Gene expression

Justin Chumbley

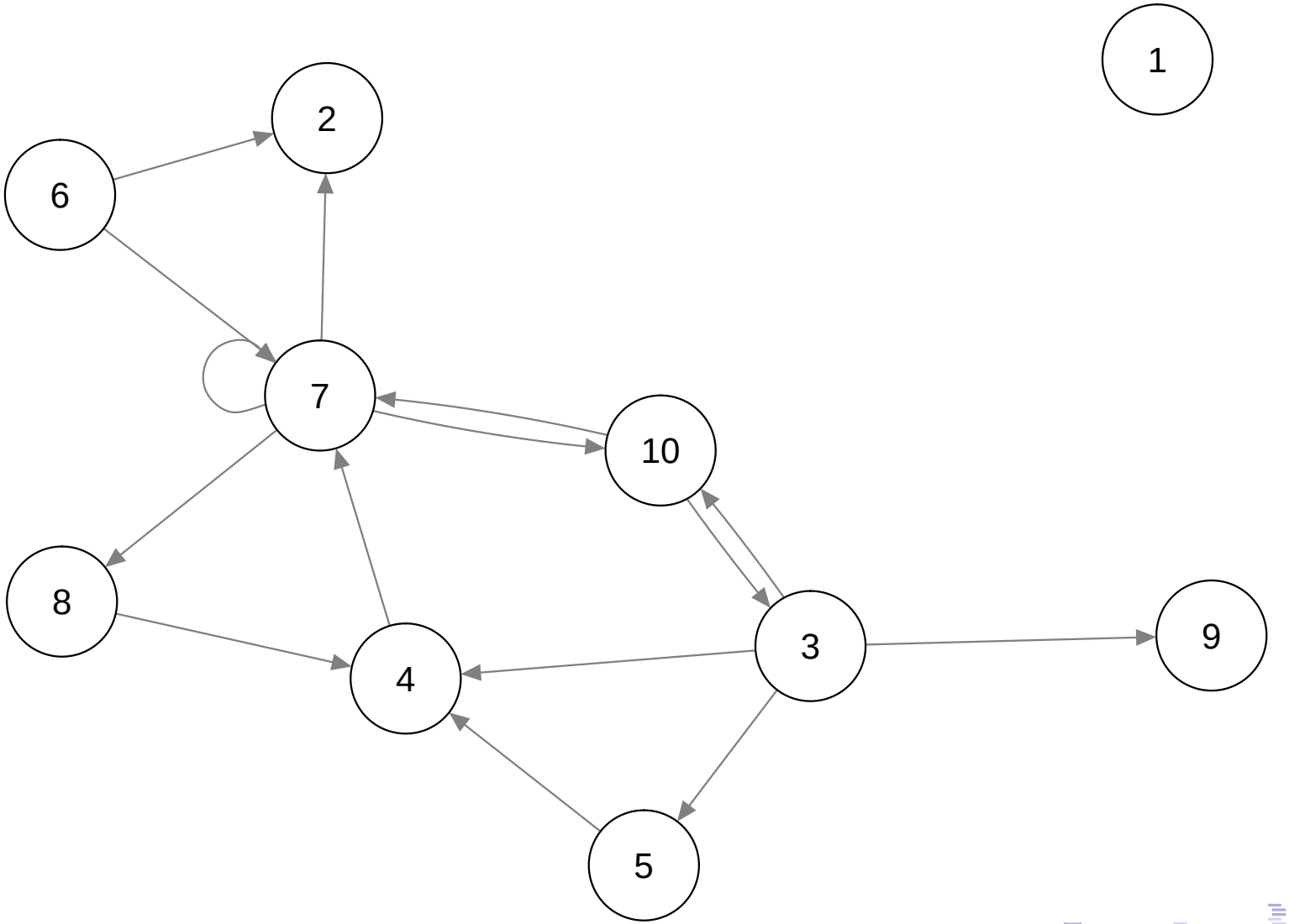
Jan 2017

Social genomics

- Social causation
- Biological causation
- Coupled systems

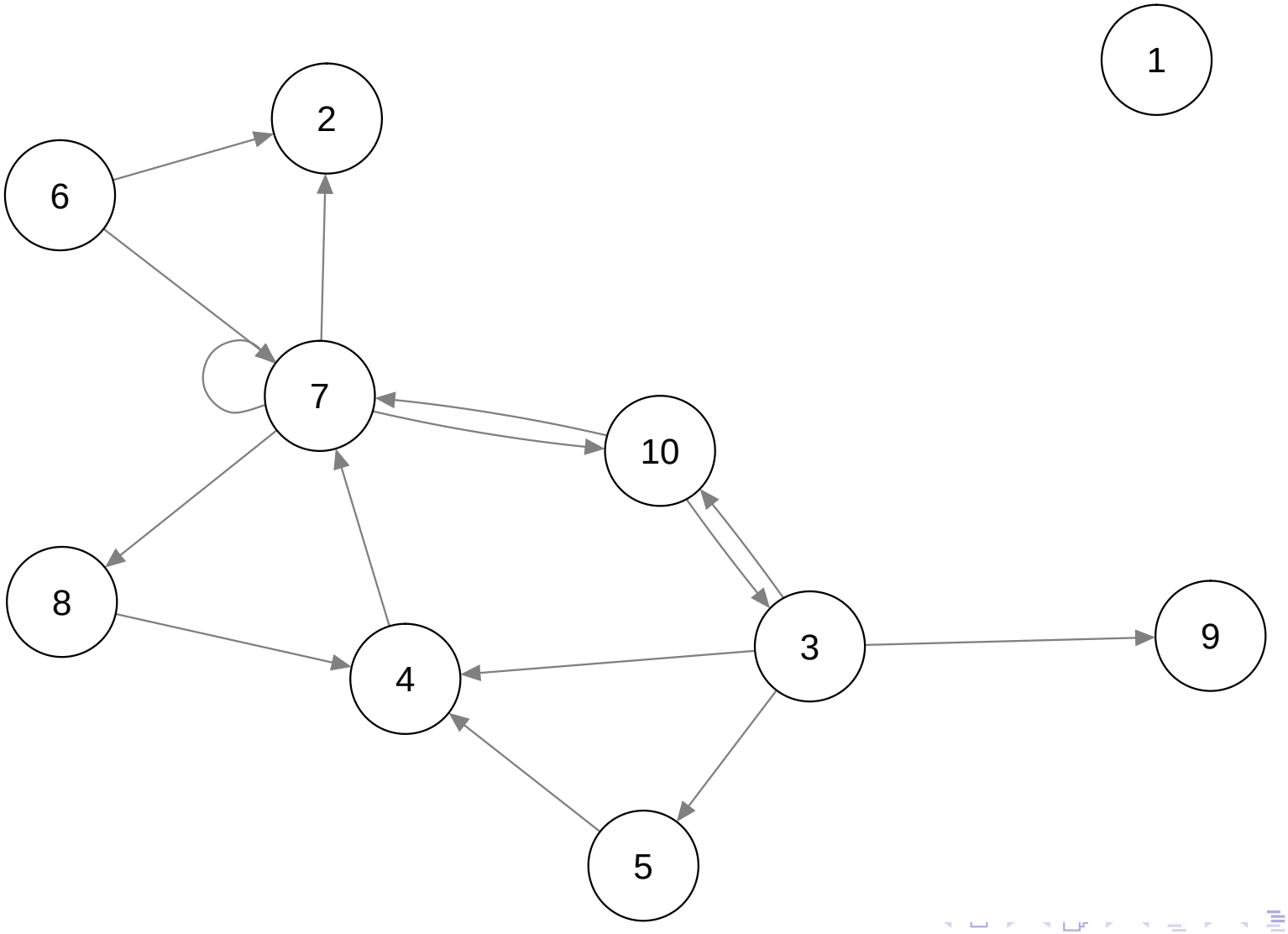
Causation

- Conditional determinism (general SEM)



Causal inference

- Fisher, Pearl

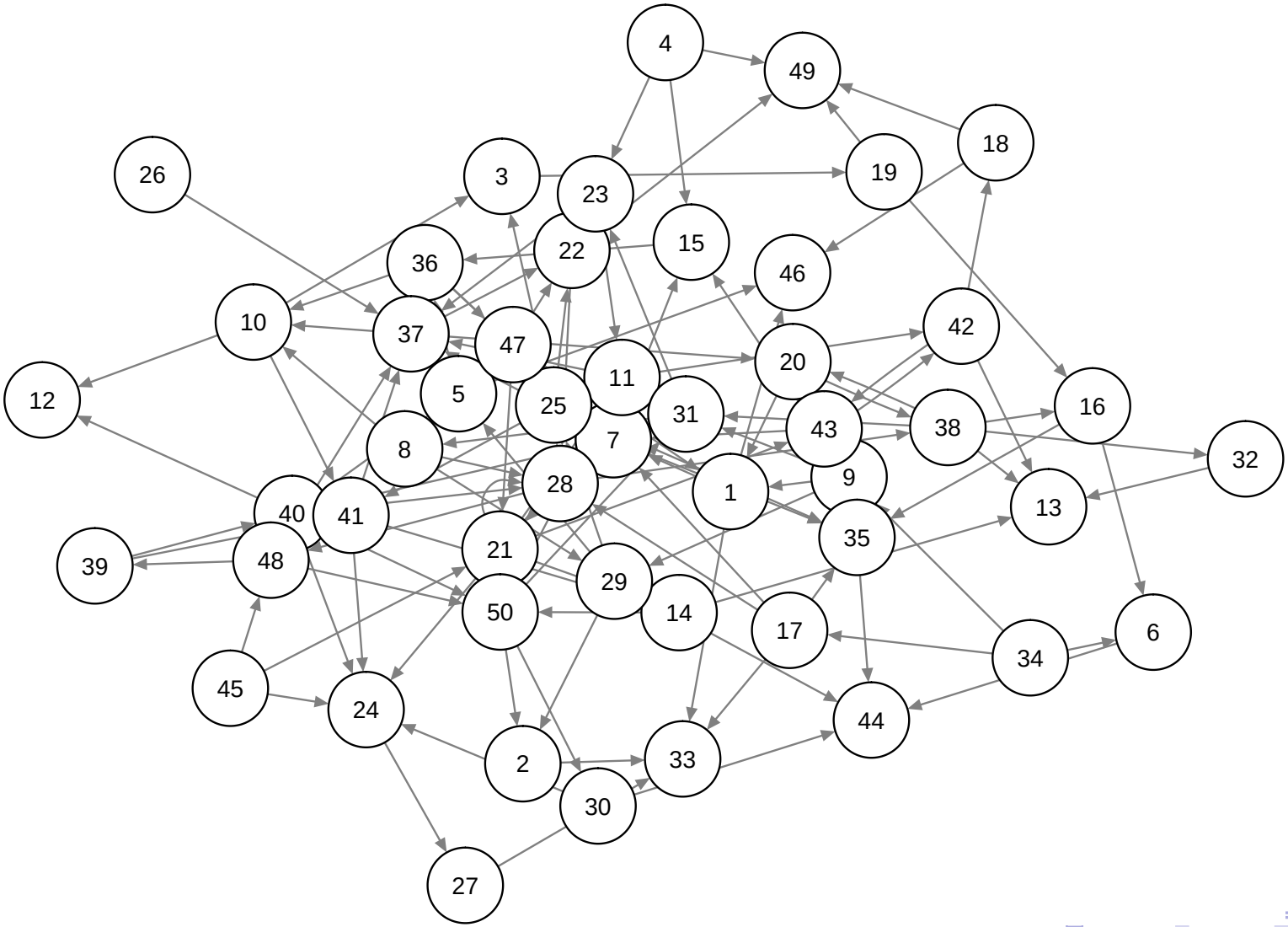


High-dimensional inference.

- High-dimensional, dependent data: $p > n$
- High-dimensional, dependent hypotheses: 2^{Ω} subsets $> p > n$.
- Overfitting/overfitting (multiple testing).
- Types of aggregate error (FWER, FDR, I^2, ...).
- Control concepts/procedures (Bonferonni, BH, Ridge).
- Model dependence structure (induced and native).
- "Effective" dimensionality: 2^{Ω} subsets $> p >? > n$.
- Subset selection/subspace projections.

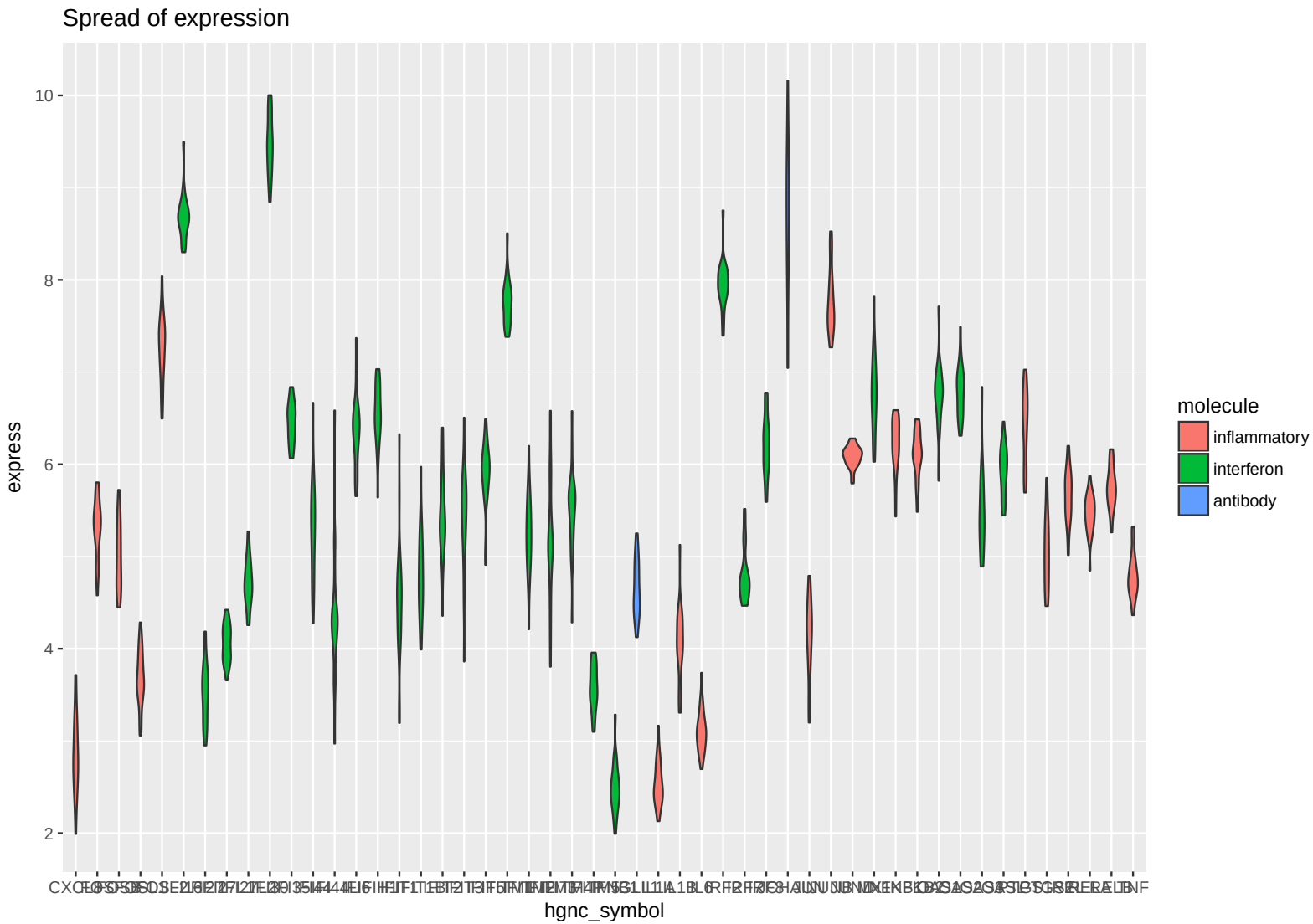
High-dimensional causal inference

- Causal identification/causal false positives.

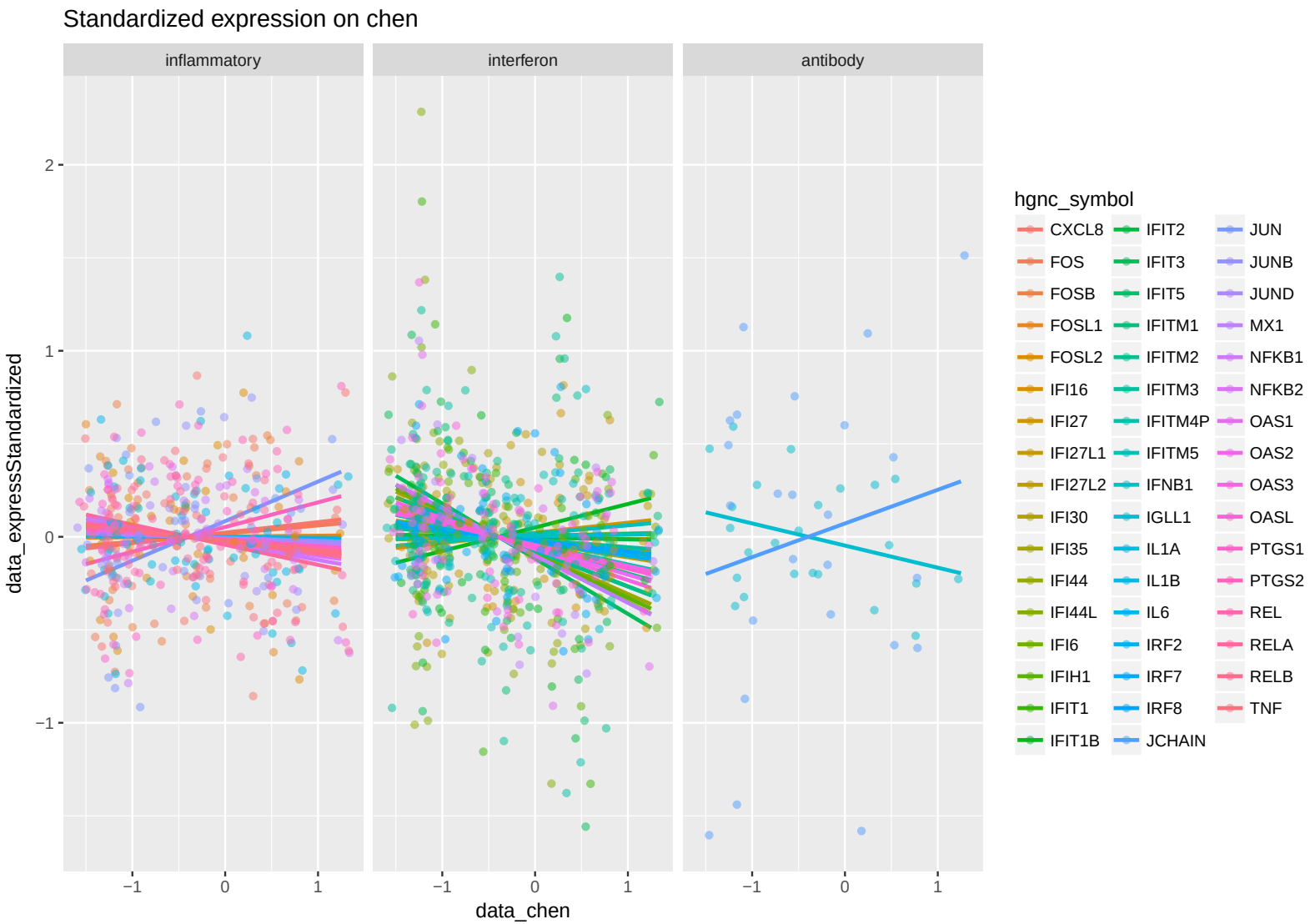


The 53 gene CTRA

- 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.


Spread of expression
express
10
8
6
4
2
hgnc_symbol
molecule
inflammatory
interferon
antibody

Mass-univariate



Mass-univariate: Issues

- False positives: bias from ommitted confounds
- False positives: overfitting
- False negatives: power

1. Condition on independent causes
2. Pool across a set
3. Tailor multiplicity correction to dependence

Mass-univariate

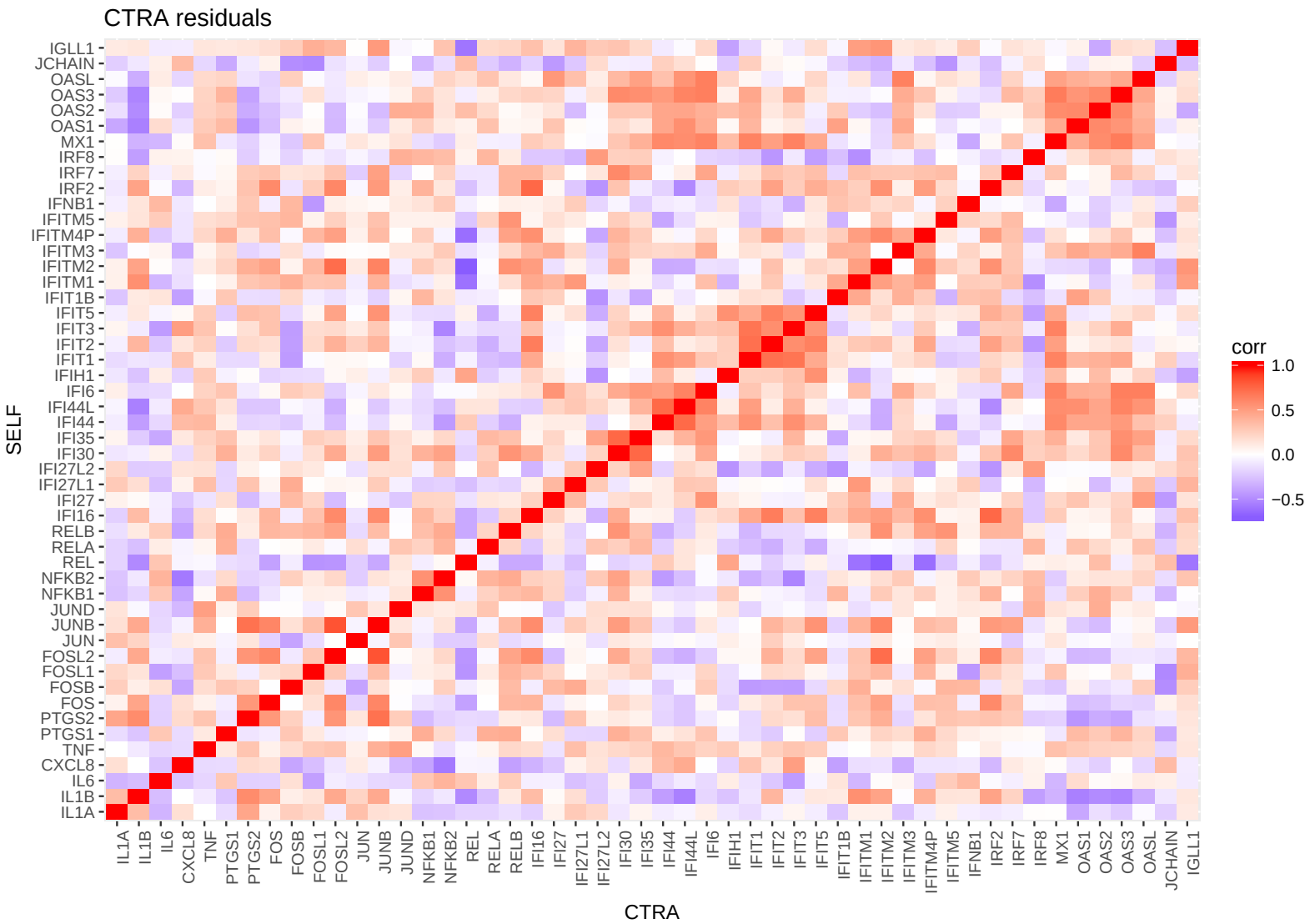
	logFC	AveExpr	t	P.Value	adj.P.Val	B
IFIH1	-0.235	6.552	-3.041	0.005	0.256	-2.112
MX1	-0.254	6.763	-2.549	0.017	0.416	-3.016
IFIT3	-0.297	5.420	-2.294	0.030	0.493	-3.451
JUN	0.213	4.115	2.006	0.055	0.496	-3.907
IFITM2	-0.191	5.241	-1.859	0.074	0.496	-4.123
RELA	-0.108	5.460	-1.768	0.088	0.496	-4.250

Table 1: Limma regression on CTRA.

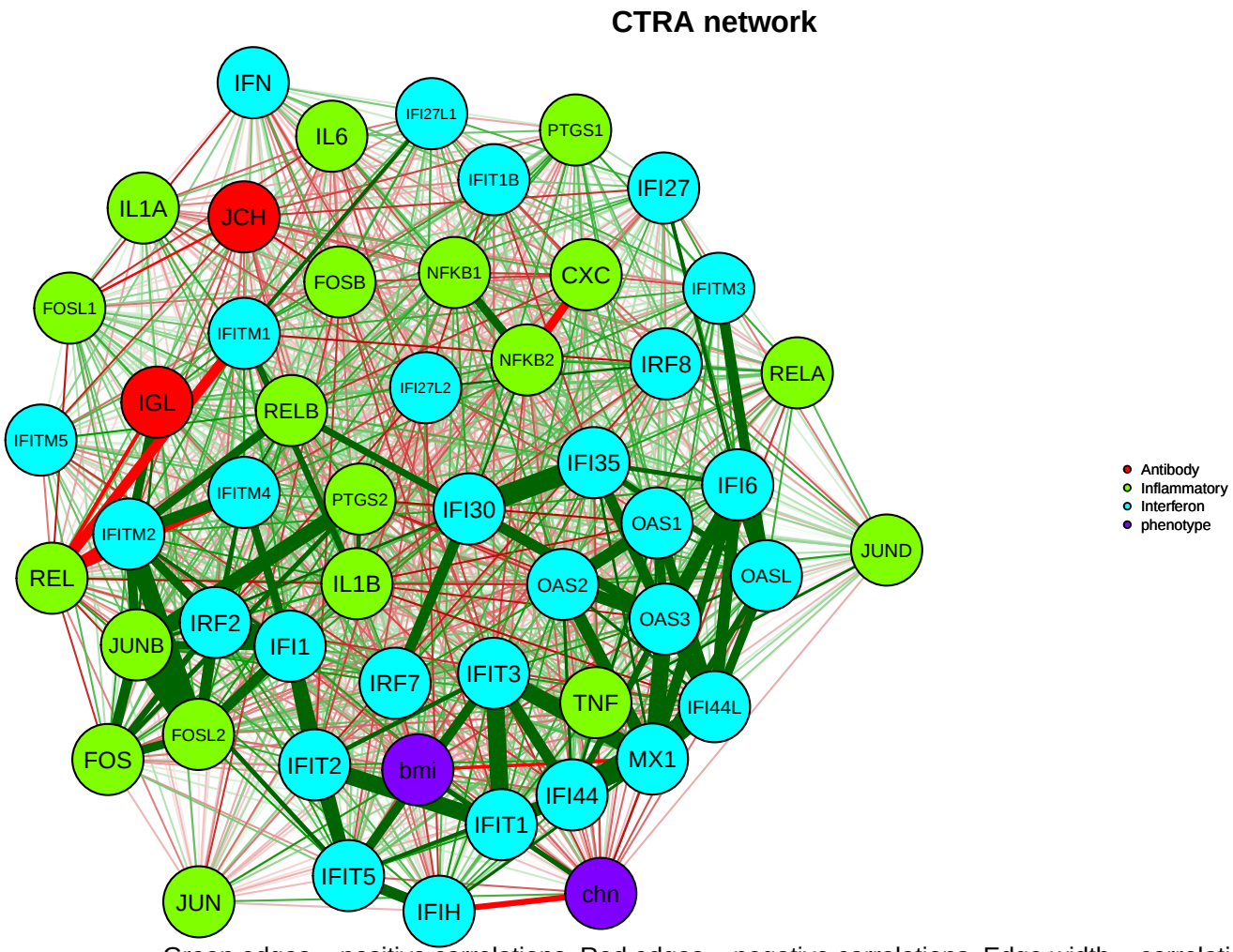
Mass-univariate (control for BMI)

	logFC	AveExpr	t	P.Value	adj.P.Val	B
IFIH1	-0.207	6.552	-2.221	0.035	0.923	-3.813
JCHAIN	0.450	8.649	1.987	0.057	0.923	-4.016
IGLL1	-0.186	4.657	-1.980	0.058	0.923	-4.022
IFITM2	-0.208	5.241	-1.672	0.106	0.923	-4.268
RELA	-0.120	5.460	-1.615	0.118	0.923	-4.310
OASL	-0.138	5.928	-1.559	0.131	0.923	-4.351

Table 2: Limma regression on CTRA.


CTRA residuals
SELF
CTRA
corr
1.0
0.5
0.0
−0.5
IGLL1
JCHAIN
OASL
OAS3
OAS2
OAS1
MX1
IRF8
IRF7
IRF2
IFNB1
IFITM5
IFITM4P
IFITM3
IFITM2
IFITM1
IFIT1B
IFIT5
IFIT3
IFIT2
IFIT1
IFIH1
IFI6
IFI44L
IFI44
IFI35
IFI30
IFI27L2
IFI27L1
IFI27
IFI16
RELB
RELA
REL
NFKB2
NFKB1
JUND
JUNB
JUN
FOSL2
FOSL1
FOSB
FOS
PTGS2
PTGS1
TNF
CXCL8
IL6
IL1B
IL1A

Marginal dependence: “co-expression”



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.

Methodological limitations

- Causality implicit (most applications experimental)
- Validity of p-value: variance inflation from false independence assumption
- Semantics of p-value
- Dependence structure a secondary nuissance

Competitive null hypothesis: the global test

- 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))

Competitive null hypothesis: the global test

- 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.)

Global test decomposition: contributions to the global test

- Global test statistic = weighted average of individual gene statistics.
- The contribution of each such gene is itself a test.
- k genes ordered in a hierarchical clustering graph: $2k - 1$ sets.
- FWER multiple correction on all $2k - 1$ sets (inheritance method)
- Technicalities: Correlation distance measure (individual gene test statistics likely similar they are strongly correlated) with average linkage clustering.

Global test: without bmi covariate

	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

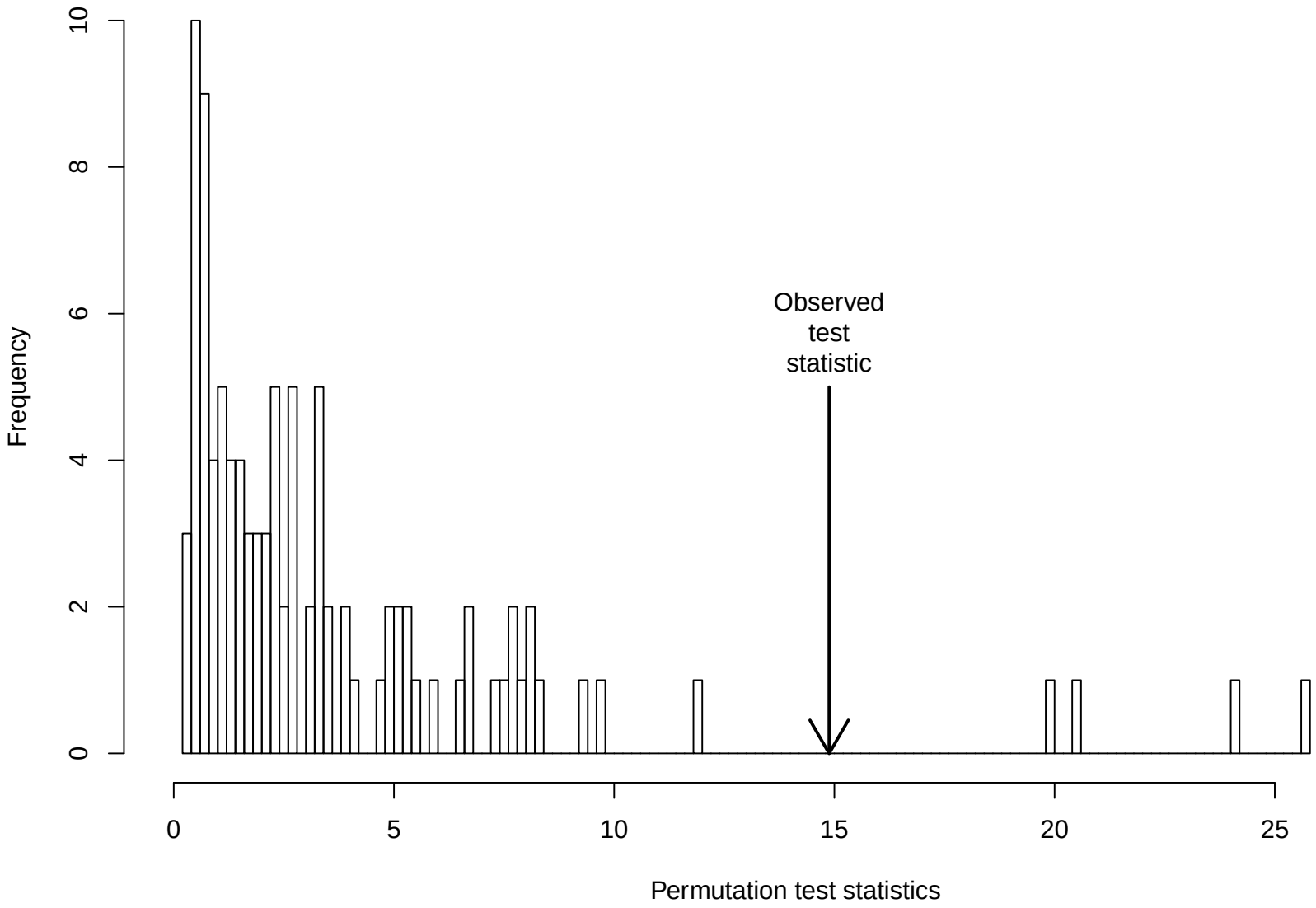
Table 3: Global Test for CTRA. The output lists the p-value of the test, the test statistic with its expected value and standard deviation under the null hypothesis. The Cov column give the number of covariates in the alternative model that are not in the null model. In the linear model the test statistic is scaled in such a way that it takes values between 0 and 100. The test statistic can be interpreted as 100 times a weighted average (partial) correlation between the covariates of the alternative and the residuals of the response.

Global test: with bmi covariate

	p-value	Statistic	Expected	Std.dev	#Cov
Inflammatory	0.780	1.524	3.897	3.180	19
Interferon	0.210	5.953	3.663	4.630	29
Antibody	0.120	9.814	4.732	5.976	2
All	0.260	4.980	3.800	4.735	50

Table 4: Global Test for CTRA controlling for bmi. The output lists the p-value of the test, the test statistic with its expected value and standard deviation under the null hypothesis. The Cov column give the number of covariates in the alternative model that are not in the null model. In the linear model the test statistic is scaled in such a way that it takes values between 0 and 100. The test statistic can be interpreted as 100 times a weighted average (partial) correlation between the covariates of the alternative and the residuals of the response.

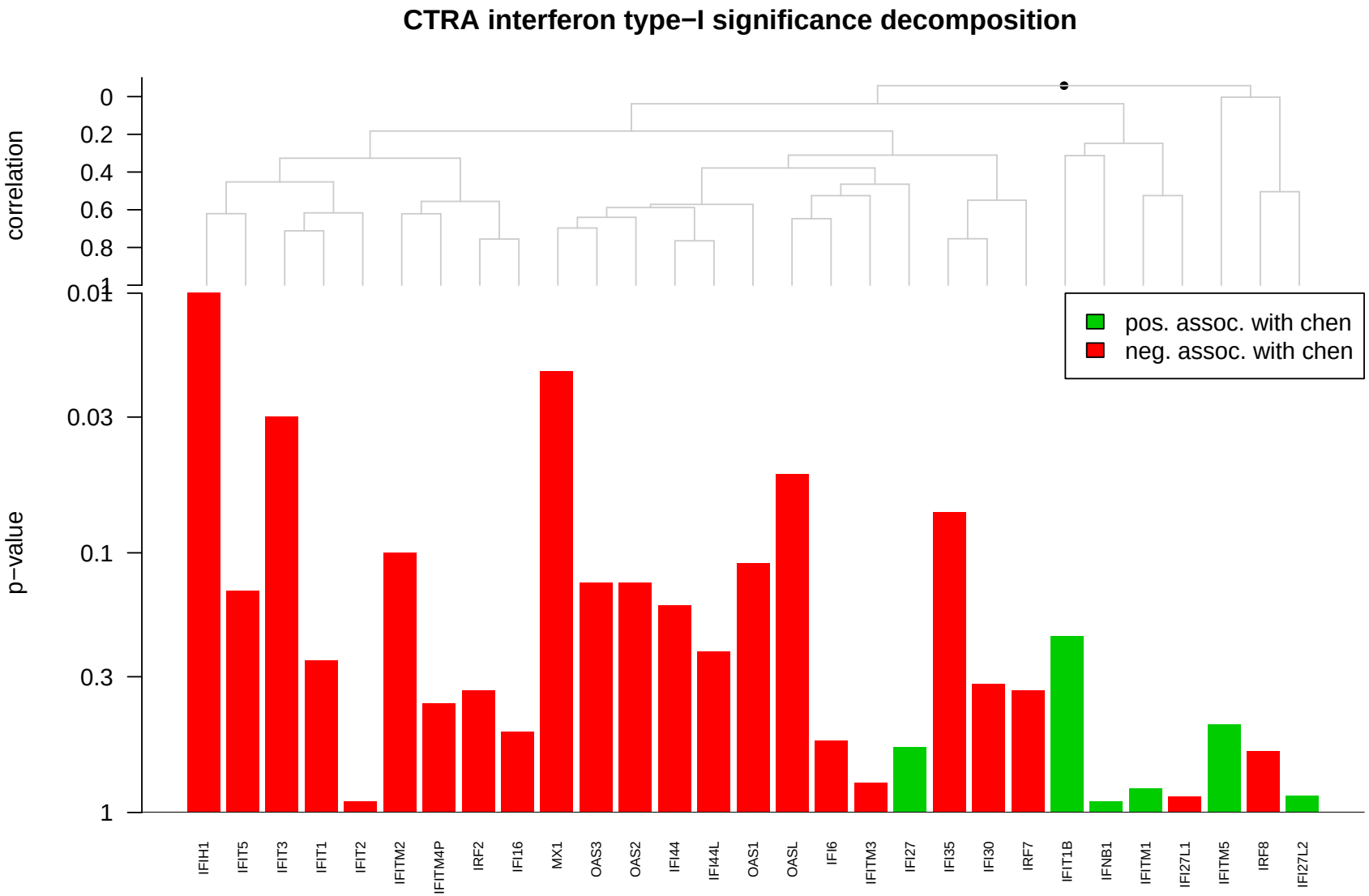
Global test without bmi: a deconstruction



Global test without bmi: a deconstruction

- For general dependence:
- Holm (1979)
- Benjamini and Yekutieli (2001)
- Special cases:
- Benjamini and Hochberg (1995) (independent/positively correlated).
- Subset structure: DAG (nodes = sets, edges = subset relations) (J. J. Goeman and Mansmann 2008).
- Subset structure: Tree (Meinshausen 2008).

Global test without bmi: a deconstruction



Global test without bmi: a deconstruction

	p-value	Statistic	Expected	Std.dev	#Cov
IFIH1	0.010	30.544	4.403	5.238	1.000
IFIT5	0.140	9.733	4.275	6.504	1.000
IFIT3	0.030	17.209	3.896	4.613	1.000
IFIT1	0.260	5.439	4.033	5.610	1.000
IFIT2	0.910	0.029	4.003	5.686	1.000
IFITM2	0.100	12.618	4.350	5.845	1.000

Table 5: Global Test: Singleton subsets of the 2k-1 subsets induced by hierarchical clustering.

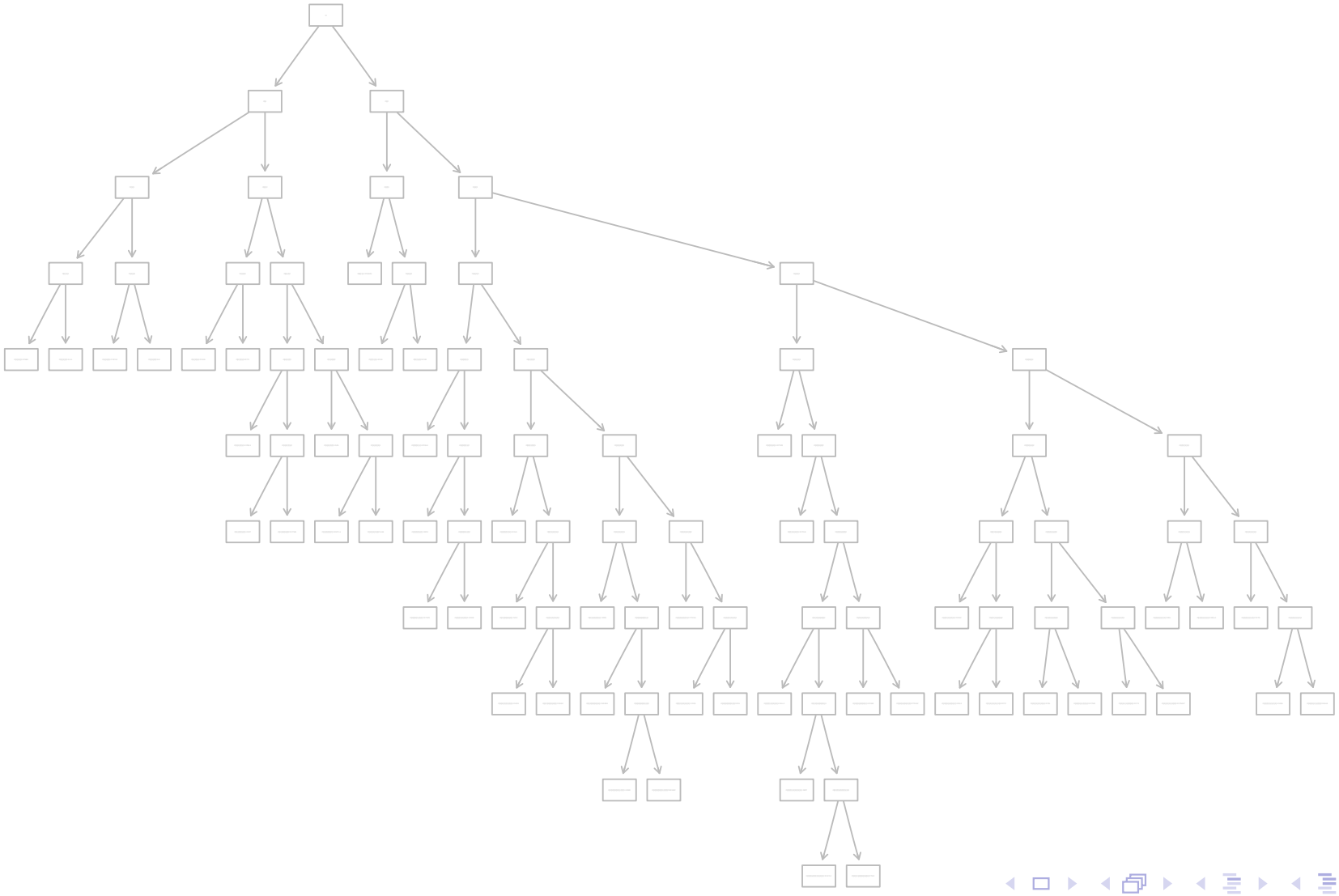
Global test without bmi: a deconstruction

	rowname	p-value	Statistic	Expected	Std.dev	#C
1	IFIH1	0.010	30.544	4.403	5.238	1.0
2	MX1	0.020	21.521	4.008	4.935	1.0
3	IFIT3	0.030	17.209	3.896	4.613	1.0
4	OASL	0.050	12.184	3.492	3.671	1.0
5	found_interferonTypeI	0.050	14.881	3.751	4.662	29.0
6	IFI35	0.070	14.151	3.856	5.200	1.0

Table 6: Global Test: Subsets induced by partitioning CTRA into inflammatory, interferon type-I and antibody.

Global test without bmi: a deconstruction

Inheritance multiplicity correction, based on tree structured hierarchical clustering.



Beyond the global test: a permutation test (without bmi confound)

- Following Ackerman, we seperately regressed each gene in the 50 CTRA on chen. The average of squared partial regression coefficients from this set provides one conventional way to quantify the aggregate, unsigned relation between the whole CTRA gene set and chen. We assessed the significance of this relation with reference to the empirically-derived null distribution arising from 100 permumations of the chen labels. This procedure yielded a p-value of $p = 0.13$.

Beyond the global test: Linear mixed effects (without bmi confound)

- 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)

Kegg Pathways

	kegg_cov_alias	p-value	Statistic	Expecte
03010	Ribosome	0.010	12.705	4.1
04950	Maturity onset diabetes of the young	0.010	9.446	4.3
00190	Oxidative phosphorylation	0.020	8.706	4.0
00730	Thiamine metabolism	0.020	10.970	3.5
05012	Parkinson's disease	0.030	7.658	4.1
05010	Alzheimer's disease	0.040	6.582	4.1

Table 7: Global Test for KEGG pathways.

Competitive null hypothesis

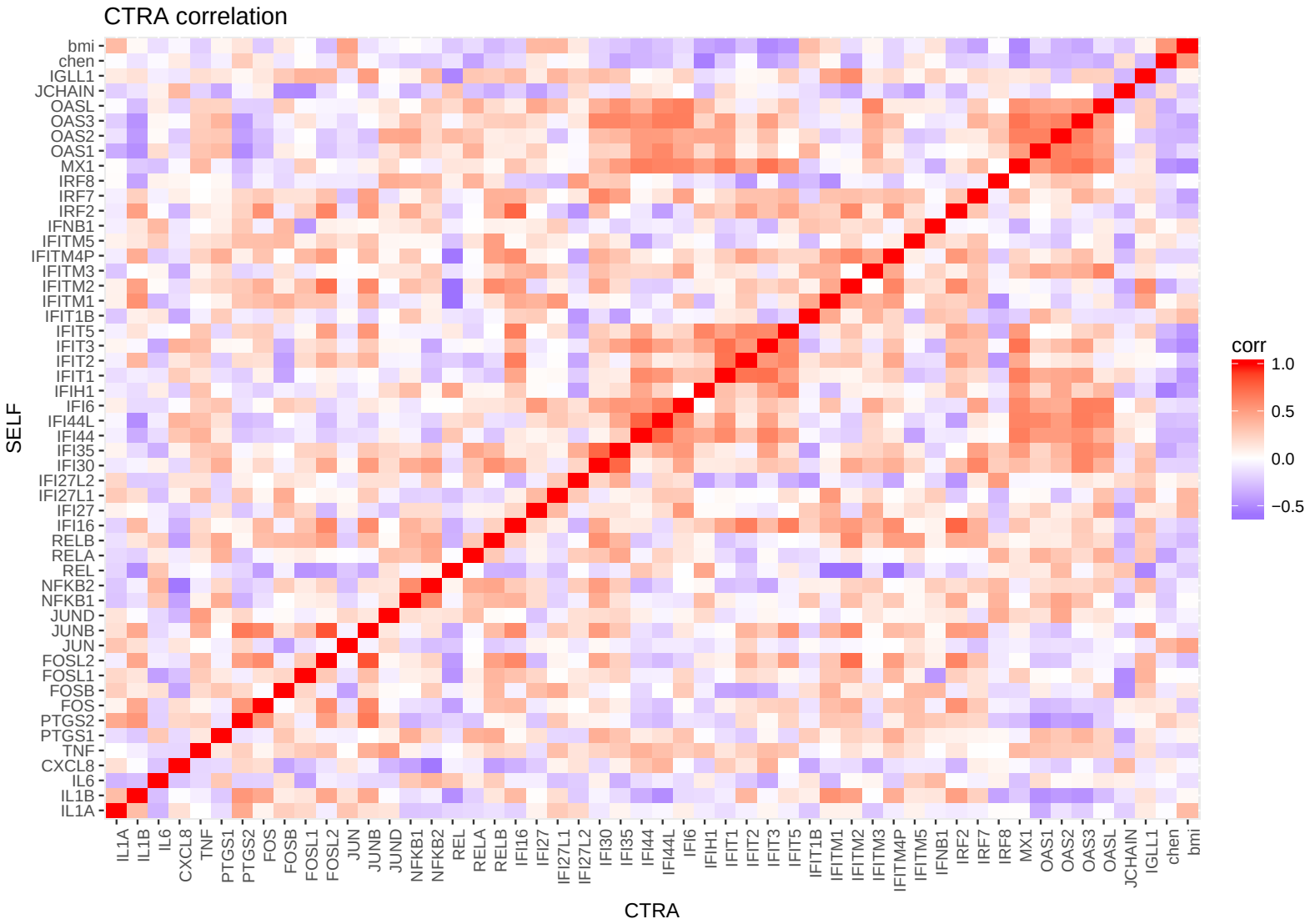
	logFC	AveExpr	t	P.Value	adj.P.Val	B
16904365	-0.235	6.552	-3.150	0.004	0.614	-2.368
16890574	-0.164	8.686	-2.903	0.007	0.614	-2.718
16752645	0.204	5.477	2.720	0.011	0.614	-2.971
16923031	-0.254	6.763	-2.620	0.014	0.614	-3.107
16802918	-0.164	5.568	-2.416	0.022	0.614	-3.375
16922275	-0.174	7.238	-2.383	0.024	0.614	-3.417

Table 8: Limma regression for GO interferon.

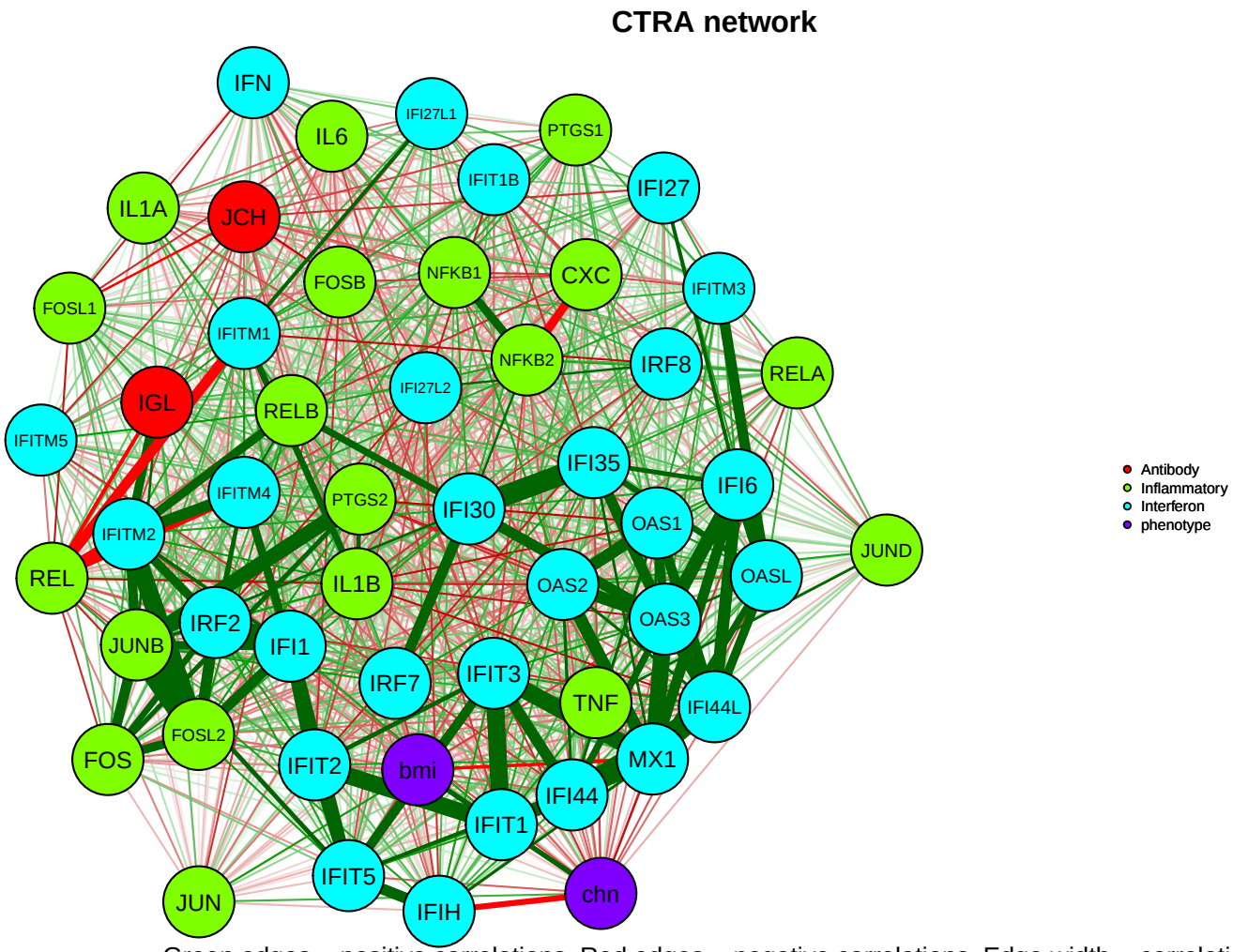
	p-value	Statistic	Expected	Std.dev	#Cov
1	0.040	19.078	4.256	5.384	357.000

Table 9: Global Test for GO interferon.

Undirected graphs

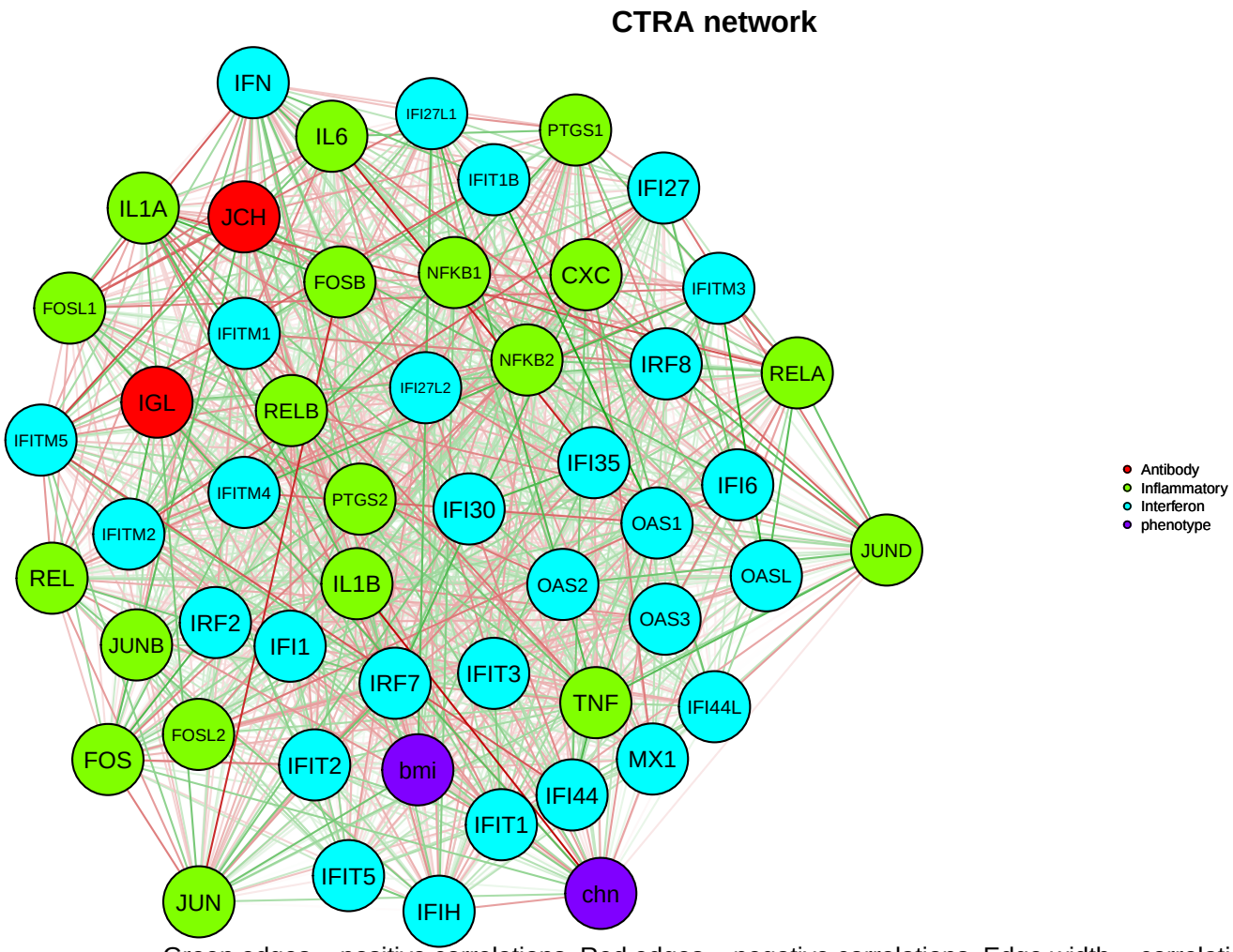


Undirected graphs



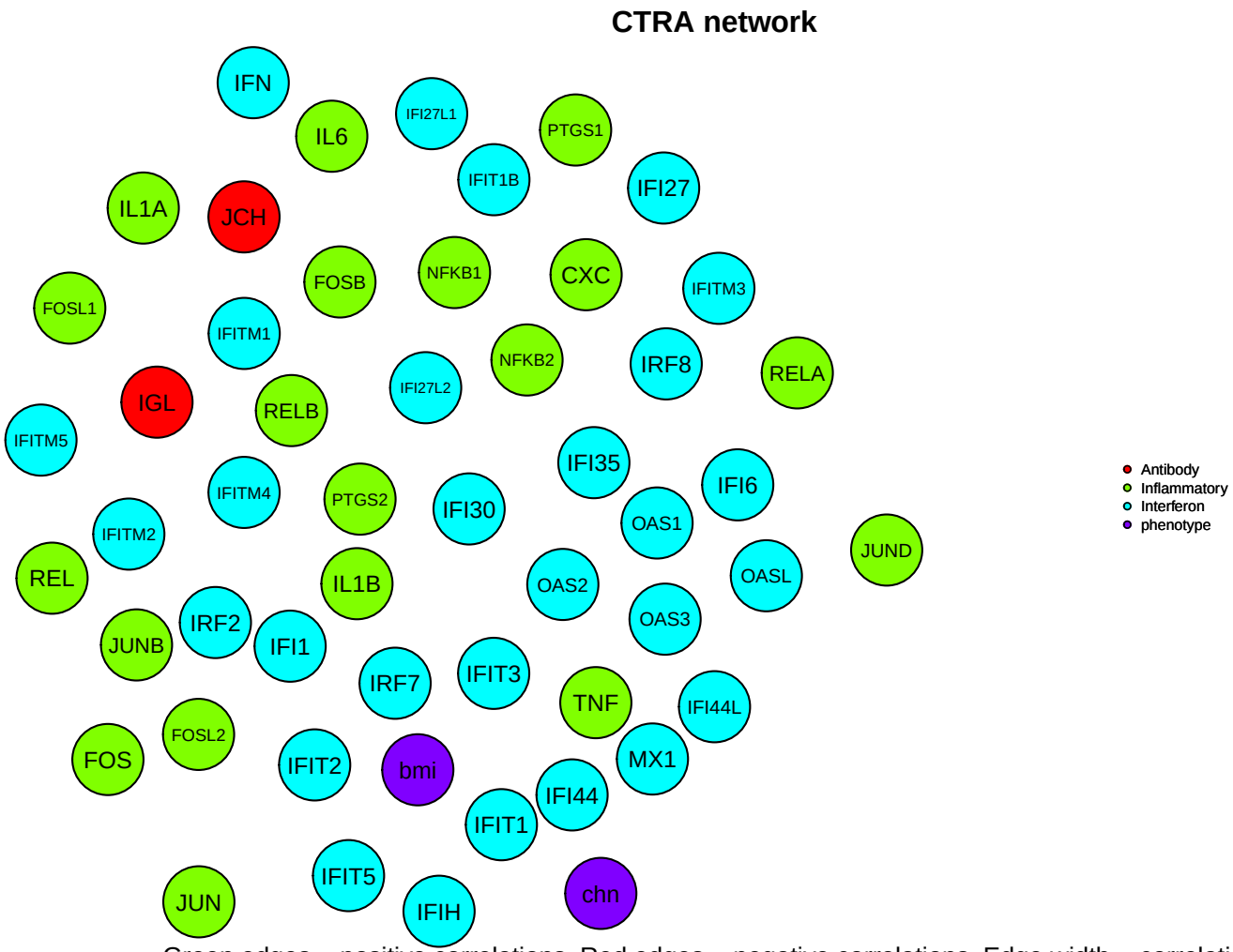
Green edges = positive correlations, Red edges = negative correlations, Edge width = correlation.
Node positioning based on a weighted version of the Fruchterman and Reingold (1991)
algorithm to place strongly correlated nodes together.

Full-conditional independence networks



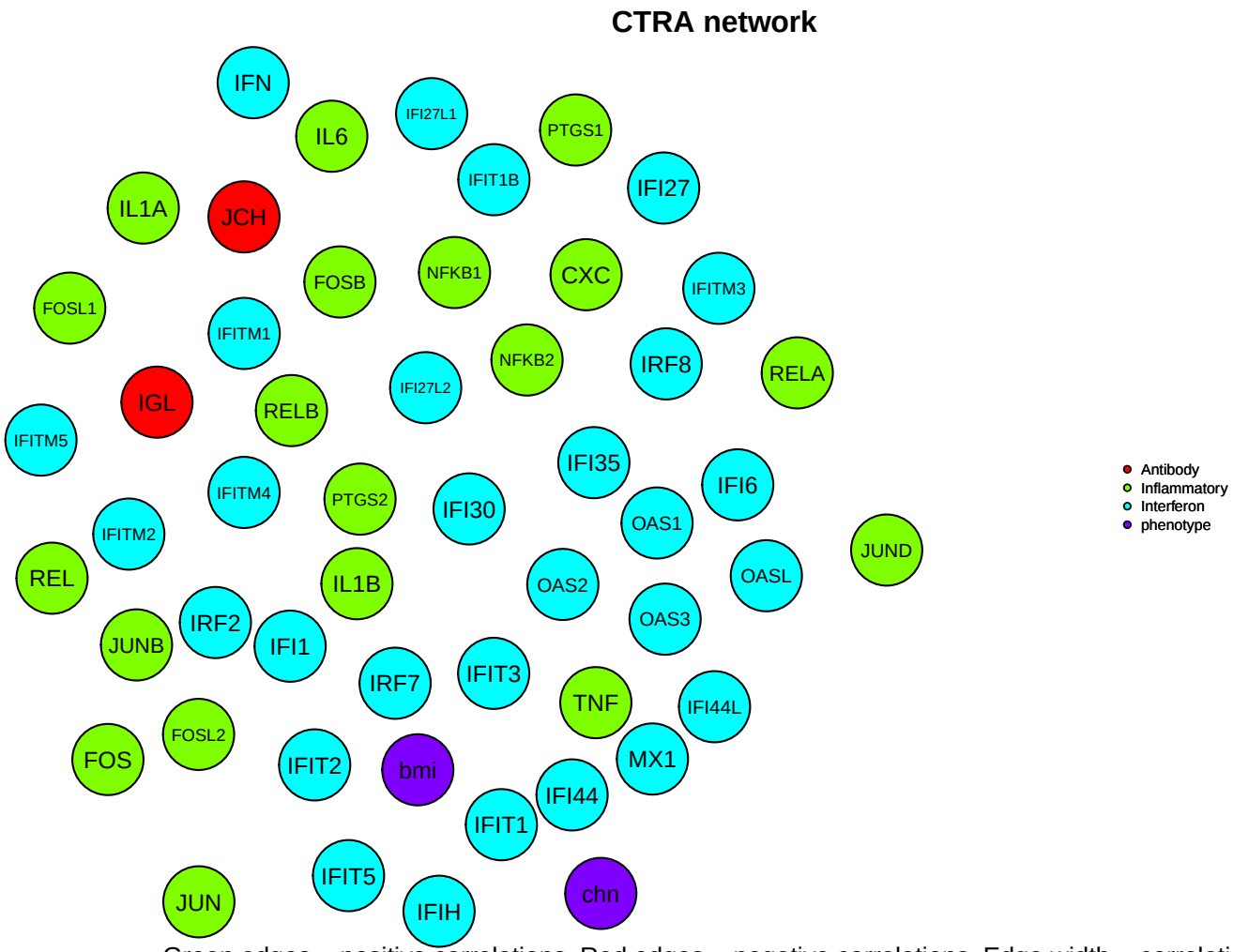
Green edges = positive correlations, Red edges = negative correlations, Edge width = correlation.
Node positioning based on a weighted version of the Fruchterman and Reingold (1991)
algorithm to place strongly correlated nodes together.

Full-conditional dependence networks



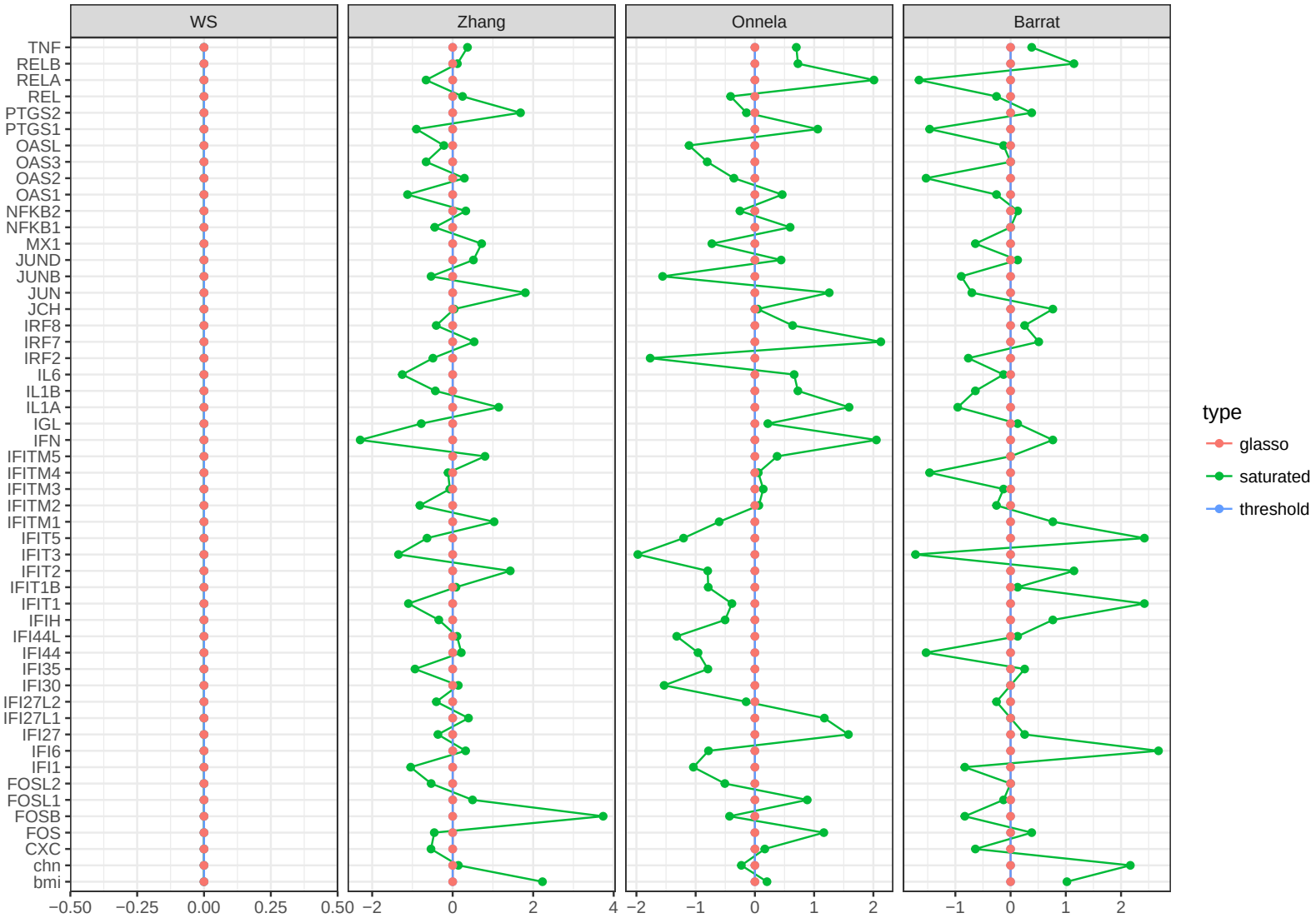
Green edges = positive correlations, Red edges = negative correlations, Edge width = correlation.
Node positioning based on a weighted version of the Fruchterman and Reingold (1991) algorithm to place strongly correlated nodes together.

Full-conditional dependence networks



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.

Full-conditional independence networks



Questions

References

Goeman, J. J., and U. Mansmann. 2008. "Multiple testing on the directed acyclic graph of gene ontology." *Bioinformatics* 24 (4). Oxford University Press: 537–44. doi:10.1093/bioinformatics/btm628.

Goeman, Jelle J., Sara A. van de Geer, and Hans C. van Houwelingen. 2006. "Testing against a high dimensional alternative." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (3). Blackwell Publishing Ltd: 477–93. doi:10.1111/j.1467-9868.2006.00551.x.

Goeman, JJ, SA Van De Geer, and F De Kort. 2004. "A global test for groups of genes: testing association with a clinical outcome." `http://bioinformatics.oxfordjournals.org/content/20/1/93.short`.

Meinshausen, Nicolai. 2008. "Hierarchical testing of variable importance." *Biometrika* 95 (2): 265–78. doi:10.1093/biomet/asn007.