Causal and statistical inference

What is missing, what is desired?

Counterfactuals?

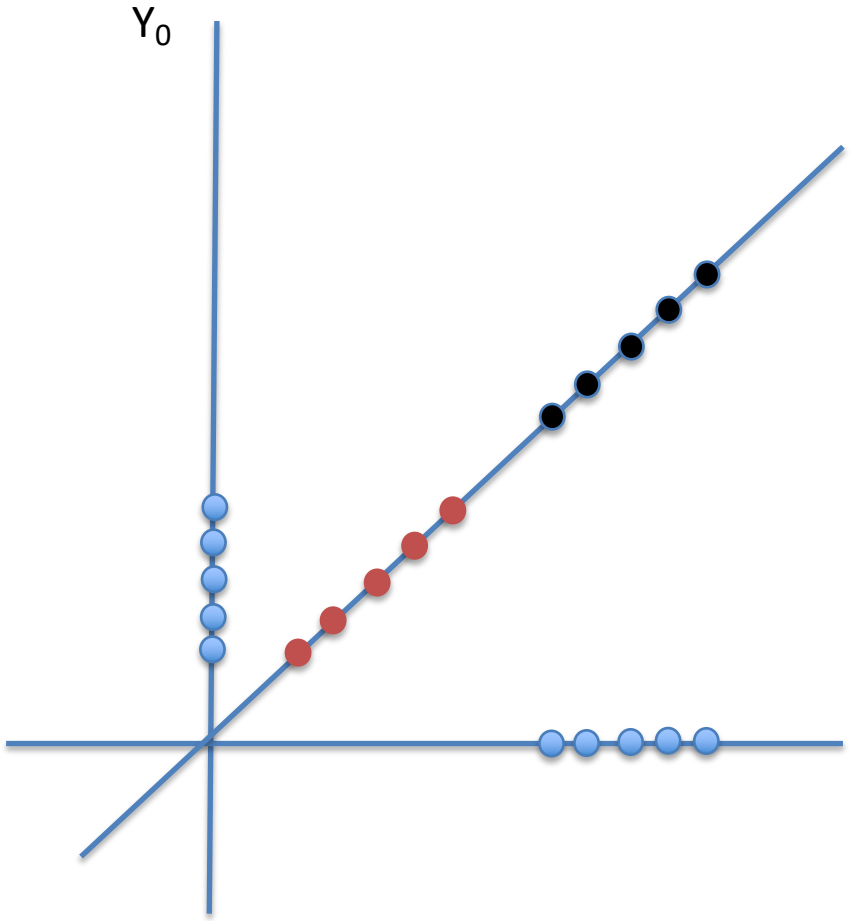
Missingness	No	Yes
No	Complete information	Finite sample causal inference (Fisher)
Yes	Standard statistical inference “to the population”	Sample & assignment selection (Neyman, Rubin)

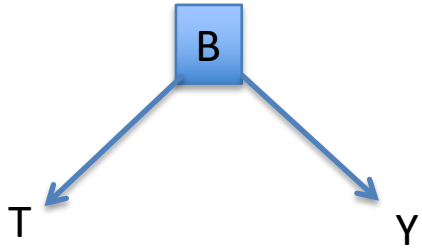
Units?

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

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

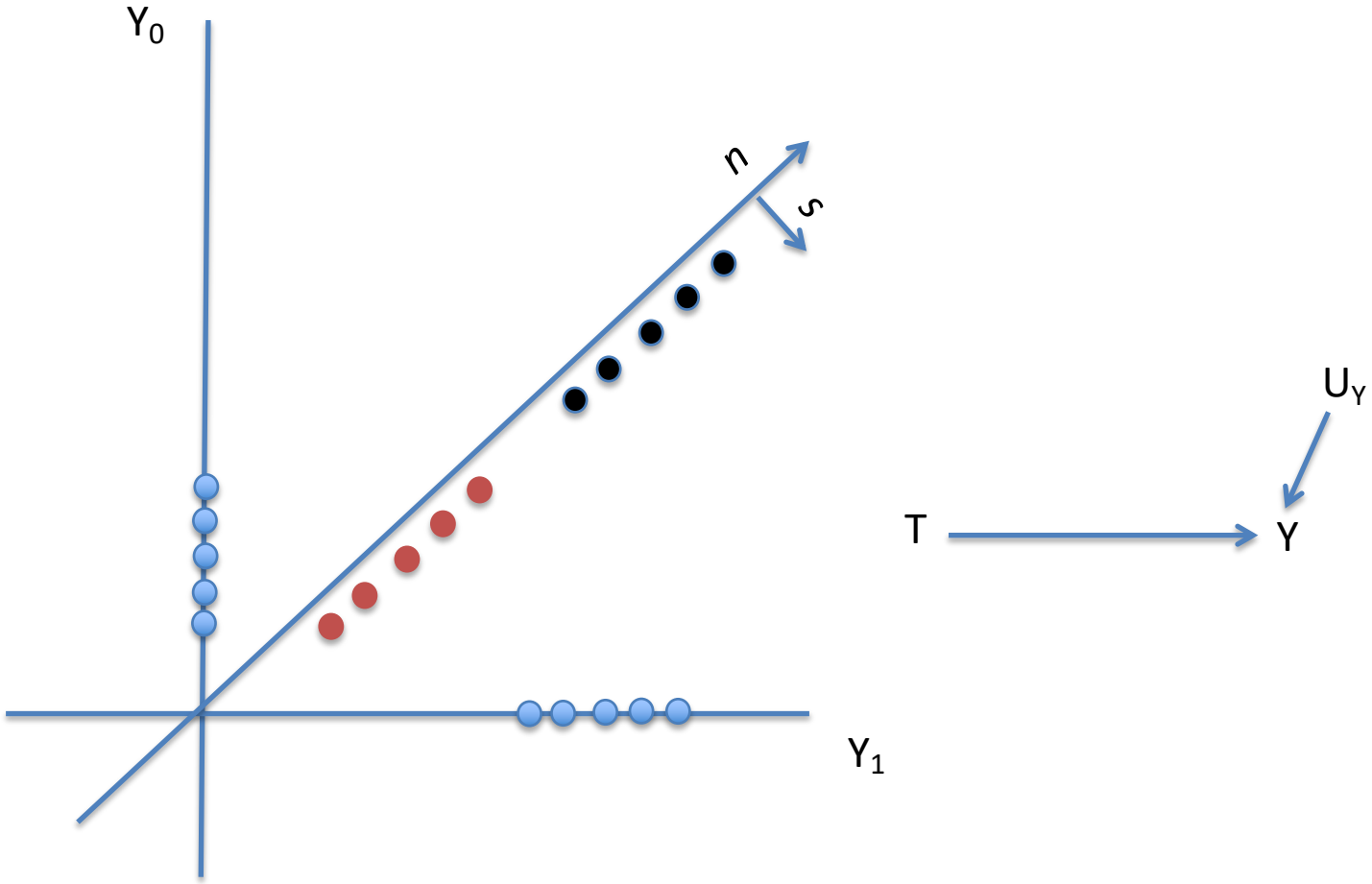
$\{X : E \perp\!\!\!\perp (Y_0, Y_1) \mid e(X)\}$





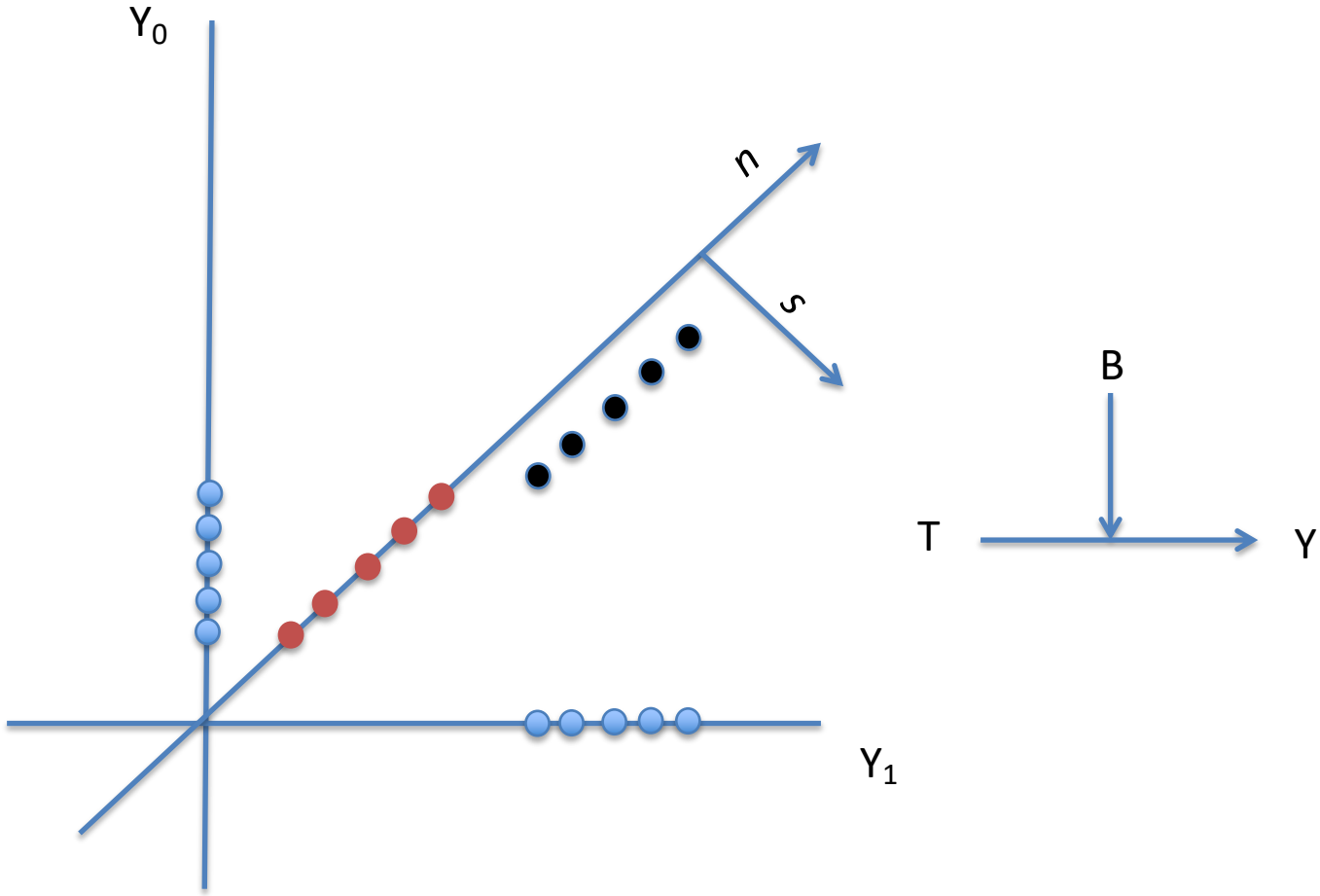
blocking can (e.g. regression, matching,…)

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



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

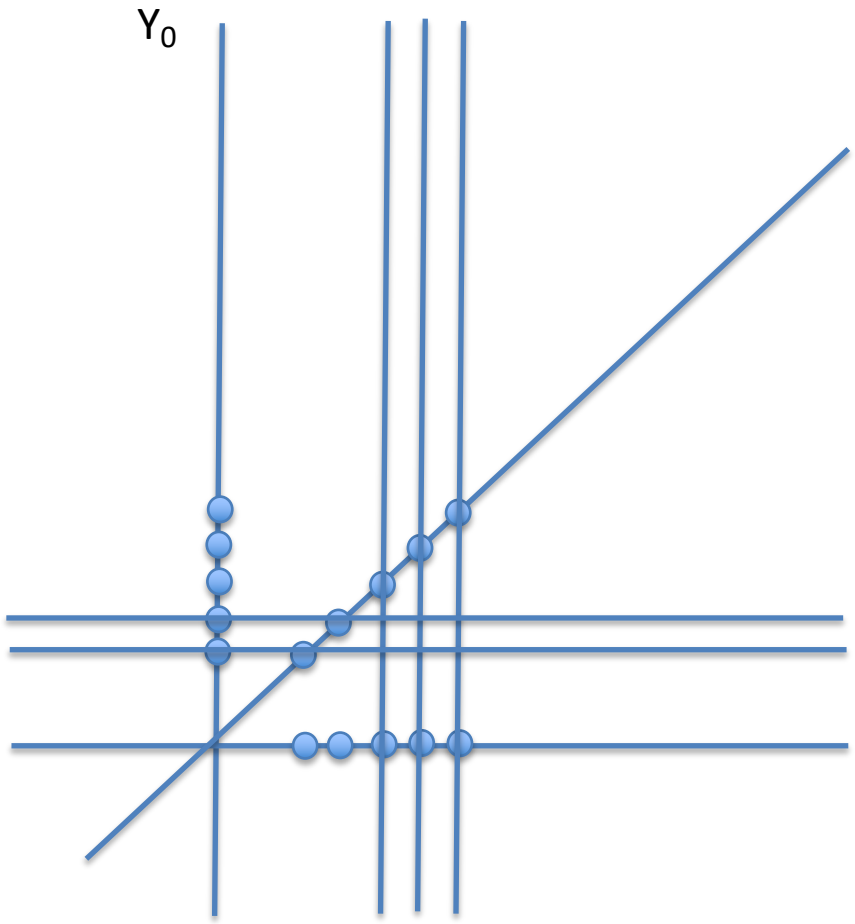
- *reduce bias*
- ***reduce noise***
- *explain heterogeneity.*

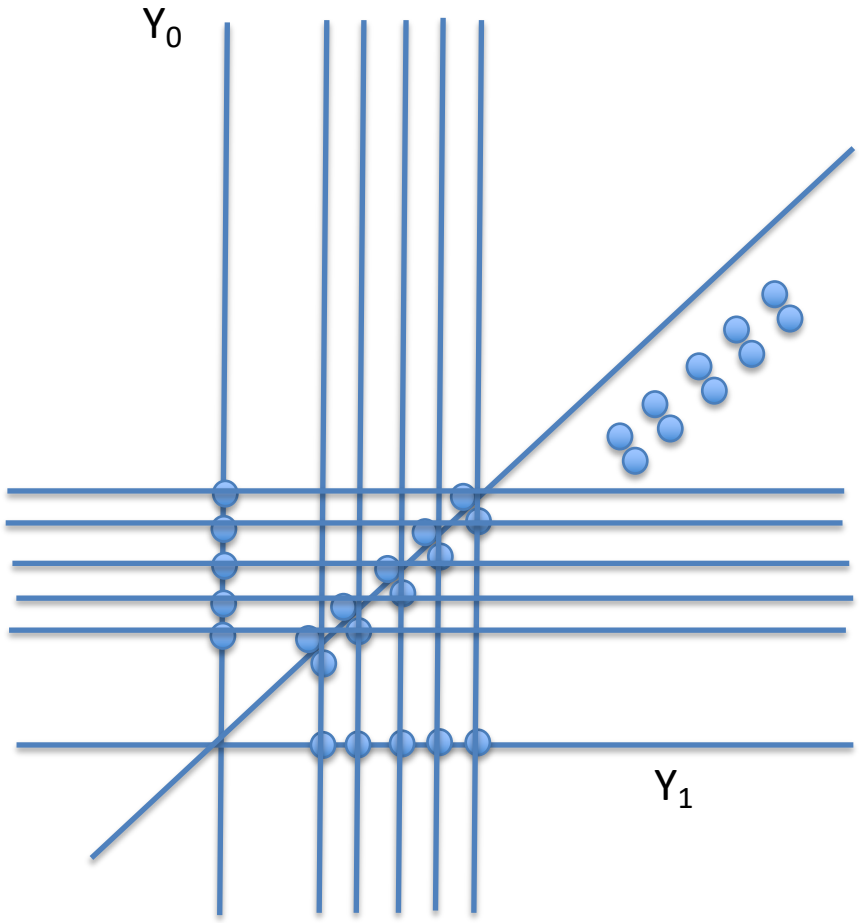


blocking can (e.g. regression, matching,…)

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

Bias (sampling & assignment)



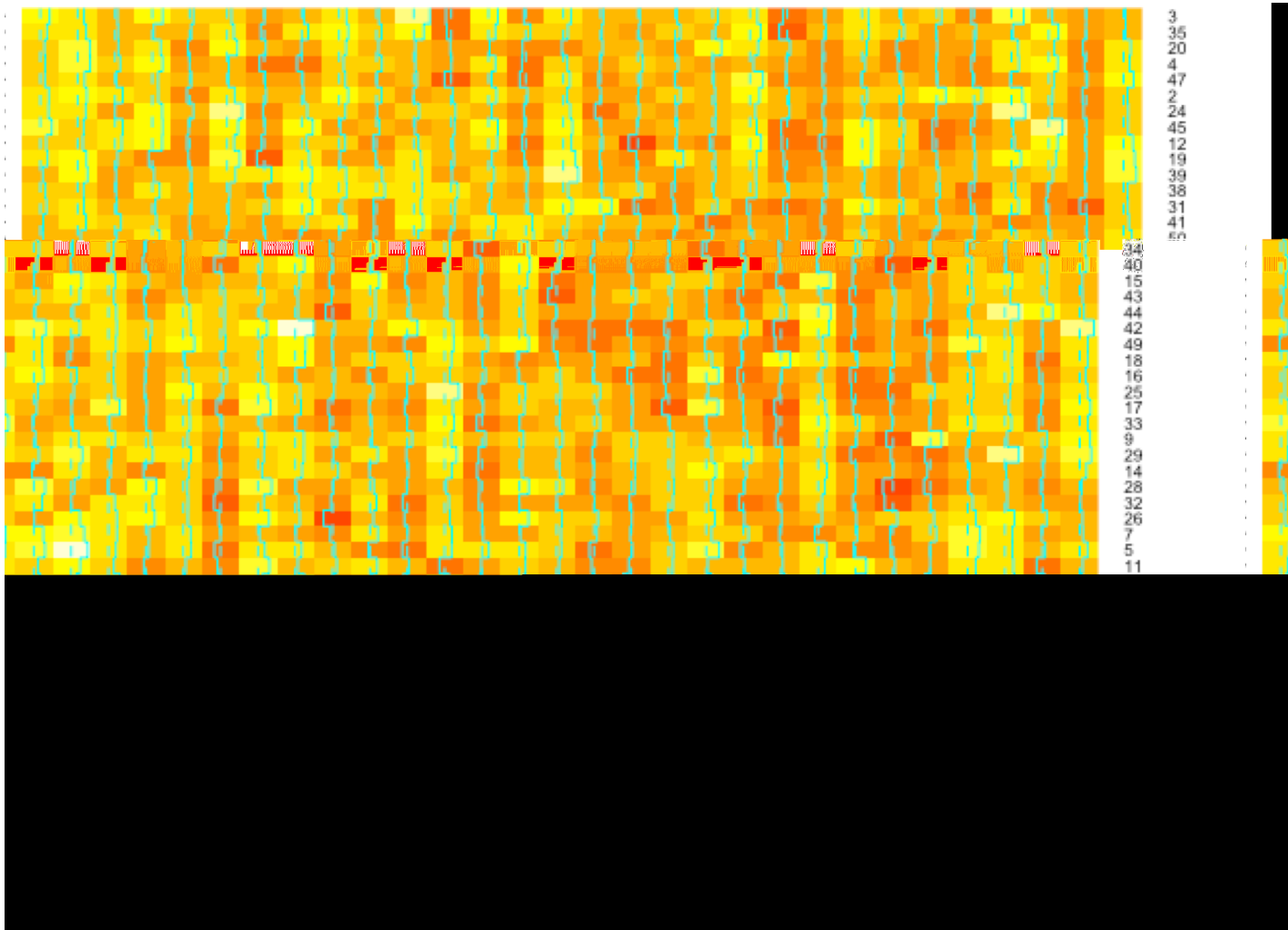


Fisher but Simpson, Neyman but Berkson
Actually, this should be a 2 x 2 table of graphs (with/without each bias).

Correlation and causation in social genomics

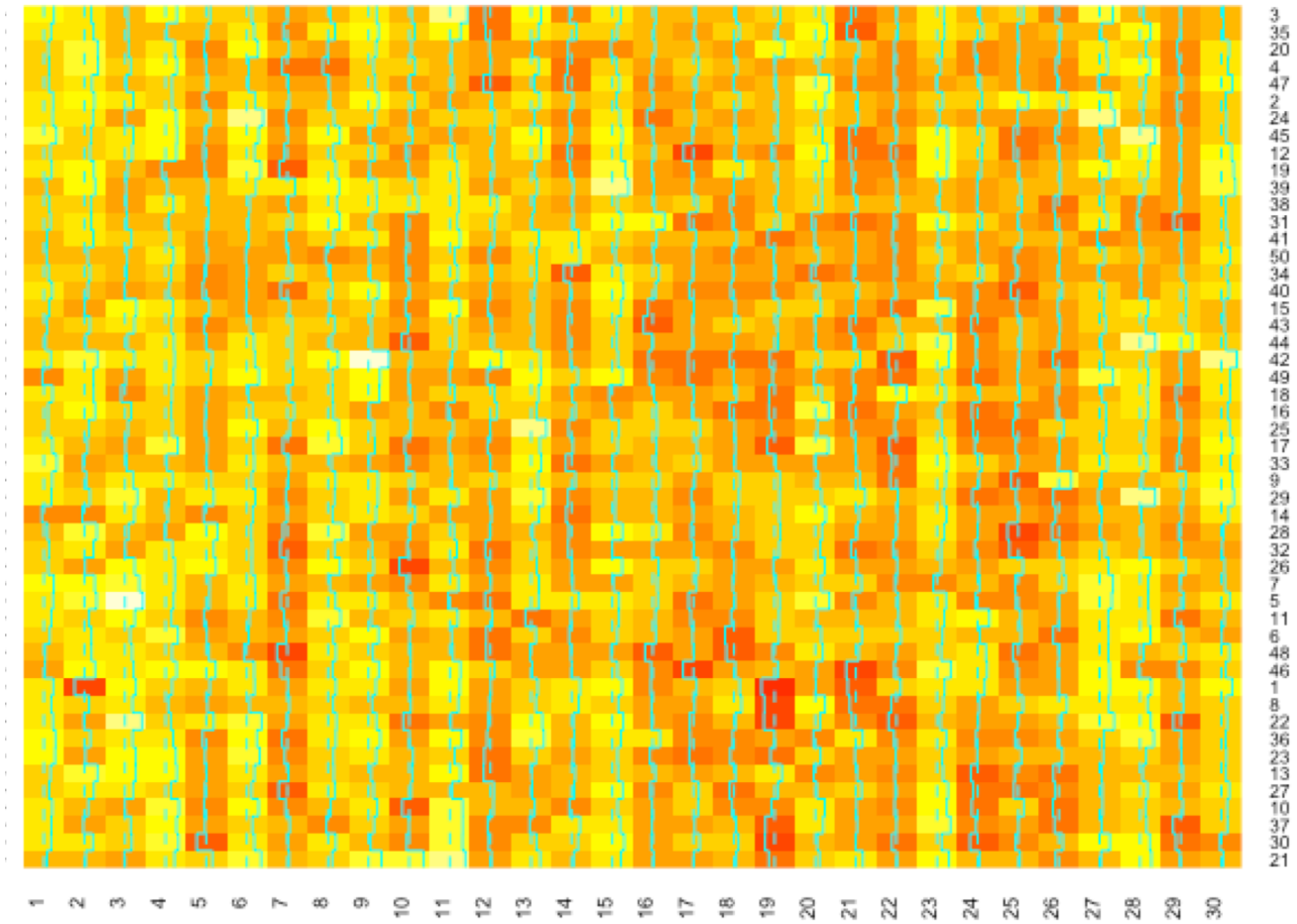
Justin Chumbley

Gene expression



Dogma of statistical inference: y ~ P

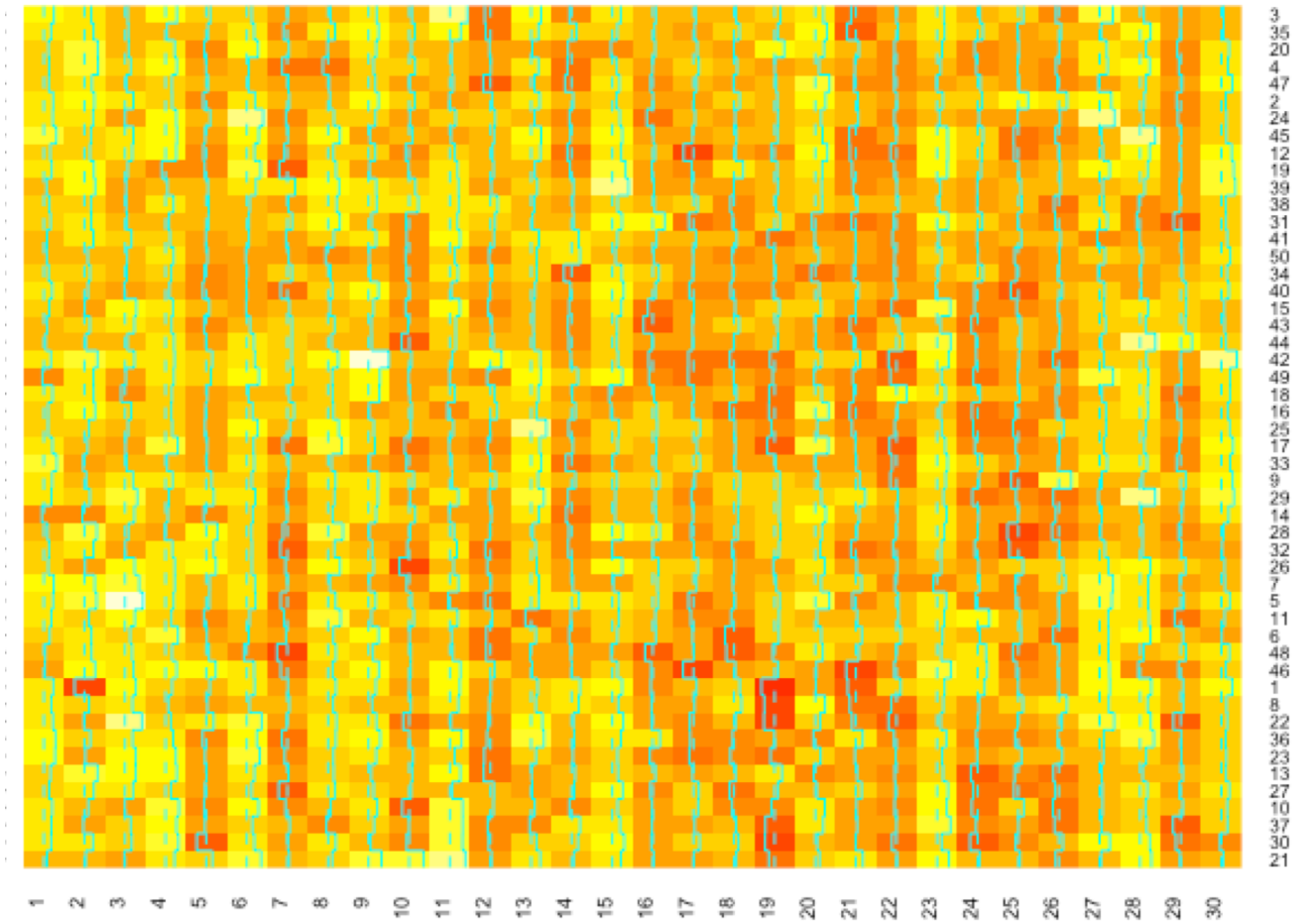
Gene expression



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

Exposure

Gene expression



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

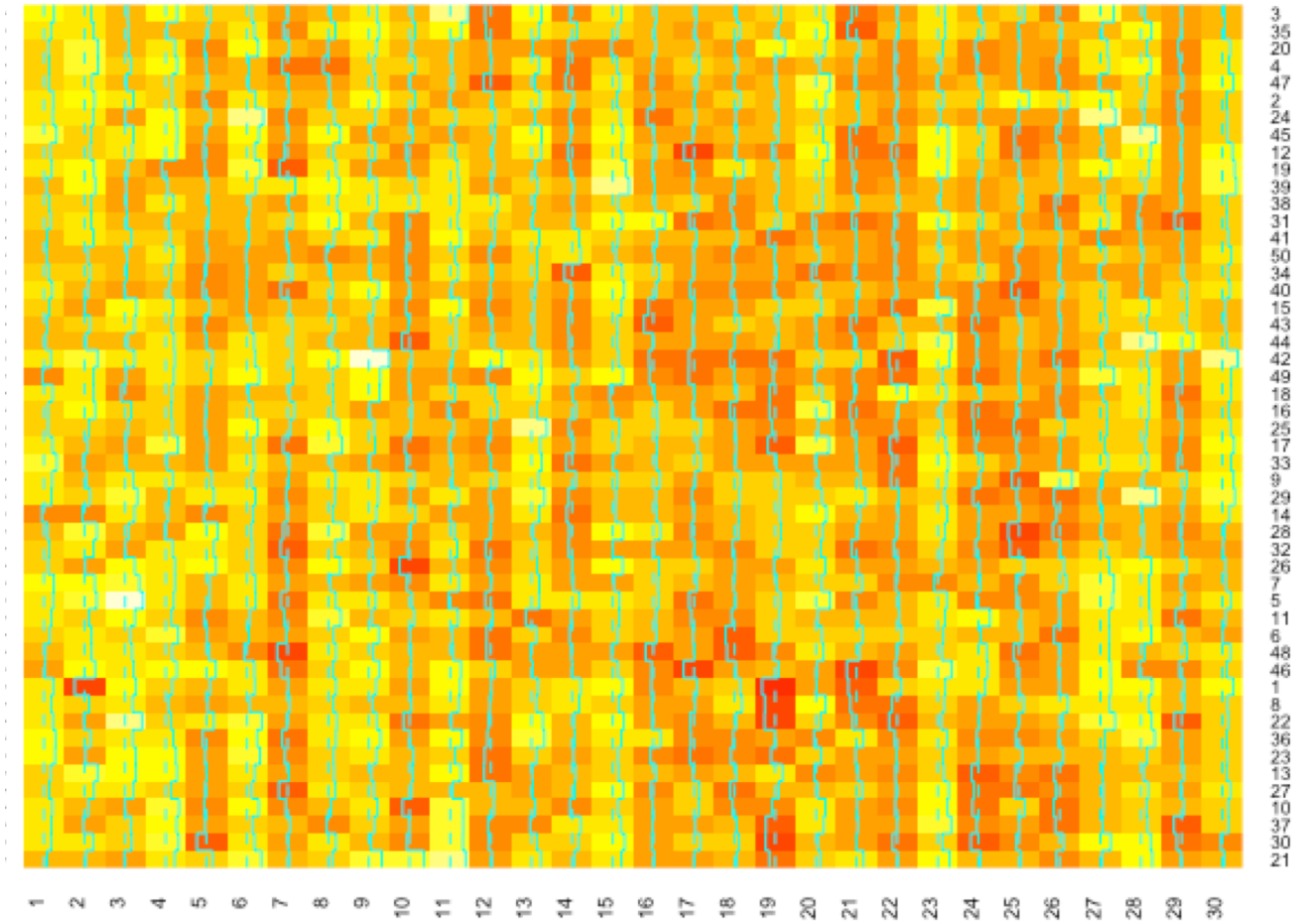
P(T)

Exposure

P(Y, T)

Gene expression

P(Y)



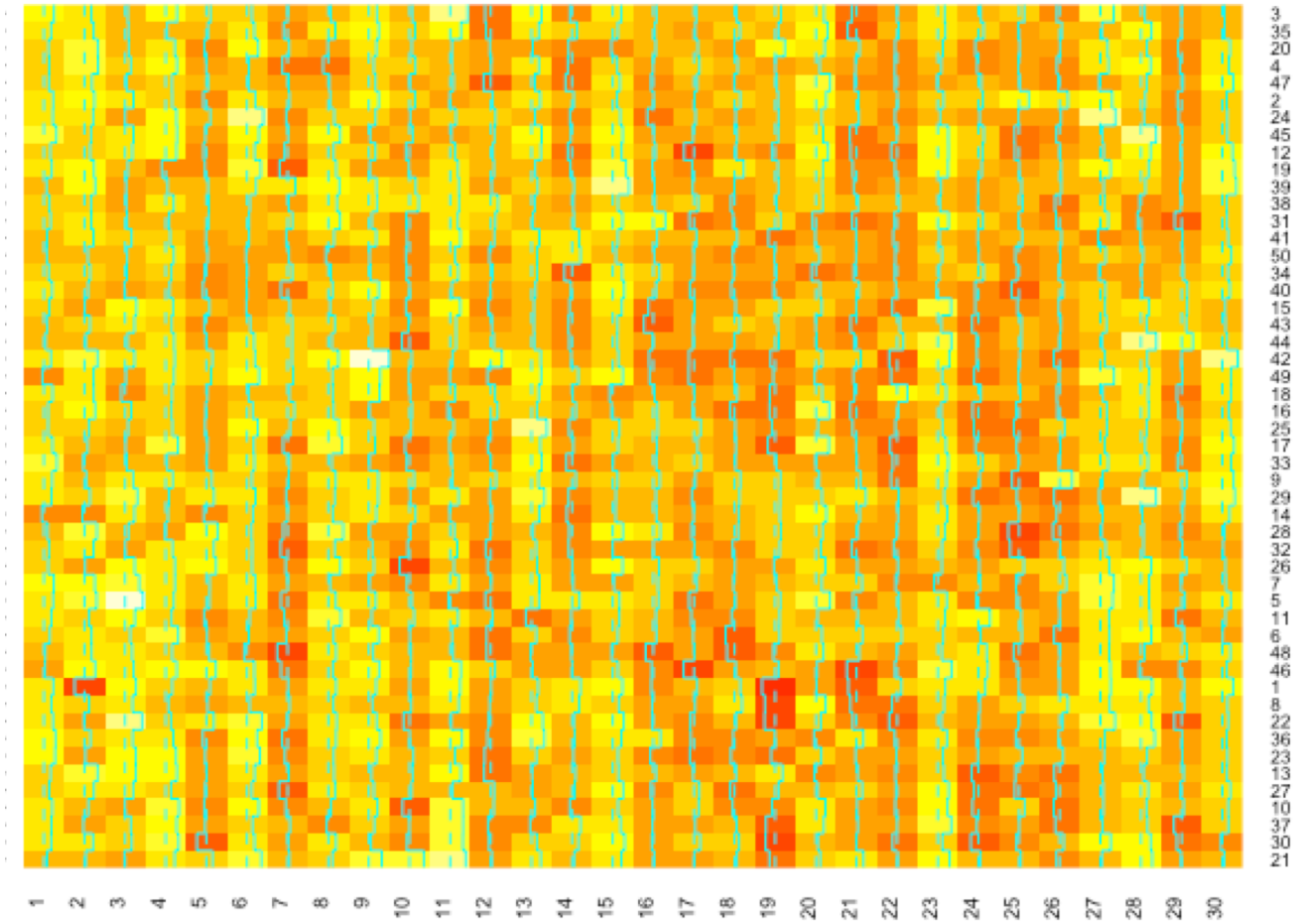
Dogma of statistical inference: (y,t) ~ P

P(T) Exposure

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

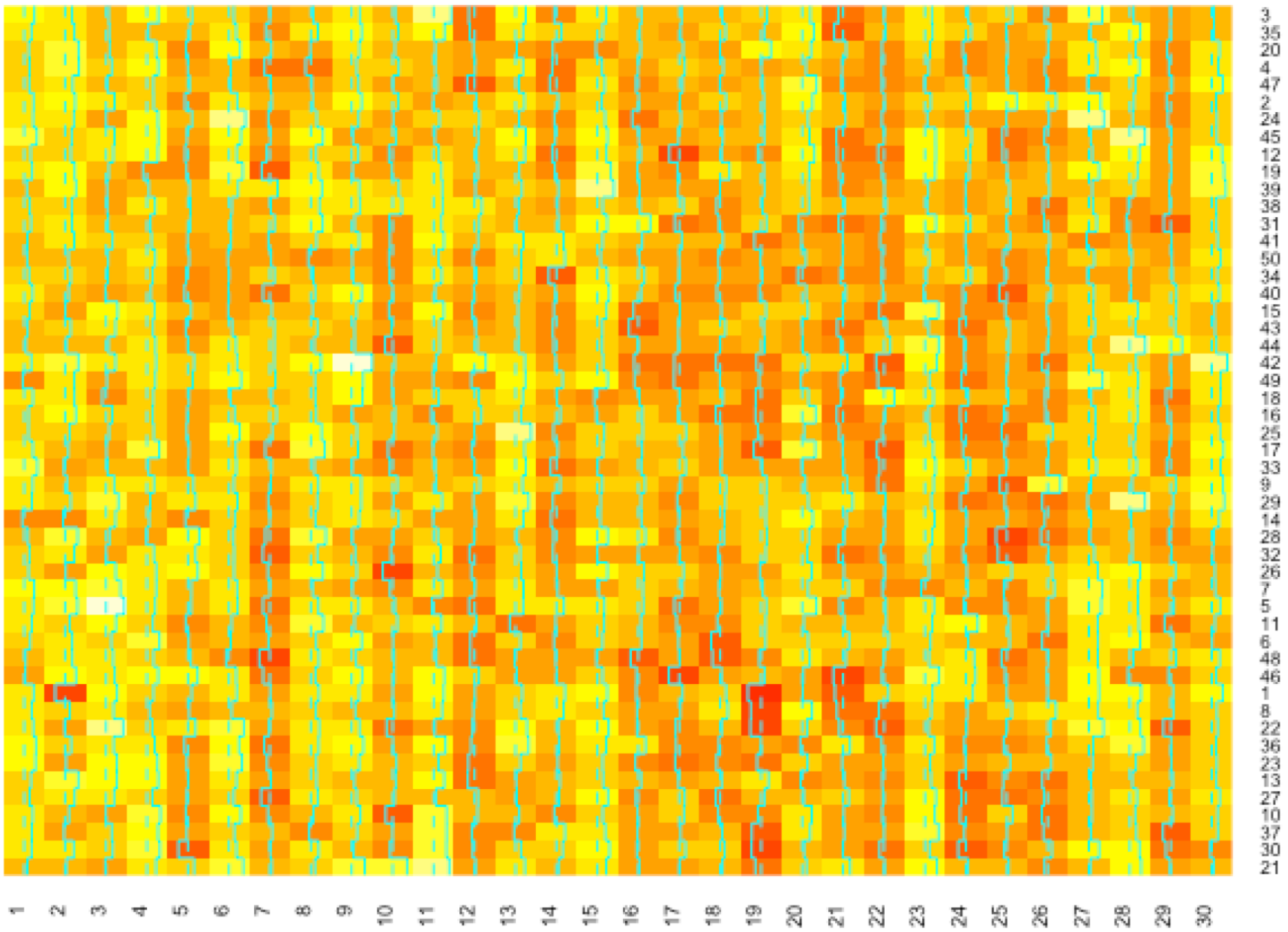
Gene expression

P(Y)



Exposure

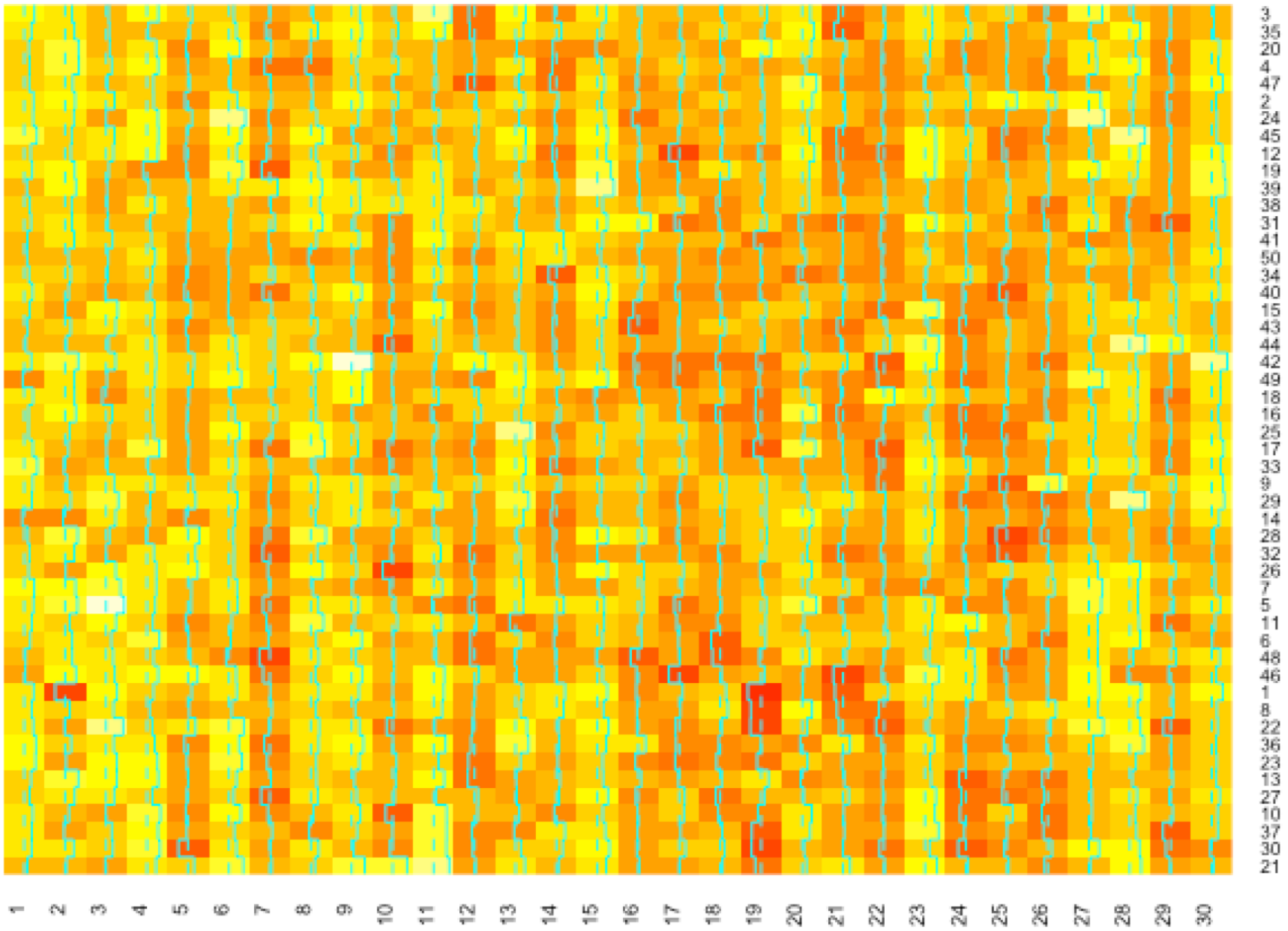
Gene expression

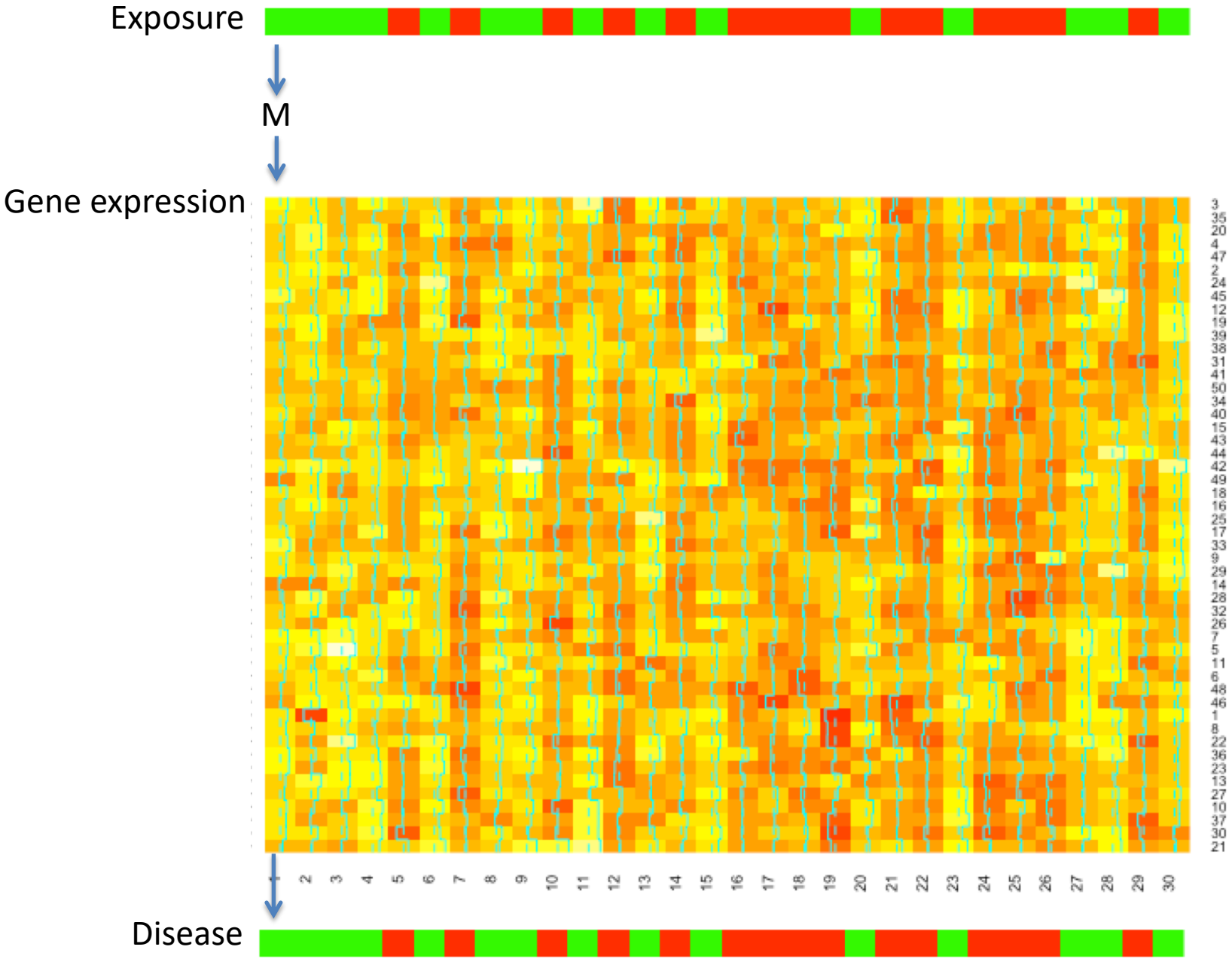


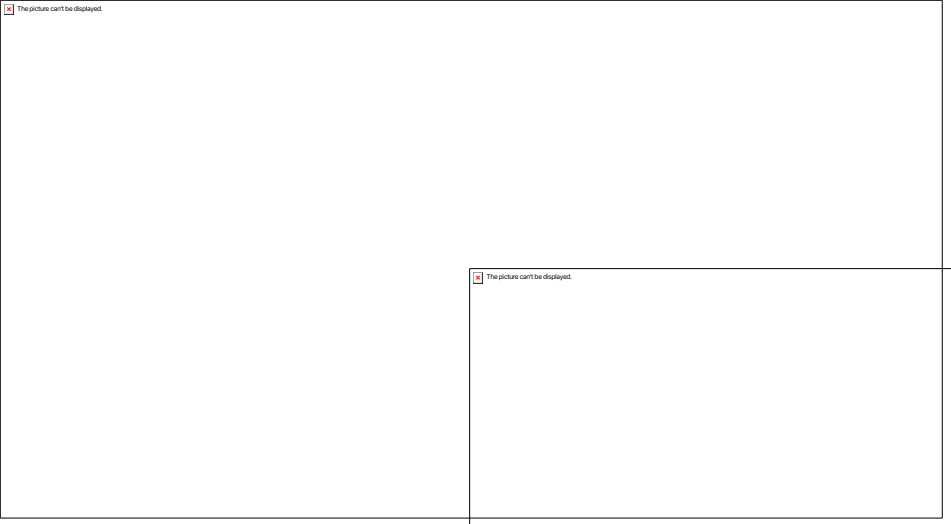
Exposure

M

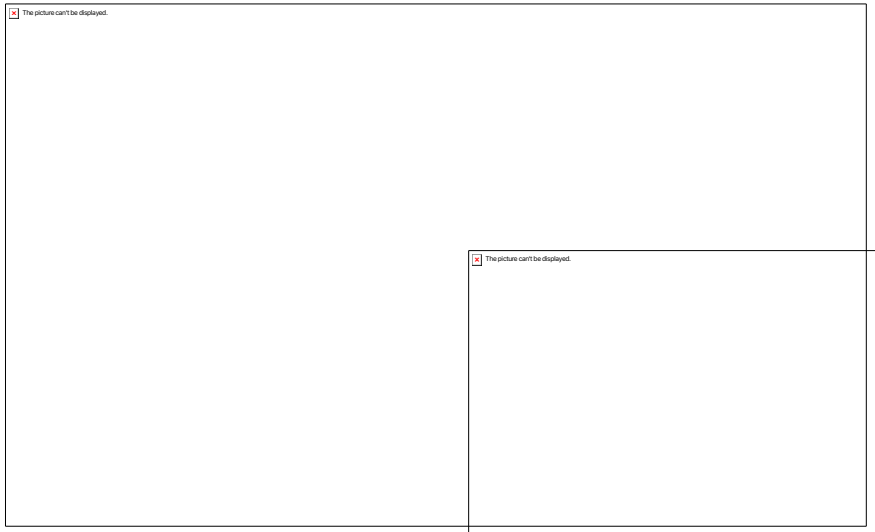
Gene expression



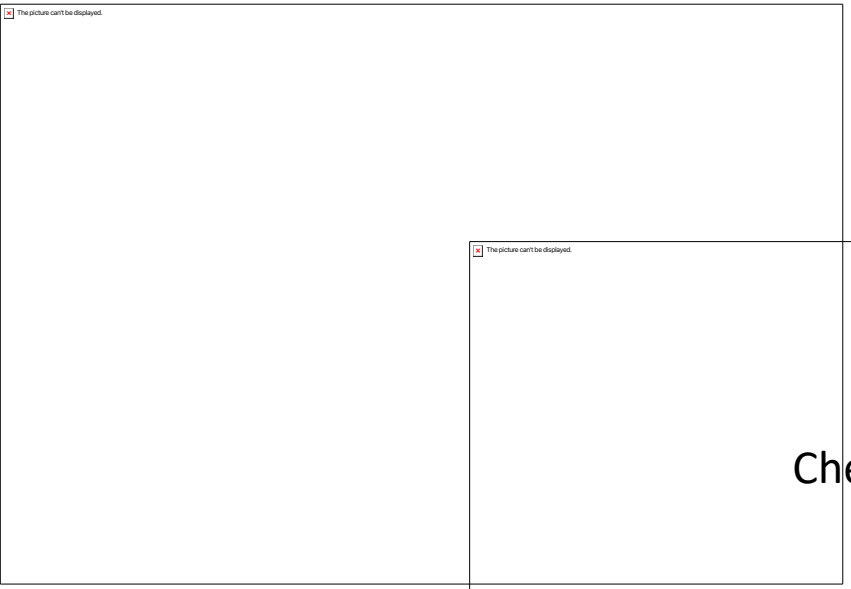

Exposure
M
Gene expression
Disease

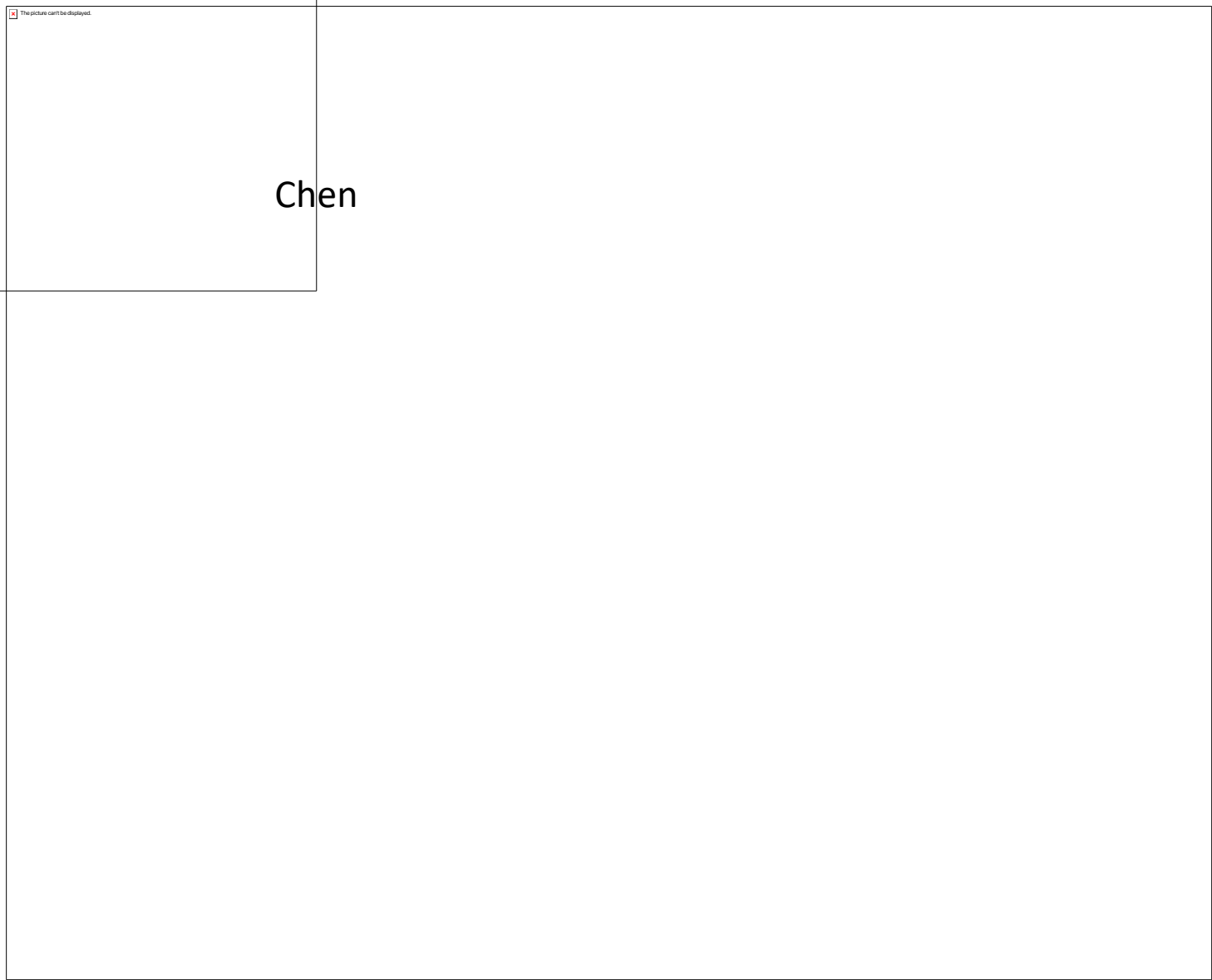

The picture can't be displayed.


The picture can't be displayed.

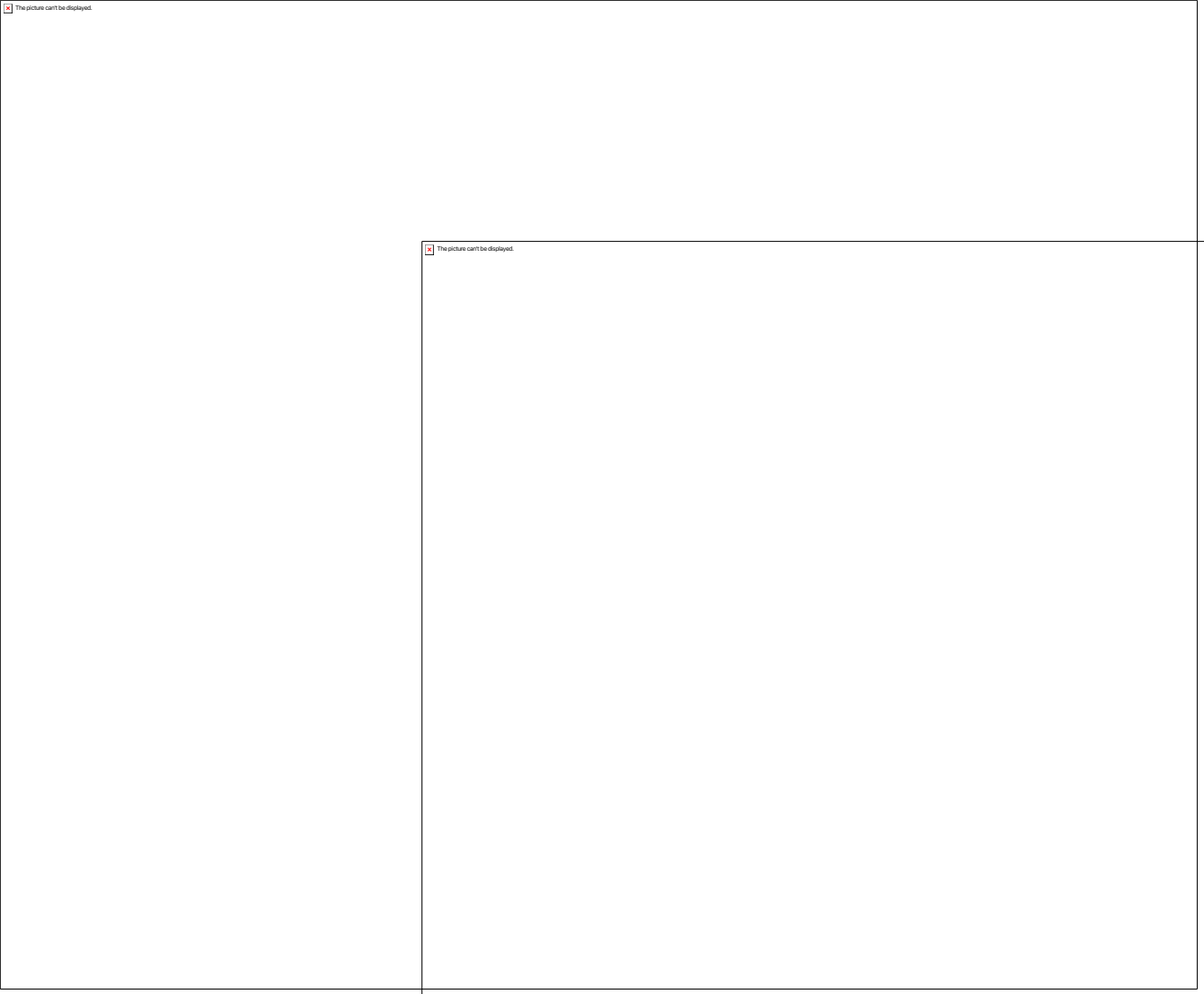

The picture can't be displayed.

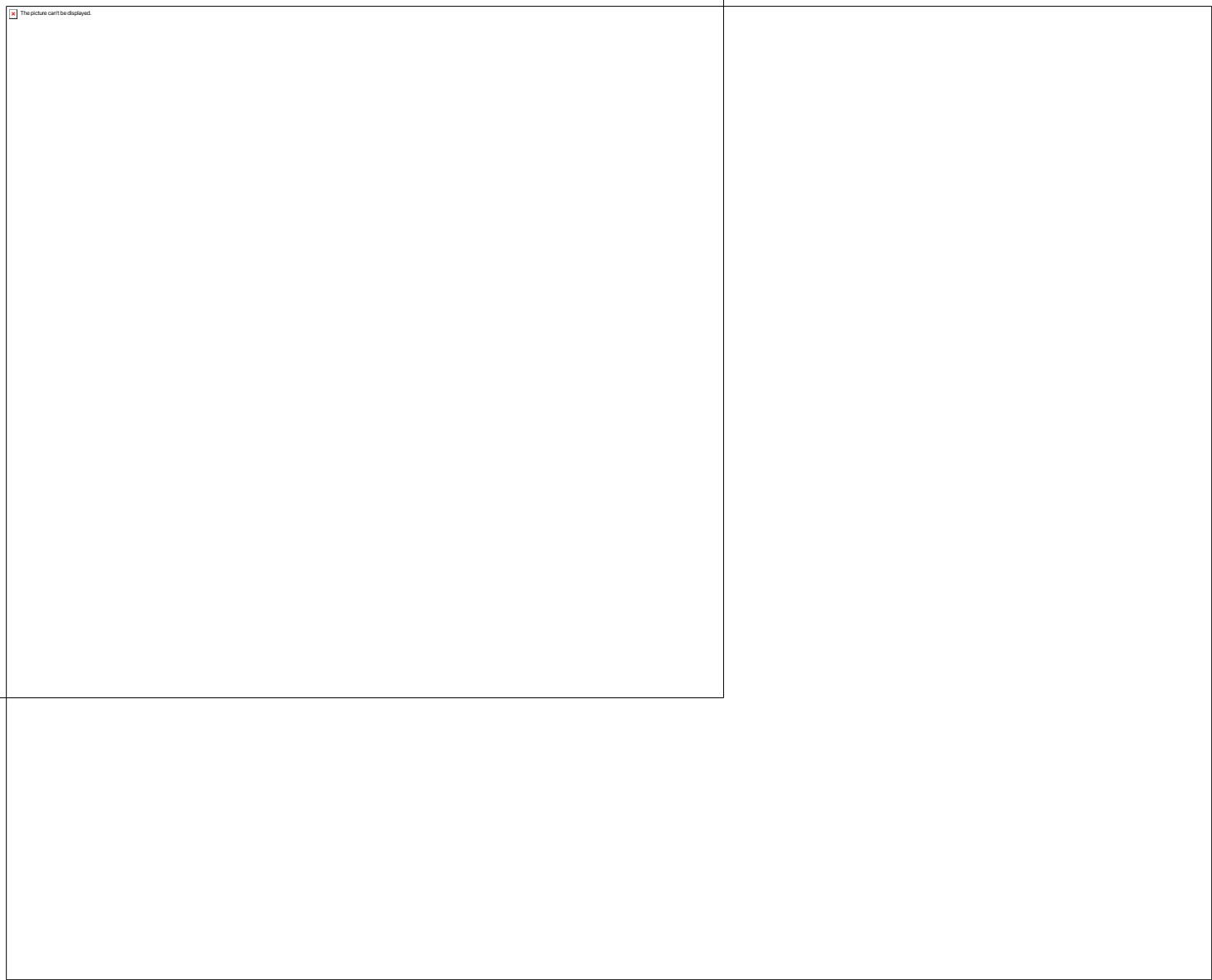

The picture can't be displayed.

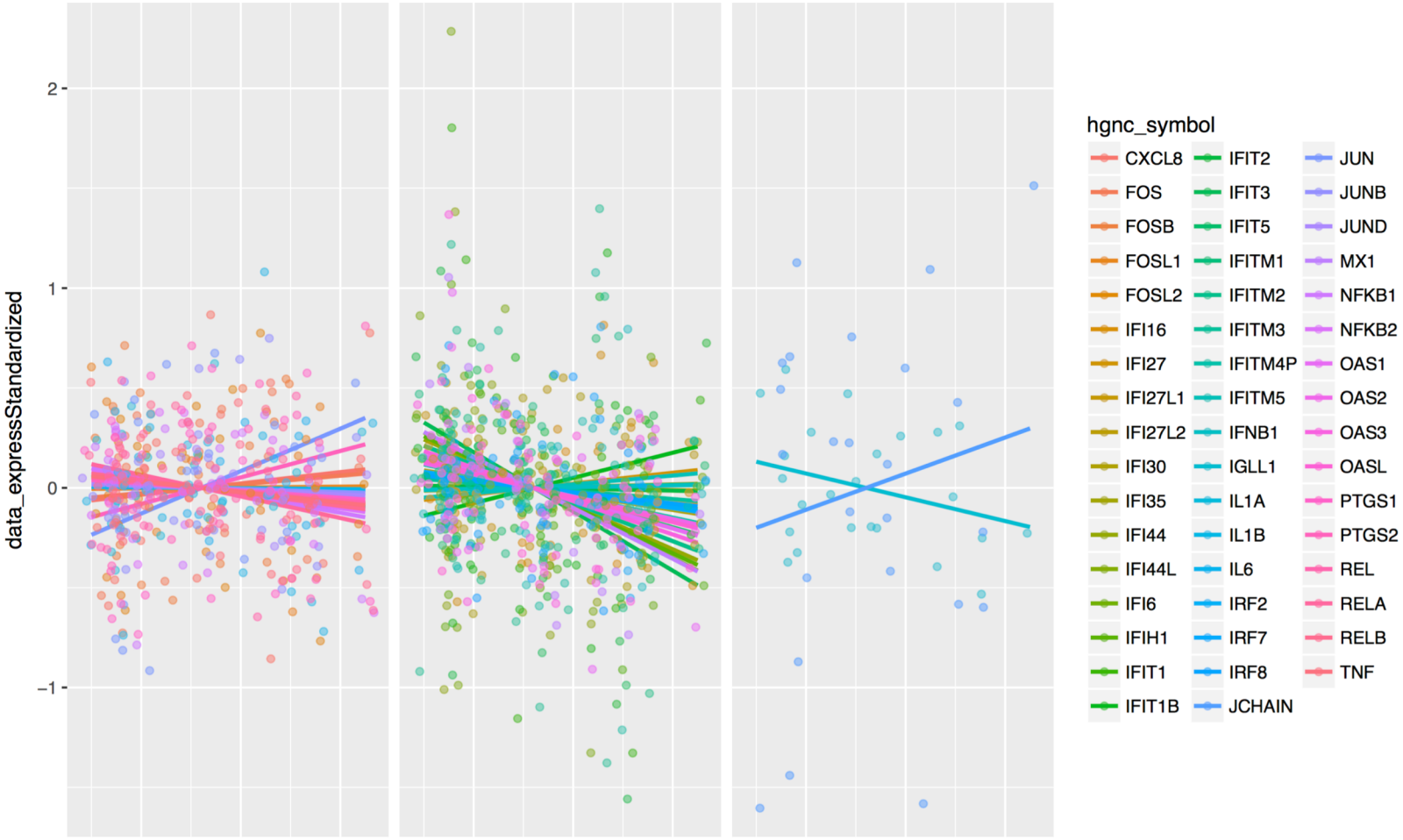

The picture can't be displayed.

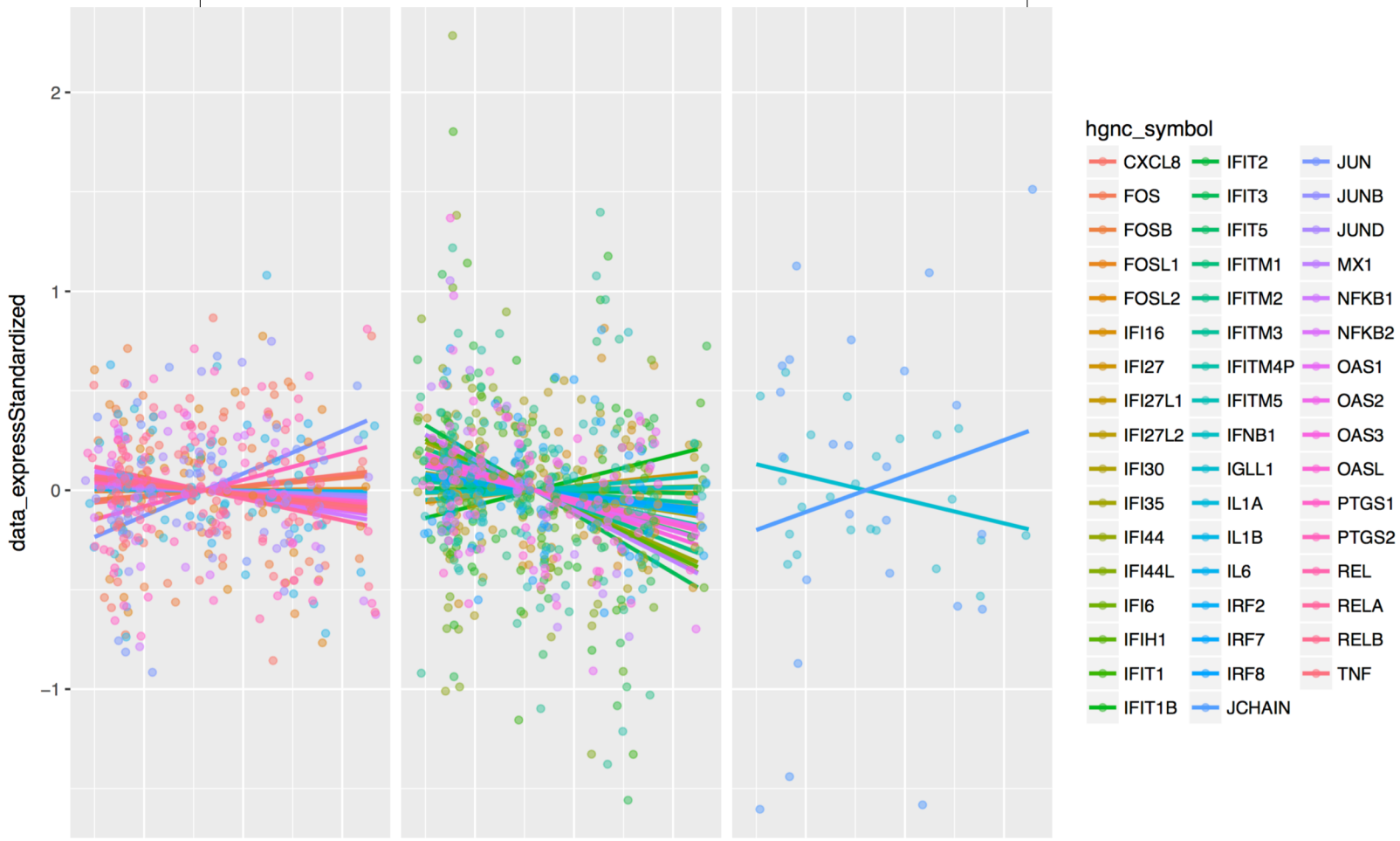

The picture can't be displayed.

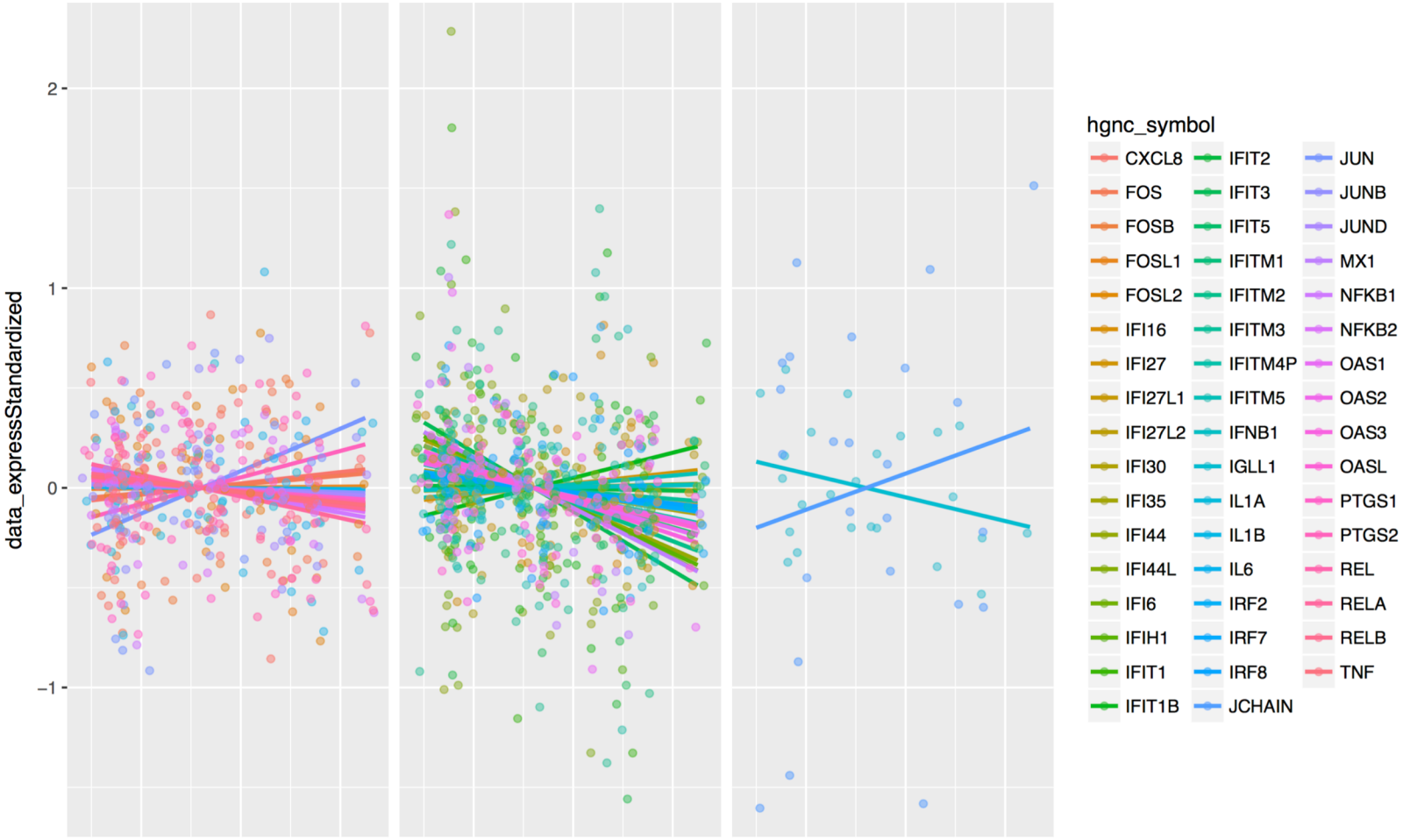
Chen

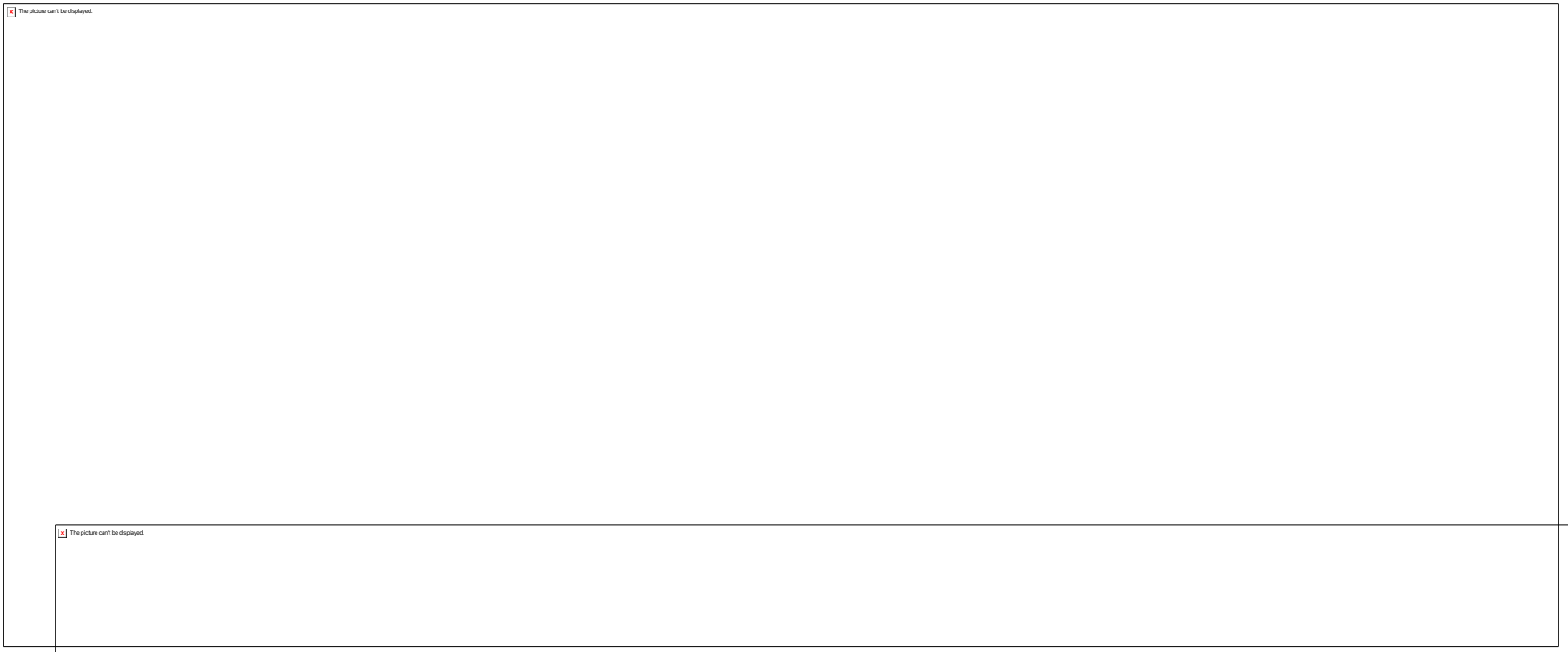

The picture can't be displayed.


The picture can't be displayed.


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

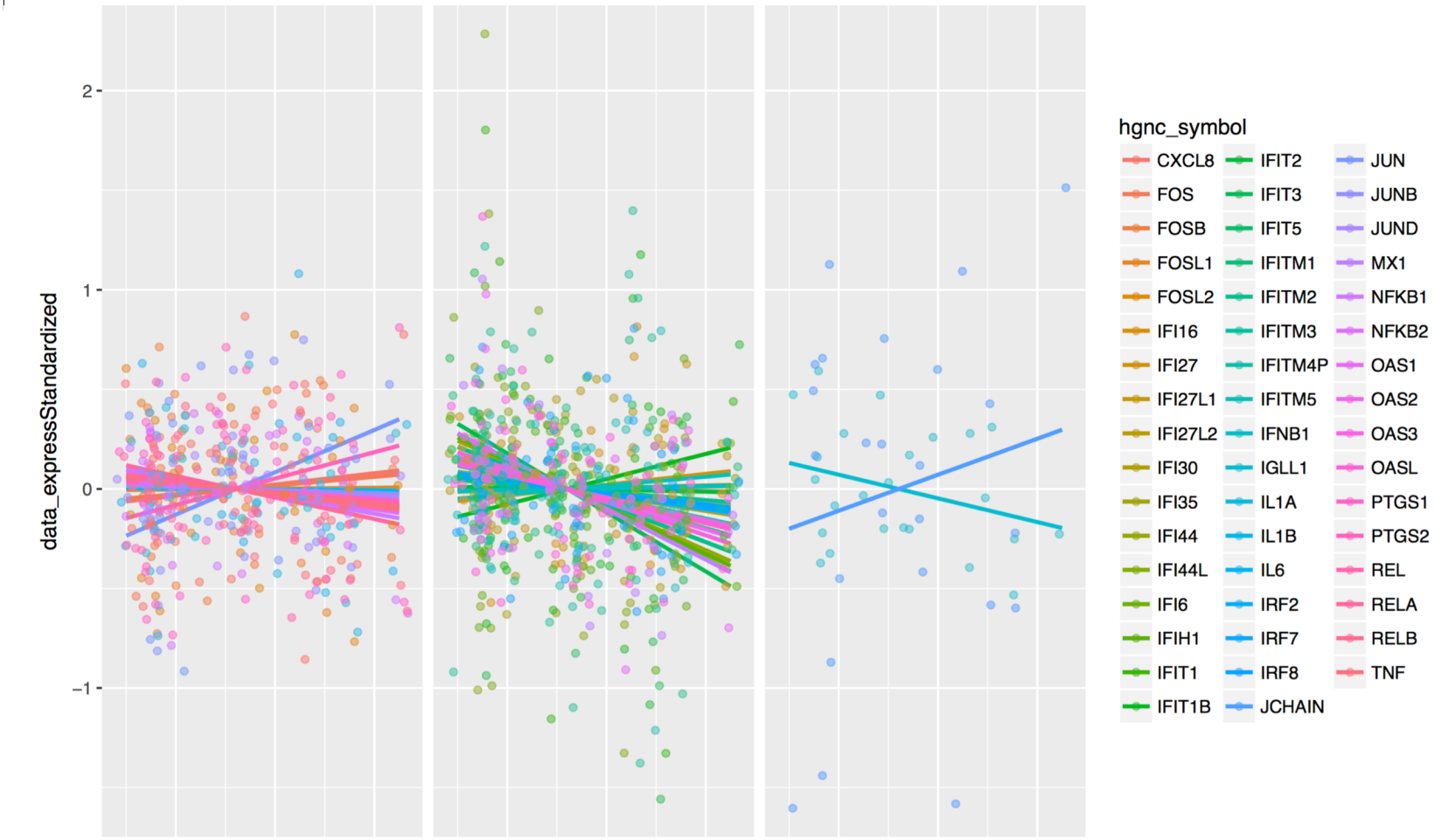

data_expressStandardized
2
1
0
−1
hgnc_symbol
CXCL8
FOS
FOSB
FOSL1
FOSL2
IFI16
IFI27
IFI27L1
IFI27L2
IFI30
IFI35
IFI44
IFI44L
IFI6
IFIH1
IFIT1
IFIT1B
IFIT2
IFIT3
IFIT5
IFITM1
IFITM2
IFITM3
IFITM4P
IFITM5
IFNB1
IGLL1
IL1A
IL1B
IL6
IRF2
IRF7
IRF8
JCHAIN
JUN
JUNB
JUND
MX1
NFKB1
NFKB2
OAS1
OAS2
OAS3
OASL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


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

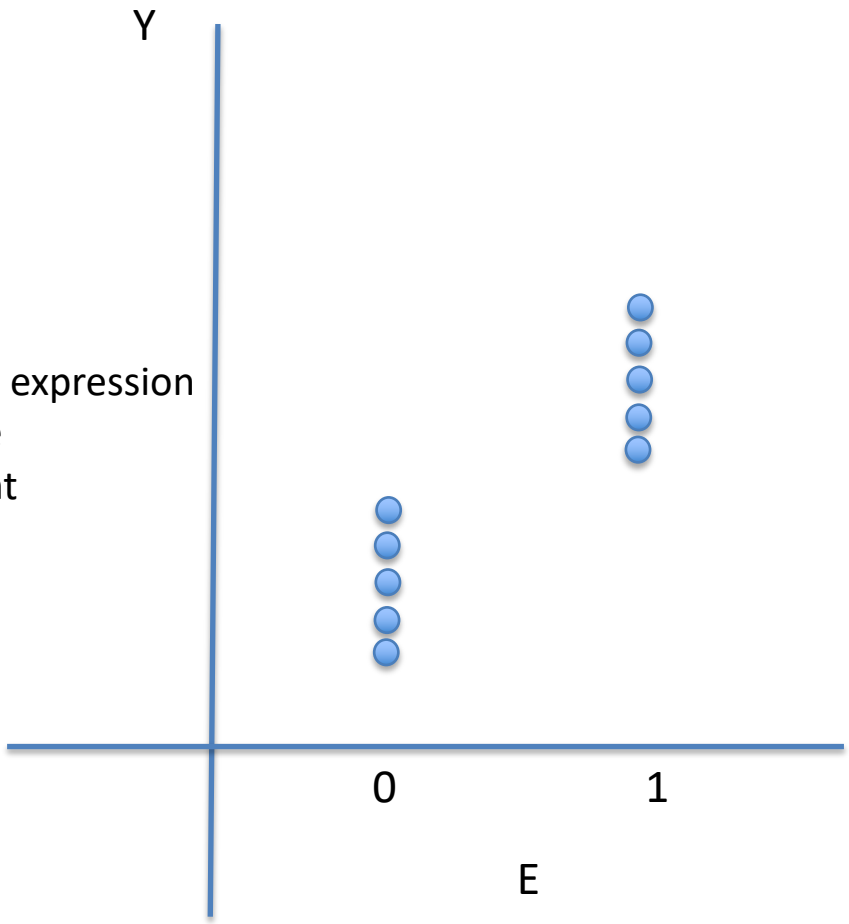

The picture can't be displayed.


The picture can't be displayed.

The picture can't be displayed.

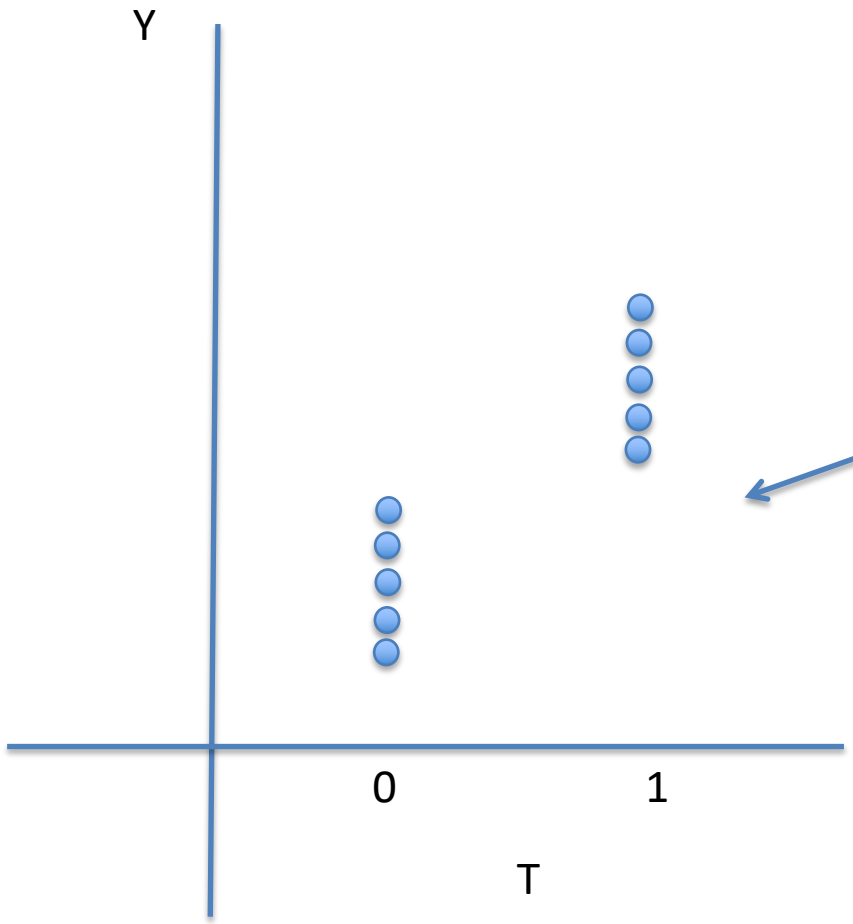


- gene expression
- N dimensional gene expression
- cell type prevalence
- clustering coefficient



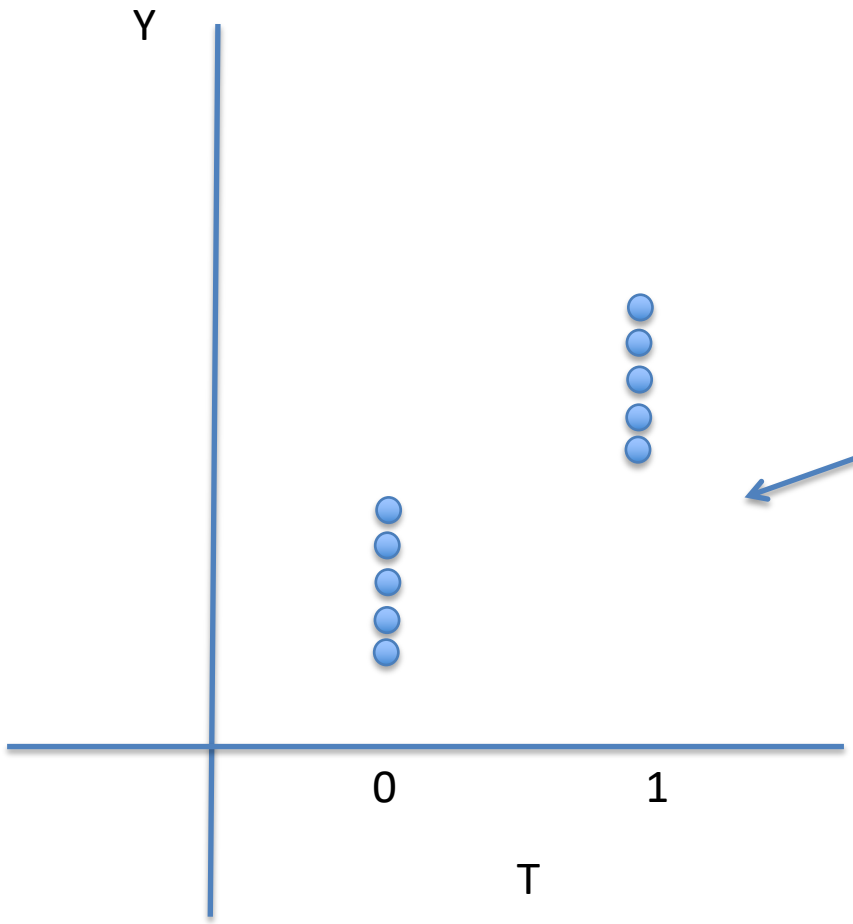
Cause (Noun): something that brings about an effect or a result

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



Probability is symmetric.
What is missing?
Missing structure.

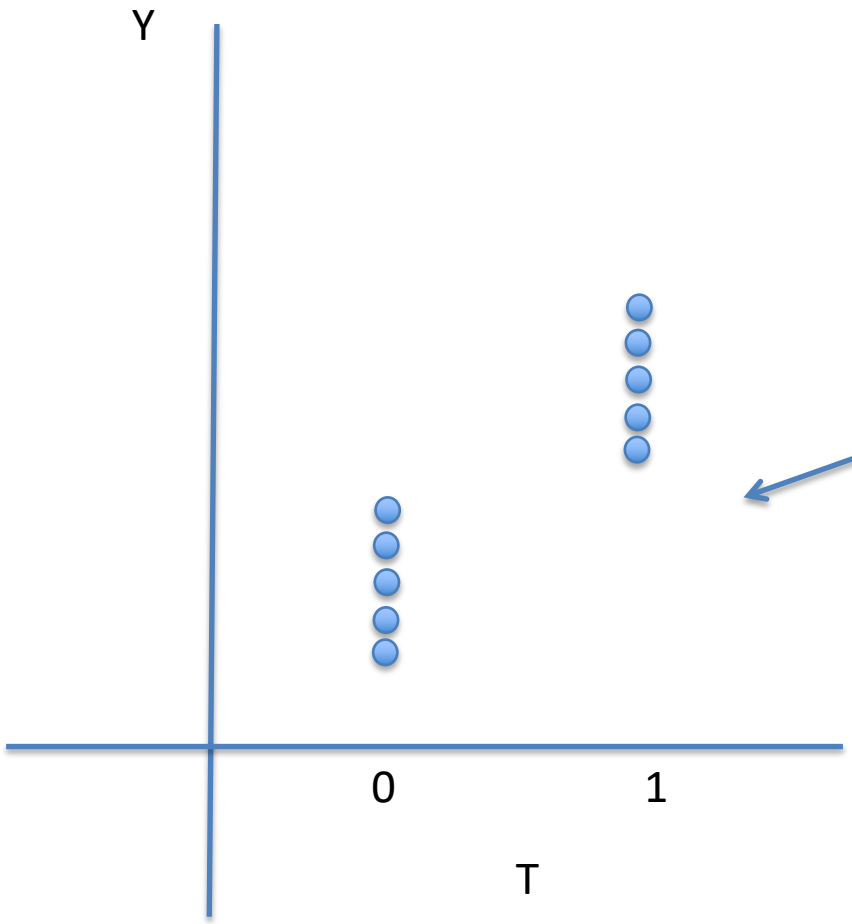
T → Y



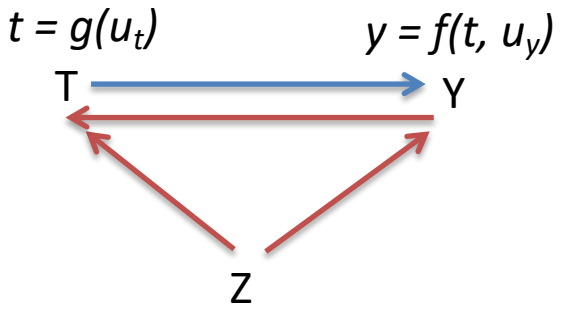
Probability is symmetric.
What is missing?
Missing structure.

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

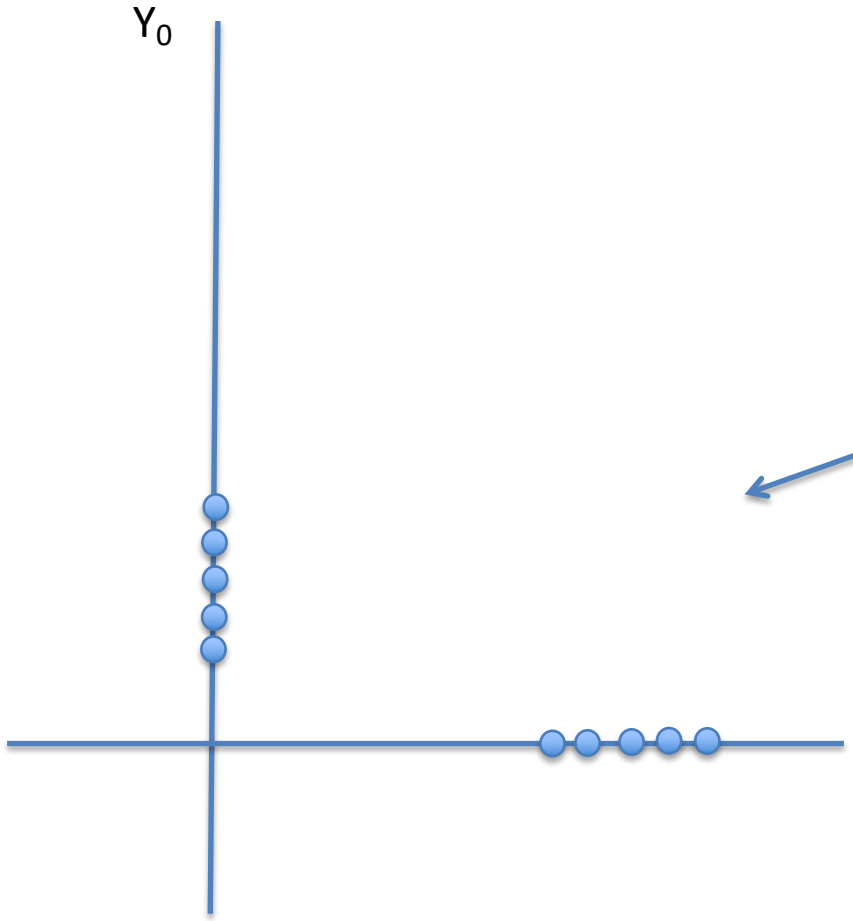
T → Y



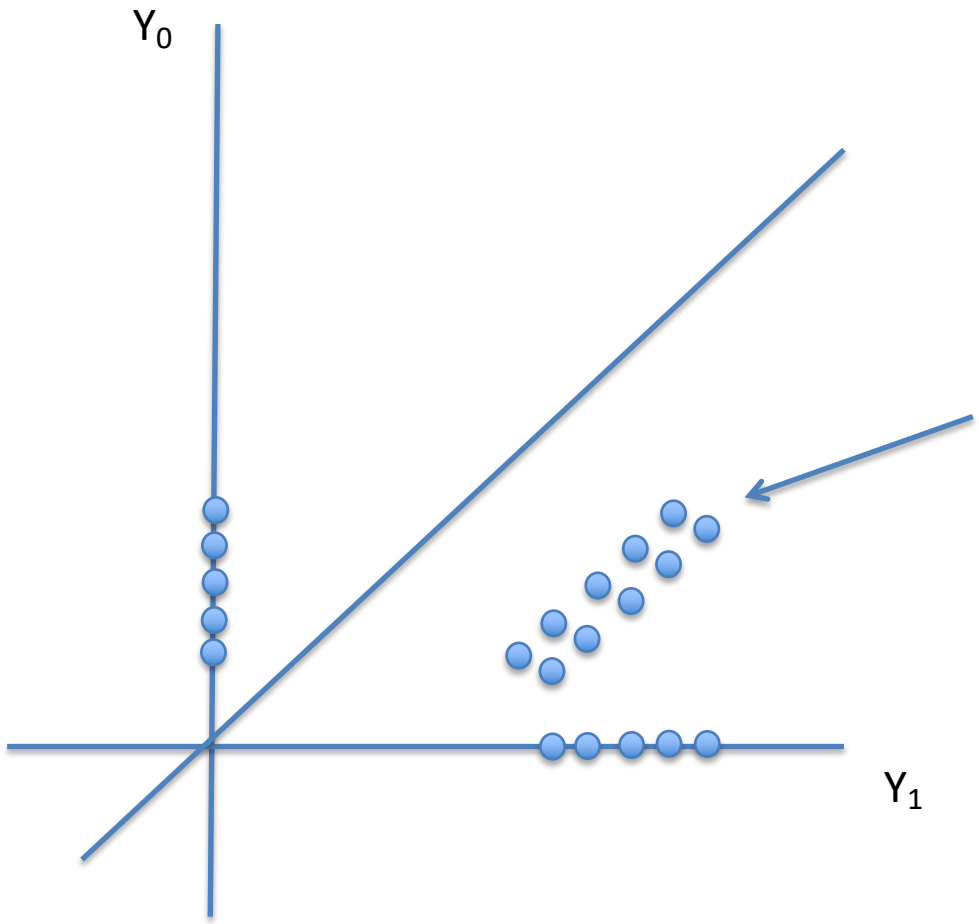
Probability is symmetric.
What is missing?
Missing structure.



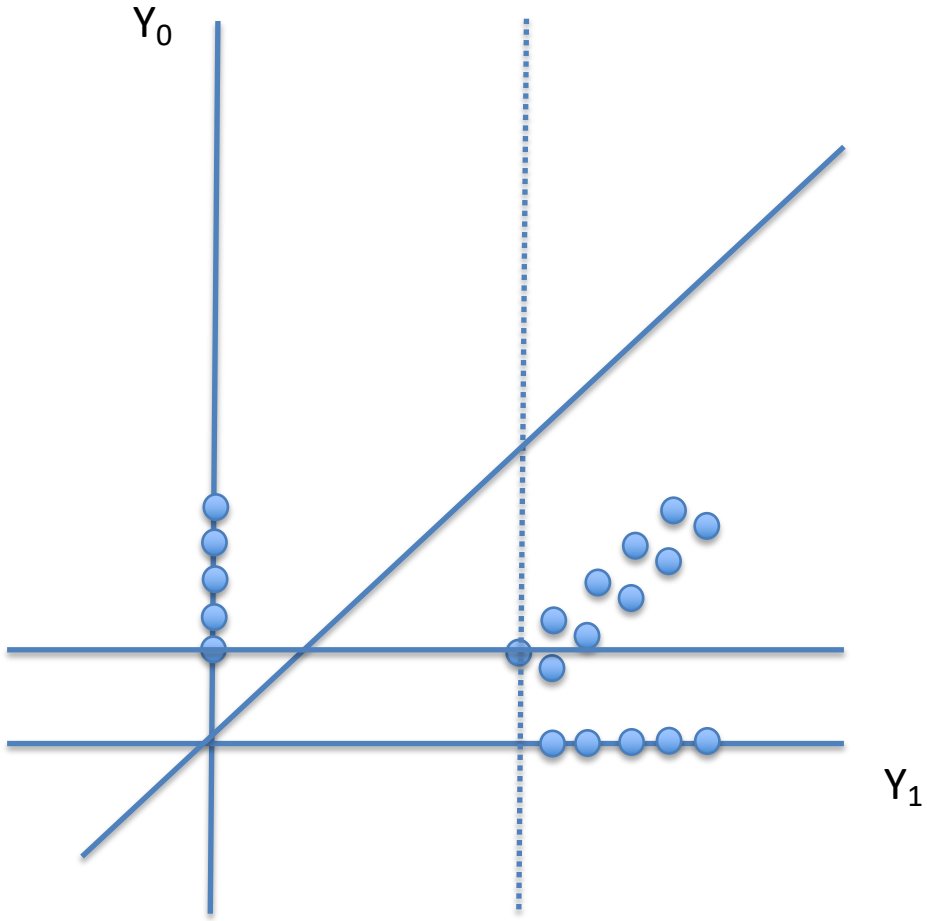
- Variables/nodes
- Included Arrows
- Excluded arrows (strong exclusion restrictions)



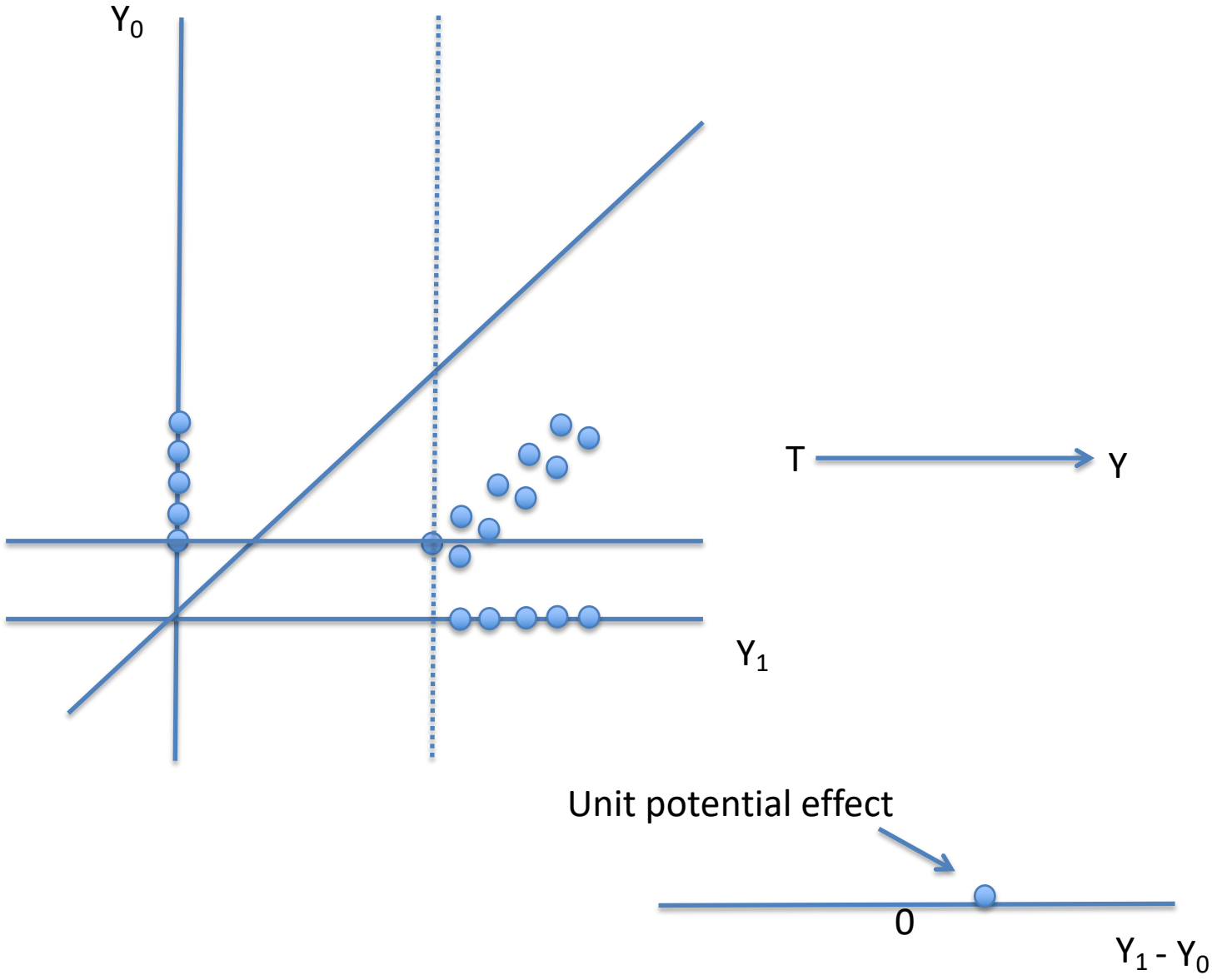
Probability is symmetric.
What is missing?

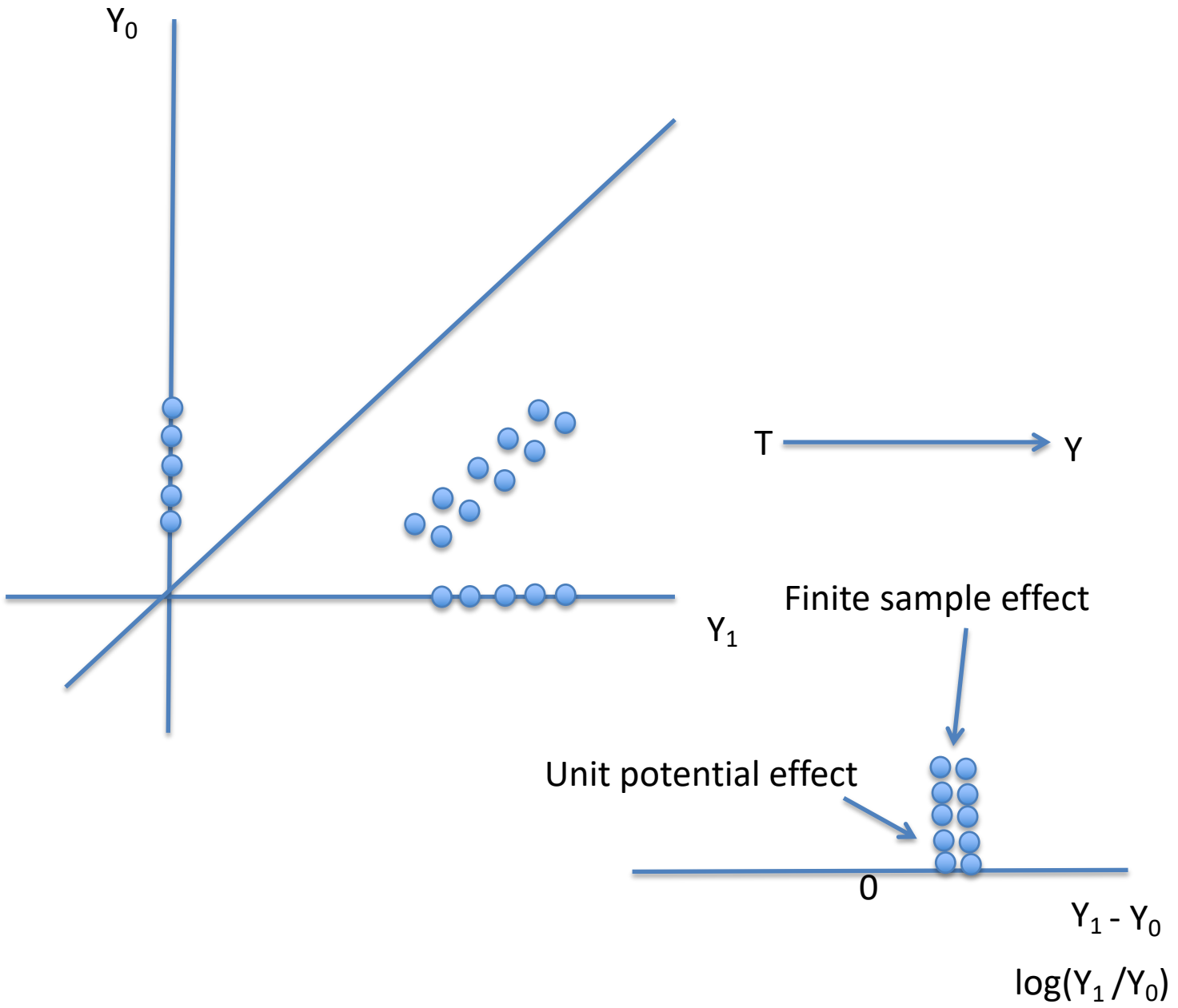


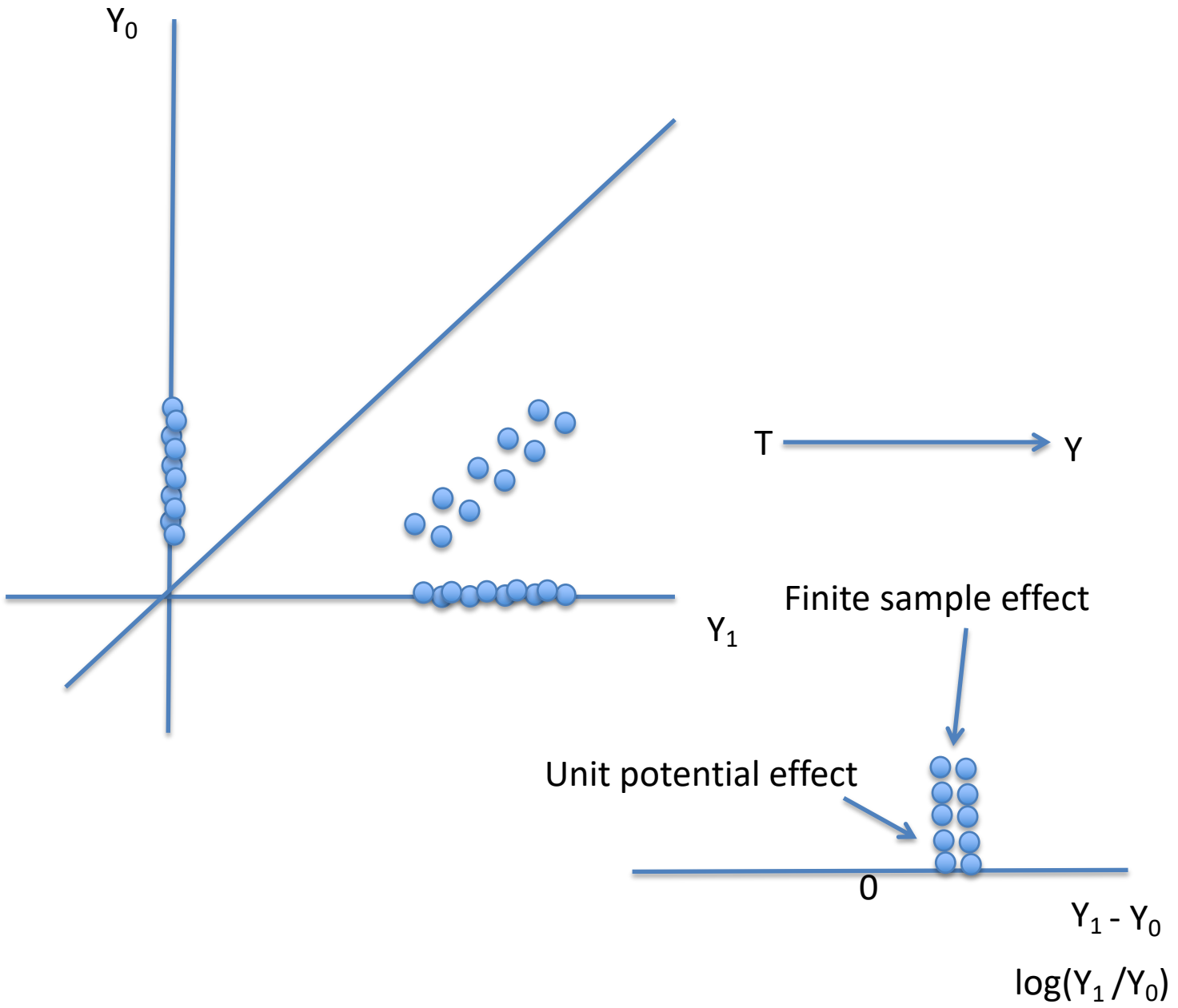
Probability is symmetric.
What is missing?
Missing counterfactual
CTRA, had you not been exposed to adversity.

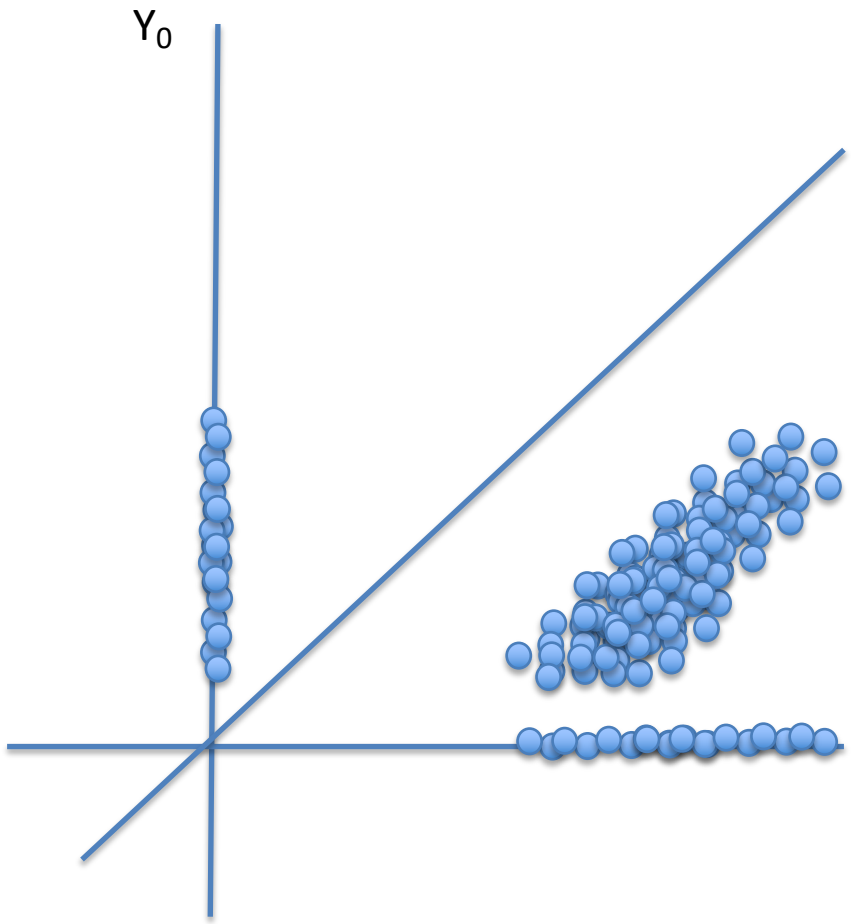


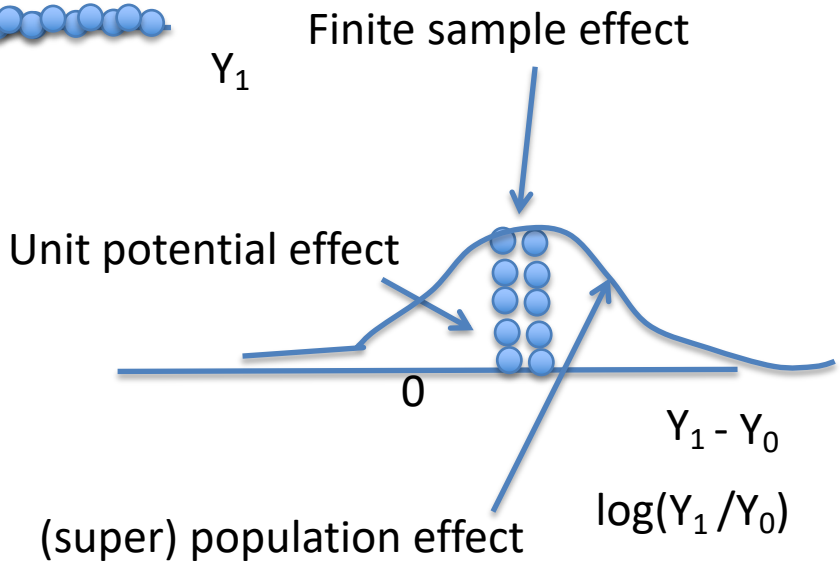
$$Y_{OBS} = Y_0 + (Y_1 - Y_0)T$$


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

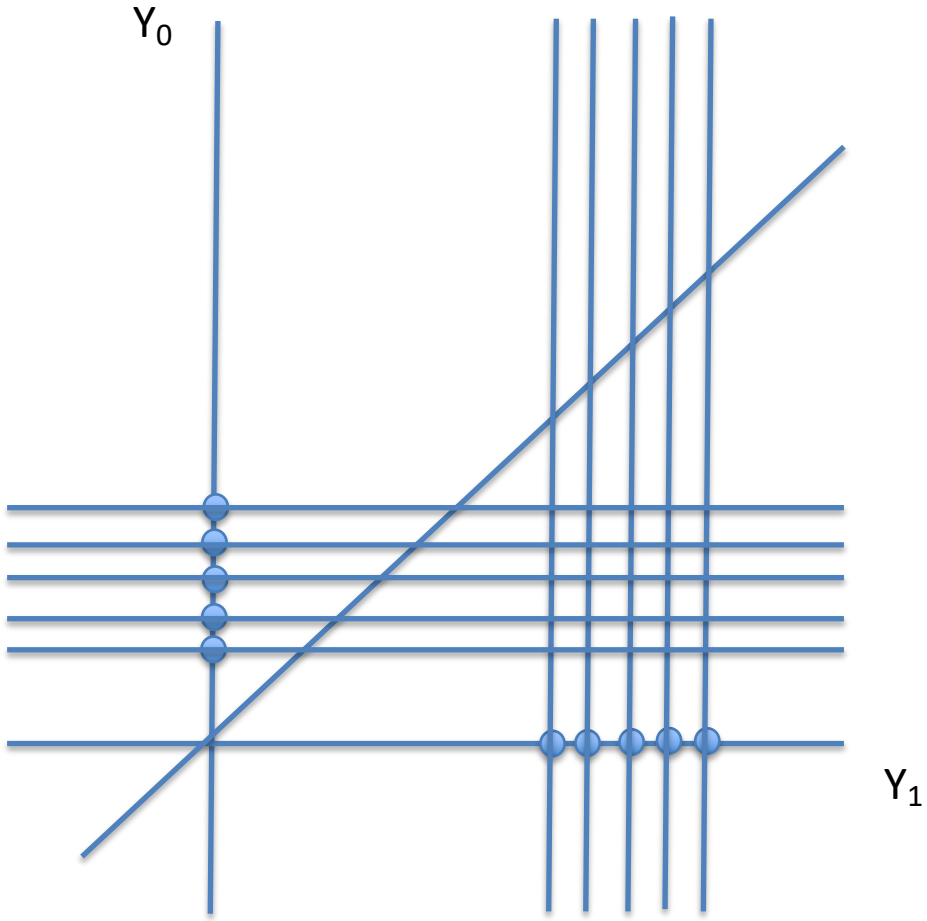

Y_0
Y_1
T
Y
Finite sample effect
Unit potential effect
0
$Y_1 - Y_0$
$\log(Y_1/Y_0)$

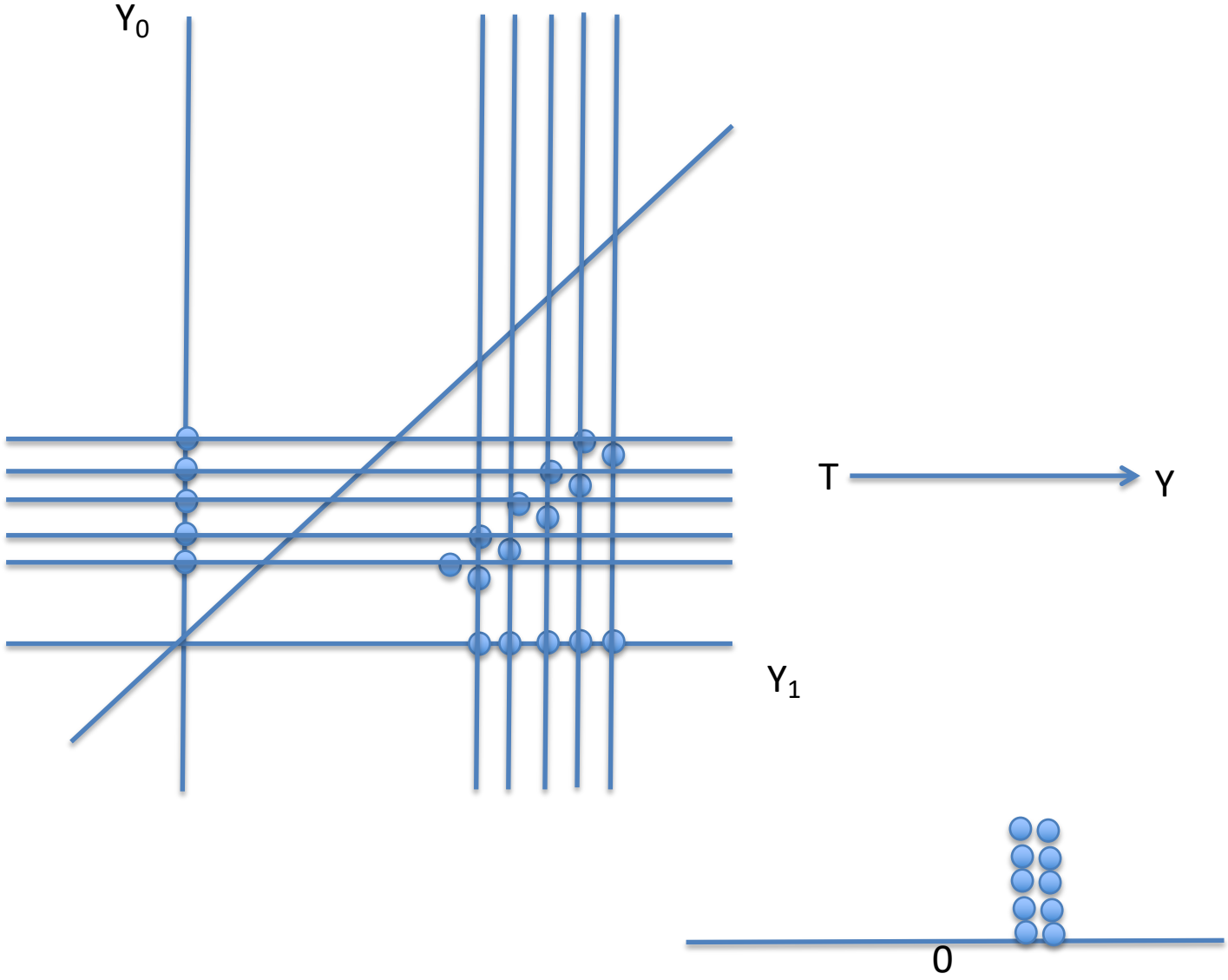

Y_0
Y_1
T
Y
Finite sample effect
Unit potential effect
0
$Y_1 - Y_0$
$\log(Y_1 / Y_0)$

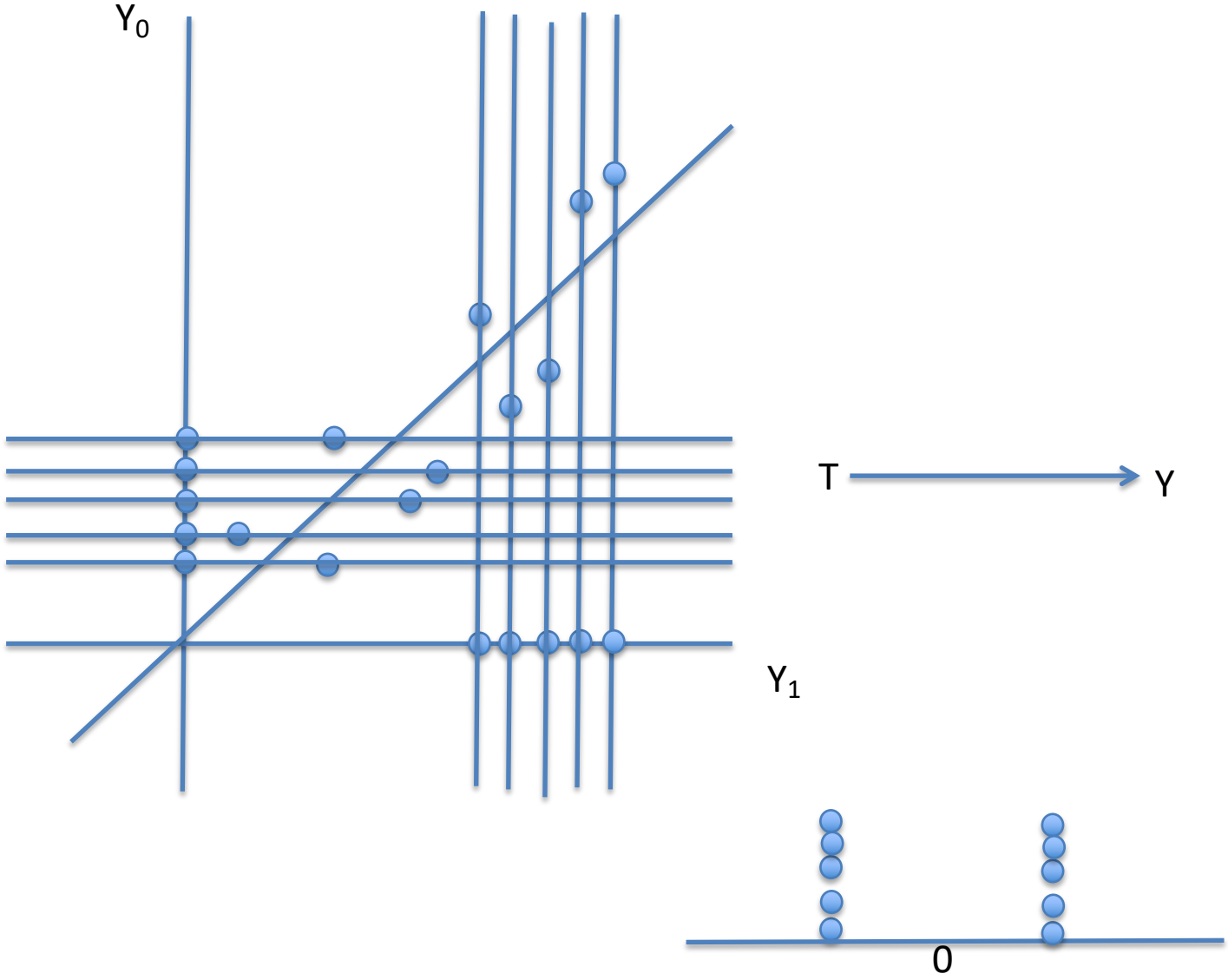


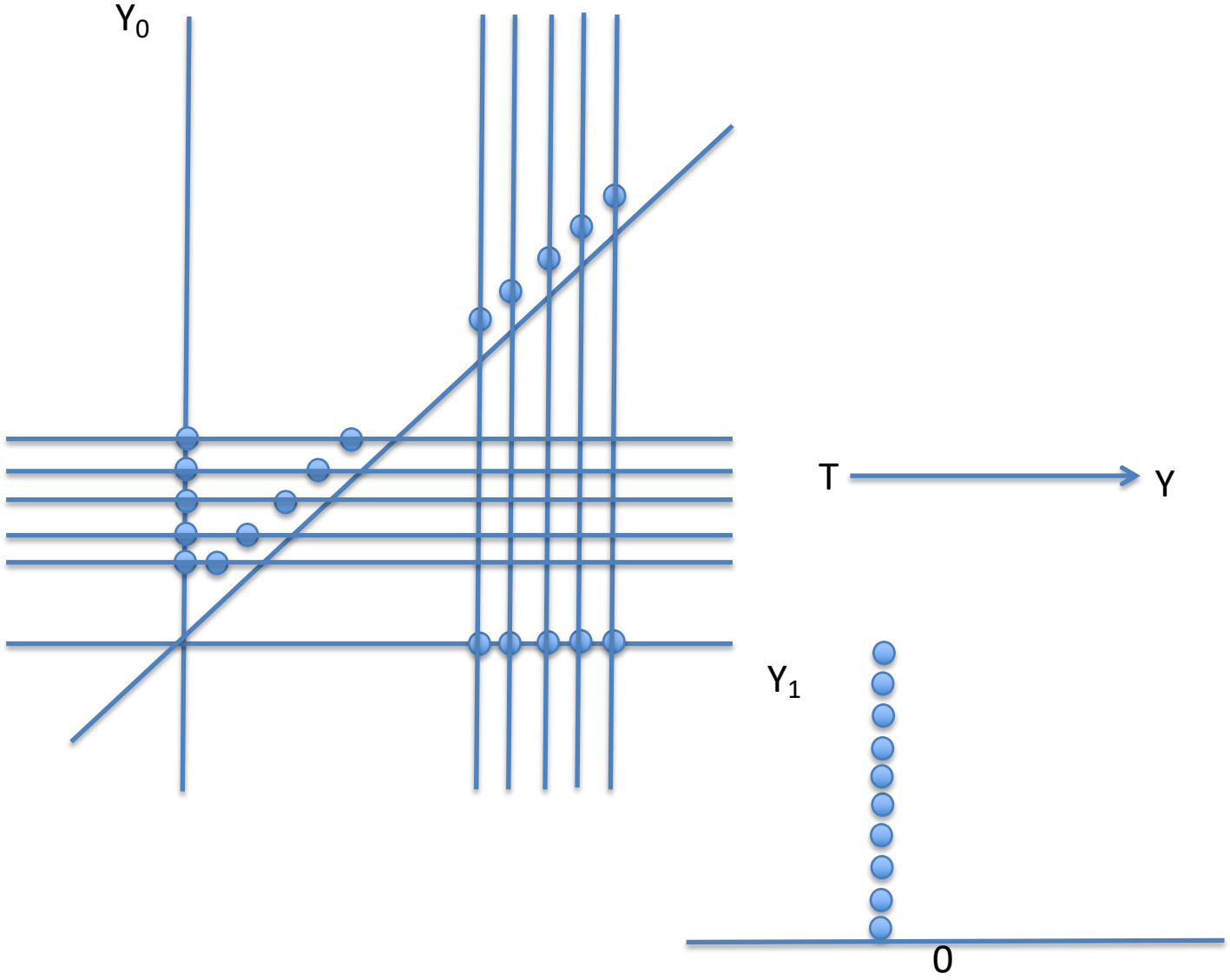


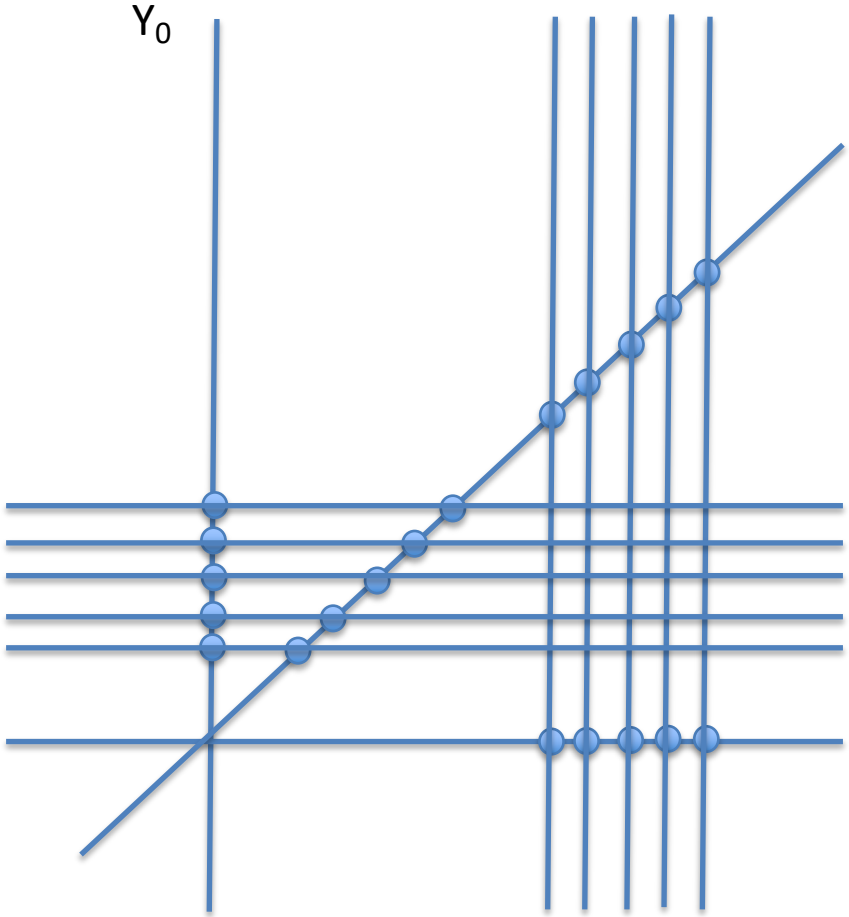
Different estimands

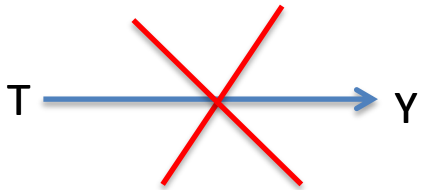

Y_0
Y_1

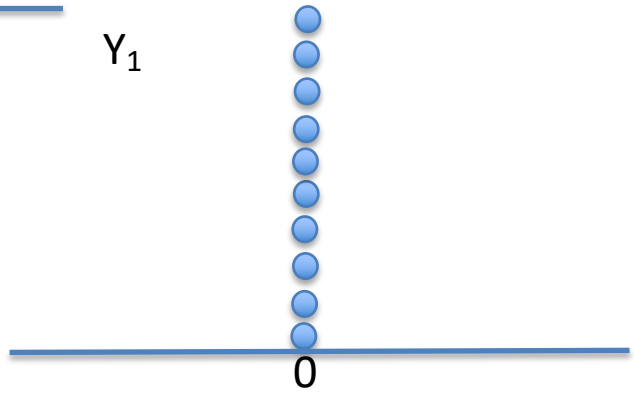

Y_0
T
Y
Y_1
0


Y_0
T
Y
Y_1
0


Y_0
T
Y
Y_1
0

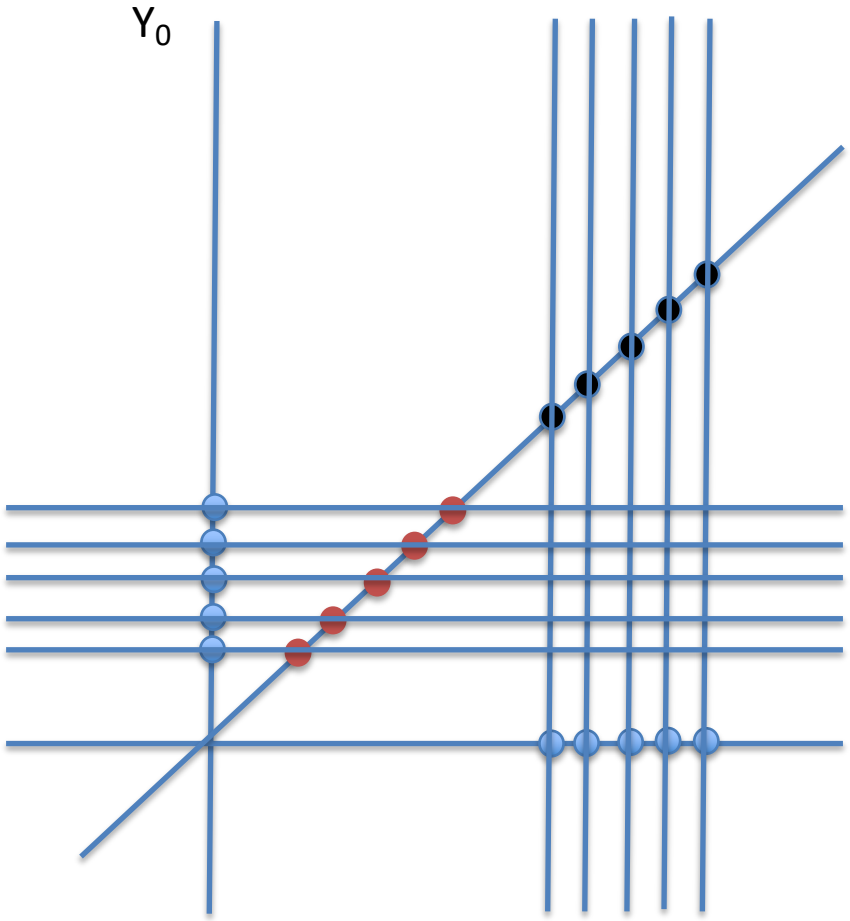


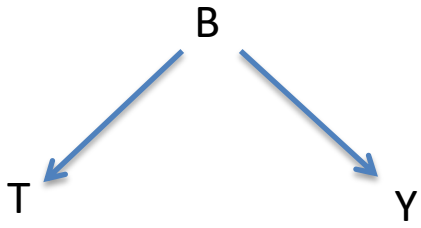


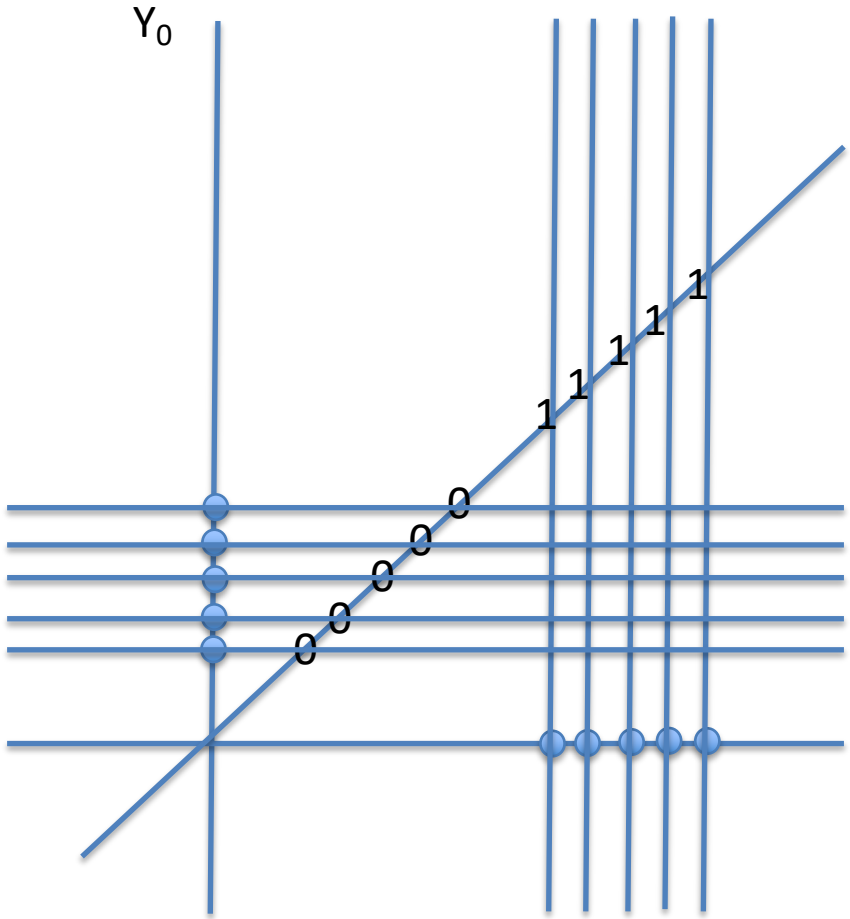


$E(Y_1-Y_0) \neq E(Y_t-Y_c)$

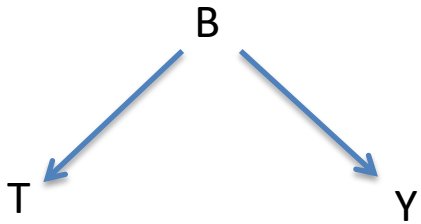
Systematic assignment selection bias: potential outcomes imbalanced across treatment groups.

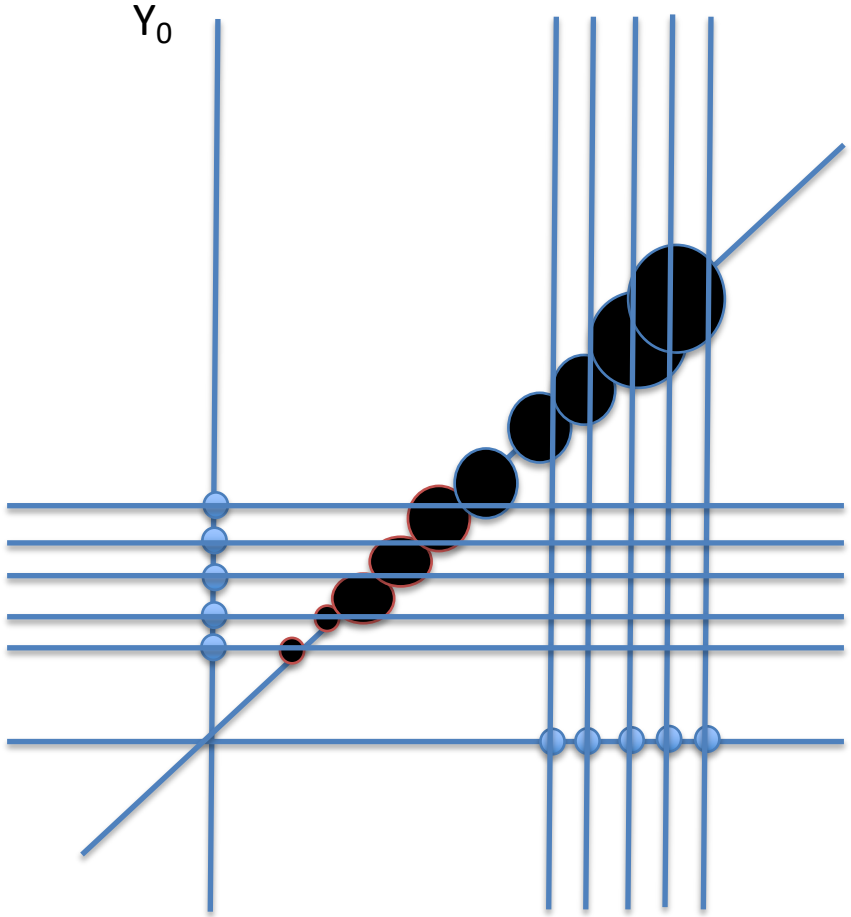


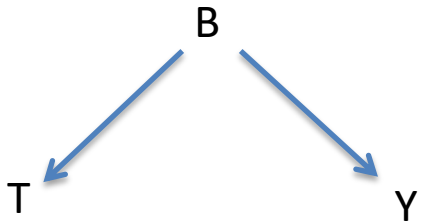




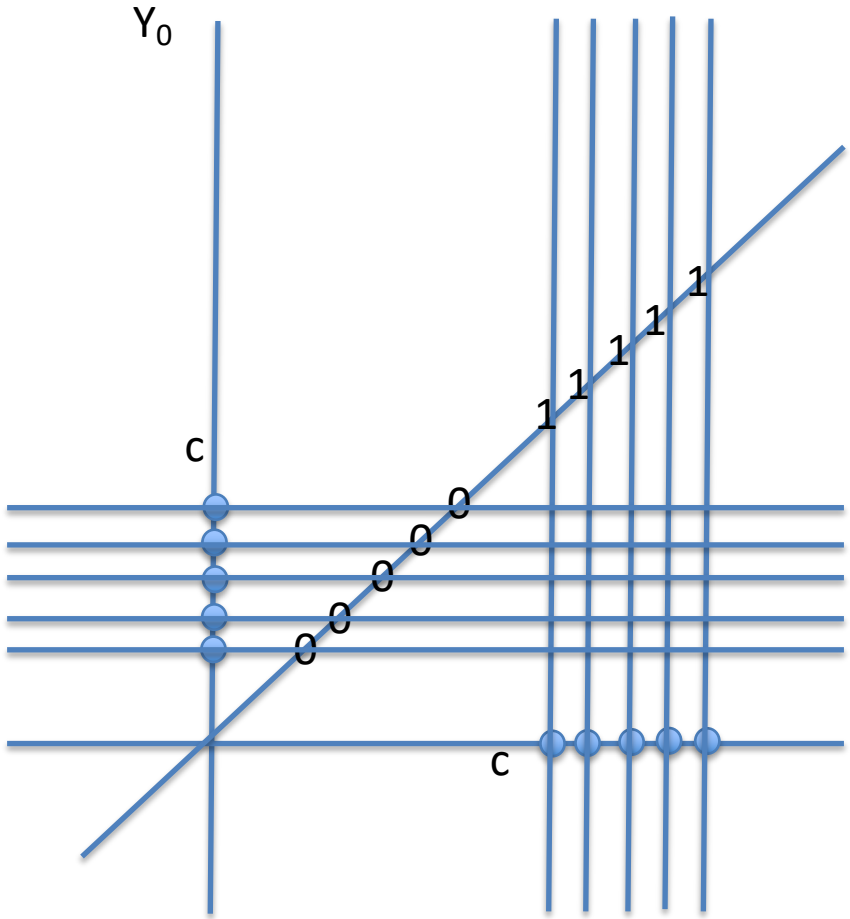
Natural propensity

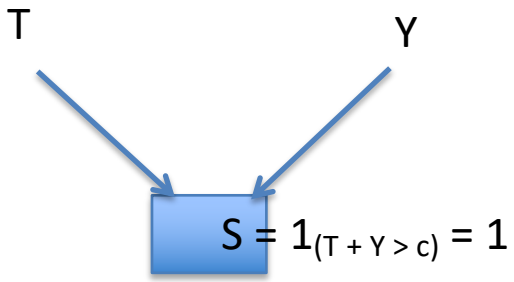


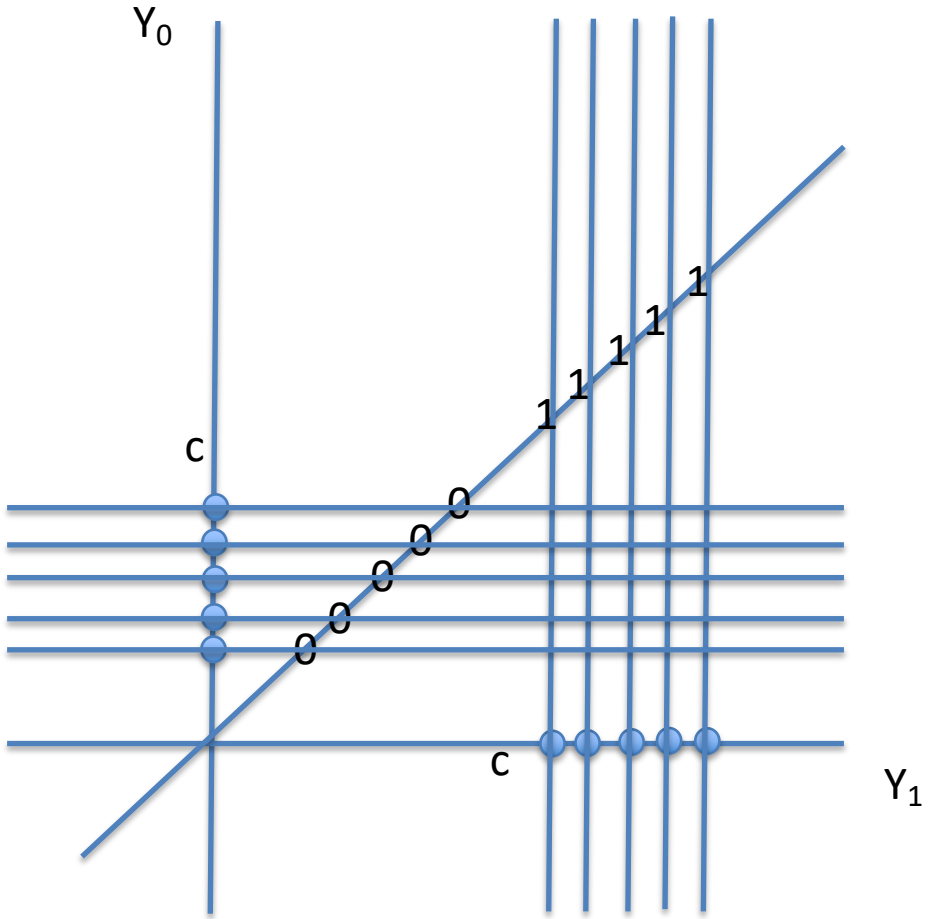




Continuous covariate.


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

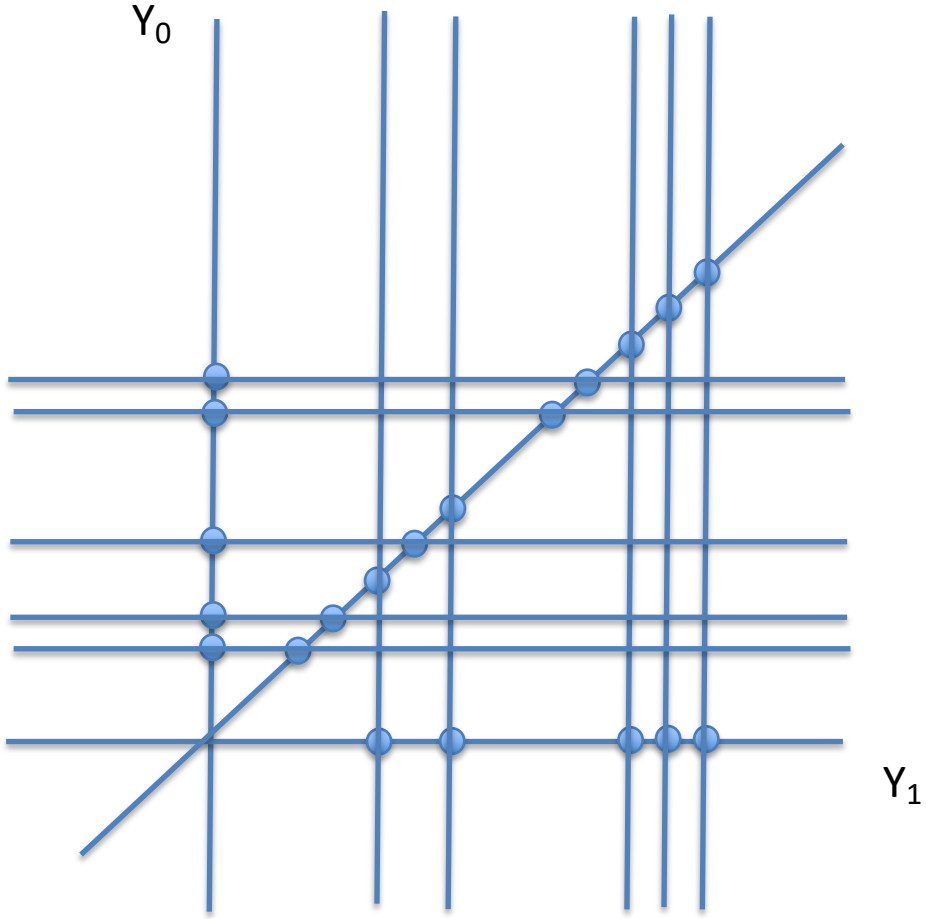

T
Y
$S = 1_{(T + Y > c)} = 1$

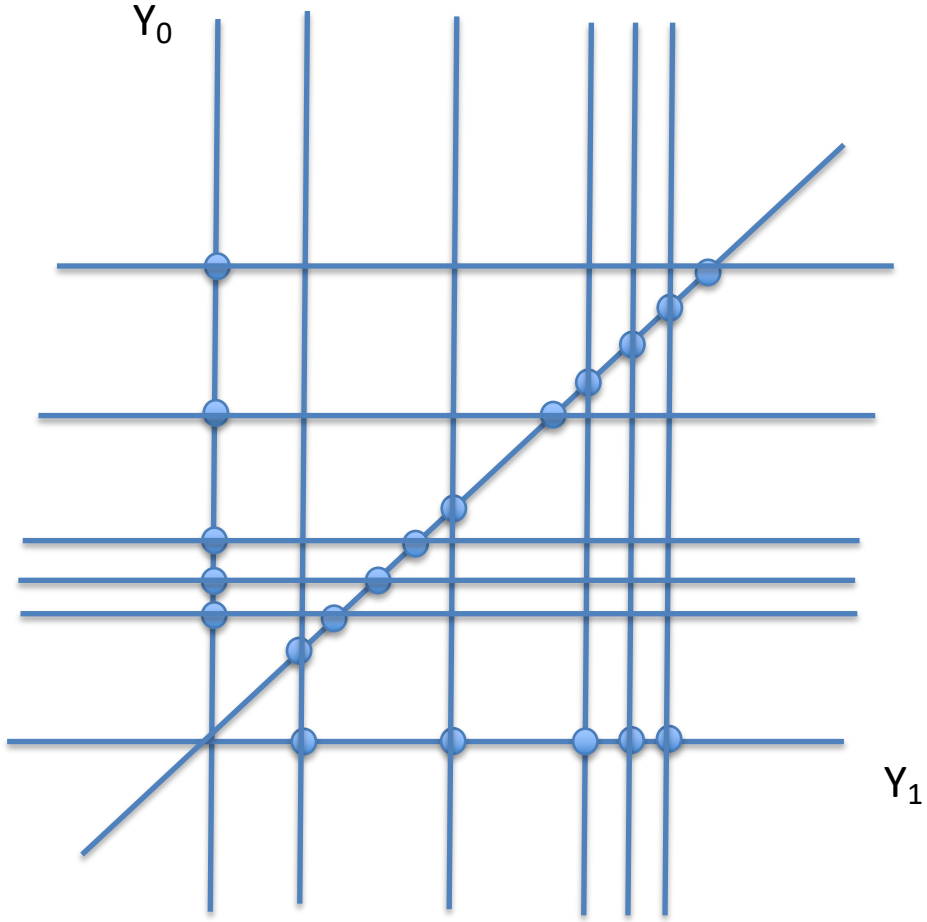


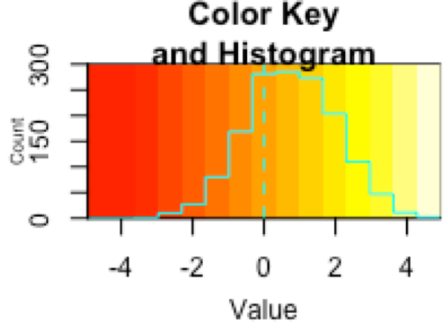
Backdoor theorem

Identification

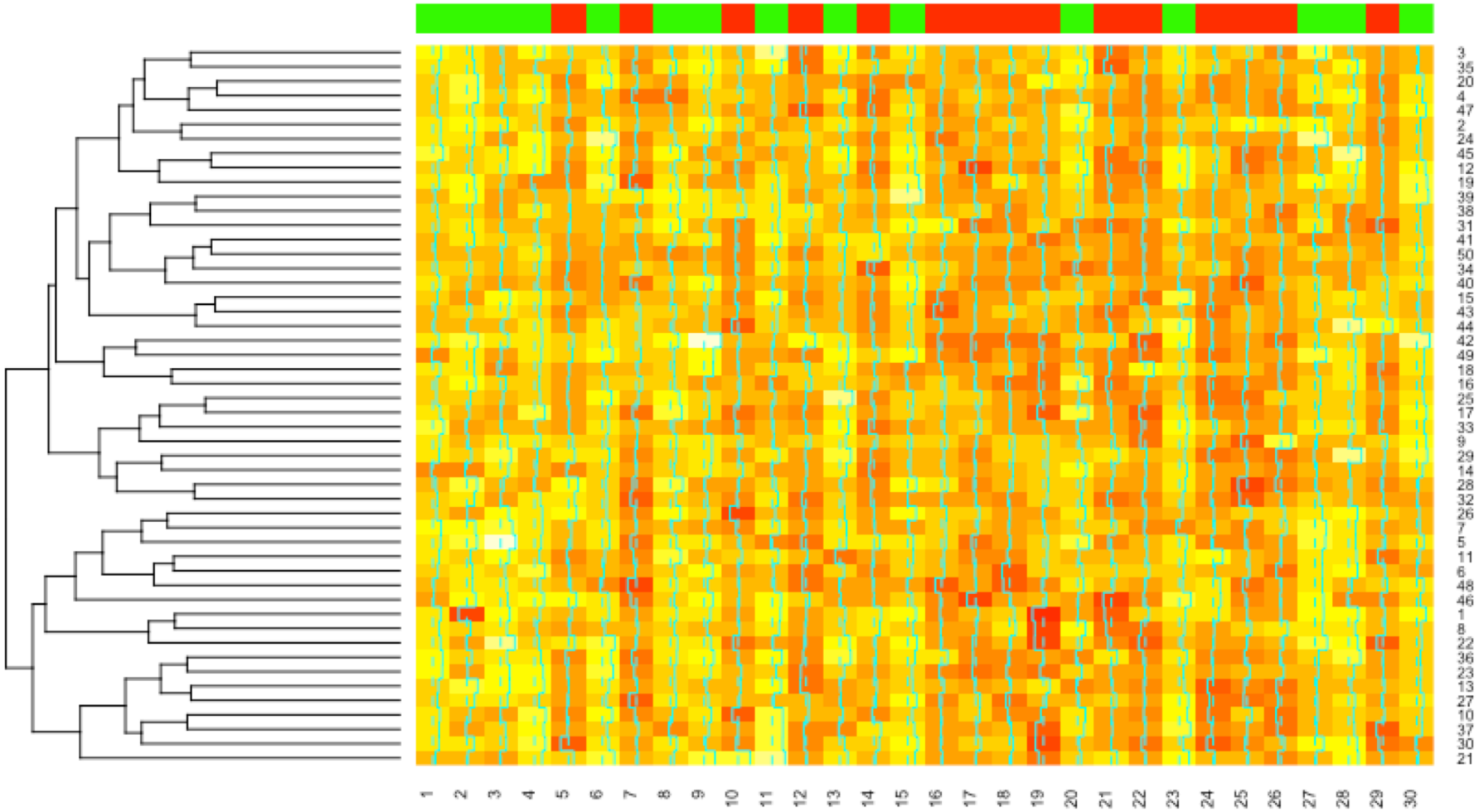
- Randomization
- Conditioning
- Instrumental variables
- Inverse-probability

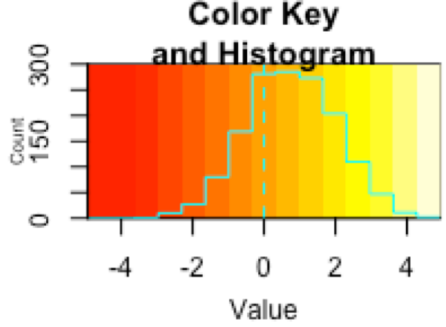

Y_0
Y_1


Y_0
Y_1

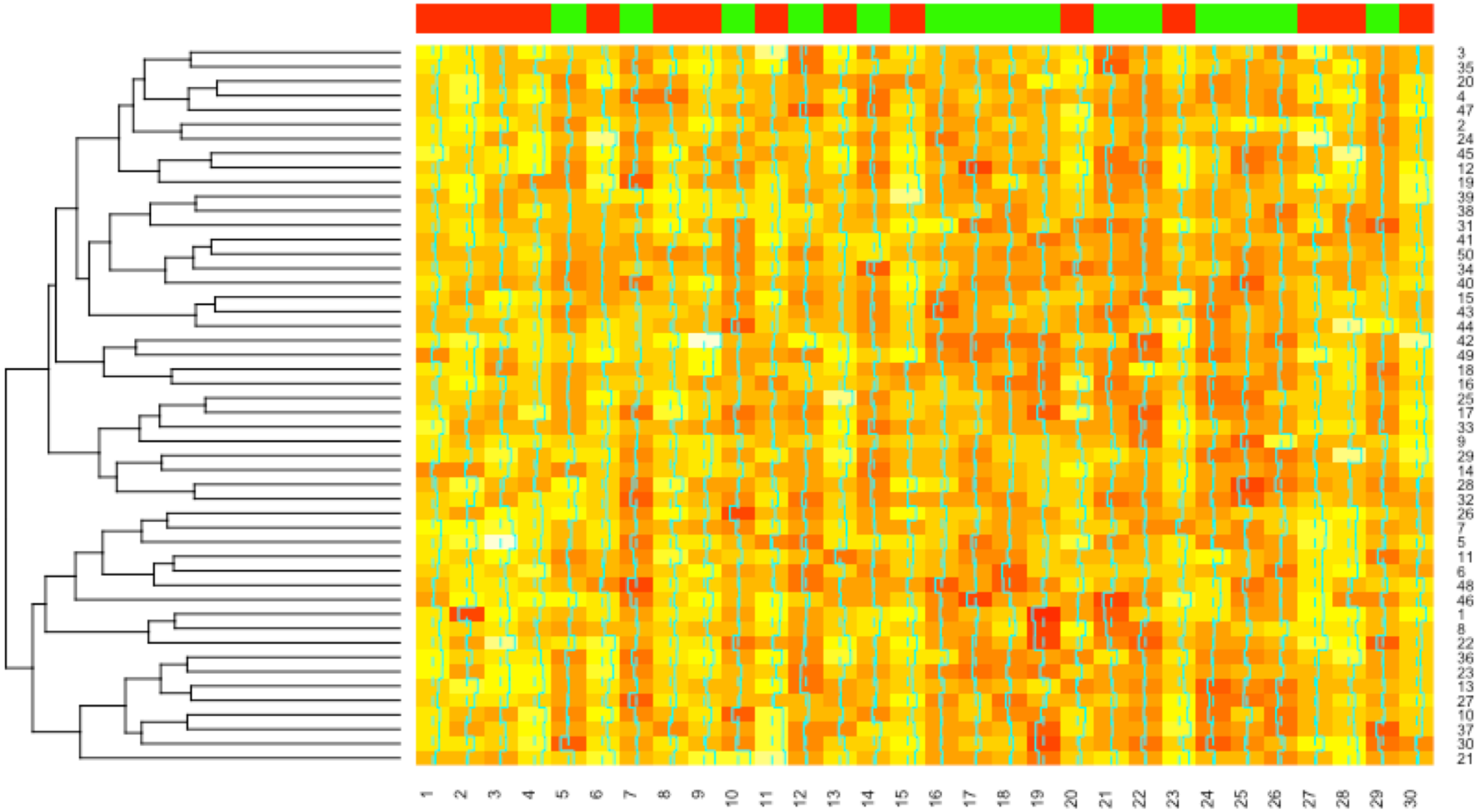


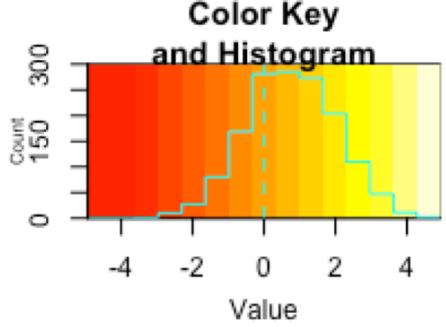
Actual observed gene expression.



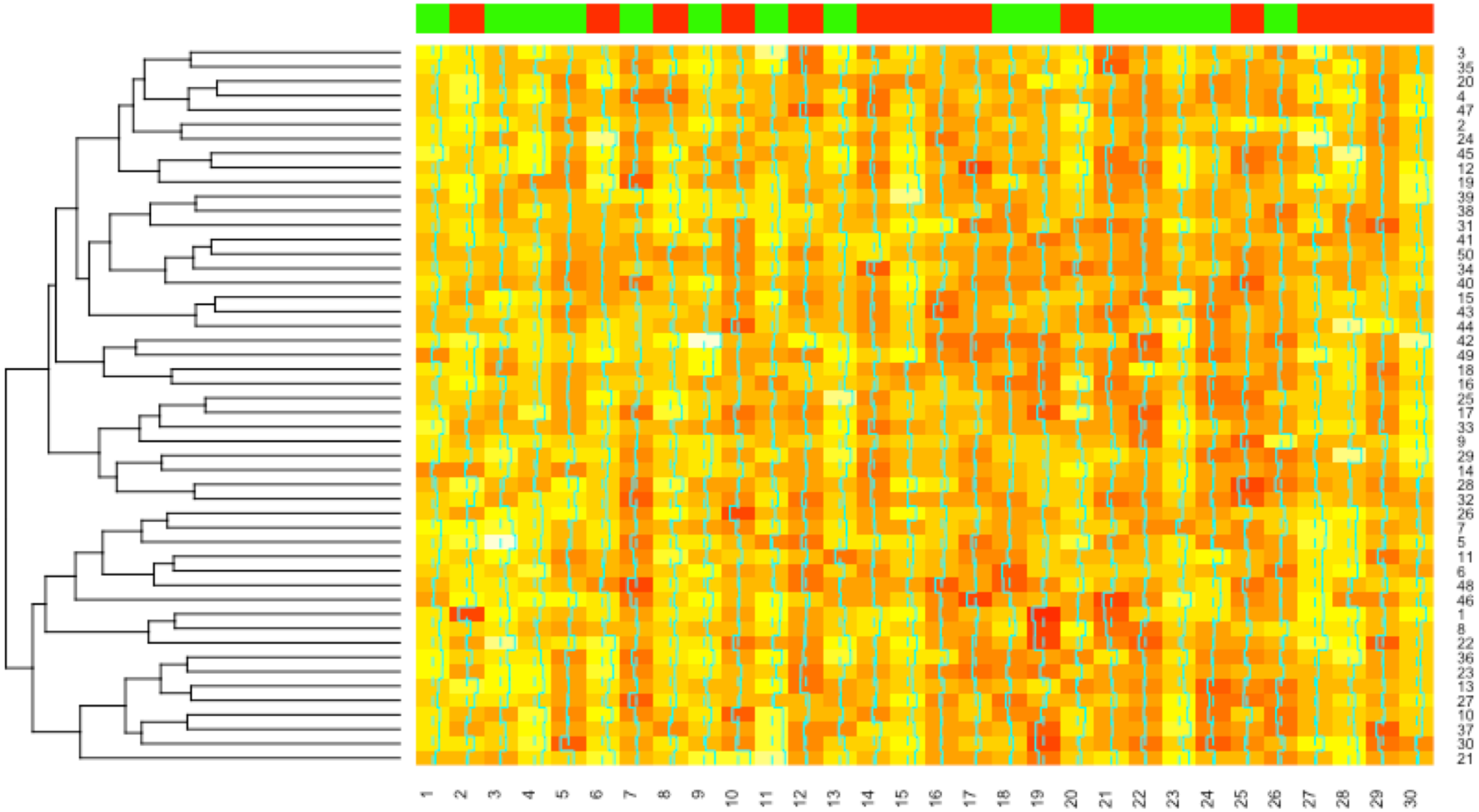


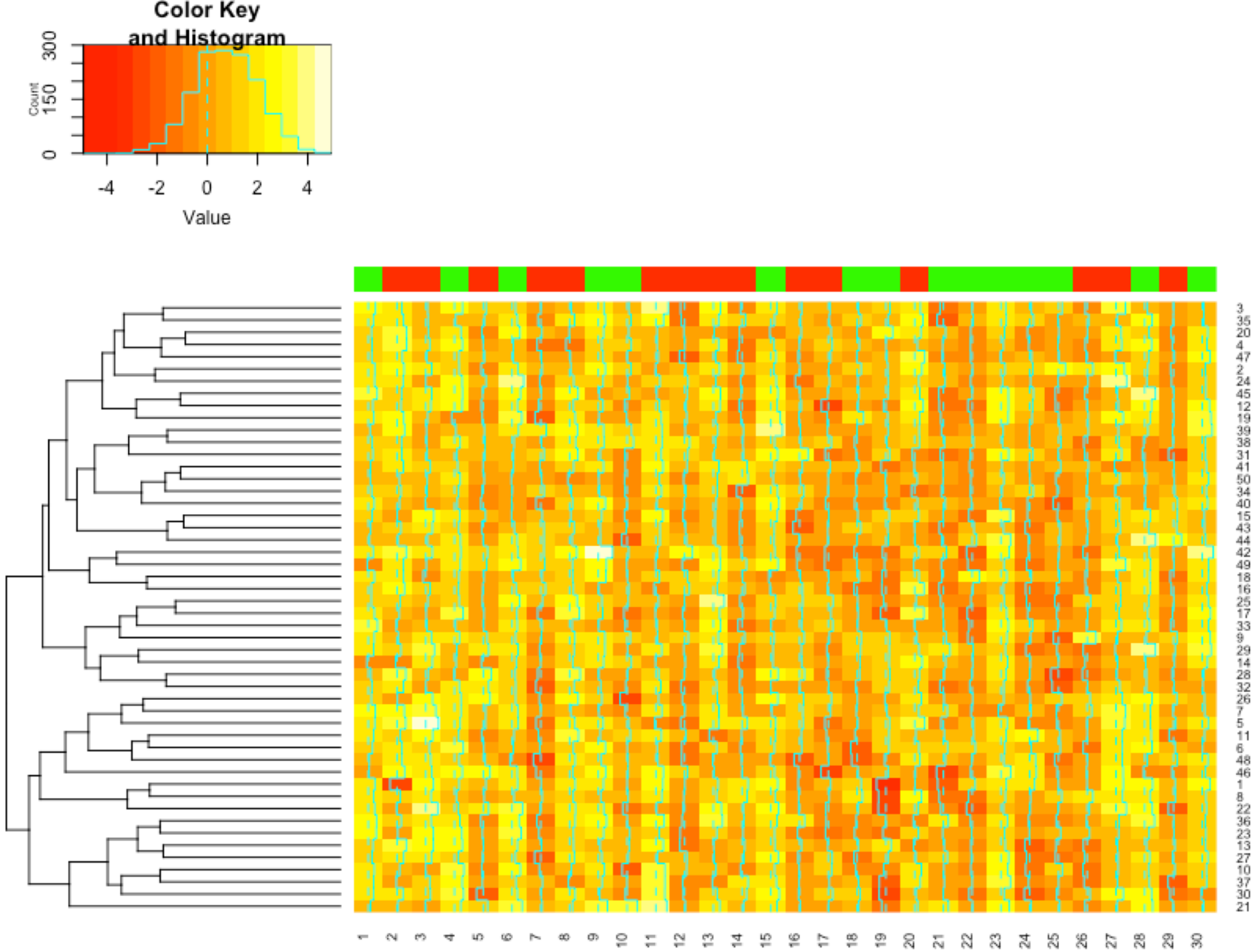
Counterfactual gene expression
under Fishers sharp imputation.

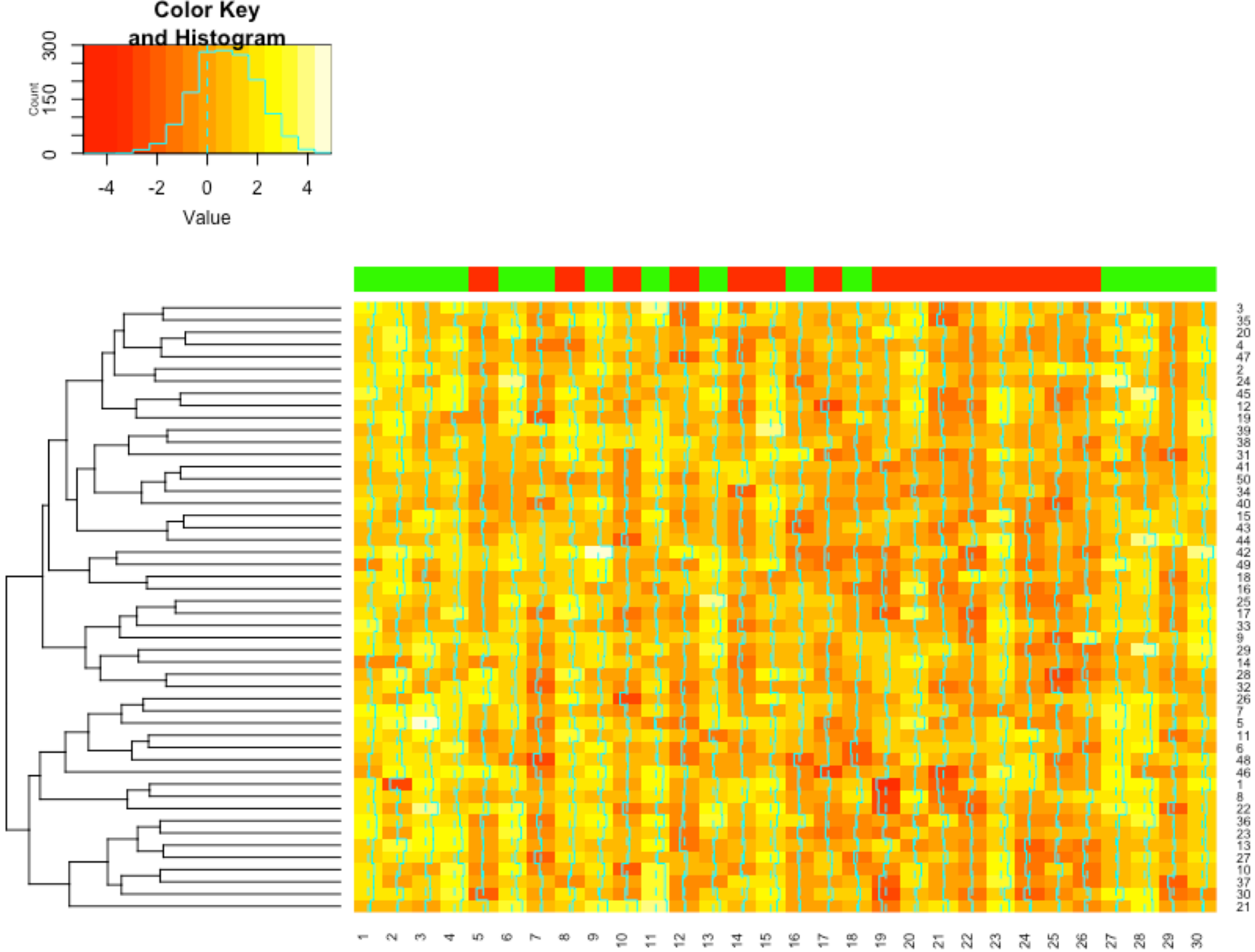


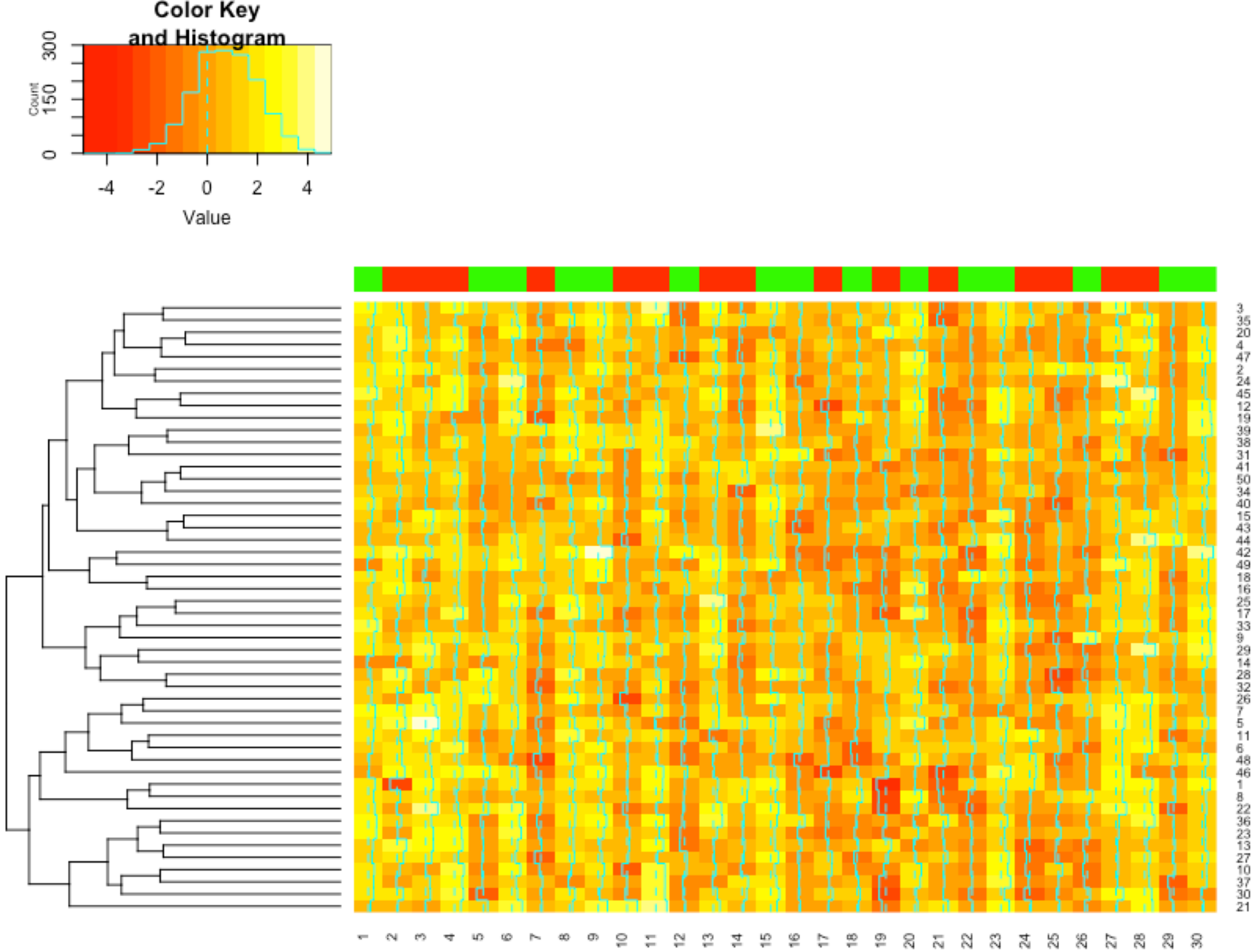


Stochastic proof by contradiction.
Gene-level (DE, MCP)
Set-level (MCP if multiple sets)
Network-level




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

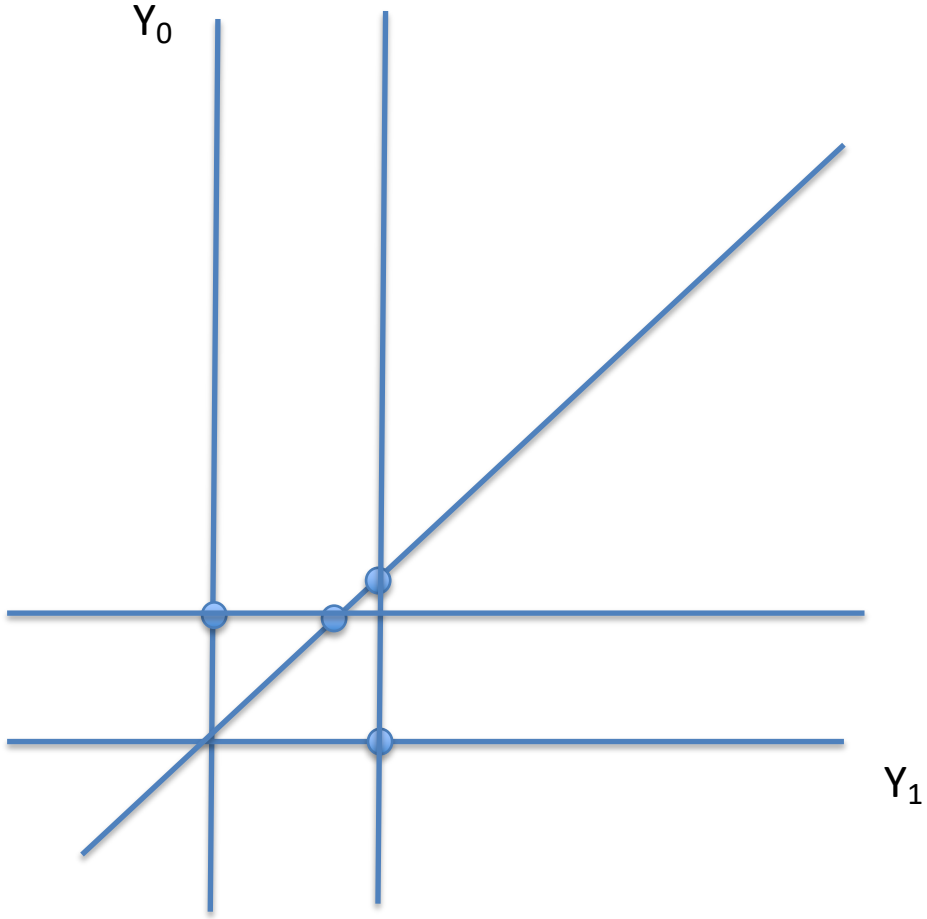
- Fisher's finite-sample
- Neyman/Rubin super-population
  - interval estimation requires additional assumptions beyond what is required for testing the causal null hypothesis
- *etc, …*

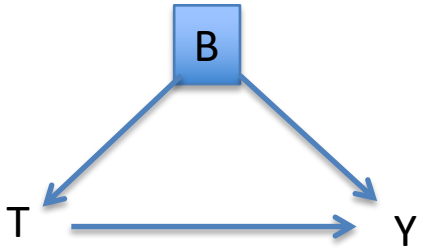
Identification

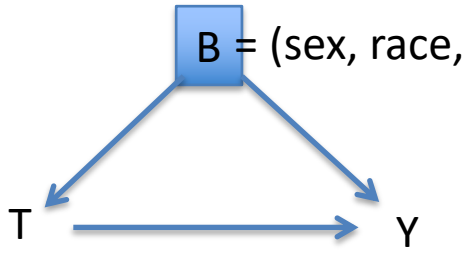
- Randomization
- Conditioning
- Instrumental variables
- Inverse-probability

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

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


Y_0
Y_1


B
T
Y

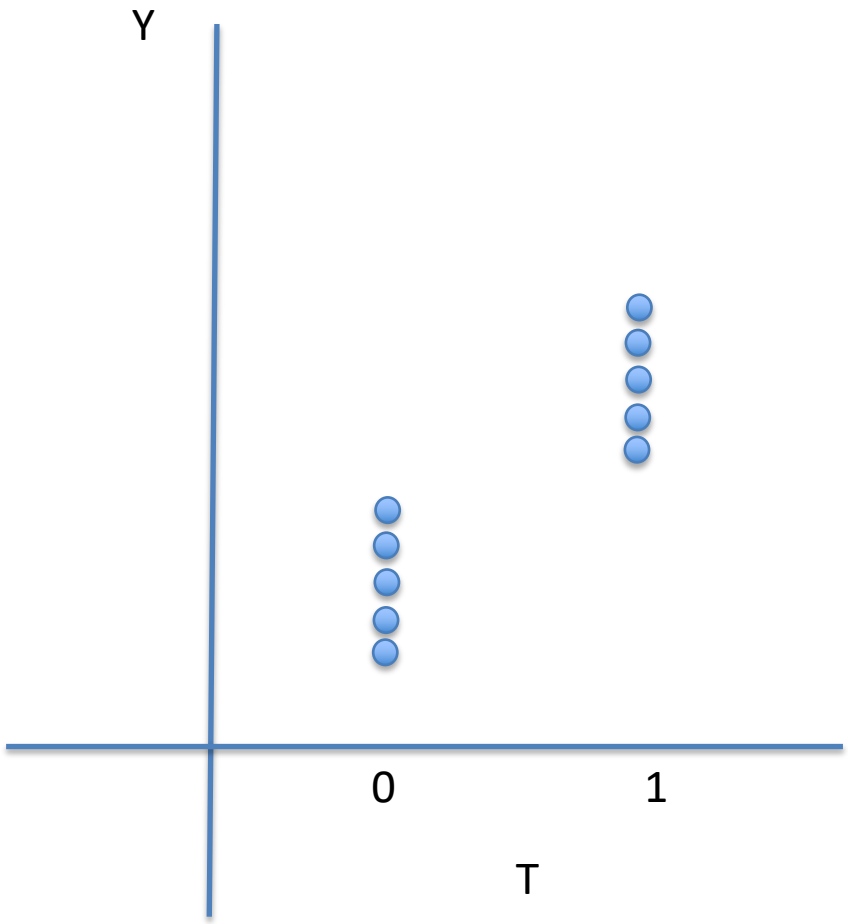


The picture can't be displayed.

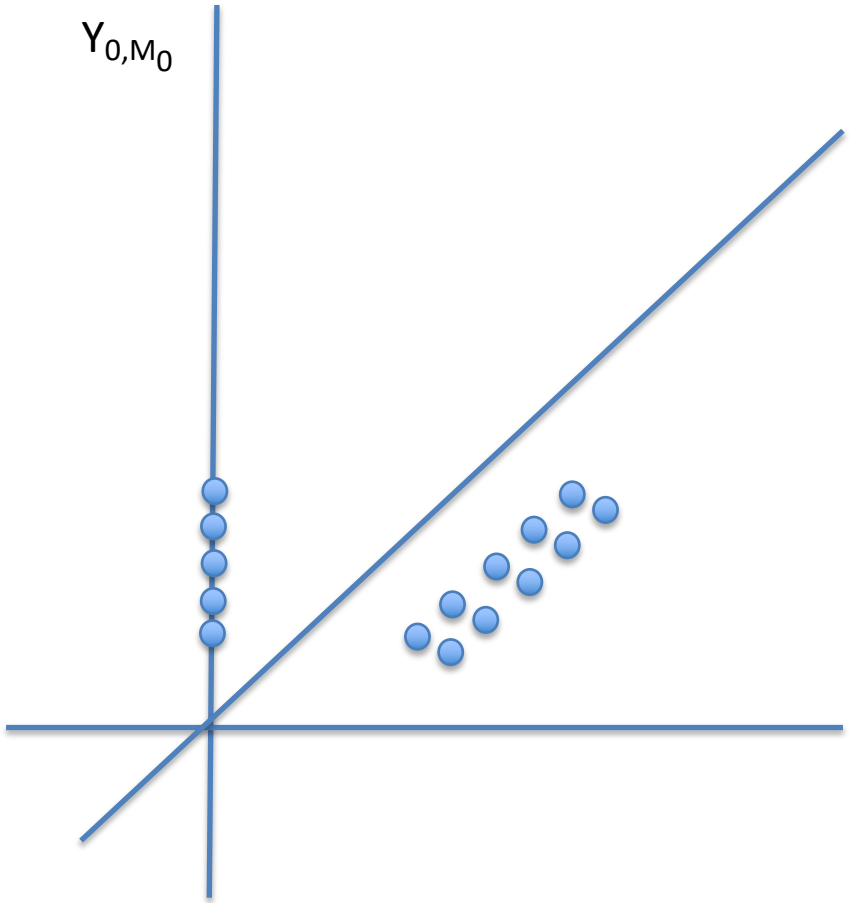
	p1	p2	p3
Inflammatory	0.787	0.556	0.801
Interferon	0.032	0.073	0.265
Antibody	0.690	0.254	0.046
All	0.033	0.084	0.285

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



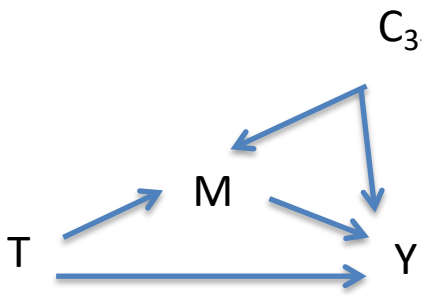
Proximal/distal
Environmental
Brain
Cognitive
Molecular – KEGG DAG/TELIS

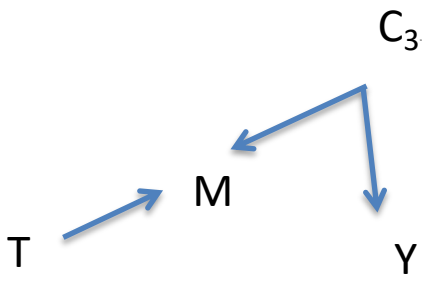


Natural variation in response of M to treatment, across units.

Exclusion restrictions.
RCT assumptions not enough.
Heterogeneity.

Exclusion restrictions.
RCT assumptions not enough.

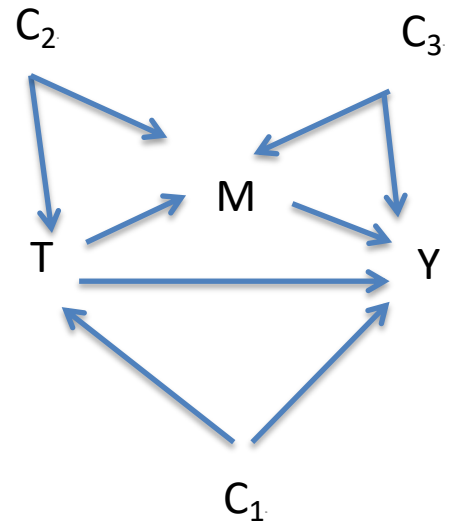


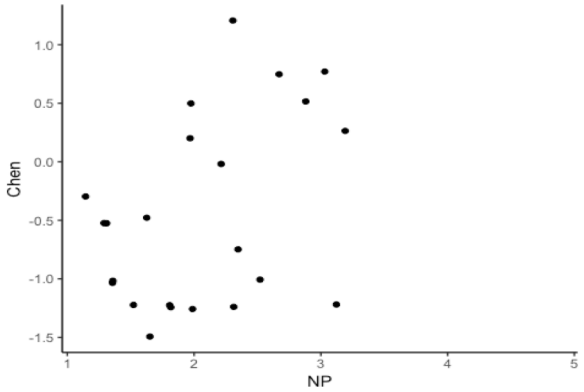


Exclusion restrictions.
Sequential ignorability.
RCT assumptions not enough.
Sequential ignorability.
Assume,
$Y = C_3$
$M = T + C_3$

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



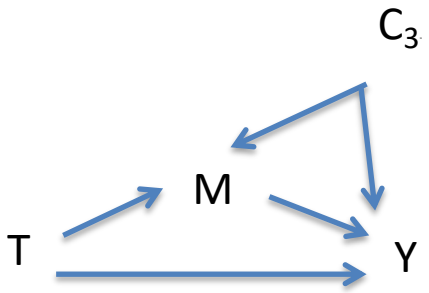
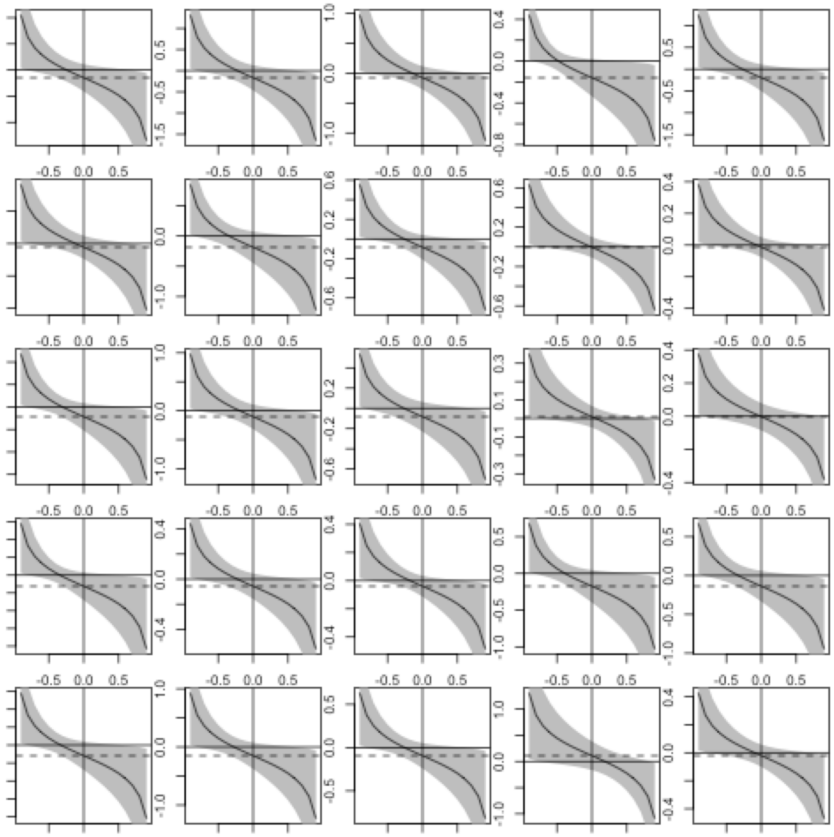


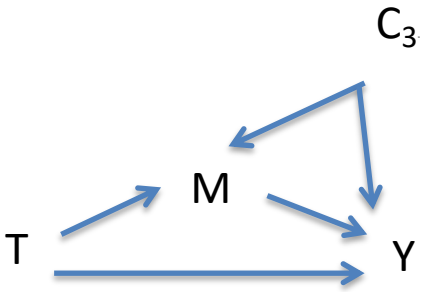
Table 3:The relation between NP, the CTRA gene set and its component subsets, with and without covariate adjustment (see Q1).

	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

Information box:
Defining and representing causation

- Modern definitions of total, and path-specific effects - direct and indirect - are general, and not tied to any specific statistical model.

- A causal variable is defined as any variable which changes the potential outcome of another variable. This idea can be interpreted as follows. First suppose we know the equations which dictate the natural directions of causation between variables in some system. Next override the equation governing one focal variable, and instead switch this variable between two different values. By definition, this focal variable is a cause of any variable which responds to this intervention (through the remaining equations). The difference between these definitions is purely notational; potential outcome definitions can easily be converted to structural definitions. Potential outcomes can be viewed as a short hand notation for general structural equations (not necessarily linear or parametric). For example, take the following trivial, linear parametric structural equation model: we can abbreviate the structural causal equations $CTRA_i(X_i = 1) = d + c + e_i$, and $CTRA_i(X_i = 0) = d + e_i$ as $CTRA_i(1)$ and $CTRA_i(0)$ respectively. Note that only one potential outcome can be observed, the other is counterfactual. Causal inference, i.e. on $CTRA_i(1) - CTRA_i(0)$, thus requires identifying conditions which justify imputing the missing counterfactual. See (Pearl 2014) to explicitly compare the structural formulation of mediation side by side with the potential outcome formulation.

- Causation is defined *ceteris parabis*, i.e. at the level of each individual "unit" subjected to intervention. Various statistical methods aim to infer population parameters of these unit-level causal effects, such as propensity score matching and nearest-neighbor matching (which often uses the Mahalanobis metric, also called Mahalanobis matching), attempt to correct for the assignment mechanism by finding control units similar to treatment units on variables which confound causal effects (implied by ceteris parabis).

Information box: what is identification?

- A parameter is said to be identified if different parameter settings of the underlying data generating process imply different distributions over observed variables. This identifiability - or lack thereof - is not a statistical problem related to the challenges of statistical inference with small samples. Pearl (2009) provides one way to think about identification. Dependence between observed variables reflects some unknown mix of causal and noncausal ("backdoor") effects. A causal effect is identified when the observed association can been adjusted somehow to remove these noncausal components. For nonparametric identification, the analyst would describe the set of assumptions that will allow us to identify a causal effect without any distributional or functional form assumptions.
- To take a famous example, randomized treatment and the SUTVA identification (Rubin 1974) together nonparametrically identify the average total effect. To identify the indirect and direct effects, additional assumptions are necessary, e.g. "sequential ignorability".
- Causal identification assumes the investigator has domain knowledge to judge the plausibility of no confounding type of assumptions which underly all mediation methods, whether under the rubric of sequential ignorability (e.g., Imai et al., 2010b), uncorrelated error terms, or graphical criterea. The assumptions identifying mediation can be stated most succinctly in the latter.
- Identification conditions can be expressed in diverse ways, e.g. judging conditional independencies among counterfactual variables, often called strong ignorability, conditional ignorability , or sequential ignorability, presents a formidable task without structural models. Efforts to replace ignorability vocabulary - with notions such as no unmeasured confounders, no unmeasured confounding, as if randomized, effectively randomly assigned, or essentially random - create ambiguity. First, the notion of a confounder varies significantly from author to author. Some define a confounder (say of the NP-CTRA relationship) as a variable that affects both NP and CTRA. Some define confounder as a variable that is associated with both NP and CTRA. Others allow for a confounder to affect NP and be associated with CTRA. Worse yet, the expression no unmeasured confounders is sometimes used to exclude the very existence of such confounders and sometimes to affirm our ability to neutralize them by controlling other variables, not necessarily confounders. Second, the interpretations have taken sequential ignorability as a starting point and consequently are overly stringent – sequential ignorability is a sufficient but not necessary condition for identifying natural effects. Weaker conditions can be articulated in a transparent and unambiguous language which provide a greater identification power and a greater conceptual clarity.

Information box: Alternatives to sequential ignorability conditions for identification

- Instrumental variables offer a very different answer from a causal mediation analysis (Keele 2015). Mechanisms based on IV have the advantage that one can allow for the possibility of unobserved confounding between the mediator and the outcome. However, to identify the indirect effect, one must assume that the direct effect is zero. The assumption that the direct effect is zero is widely referred to as the exclusion restriction (Angrist, Imbens, & Rubin, 1996). Thus, one must assume that there is only an indirect effect, which implies that the effect of the treatment is entirely mediated. Under this form of mechanism, we must assume that the effect of a NP only works through Chen: There cannot be any other mechanisms for the intervention.
- Statistically "controlling" for M in the analysis (by including M in the regression equation) does not physically disable the paths going through M ; it merely matches samples with equal M values, and thus induces spurious correlations among other factors in the analysis, see (Pearl 2014). This can be readily shown using classical path-tracing rules. Such dependence cannot be detected by statistical means, so theoretical knowledge must be invoked to identify the sources of these correlations and control for common causes (so called "confounders") of M and CTRA whenever they are observable. This approach to mediation has two major drawbacks. One (mentioned above) is its reliance on the untested assumption of uncorrelated errors, and the second is its reliance on linearity and, in particular, on a property of linear systems called effect constancy (or no interaction): The effect of one variable on another is independent of the level at which we hold a third. This property does not extend to nonlinear systems; in such systems, the level at which we control M would in general modify the effect of T on CTRA. For example, if the output CTRA requires both T and M to be present, then holding M at zero would disable the effect of T on CTRA , while holding M at a high value would enable the latter.

Information Box: Modern mediation

- Although one could define mediation statistically, we follow the causal definition.
- The conventional mediation analysis entails fitting a set of linear regression models: ``mediation effects'' are defined in terms of these estimated model parameters. One problem with *defining* mediation in terms of statistical changes induced by adding a third mediator variable into a regression equation, is that mediation is inherently a causal notion hence should not be defined in statistical terms. Modern approaches therefore define mediation in terms of potential outcomes, or equivalently causal graphs. In the language of the latter, a mediator is then an intermediate variable that lies on the causal path from the treatment to the outcome. This definition is grounded in the notion of a causal path and emphasizes the difference between "fixing a variable" and "statistically adjusting for" (conditioning on) a variable as in regression.
- To illustrate our measure of ACME more formally, consider a binary measure of negative parenting, a variable we call t which takes 0 or 1. We will now define indirect effect of NP - via mediator Chen M - within the modern framework. $M_i(t)$ is the effect of NP on Chen for subject i under treatment (NP) status t. Let $CTRA_i(t, m)$ denote the potential outcome if NP and Chen took values t, m respectively. We only observe one of these potential outcomes $CTRA_i(t_i, M_i(t_i))$, where $M_i(t_i)$ is the observed value of Chen at the observed NP level t_i. $CTRA_i(t, M_i(t))$ is the effect of t on CTRA, which in general and be transmitted both indirectly, through $M_i(t)$, and "directly" (i.e. not through M but possibly through some independent mediators). Let the total causal effect for unit (subject) i be
- $\tau_i = CTRA_i(1, M_i(1)) - CTRA_i(0, M_i(0))$
- and the unit-level indirect effect be
- $\delta_i = CTRA_i(t, M_i(1)) - CTRA_i(t, M_i(0))$.
- This latter relates to the following counterfactual question: how would CTRA change in this indidividual if we were to physically (counterfactually) change Chen's value under $t = 0$ (no negative parenting) to that under $t = 1$ (negative parenting), while keeping NP at its observed value t? Because these two values of Chen would naturally occur as responses to changes in NP, this quantity formalizes the notion of a causal mechanism that the causal effect of the treatment is transmitted through changes in the mediator of interest. Similarly, we define the unit direct effect, corresponding to all other possible causal mechanisms (sometimes refered to en masse as the "direct effect"), as:
- $\gamma_i = CTRA_i(1, M_i(t)) - CTRA_i(0, M_i(t))$.
- The counterfactual question here is: how would CTRA respond to NP change $T_i = 0$ to $T_i = 1$, if (counterfactually) Chen was held constant? Mediation analysis creates an identification problem. The quantity $CTRAi(1, M_i(0))$, for example, is unobservable, but to estimate the mediation effect we need assumptions which link this unobserved counterfactual to observed quantities. We examine these assumptions. Such definitions can easily be extended to continuous treatments (NP not binary) (Imai, Keele, and Yamamoto 2010).