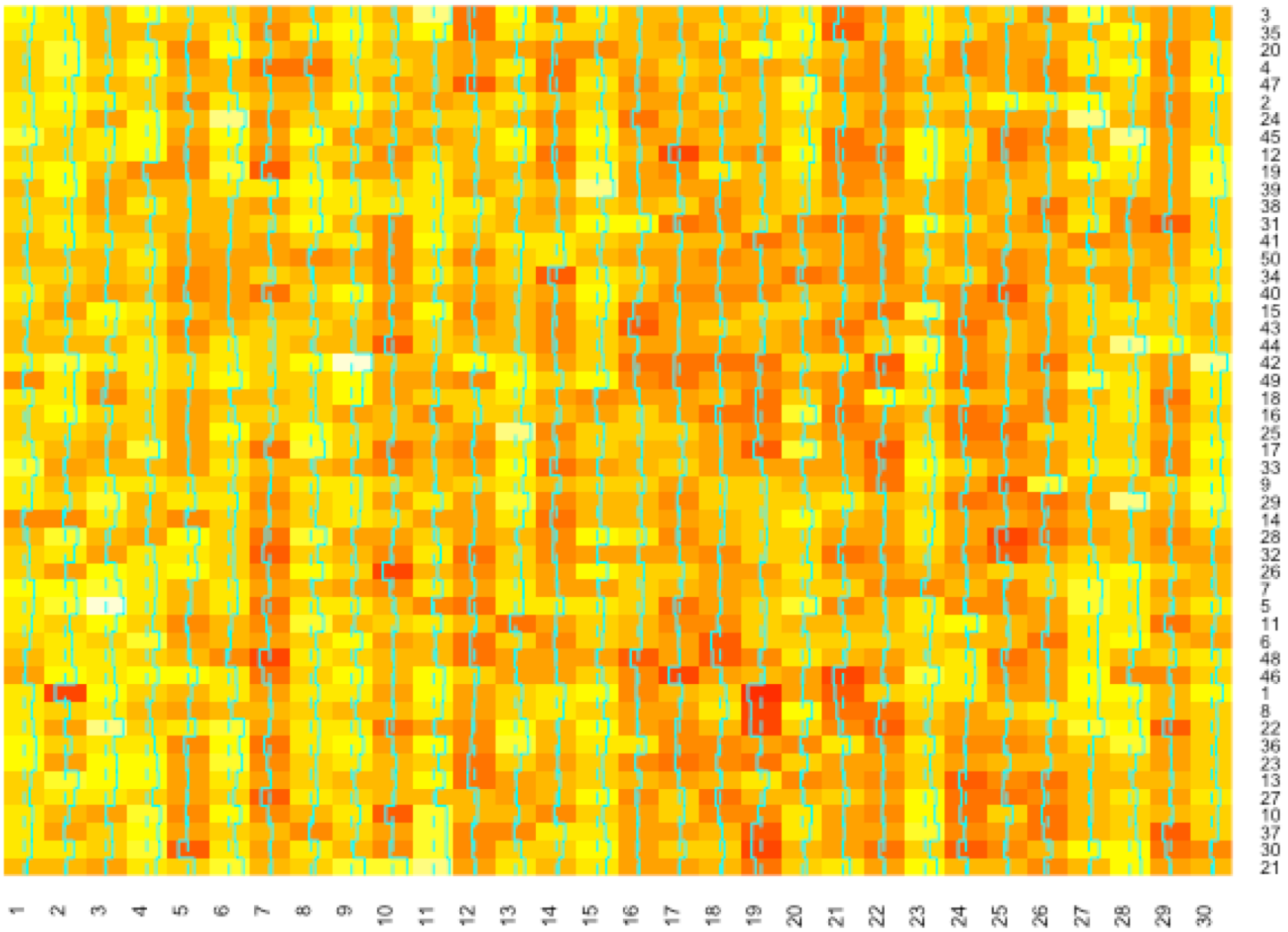
Gene expression

P(Y)



Exposure

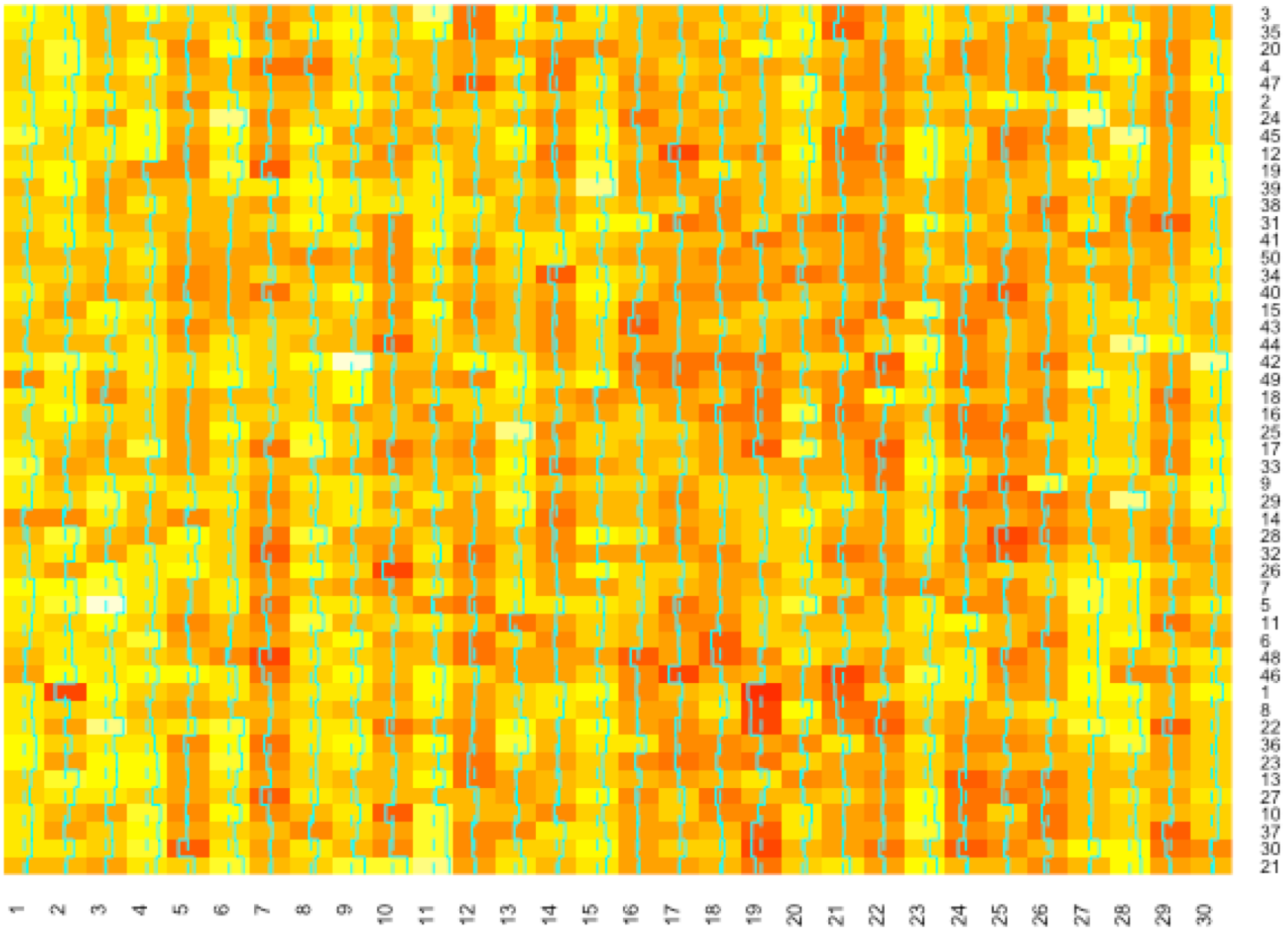
Gene expression

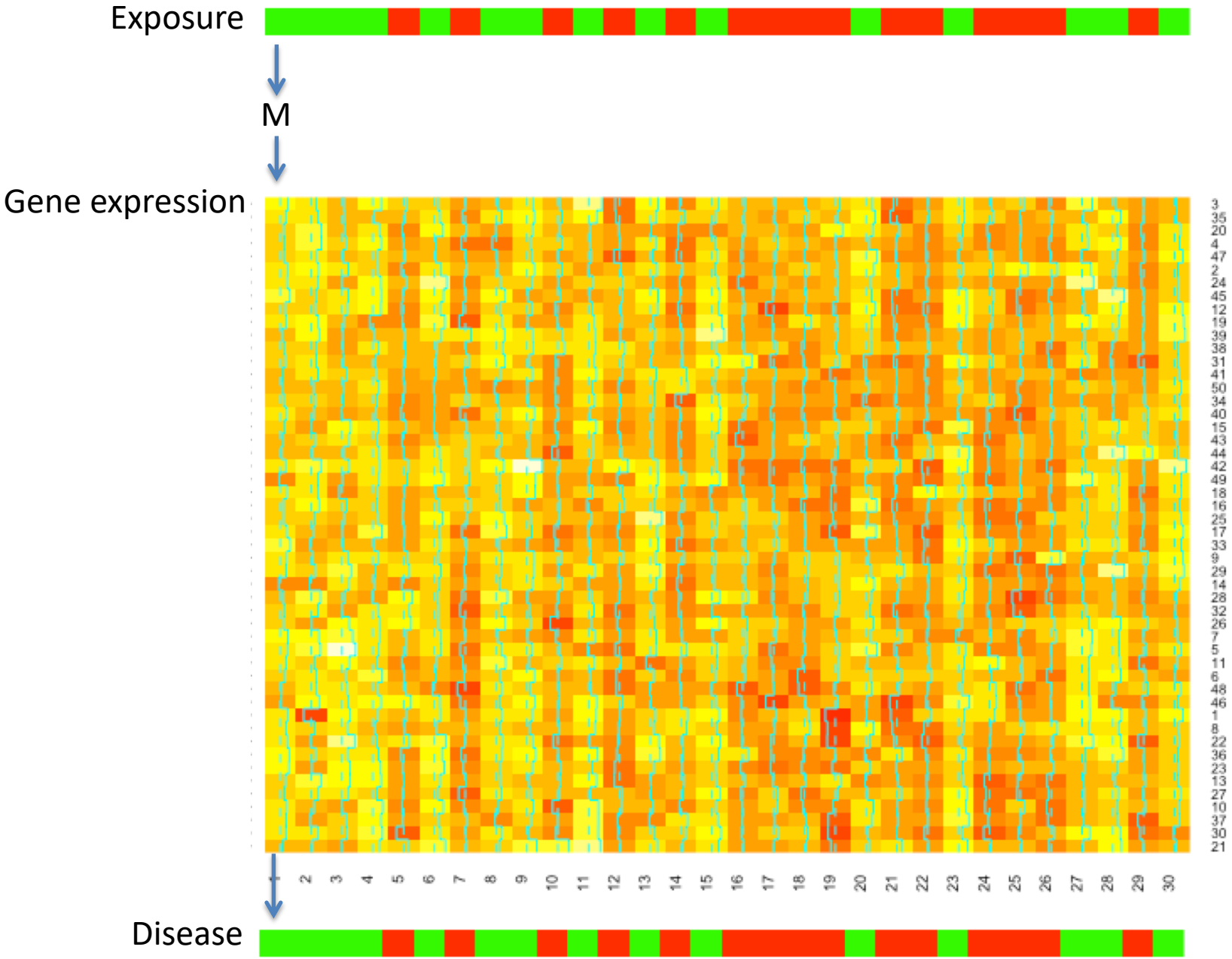


Exposure

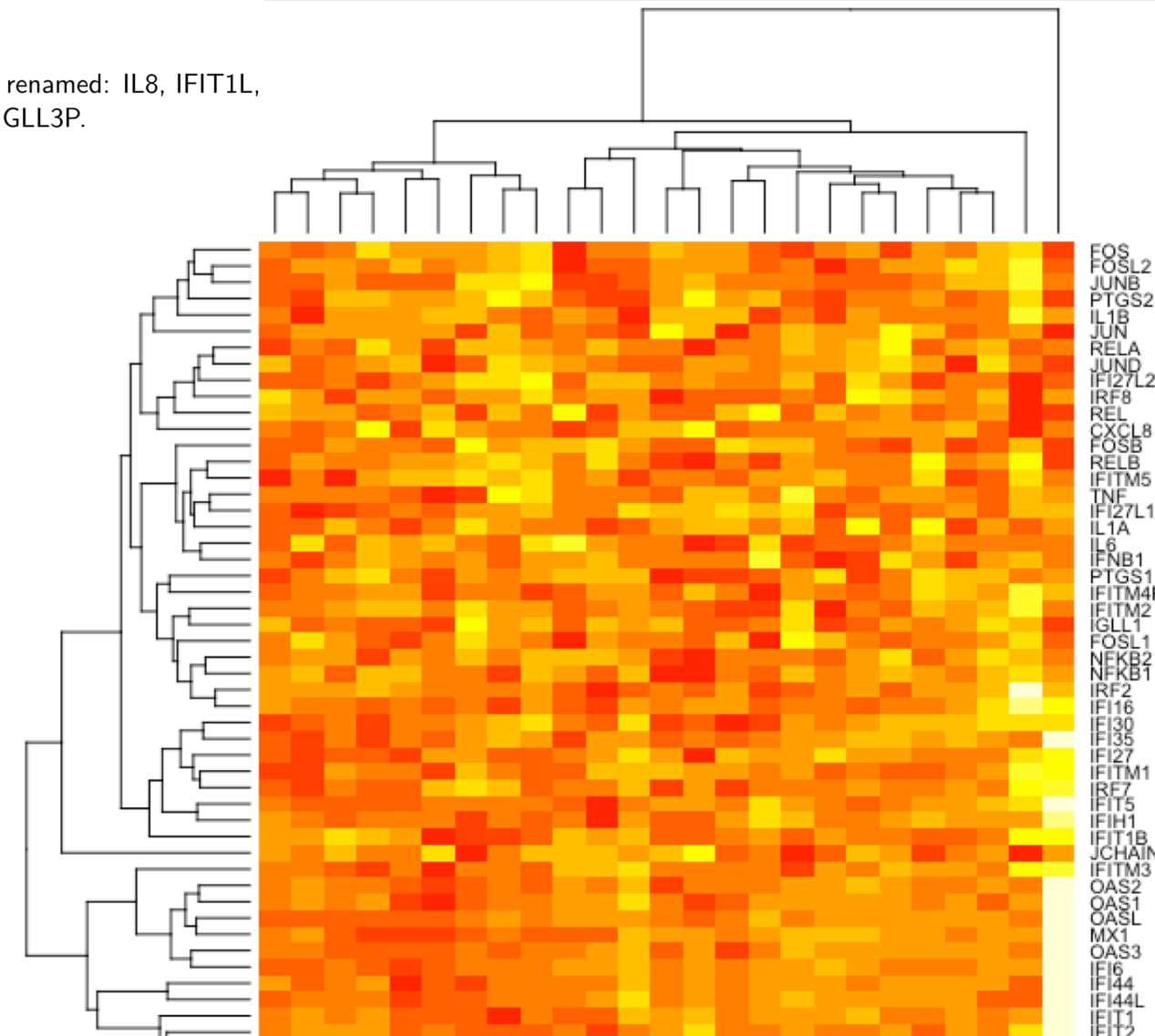
M

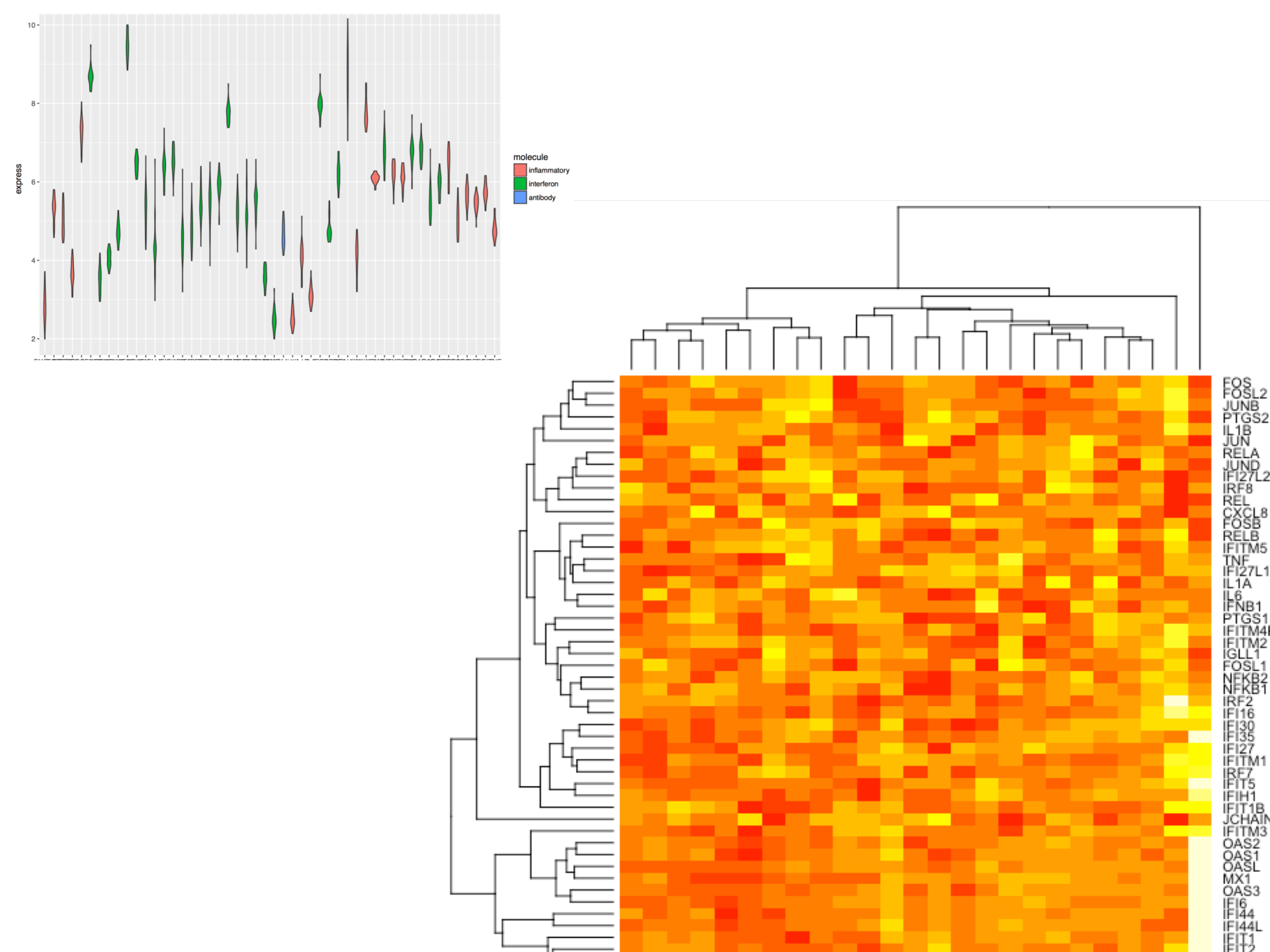
Gene expression

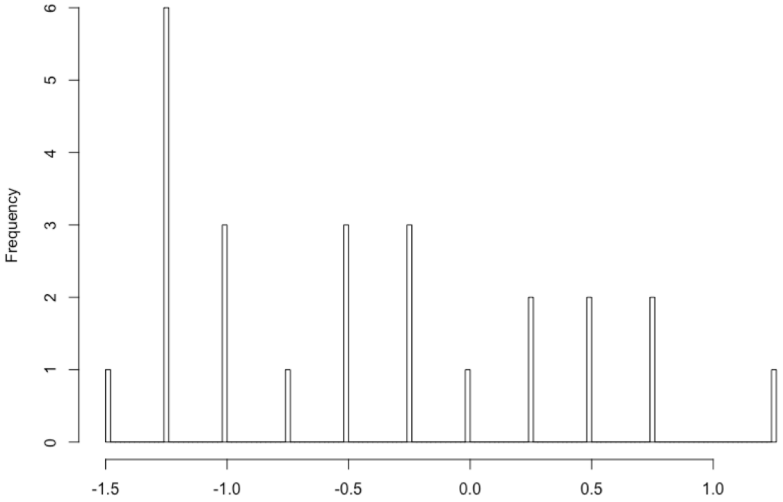



Exposure
M
Gene expression
Disease

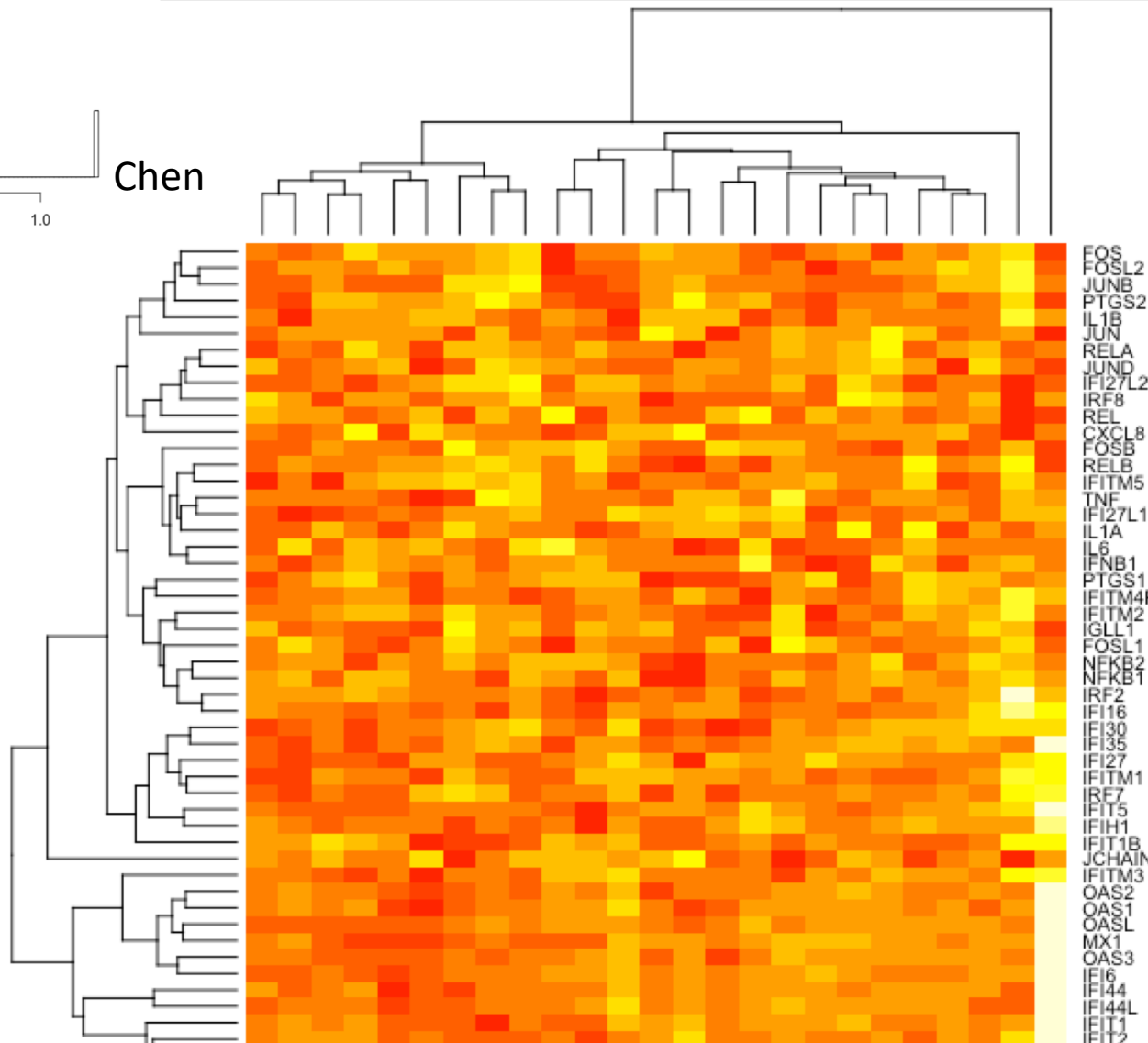
- 50 of the 53 CTRA were on our chip.
- Inflamatory: IL1A, IL1B, IL6, CXCL8, TNF, PTGS1, PTGS2, FOS, FOSB, FOSL1, FOSL2, JUN, JUNB, JUND, NFKB1, NFKB2, REL, RELA, RELB .
- Interferon type-I: IFI16, IFI27, IFI27L1, IFI27L2, IFI30, IFI35, IFI44, IFI44L, IFI6, IFIH1, IFIT1, IFIT2, IFIT3, IFIT5, IFIT1B, IFITM1, IFITM2, IFITM3, IFITM4P, IFITM5, IFNB1, IRF2, IRF7, IRF8, MX1, OAS1, OAS2, OAS3, OASL.
- Antibody: JCHAIN, IGLL1.
- Note that 4 of the original 53 CTRA have been renamed: IL8, IFIT1L, IGJ, IGLL3 are now CXCL8, IFIT1B, JCHAIN, IGLL3P.



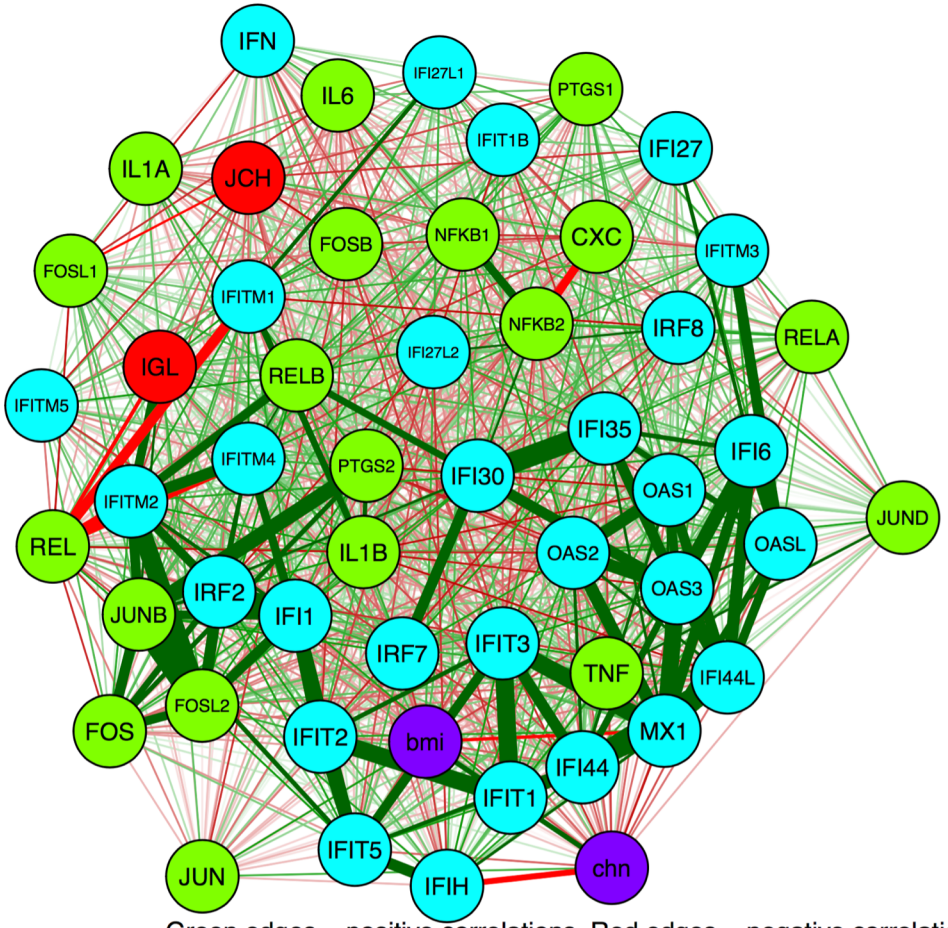

express
10
8
6
4
2
molecule
inflammatory
interferon
antibody
FOS
FOSL2
JUNB
PTGS2
IL1B
JUN
RELA
JUND
IFI27L2
IRF8
REL
CXCL8
FOSB
RELB
IFITM5
TNF
IFI27L1
IL1A
IL6
IFNB1
PTGS1
IFITM4
IFITM2
IGLL1
FOSL1
NFKB2
NFKB1
IRF2
IFI16
IFI30
IFI35
IFI27
IFITM1
IRF7
IFIT5
IFIH1
IFIT1B
JCHAIN
IFITM3
OAS2
OAS1
OASL
MX1
OAS3
IFI6
IFI44
IFI44L
IFIT1


Frequency
0
1
2
3
4
5
6
-1.5
-1.0
-0.5
0.0
0.5
1.0

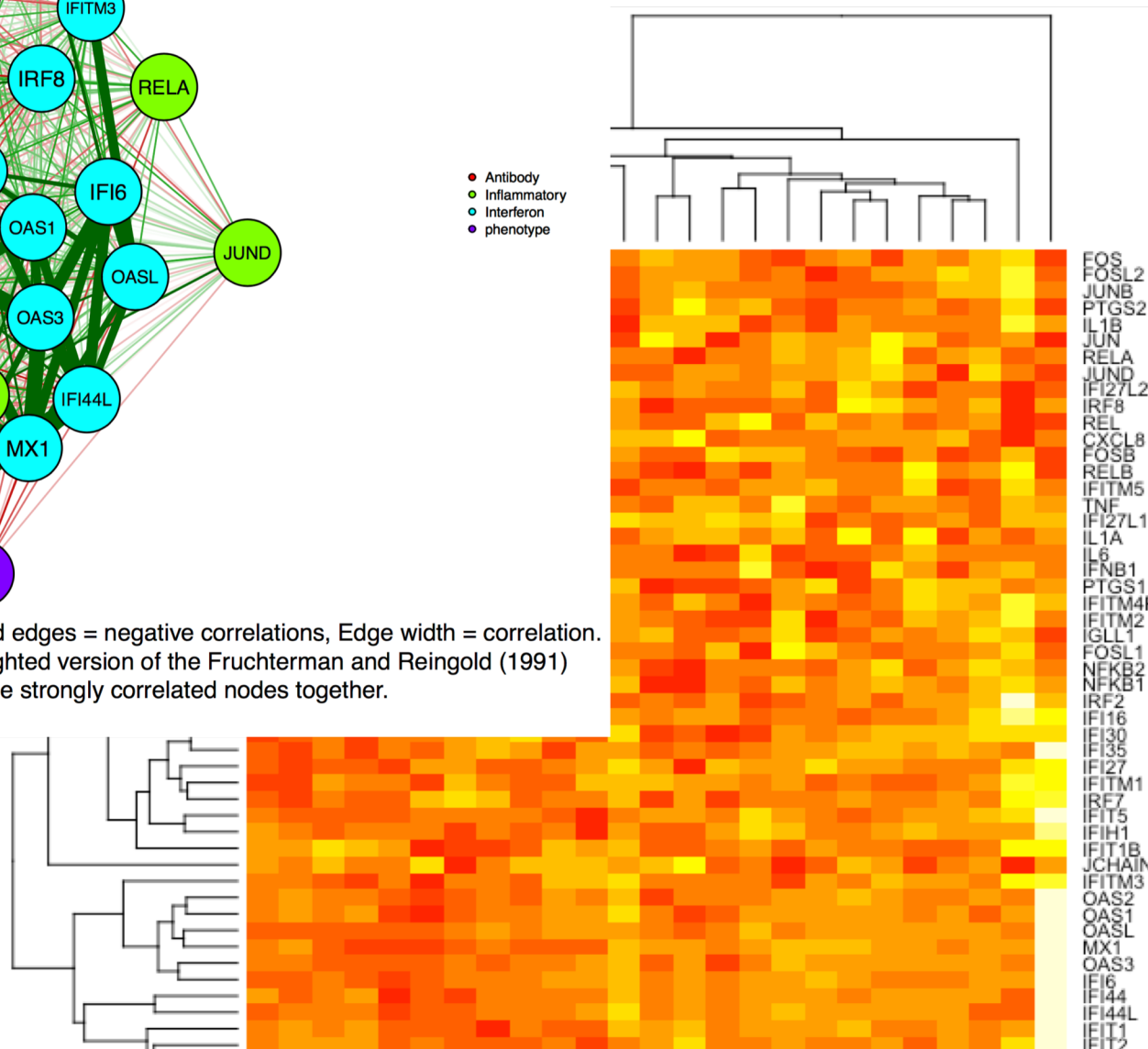
Chen

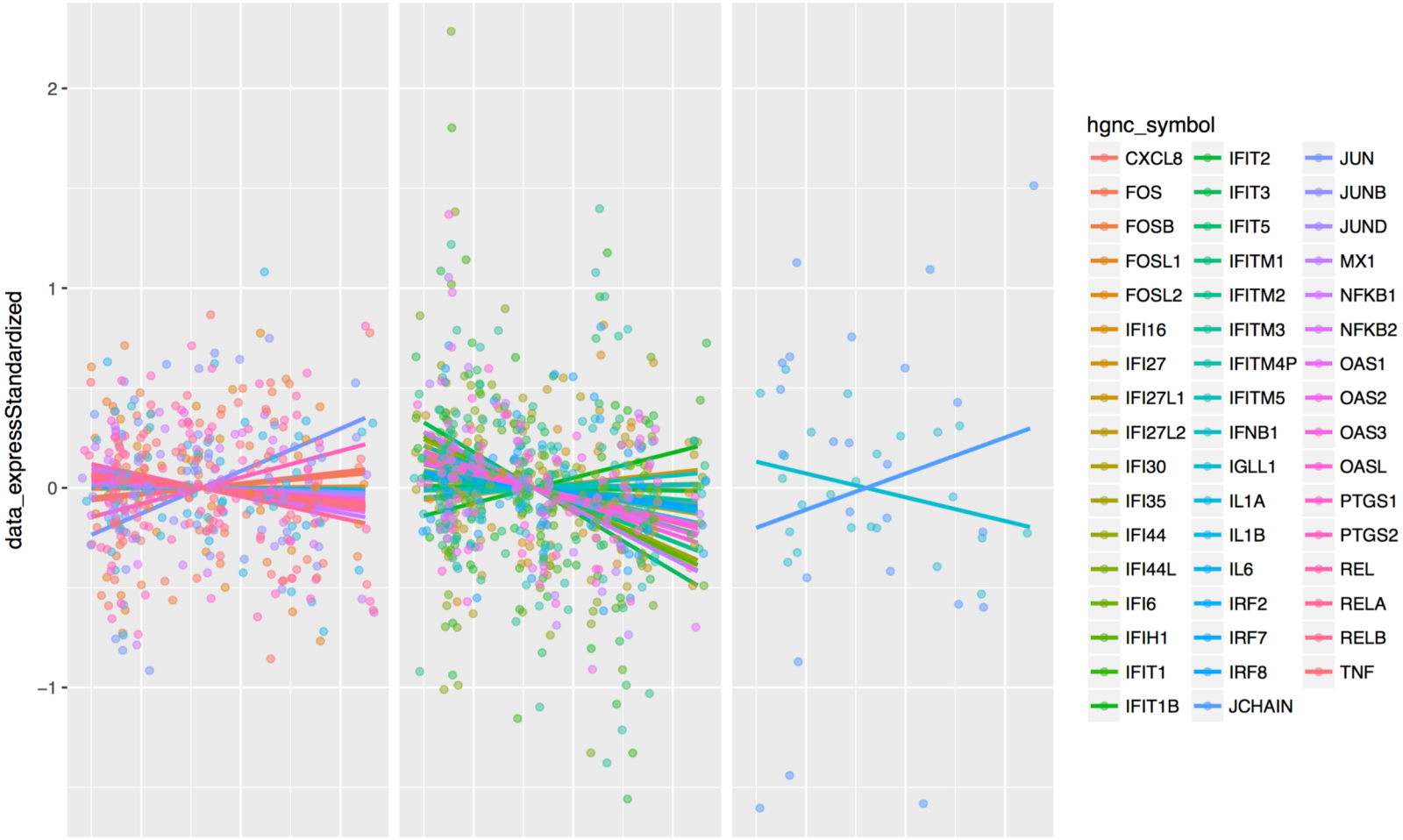

FOS
FOSL2
JUNB
PTGS2
IL1B
JUN
RELA
JUND
IFI27L2
IRF8
REL
CXCL8
FOSB
RELB
IFITM5
TNF
IFI27L1
IL1A
IL6
IFNB1
PTGS1
IFITM4
IFITM2
IGLL1
FOSL1
NFKB2
NFKB1
IRF2
IFI16
IFI30
IFI35
IFI27
IFITM1
IRF7
IFIT5
IFIH1
IFIT1B
JCHAIN
IFITM3
OAS2
OAS1
OASL
MX1
OAS3
IFI6
IFI44
IFI44L
IFIT1

CTRA network



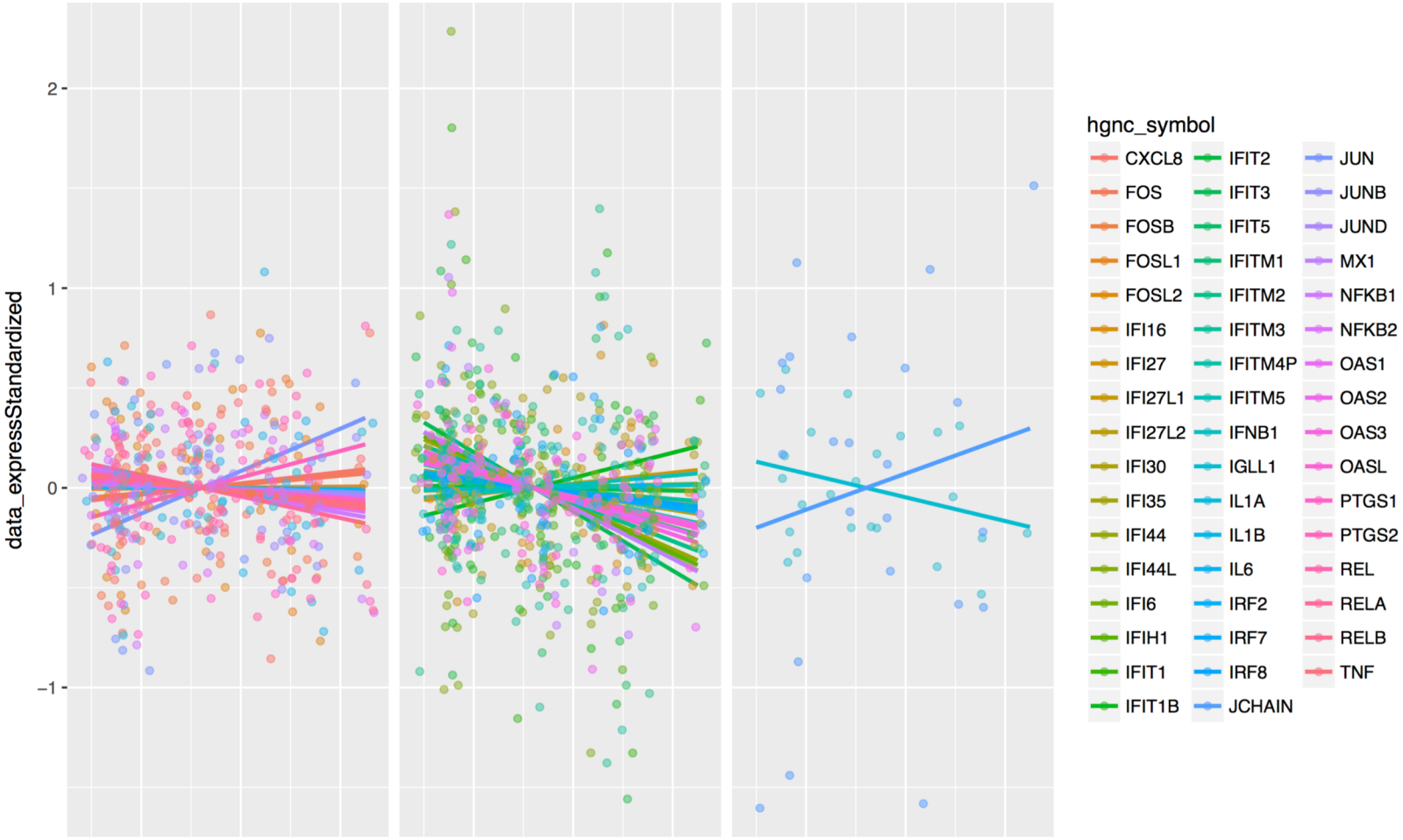
Green edges = positive correlations, Red edges = negative correlations, Edge width = correlation.
Node positioning based on a weighted version of the Fruchterman and Reingold (1991) algorithm to place strongly correlated nodes together.



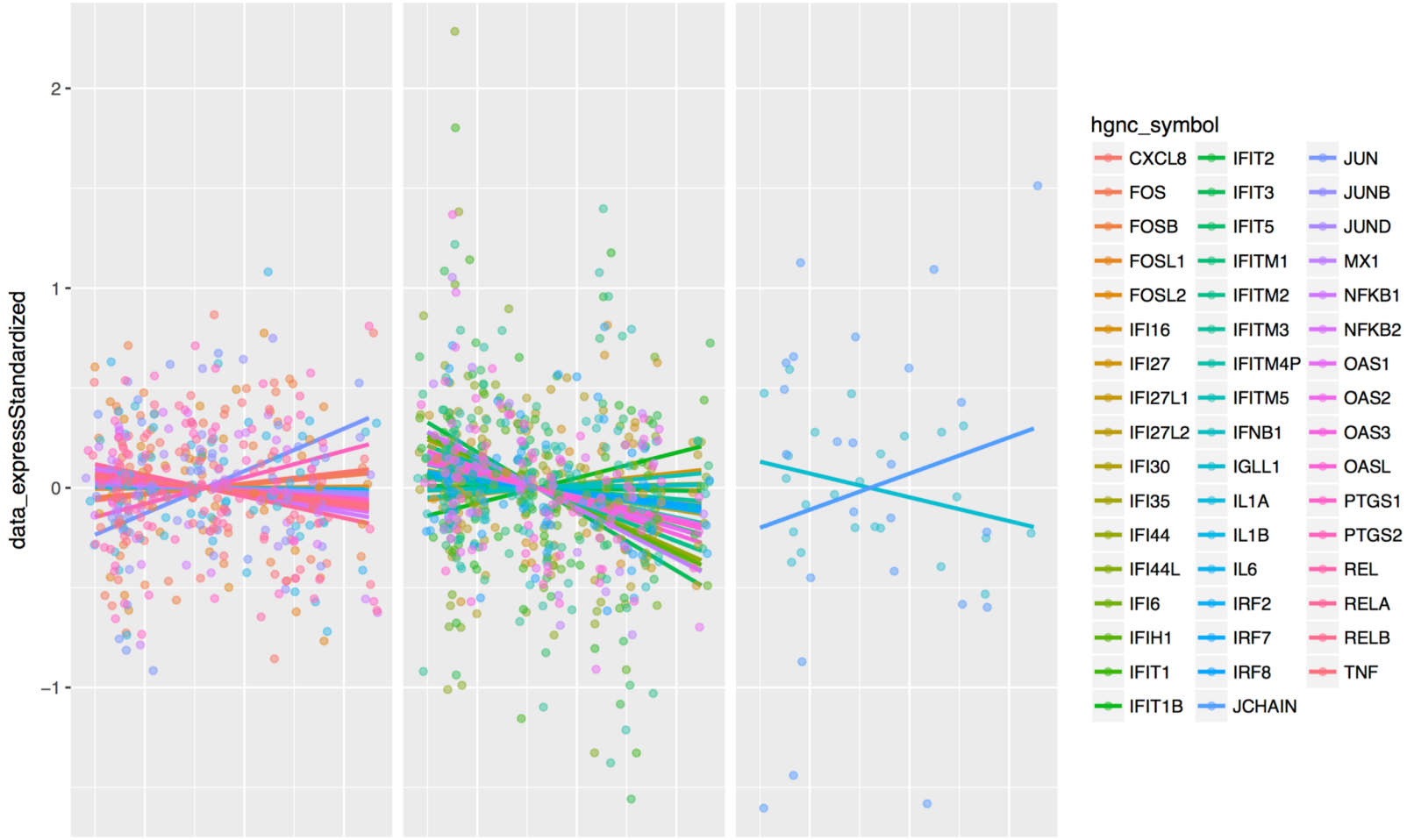

data_expressStandardized
2
1
0
−1
hgnc_symbol
CXCL8
FOS
FOSB
FOSL1
FOSL2
IFI16
IFI27
IFI27L1
IFI27L2
IFI30
IFI35
IFI44
IFI44L
IFI6
IFIH1
IFIT1
IFIT1B
IFIT2
IFIT3
IFIT5
IFITM1
IFITM2
IFITM3
IFITM4P
IFITM5
IFNB1
IGLL1
IL1A
IL1B
IL6
IRF2
IRF7
IRF8
JCHAIN
JUN
JUNB
JUND
MX1
NFKB1
NFKB2
OAS1
OAS2
OAS3
OASL
PTGS1
PTGS2
REL
RELA
RELB
TNF

	logFC	AveExpr	t	P.Value	adj.P.Val
IFIH1	-0.235	6.552	-3.041	0.005	0.256
MX1	-0.254	6.763	-2.549	0.017	0.416
IFIT3	-0.297	5.420	-2.294	0.030	0.493
JUN	0.213	4.115	2.006	0.055	0.496
IFITM2	-0.191	5.241	-1.859	0.074	0.496
RELA	-0.108	5.460	-1.768	0.088	0.496

Table 1: Limma regression on CTRA.



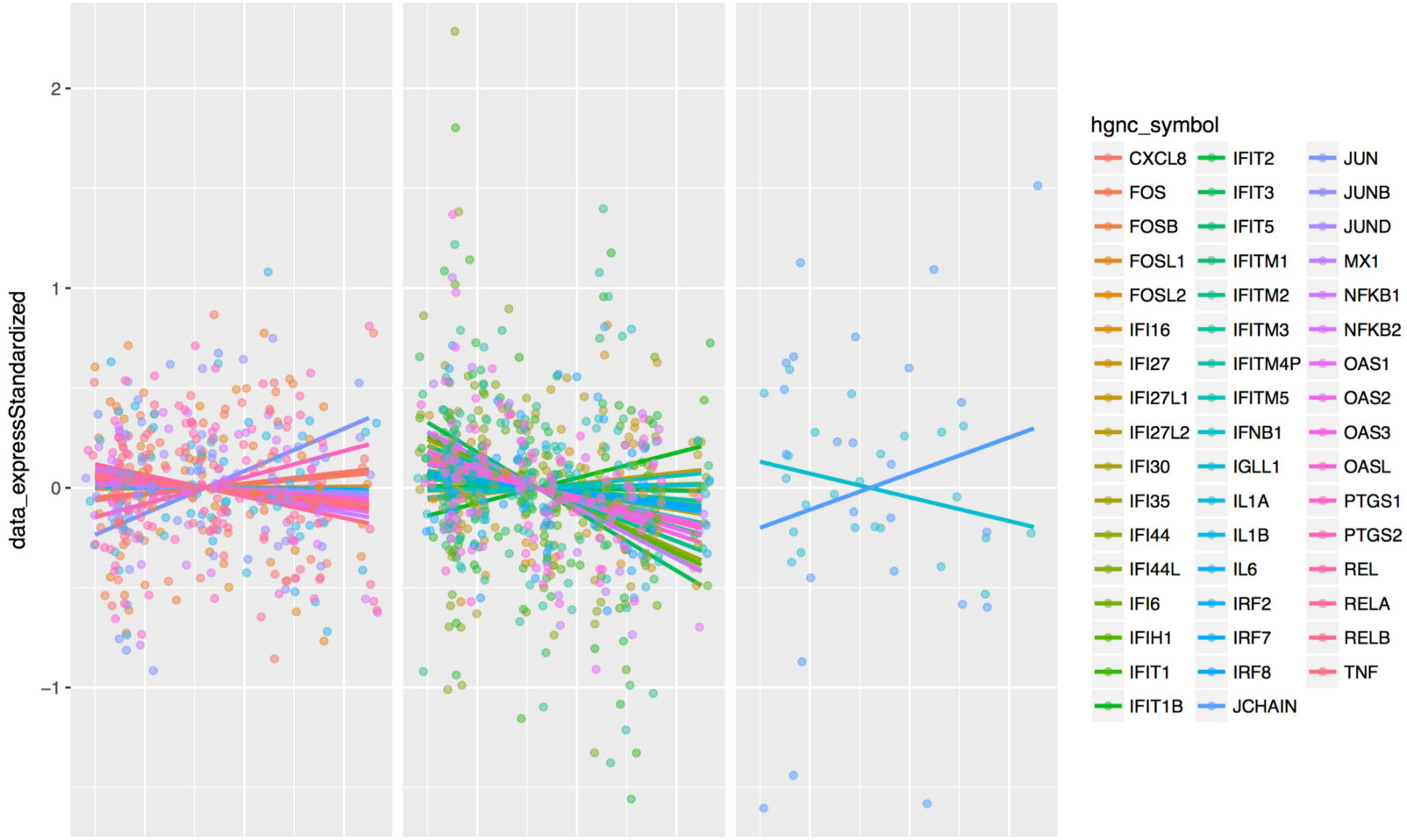
	p-value	Statistic	Expected	Std.dev	#Cov
Inflammatory	0.660	1.833	4.306	3.854	19
Interferon	0.050	14.881	3.751	4.662	29
Antibody	0.620	2.177	4.648	5.272	2
All	0.070	13.161	3.830	4.673	50



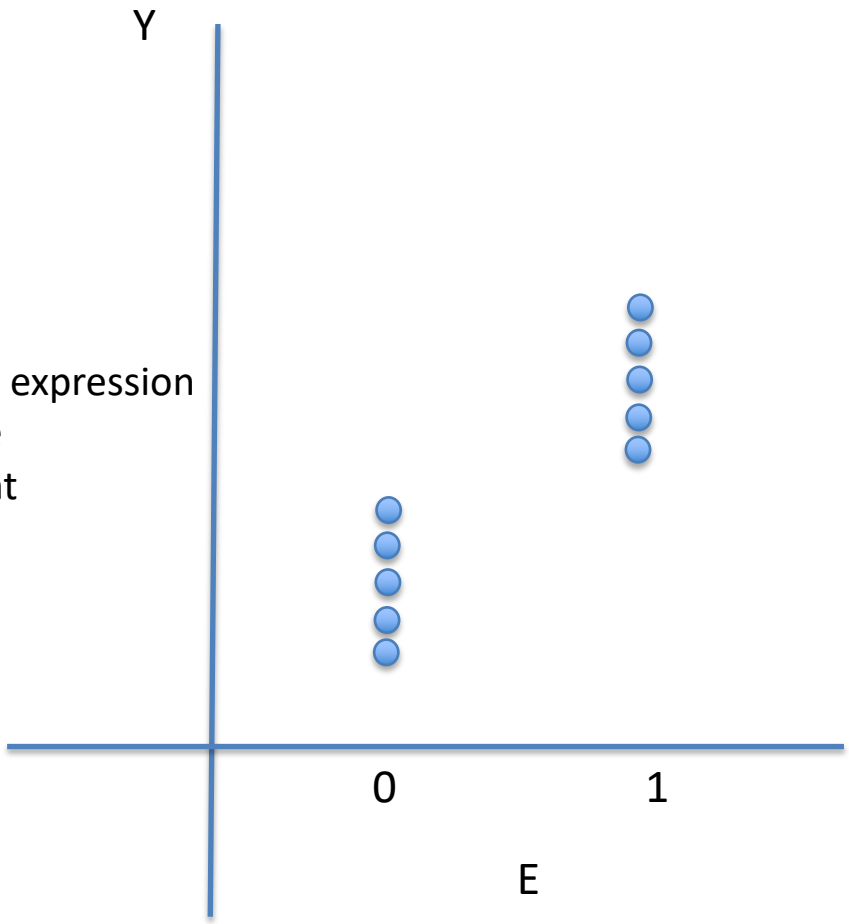
- A score test for nested parametric models (like the likelihood ratio test, but not parametrization-invariant).
- Optimal in a neighborhood of the null hypothesis.
- Handles p>n alternatives.
- (J. Goeman, Geer, and Kort 2004, Jelle J. Goeman, Geer, and Houwelingen (2006))

- H_0 global test: no gene (covariate) is associated with the response.
- H_1 at least one gene is associated
- Linear regression model: accomodates linear confounds.
- Power: tailored toward alternatives with many small regression coefficients of the same sign.
- (Model assumes random coefficients positively correlated, a priori.)

- We infered the fixed effect of chen on gene expression in a multilevel linear mixed model with independent random intercepts for both participant and CTRA gene (CI = -0.1860669, -0.0026783)

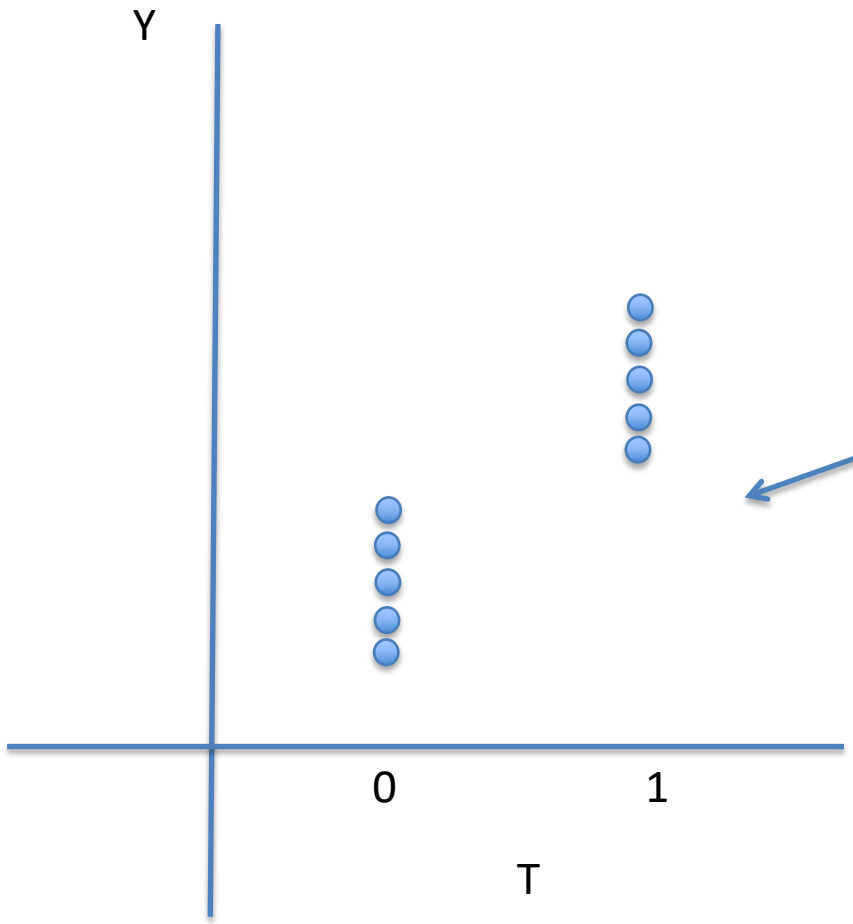


- gene expression
- N dimensional gene expression
- cell type prevalence
- clustering coefficient



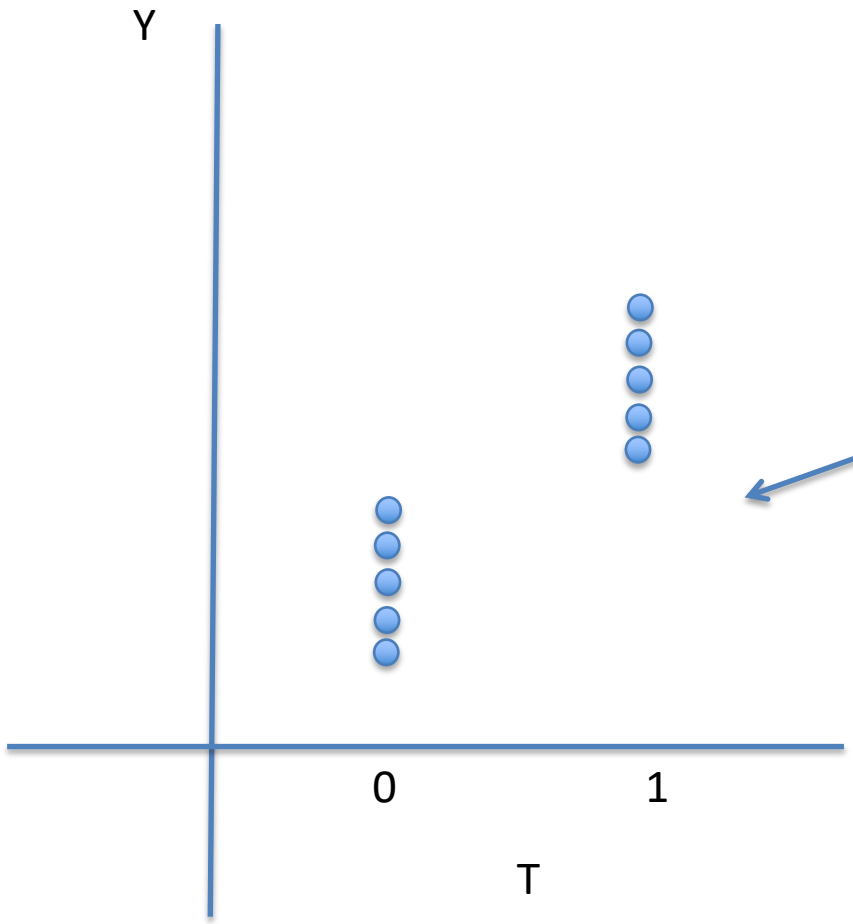
Cause (Noun): something that brings about an effect or a result

Effect (Noun): something that inevitably follows a … a cause



Probability is symmetric.
What is missing?
Missing structure.

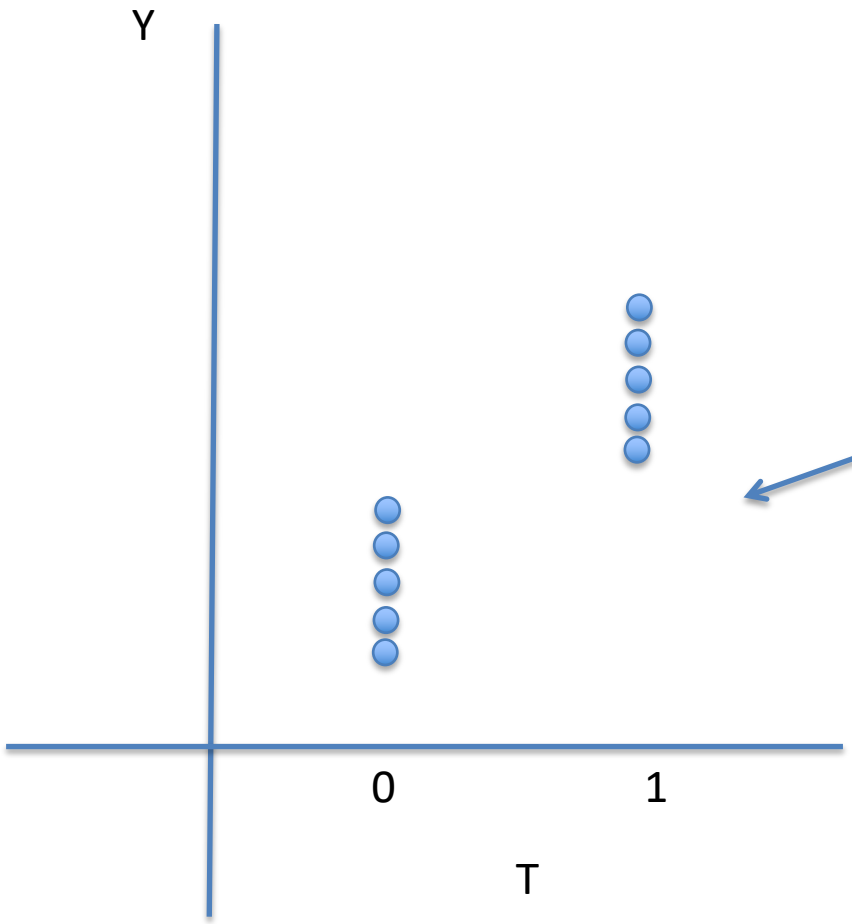
T → Y



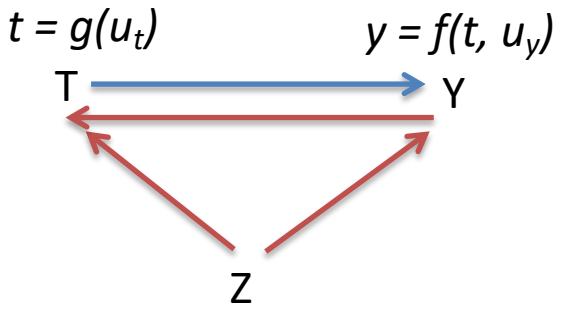
Probability is symmetric.
What is missing?
Missing structure.

$t = g(u_t)$ $y = f(t, u_y)$

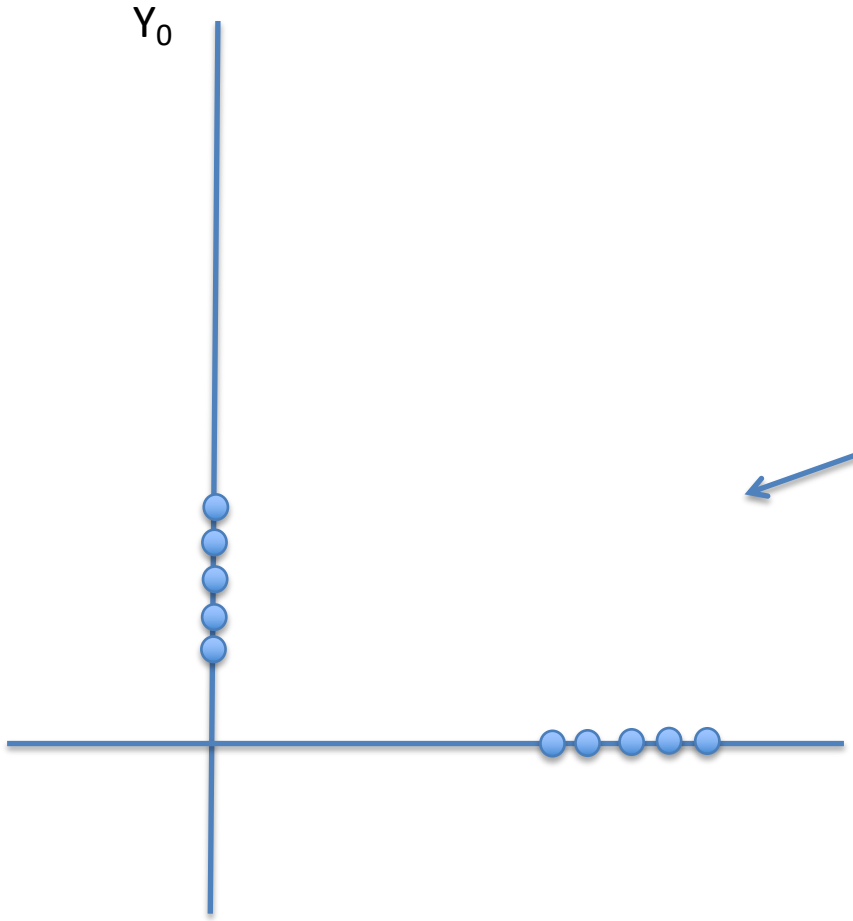
T → Y



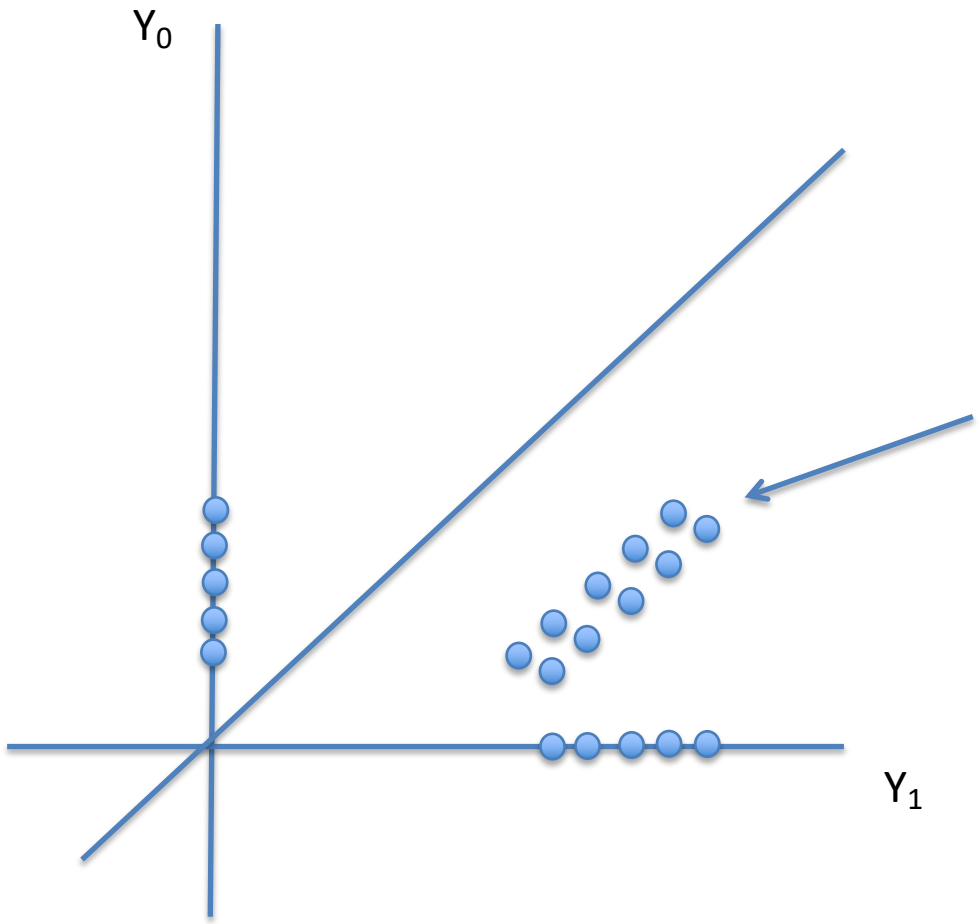
Probability is symmetric.
What is missing?
Missing structure.



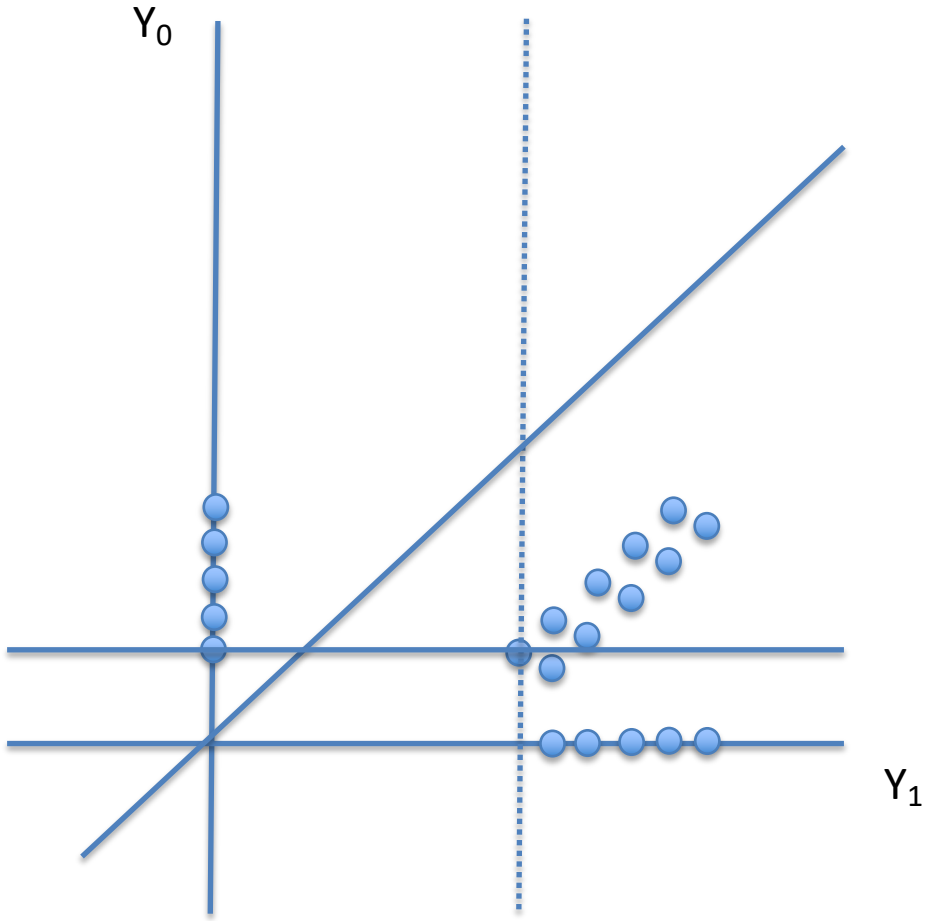
- Variables/nodes
- Included Arrows
- Excluded arrows (strong exclusion restrictions)



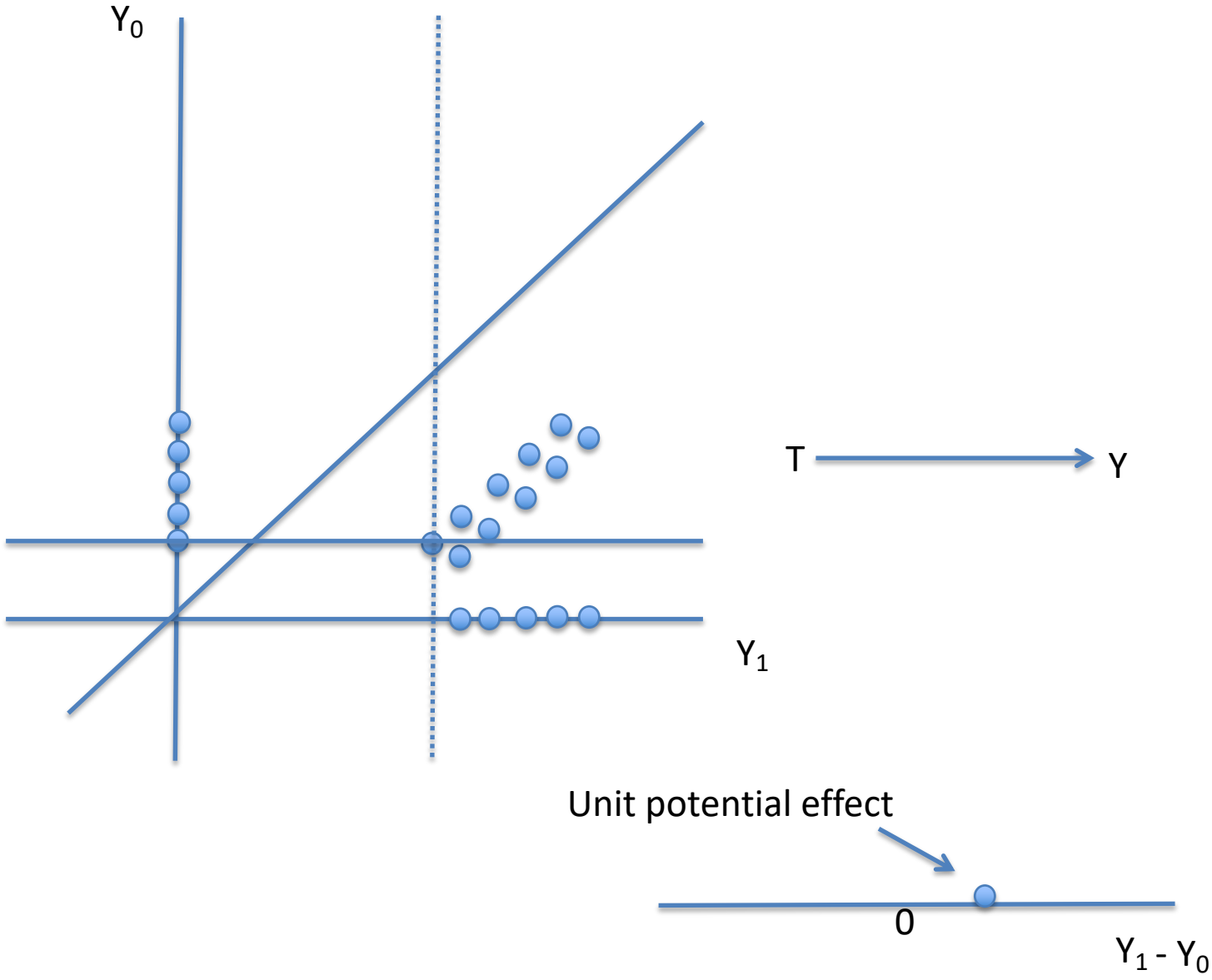
Probability is symmetric.
What is missing?

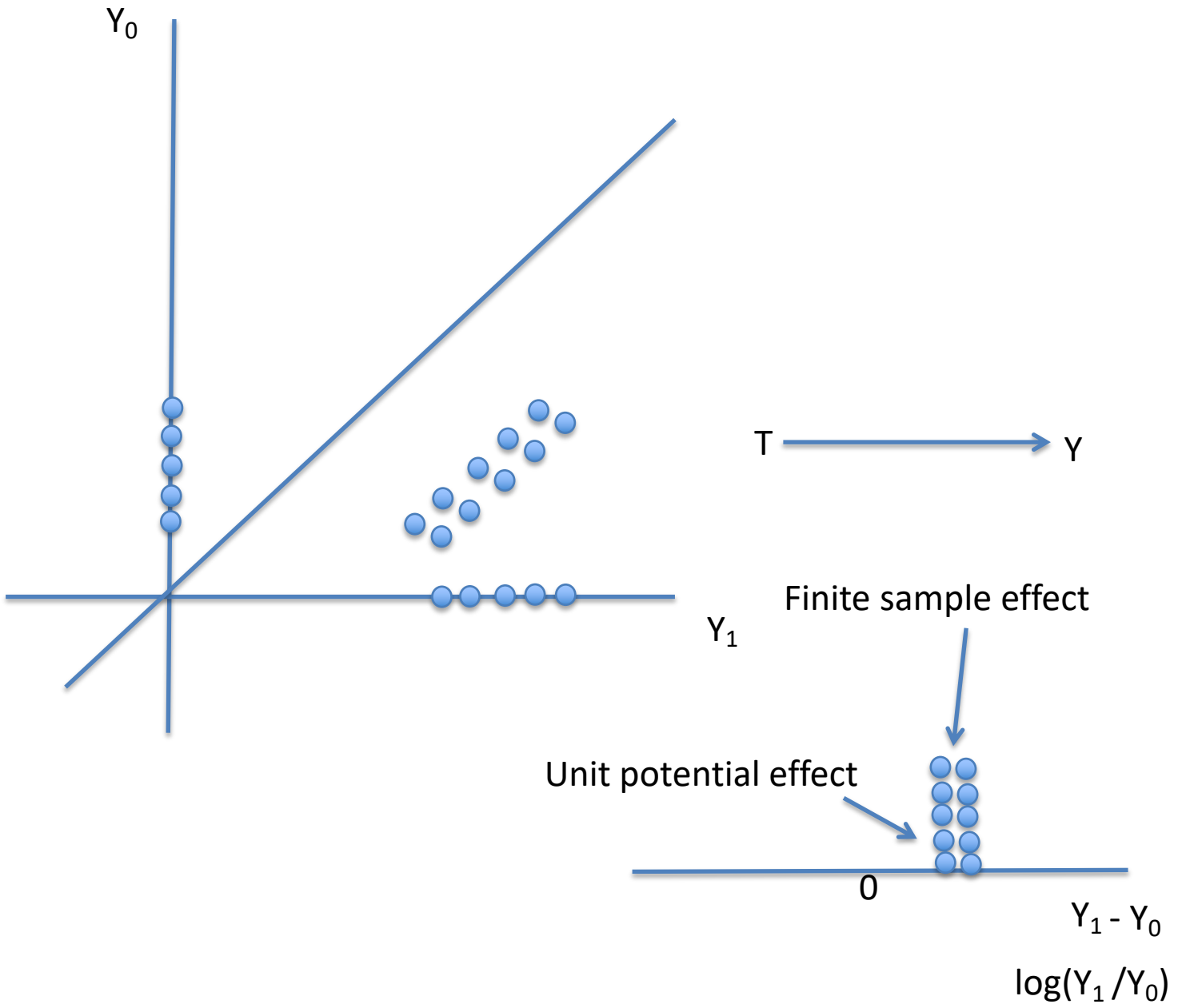


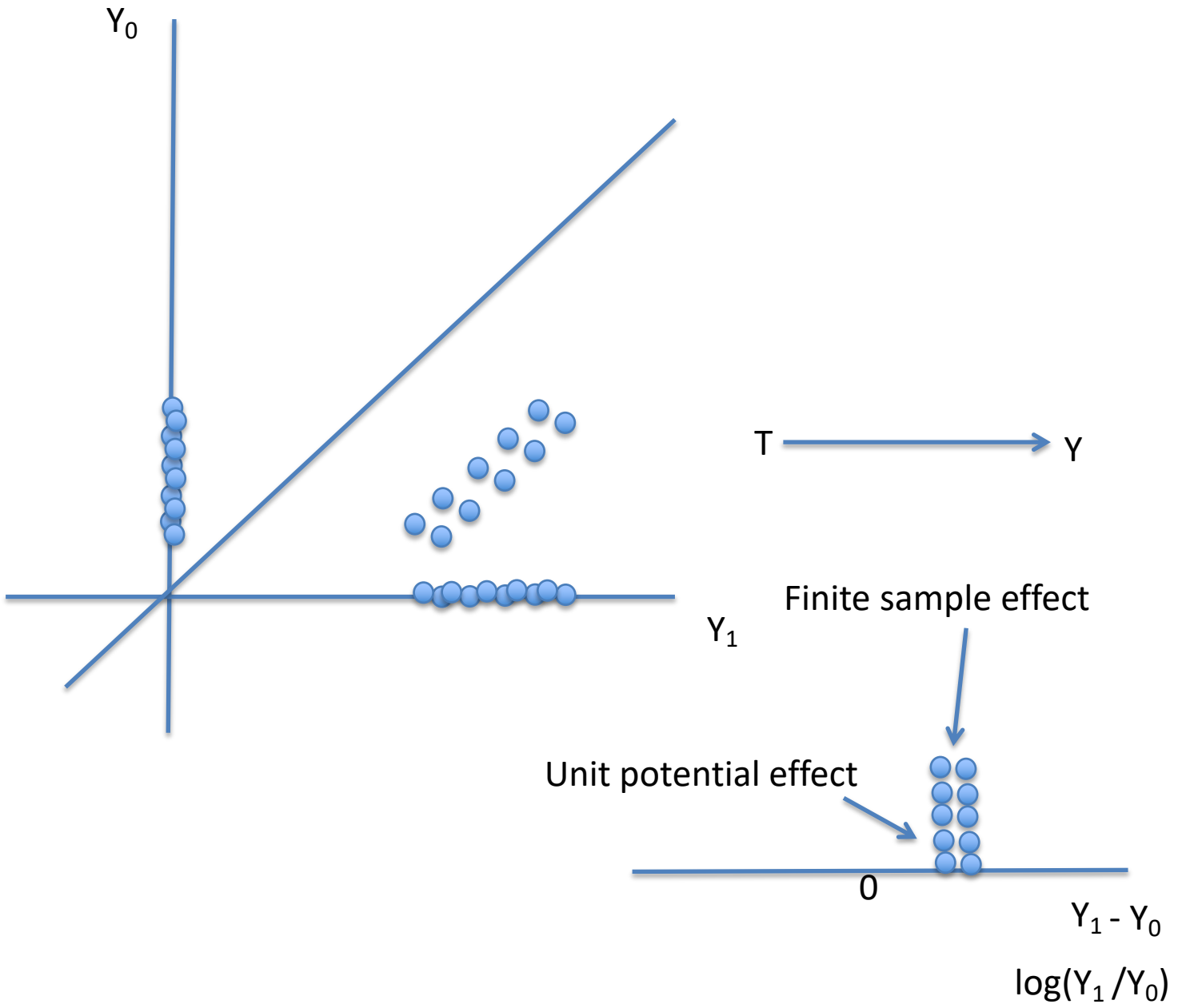
Probability is symmetric.
What is missing?
Missing counterfactual
CTRA, had you not been exposed to adversity.

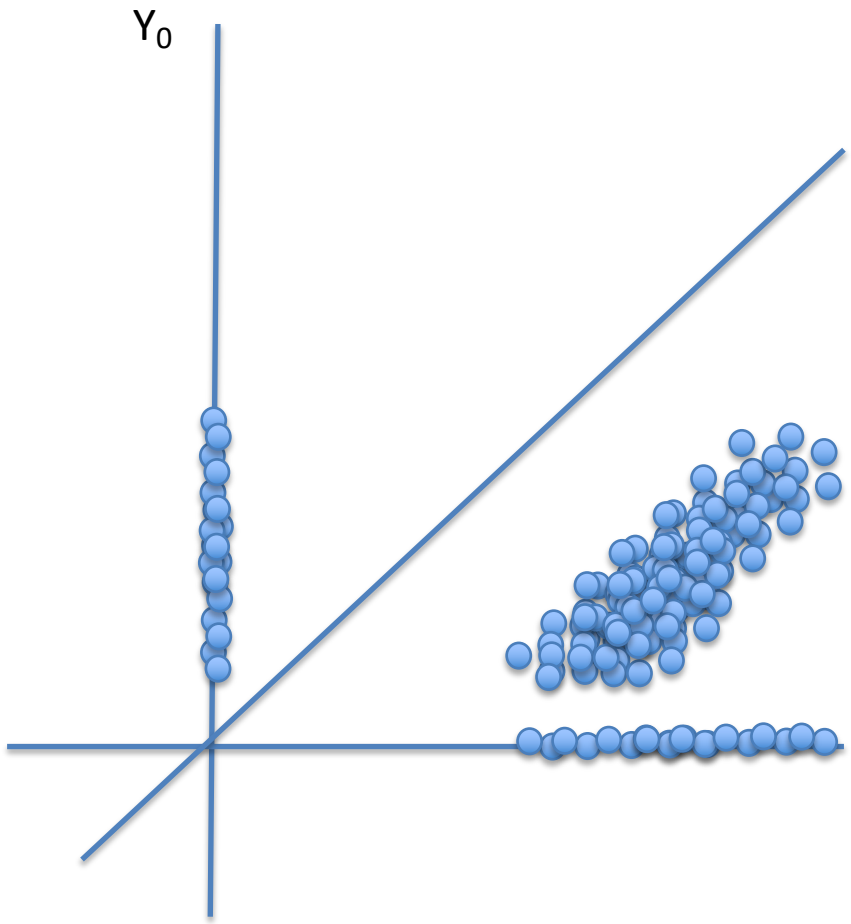


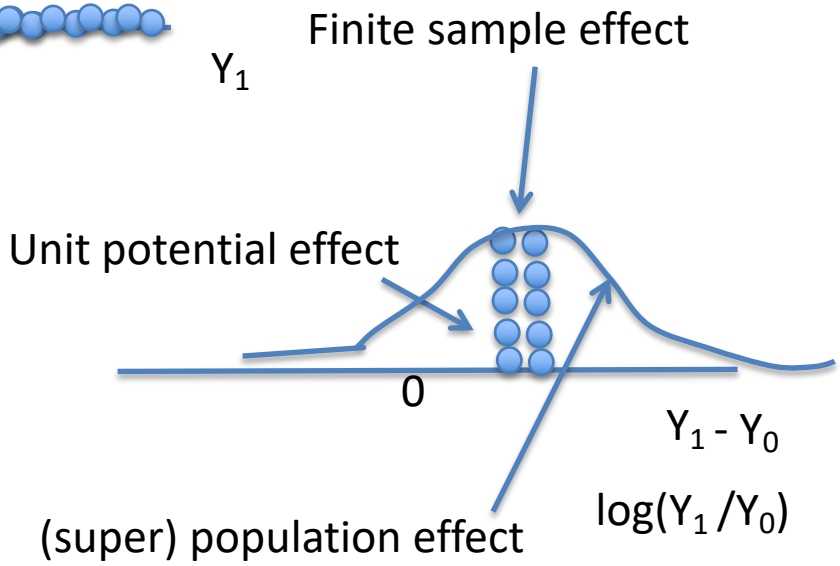
$$Y_{OBS} = Y_0 + (Y_1 - Y_0)T$$


Y_0
Y_1
T
Y
Unit potential effect
0
$Y_1 - Y_0$

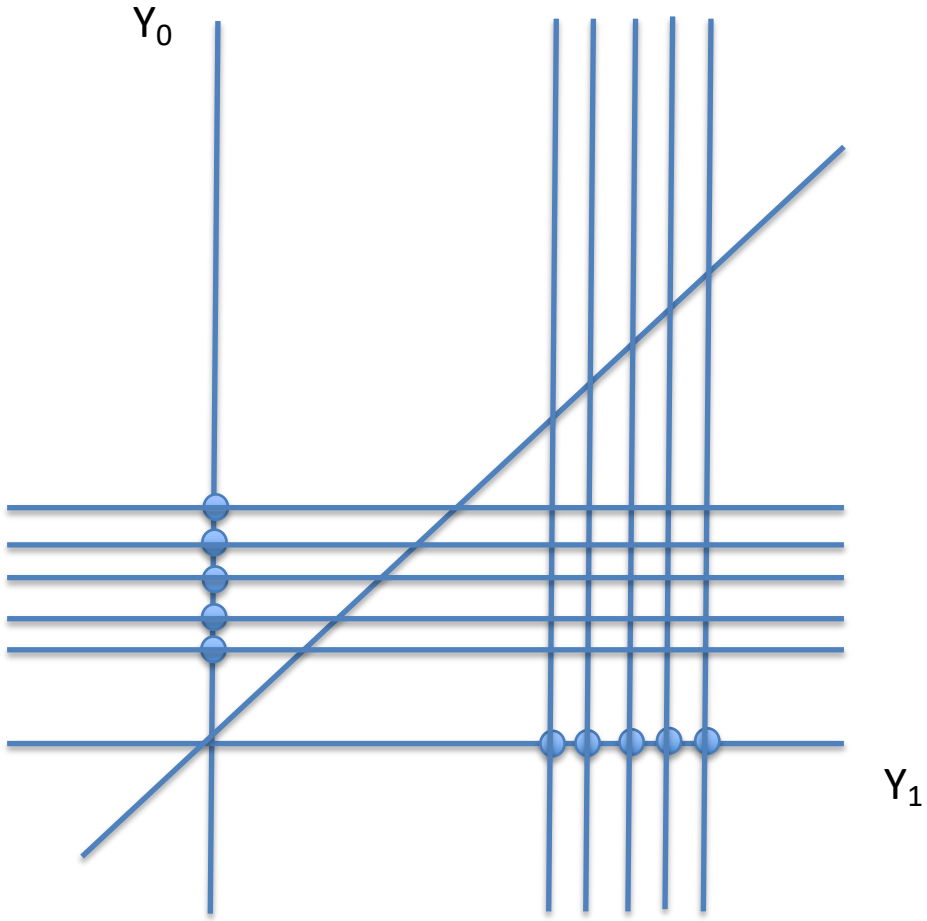

Y_0
Y_1
T
Y
Finite sample effect
Unit potential effect
0
$Y_1 - Y_0$
$\log(Y_1/Y_0)$

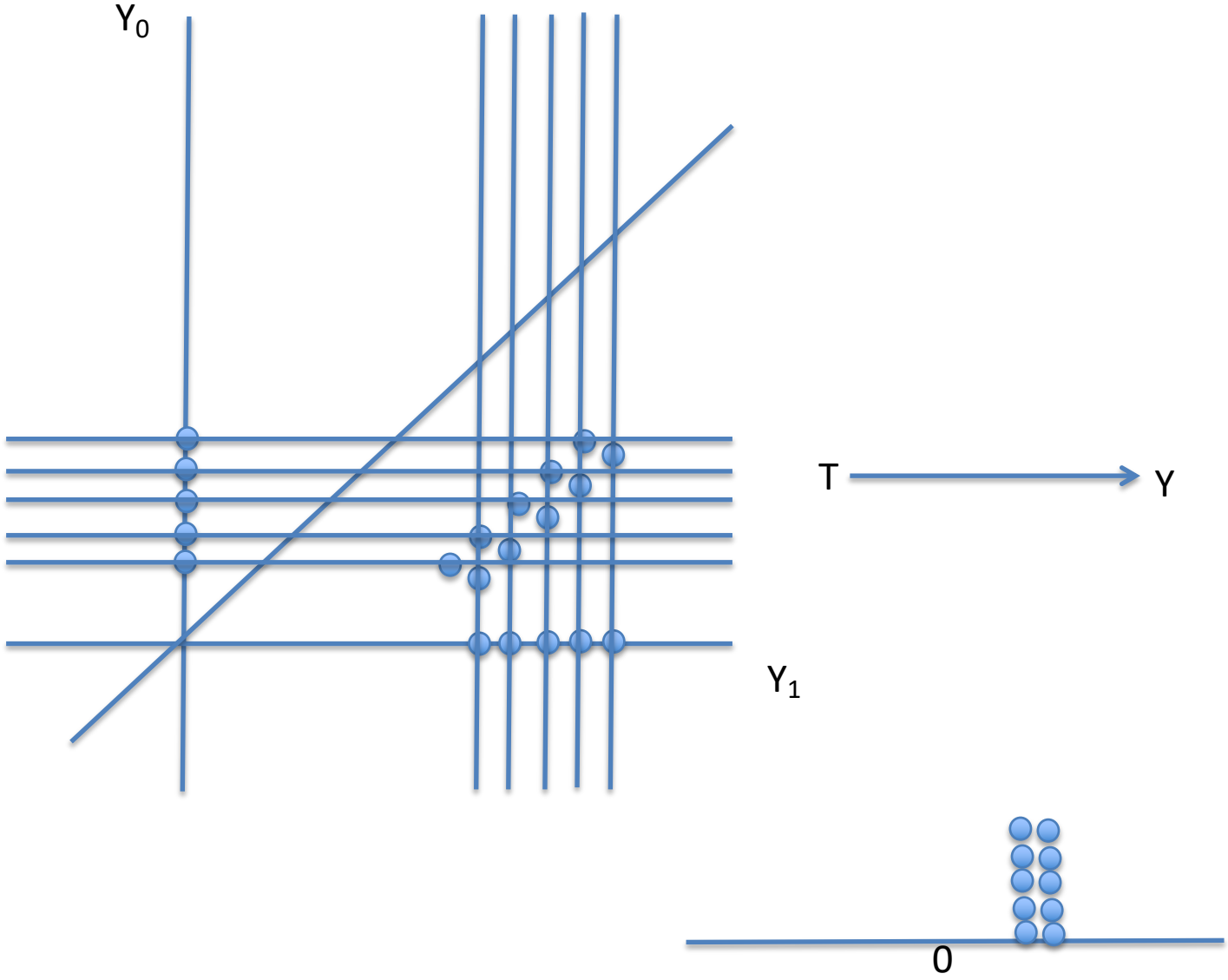

Y_0
Y_1
T
Y
Finite sample effect
Unit potential effect
0
$Y_1 - Y_0$
$\log(Y_1 / Y_0)$

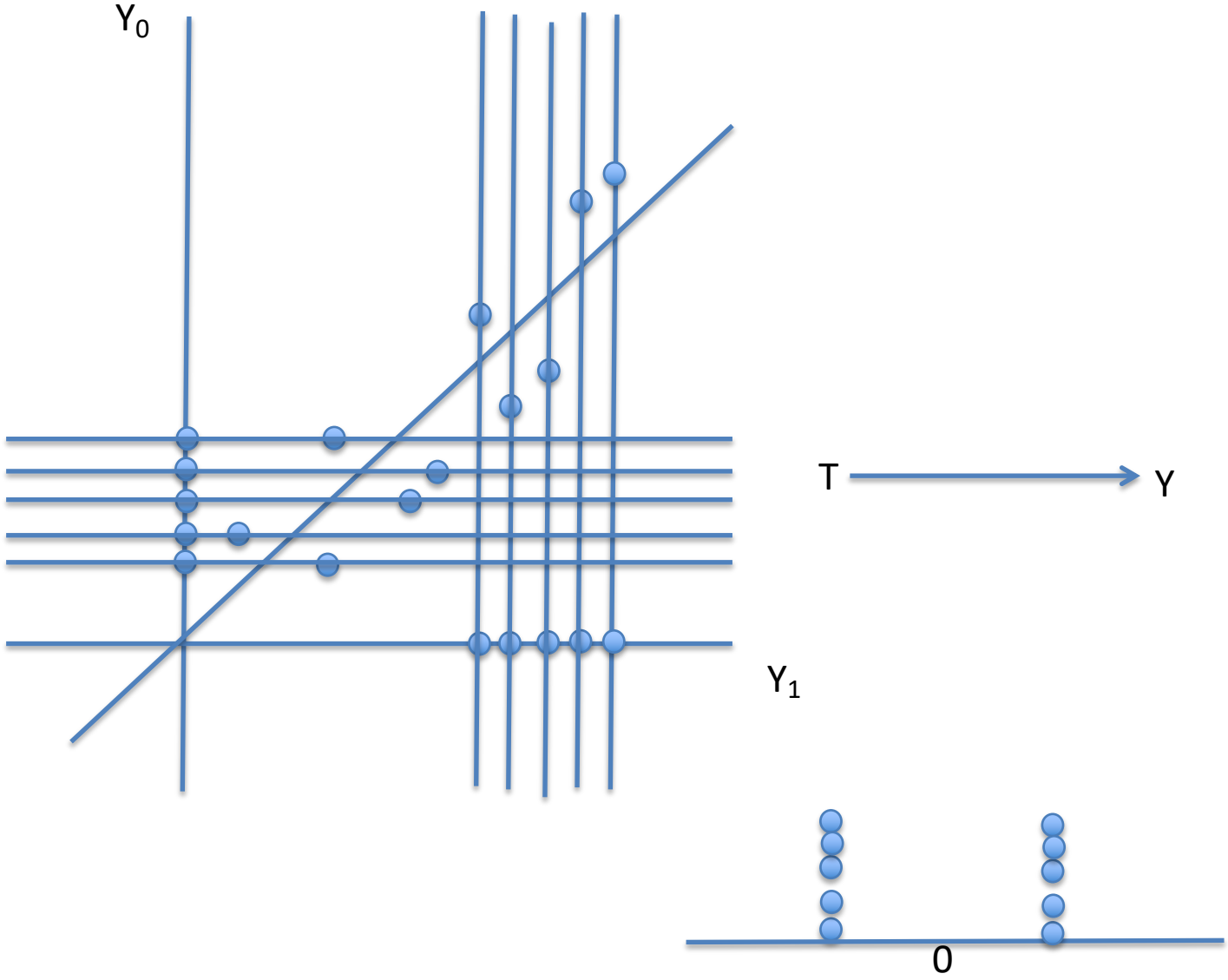


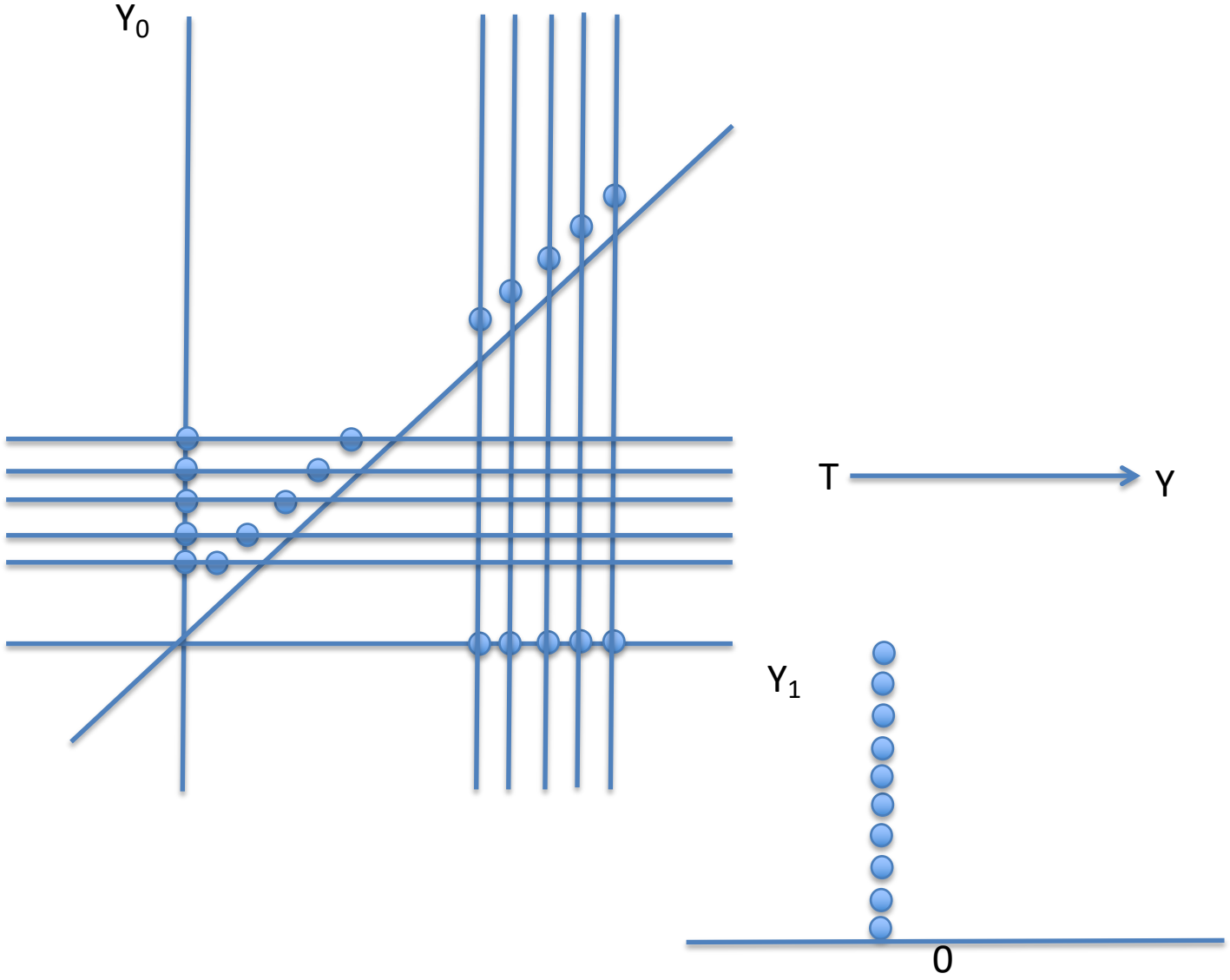


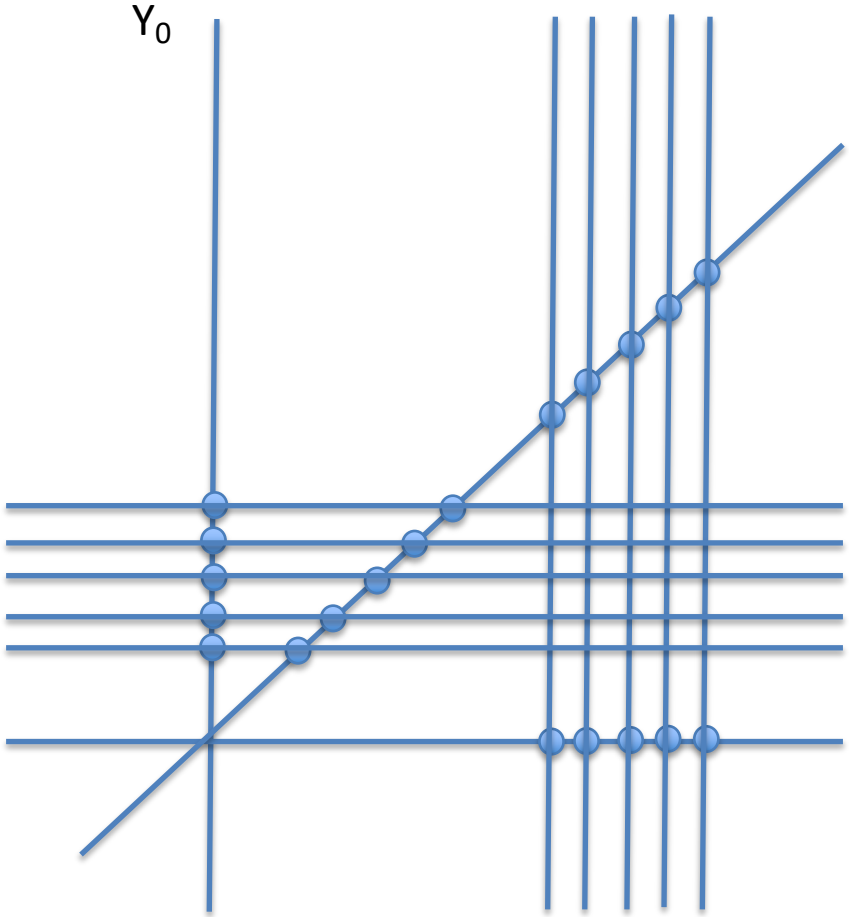
Different estimands

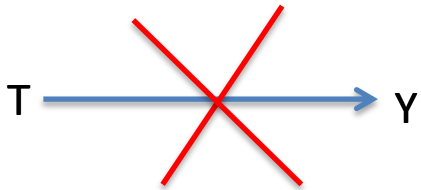

Y_0
Y_1

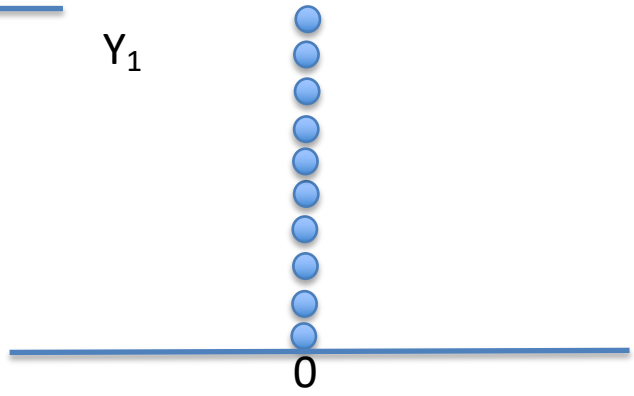

Y_0
T
Y
Y_1
0


Y_0
T
Y
Y_1
0


Y_0
T
Y
Y_1
0

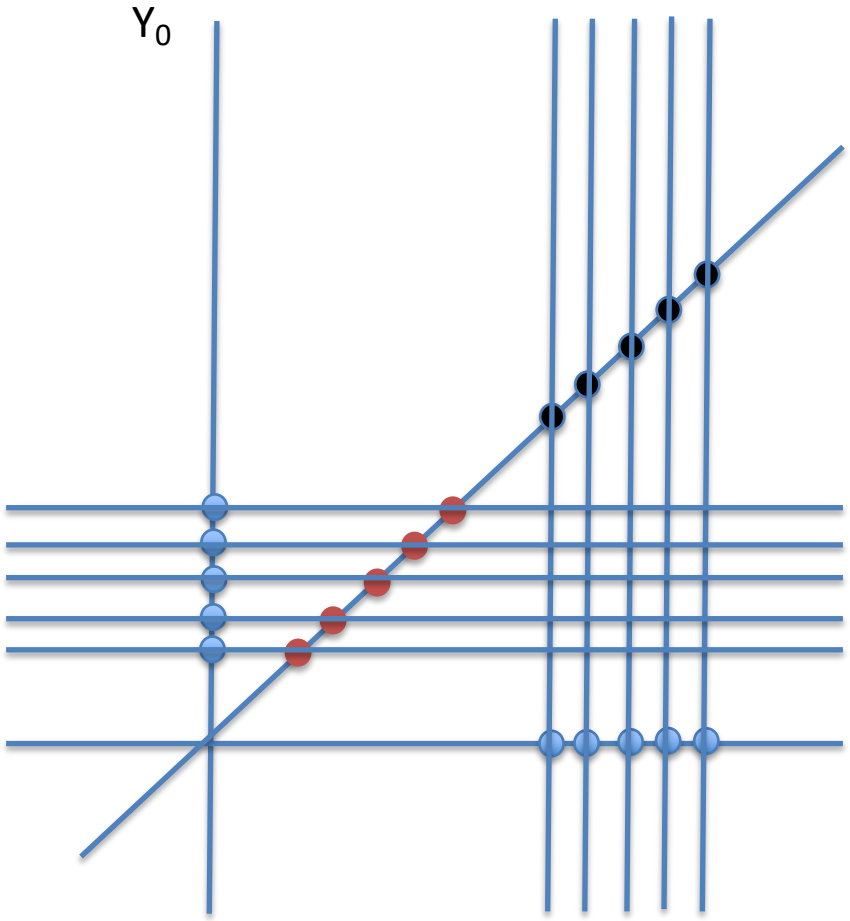


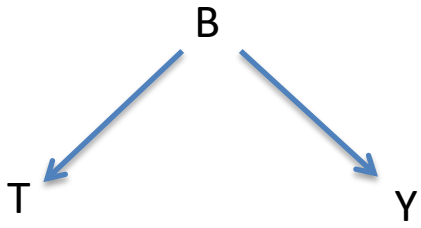


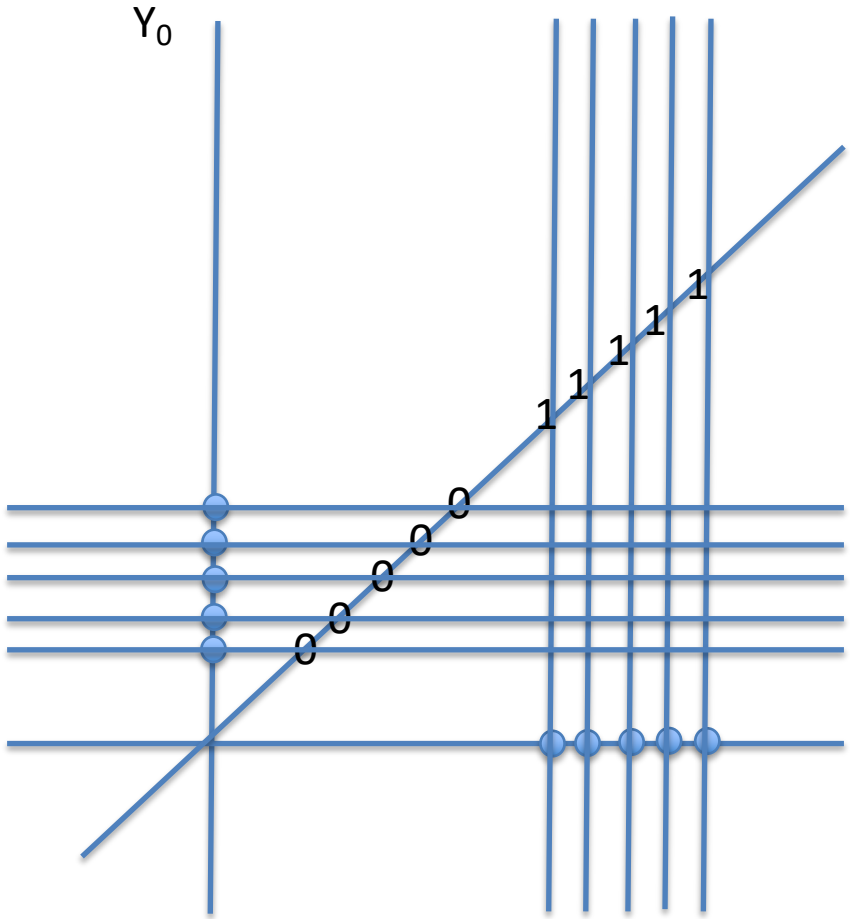


$E(Y_1-Y_0) \neq E(Y_t-Y_c)$

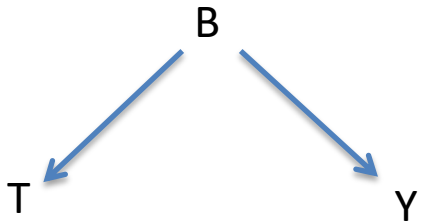
Systematic assignment selection bias: potential outcomes imbalanced across treatment groups.

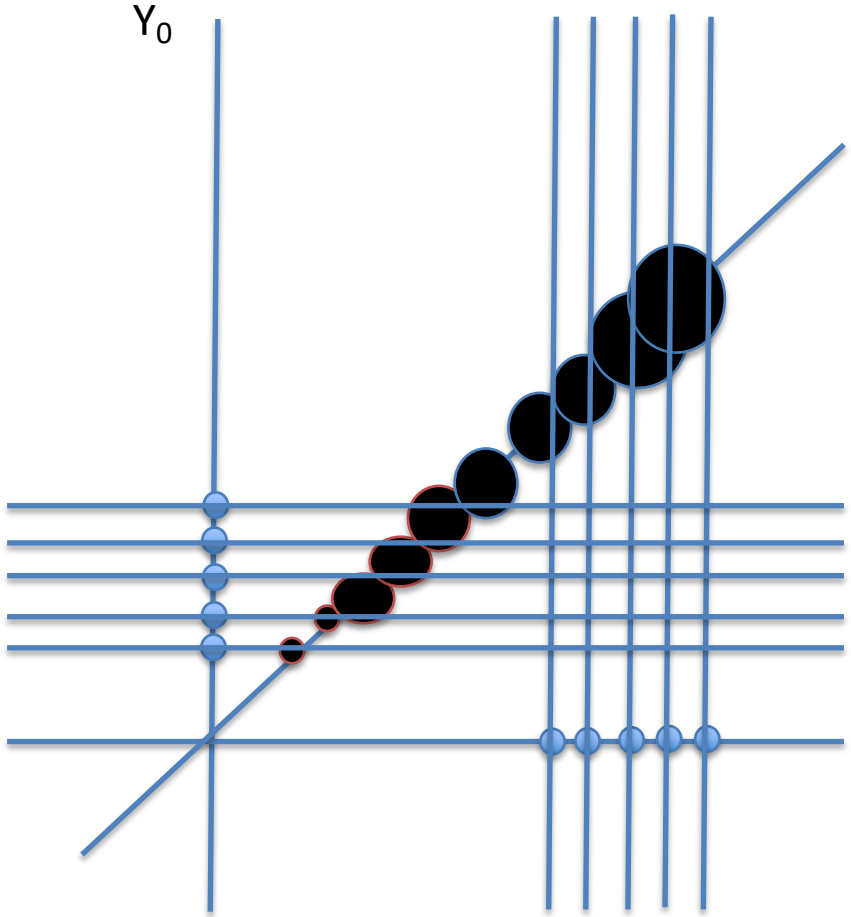


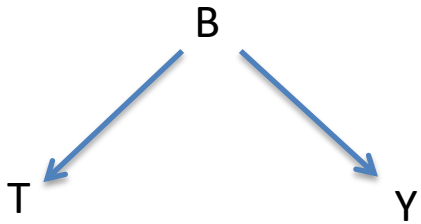




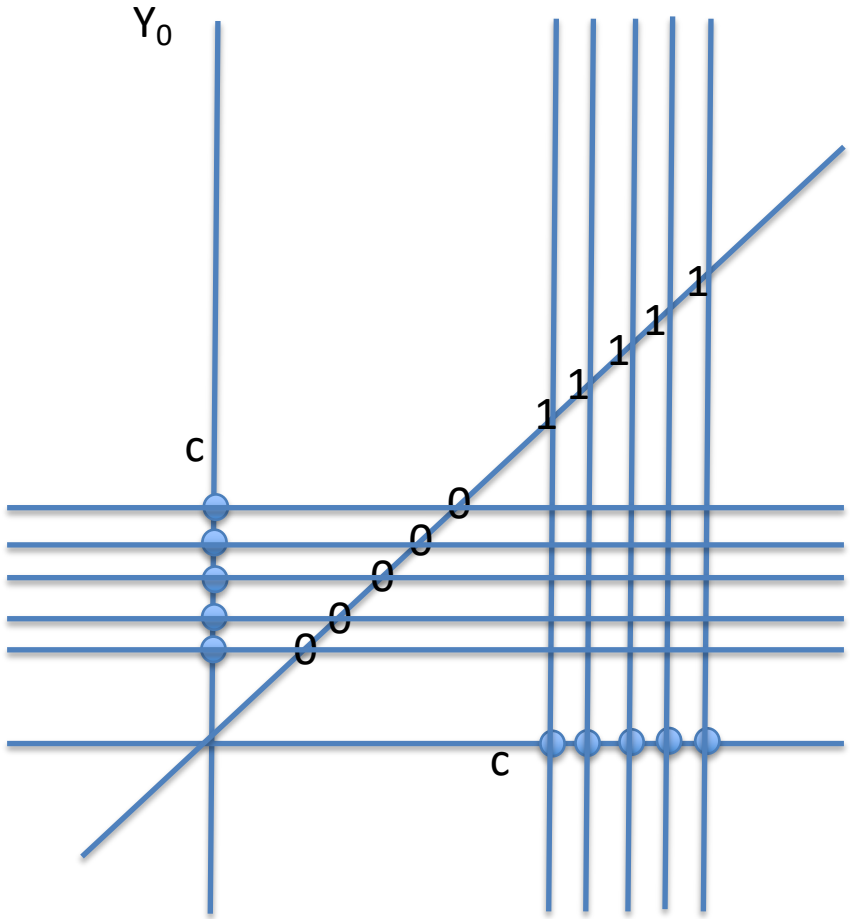
Natural propensity

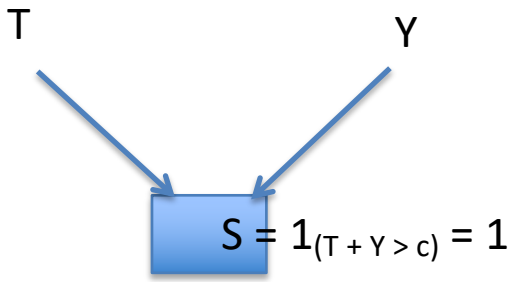


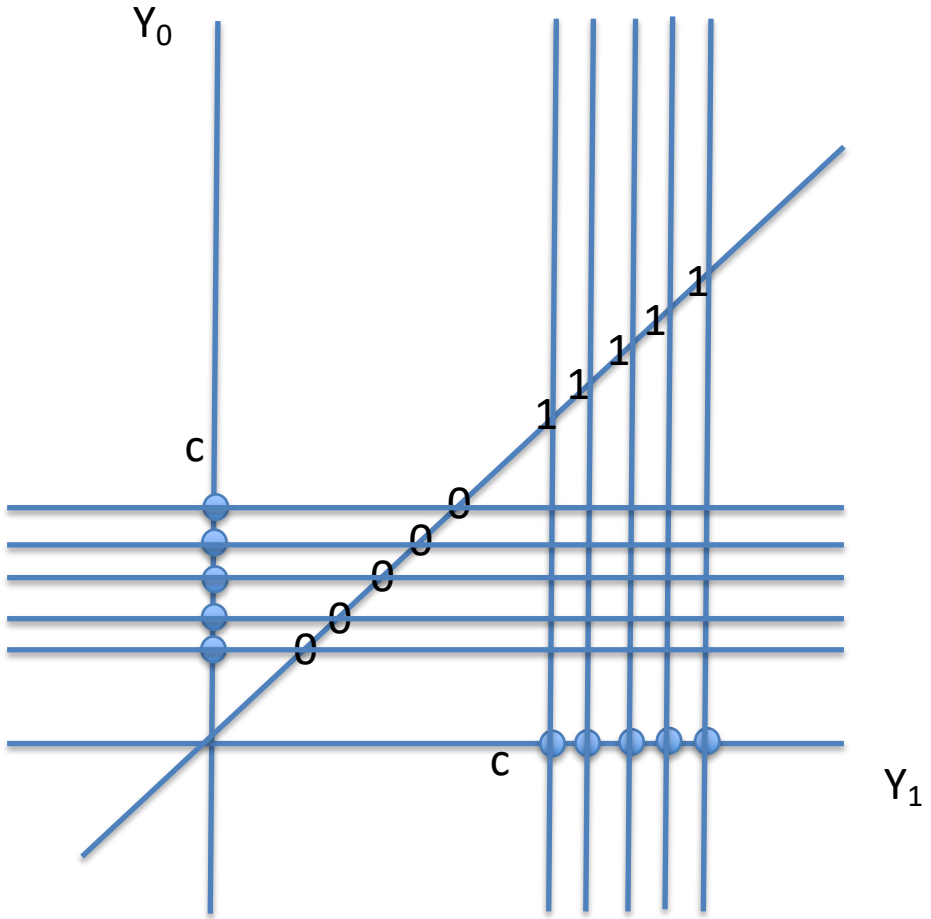




Continuous covariate.


Y_0
c
0
0
0
0
0
1
1
1
1
1
c
Y_1

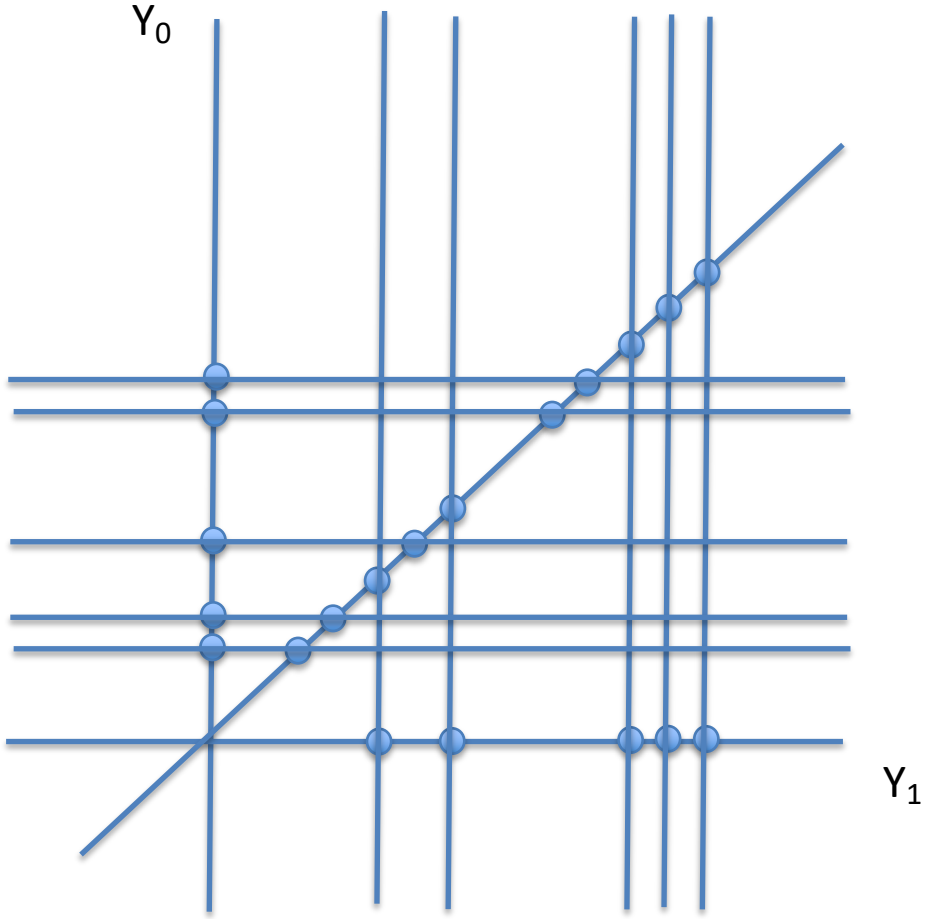

T
Y
$S = 1_{(T + Y > c)} = 1$

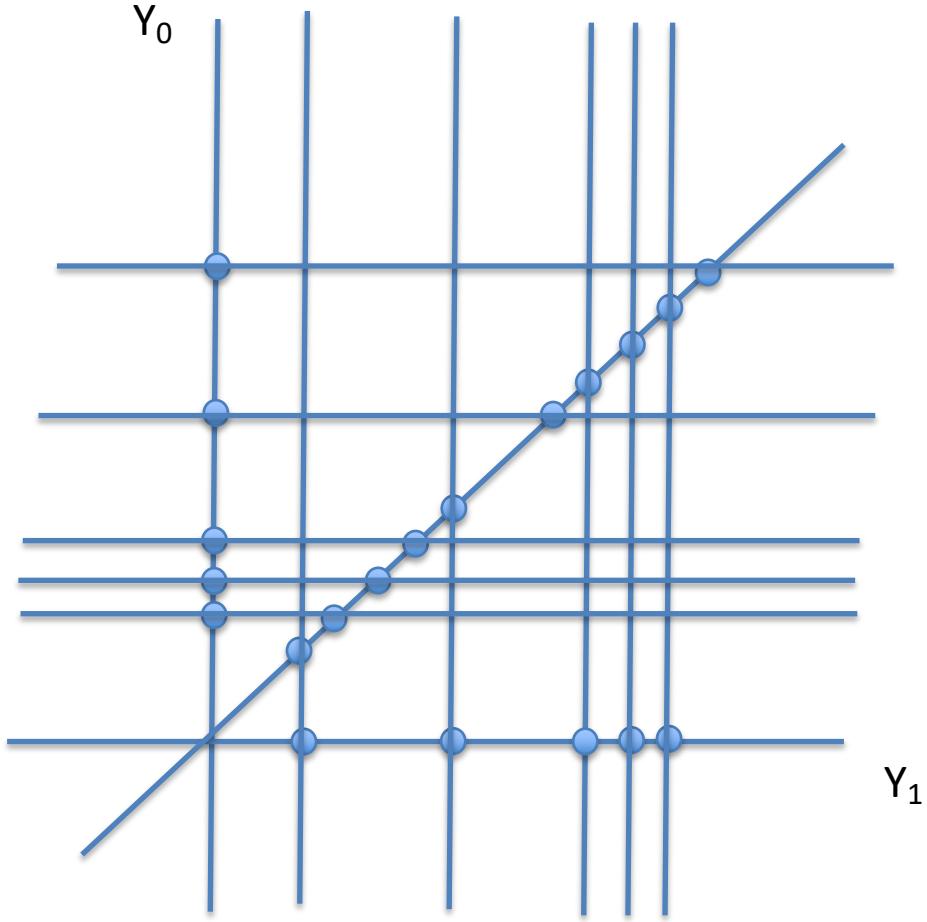


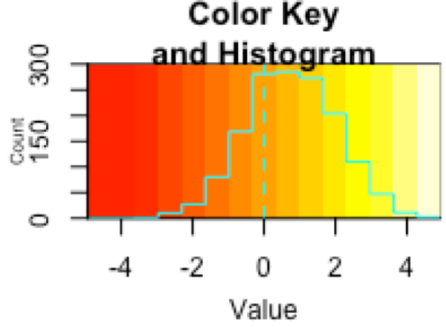
Backdoor theorem

Identification

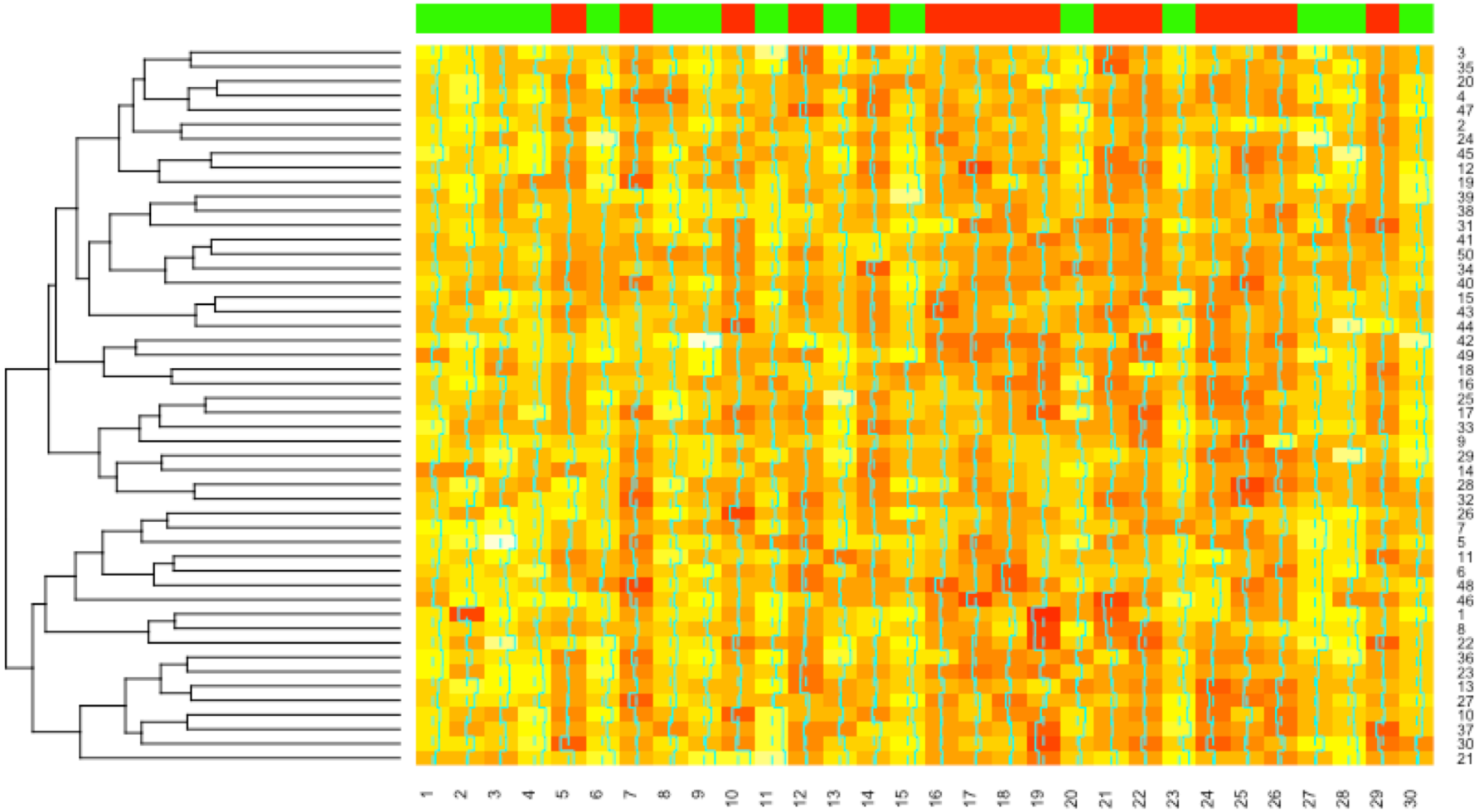
- Randomization
- Conditioning
- Instrumental variables
- Inverse-probability

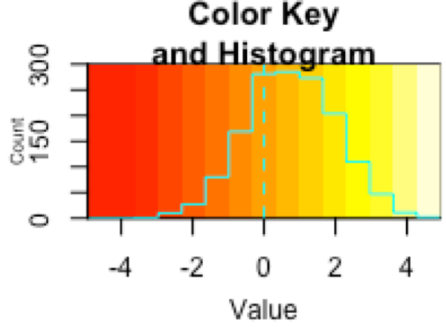

Y_0
Y_1


Y_0
Y_1

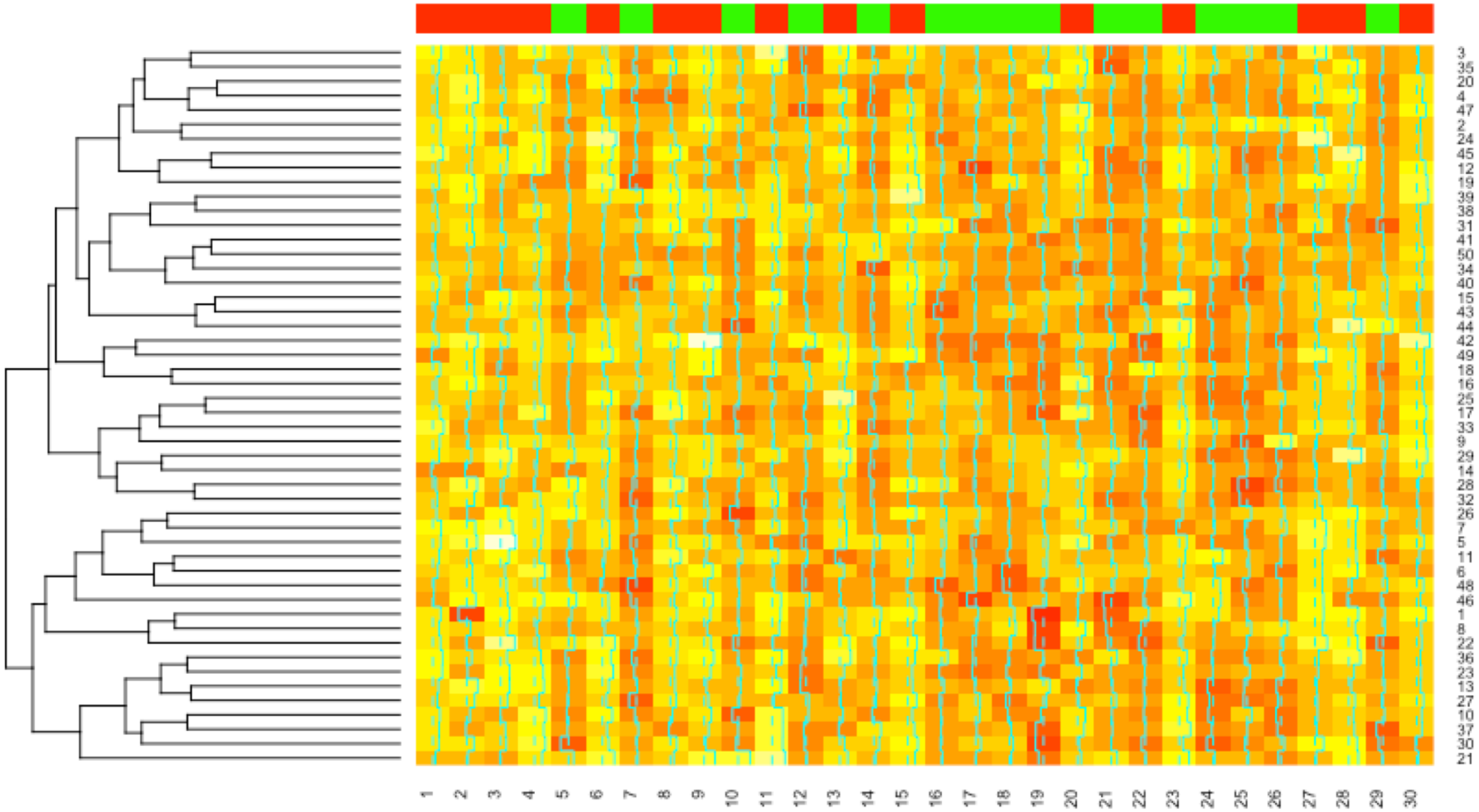


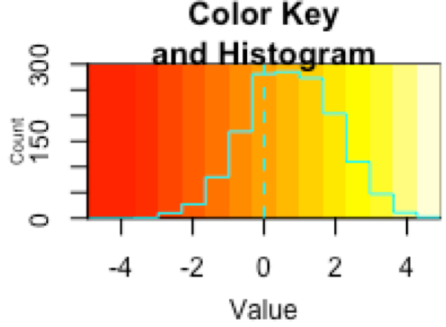
Actual observed gene expression.



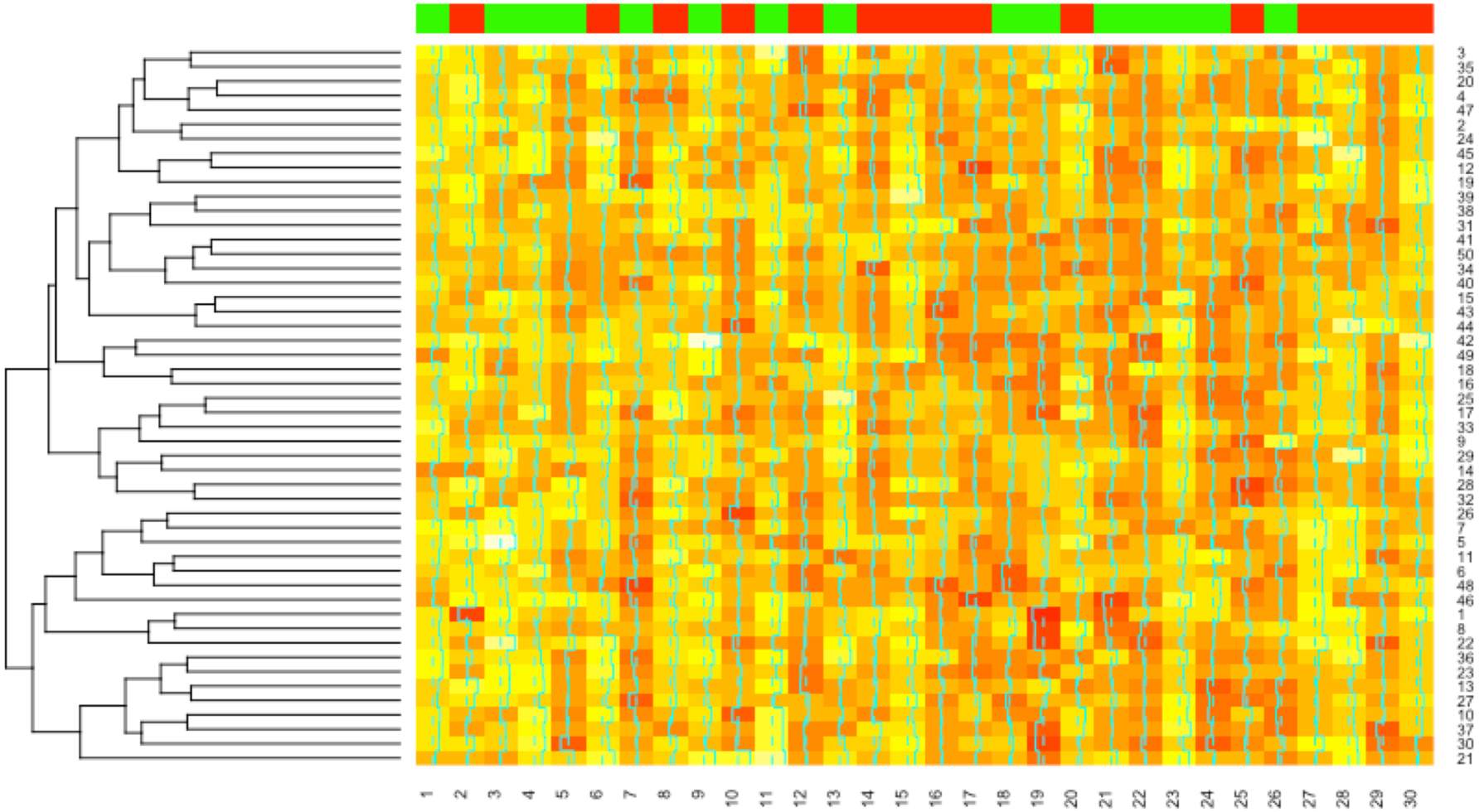


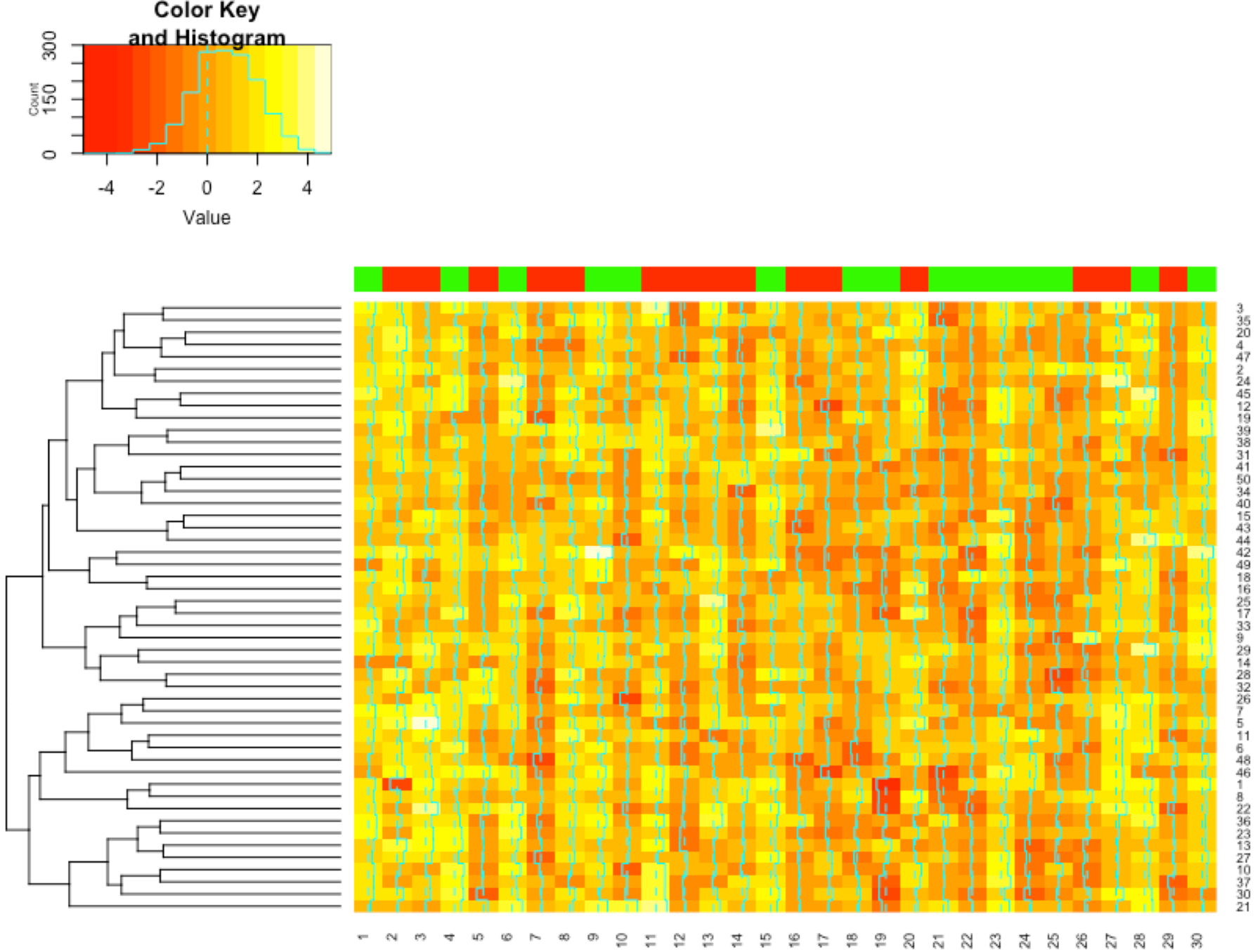
Counterfactual gene expression
under Fishers sharp imputation.

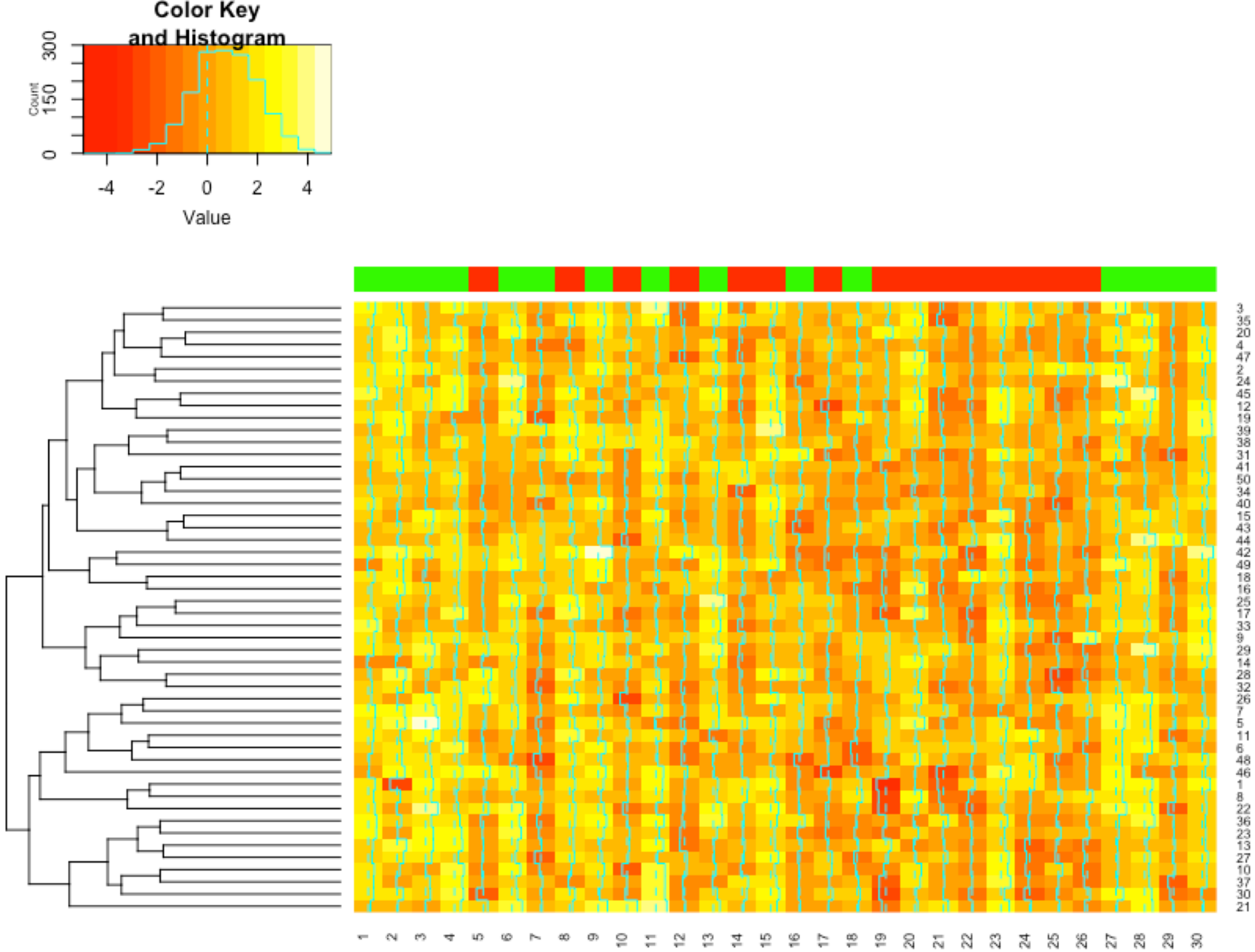


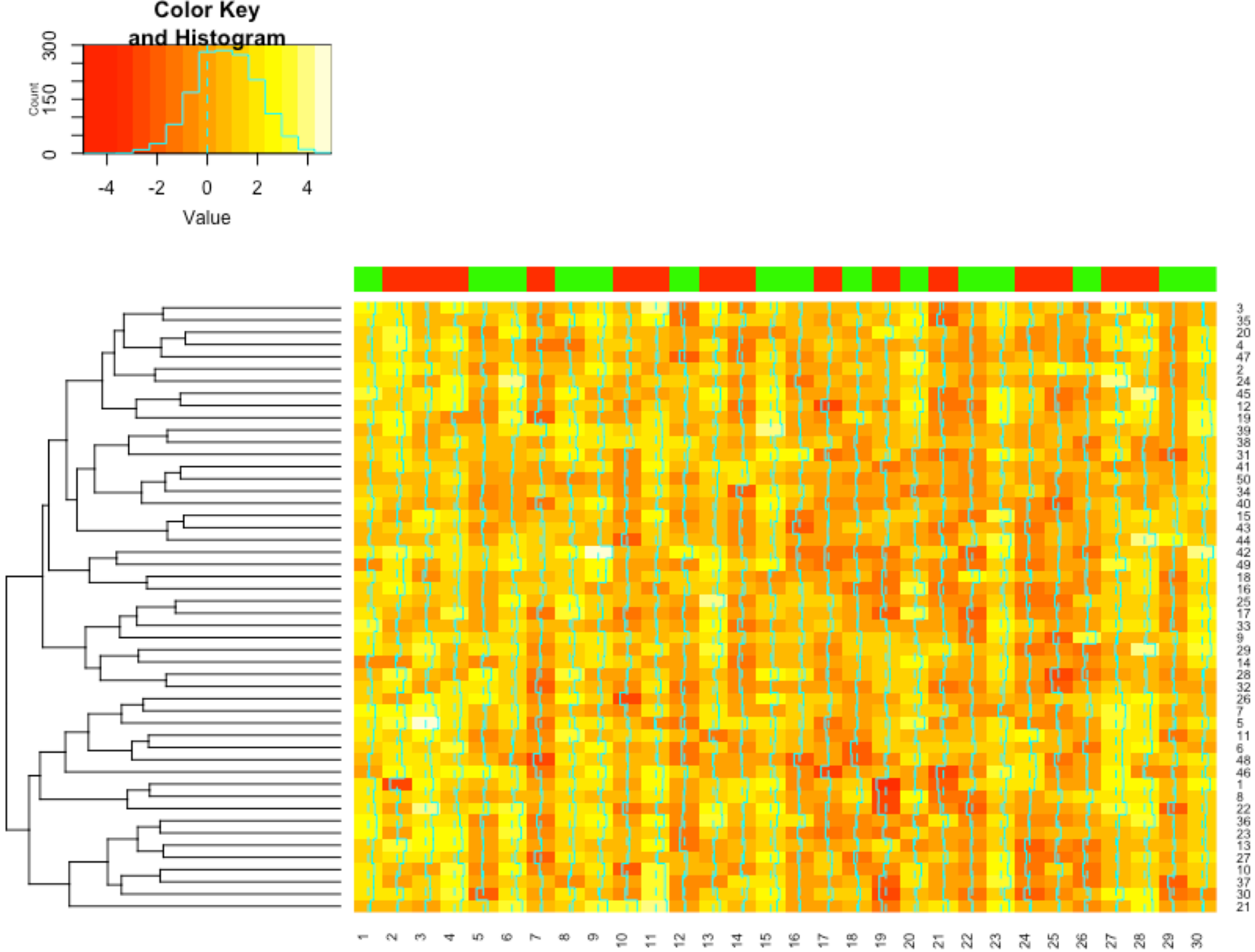


Stochastic proof by contradiction.
Gene-level (DE, MCP)
Set-level (MCP if multiple sets)
Network-level




Color Key
and Histogram
Count
300
150
0
-4
-2
0
2
4
Value
3
35
20
4
47
2
24
45
12
19
39
38
31
41
50
34
40
15
43
44
42
49
18
16
25
17
33
9
29
14
28
32
26
7
5
11
6
48
46
1
8
22
36
23
13
27
10
37
30
21
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30


Color Key
and Histogram
Count
300
150
0
-4
-2
0
2
4
Value
3
35
20
4
47
2
24
45
12
19
39
38
31
41
50
34
40
15
43
44
42
49
18
16
25
17
33
9
29
14
28
32
26
7
5
11
6
48
46
1
8
22
36
23
13
27
10
37
30
21
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30


Color Key
and Histogram
Count
300
150
0
-4
-2
0
2
4
Value
3
35
20
4
47
2
24
45
12
19
39
38
31
41
50
34
40
15
43
44
42
49
18
16
25
17
33
9
29
14
28
32
26
7
5
11
6
48
46
1
8
22
36
23
13
27
10
37
30
21
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

Variants

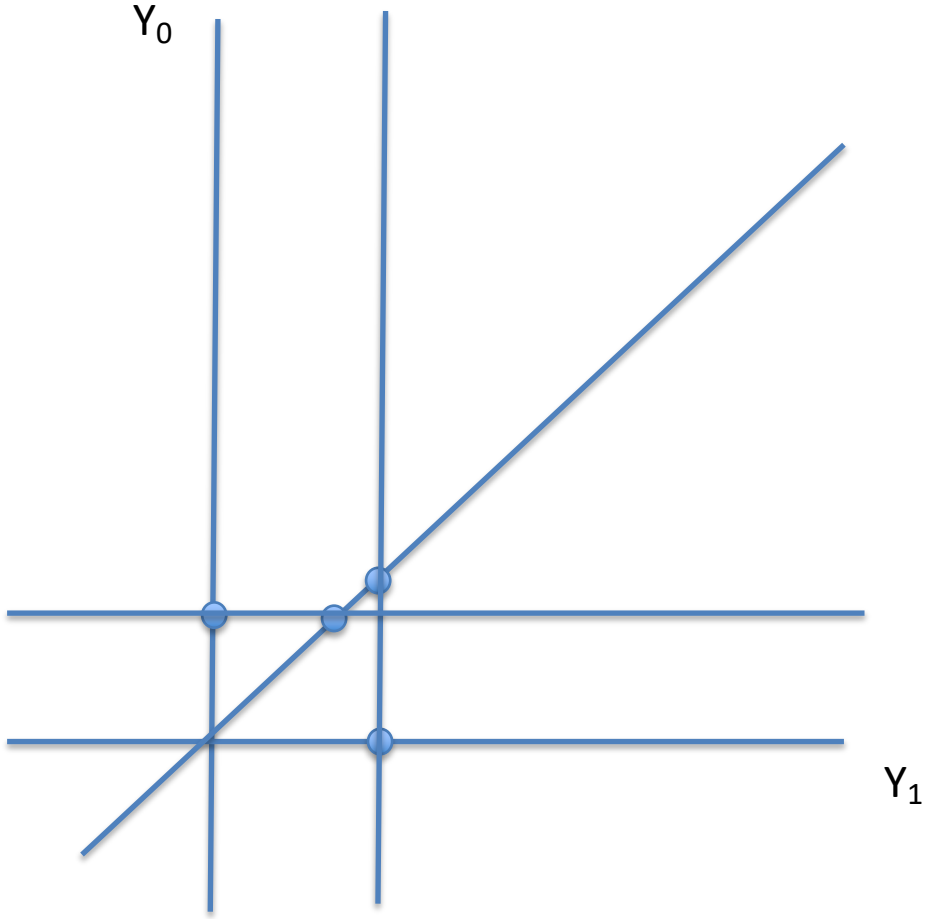
- Fisher's finite-sample
- Neyman/Rubin super-population
  - interval estimation requires additional assumptions beyond what is required for testing the causal null hypothesis
- *etc, …*

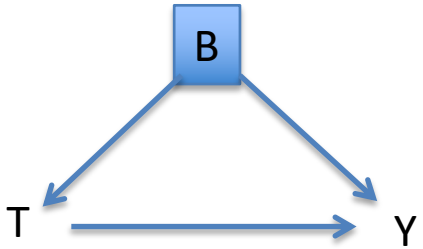
Identification

- Randomization
- Conditioning
- Instrumental variables
- Inverse-probability

$\{X : T \perp\!\!\!\perp (Y_0, Y_1) \mid X\}$

$\{X : T \perp\!\!\!\perp (Y_0, Y_1) \mid e(X)\}$


Y_0
Y_1


B
T
Y

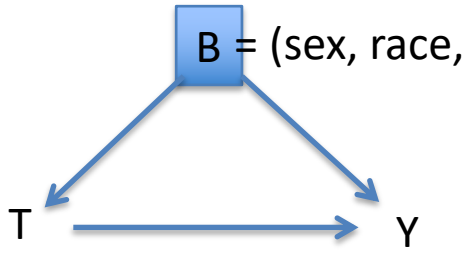


Table 1a: Sex and race

	Black	Mixed	White
F	3	1	7
M	6	0	8

Table 1b: Sex and BMI

	1st	2nd	3rd	4th
F	4	2	3	2
M	3	4	3	4

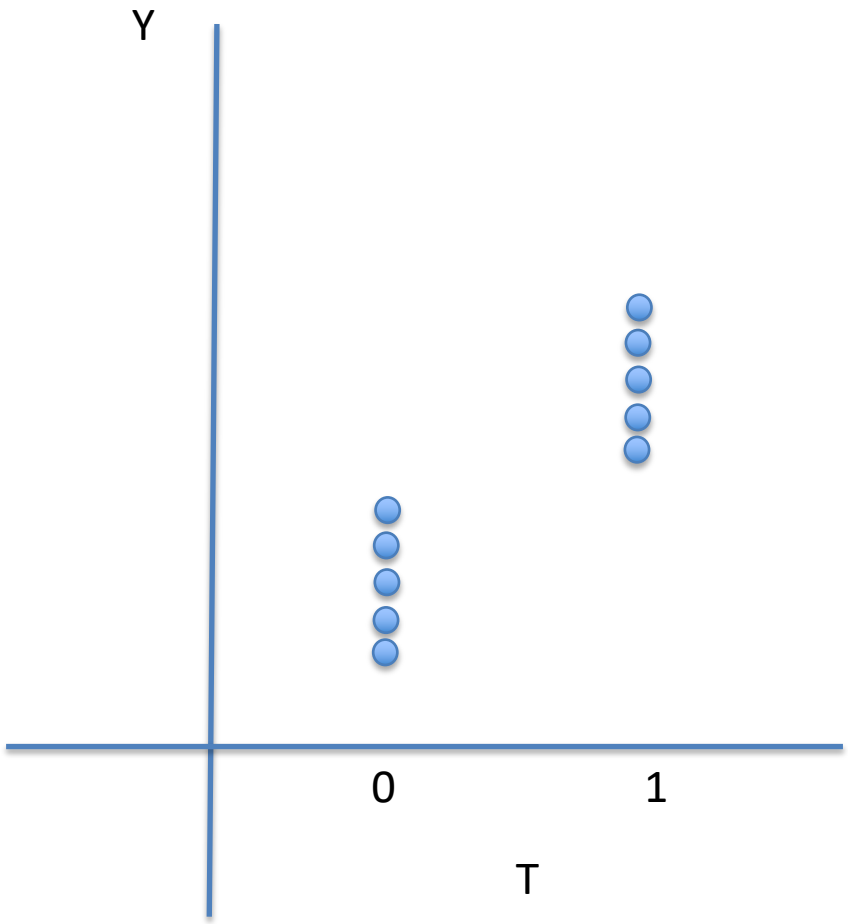
Table 1c: Race and BMI

	1st	2nd	3rd	4th
Black	2	1	2	4
Mixed	0	0	1	0
White	5	5	3	2

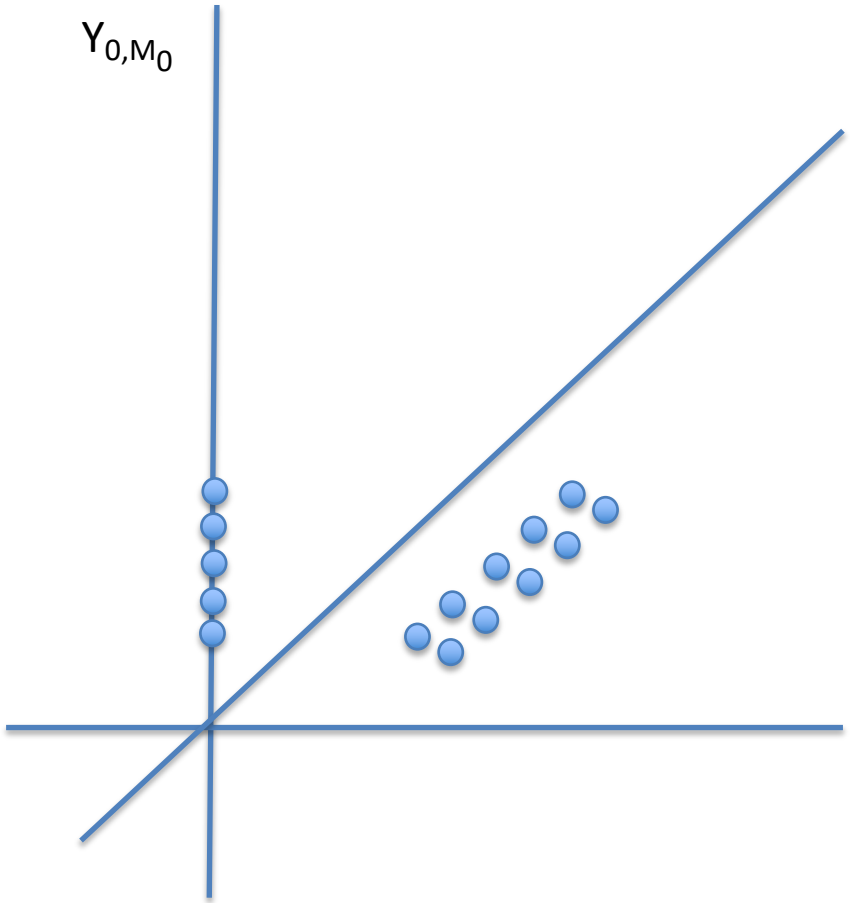
	p1	p2	p3
Inflammatory	0.787	0.556	0.801
Interferon	0.032	0.073	0.265
Antibody	0.690	0.254	0.046
All	0.033	0.084	0.285

Table 2: Our analysis of the relation between Chen, the CTRA gene set and its component subsets, with and without covariate adjustment (see Q1). This table presents p-values from a global score test on a hyperparameter in an empirical Bayesian model, and is an alternative to classical tests of a point null hypothesis against a high dimensional alternative, even when the number of genes exceeds the number of samples. This test has optimal expected power in the neighbourhood of the null hypothesis. We used a permutation null distribution which requires the assumption that there is no relationship between gene expressions on the one hand, the covariates (bmi, sex and race) and the censoring mechanism on the other hand: permuting destroys these associations. The main advantage of the permutation-based P-value is that it gives an 'exact' P-value, which is guaranteed to keep the alpha level provided enough permutations are used. This is especially useful for smaller sample sizes like ours, where we may not trust the normality of the distribution of our score statistic. Note that a significant global test does not mean that every interferon gene is associated with Chen. It means that the subjects with similar Chen have relatively similar CTRA interferon expression profile. It also means that there is potential to predict Chen from interferon gene expression.

Reality check



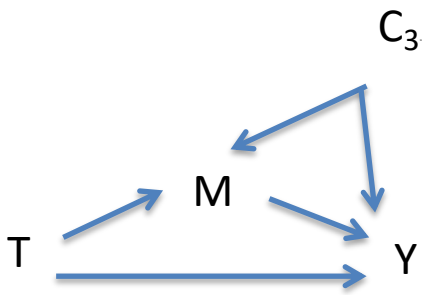
Proximal/distal
Environmental
Brain
Cognitive
Molecular – KEGG DAG/TELIS

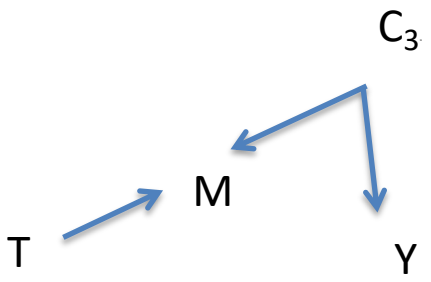


Natural variation in response of M to treatment, across units.

Exclusion restrictions.
RCT assumptions not enough.
Heterogeneity.

Exclusion restrictions.
RCT assumptions not enough.

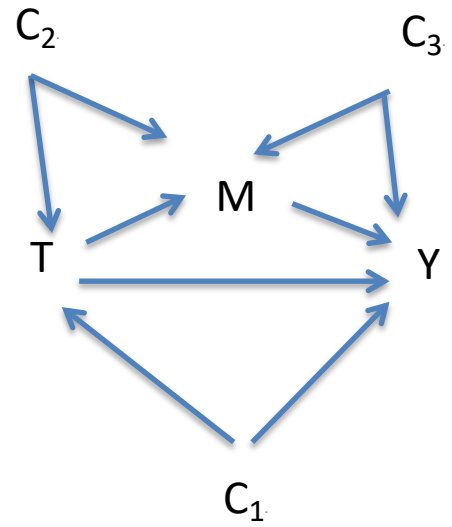


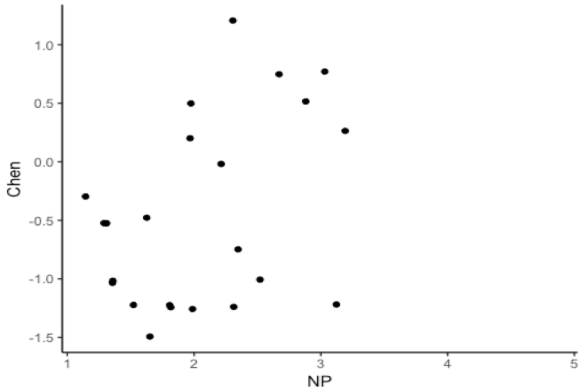


Exclusion restrictions.
Sequential ignorability.
RCT assumptions not enough.
Sequential ignorability.
Assume,
$Y = C_3$
$M = T + C_3$

Adjusted treatment effect.
Compare T = 0/1 for matched units M = 0.

$Y = C_3 = M - T = 0$
$Y = C_3 = M - T = -1$


C2
C3
M
T
Y
C1



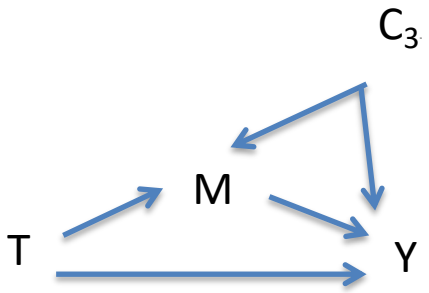
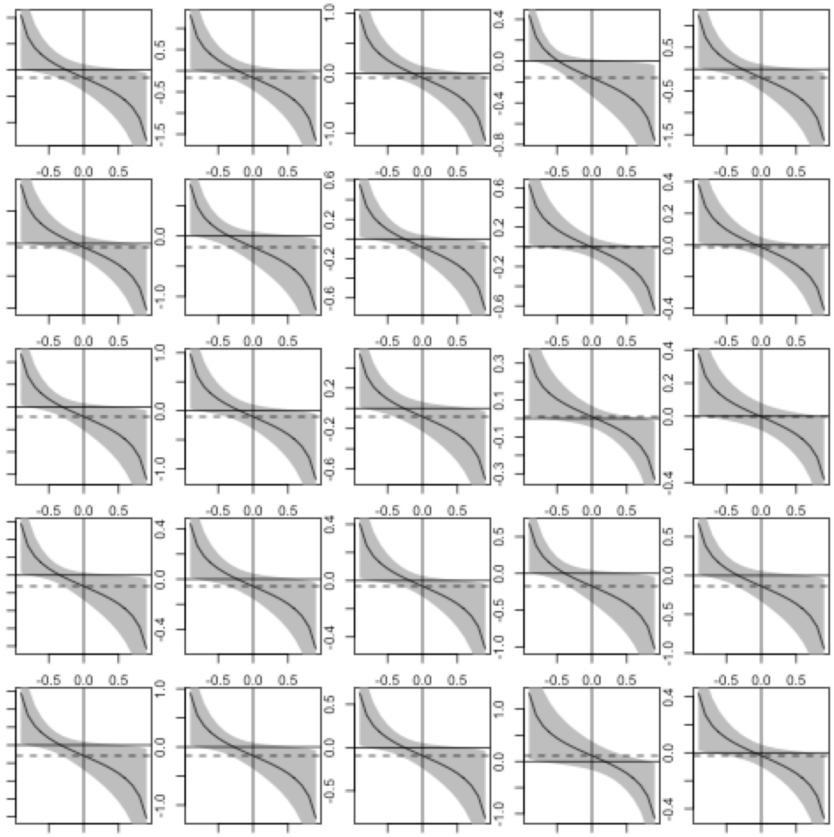


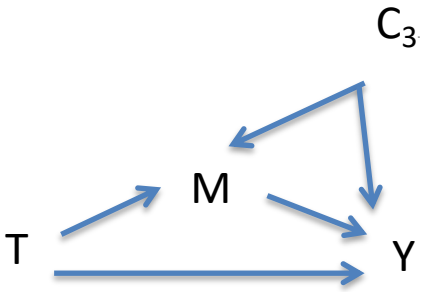
Table 3:The relation between NP, the CTRA gene set and its component subsets, with and without covariate adjustment (see Q1).

	p1	p2	p3
Inflammatory	0.168	0.950	0.872
Interferon	0.061	0.228	0.107
Antibody	0.604	0.915	0.915
All	0.030	0.200	0.091

$$
\begin{aligned}
T_i &= U_{T_i} \\
M_i &= a_M + b_{M,T}T_i + c_{M,Z}Z + U_{M_i} \\
Y_i &= a_Y + b_{Y,T}T_i + c_{Y,Z}Z + d_{Y,M}M + U_{Y_i}
\end{aligned}
$$

These equations satisfies sequential ignorability if $Cov(U_{M_i}, U_{Y_i}|Z_i) = Cov(U_{T,i}, U_{M,i}|Z_i) = Cov(U_{T_i}, U_{Y_i}|Z_i) = 0$ (Imai, Keele, and Yamamoto 2010; Pearl 2014).





ACME versus latent endogeneity: correlated innovations U_{CHEN}, U_{CTRA}

Reality check

- Mediation in feedback systems
- Beyond conditioning
  - Omitted confounds
  - No treatment variation within-strata
  - Artifactual selection (marginal structural models, etc).
- Methods from life-course epidemiology, statistics, engineering….
  - Marginal structural causal models
  - Causal structure learning.

Thank you