BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 21 2012, pages 2819–2823
doi:10.1093/bioinformatics/bts523

Systems biology
Advance Access publication September 3, 2012

Causal stability ranking

Daniel J. Stekhoven[1,2,3,*], Izabel Moraes[4], Gardar Sveinbjörnsson[1], Lars Hennig[4,5], Marloes H. Maathuis[1] and Peter Bühlmann[1,3]

[1]Seminar for Statistics, Department of Mathematics, [2]Life Science Zurich PhD Program on Systems Biology of Complex Diseases, ETH Zurich, 8092 Zurich, Switzerland, [3]Competence Center for Systems Physiology and Metabolic Diseases, 8092 Zurich, Switzerland, [4]Uppsala BioCenter, Department of Plant Biology and Forest Genetics, Swedish University of Agricultural Sciences and Linnean Center for Plant Biology, 750 07 Uppsala, Sweden and [5]Plant Biotechnology, Department of Biology, ETH Zurich, 8092 Zurich, Switzerland

Associate Editor: Olga Troyanskaya

ABSTRACT

Genotypic causes of a phenotypic trait are typically determined via randomized controlled intervention experiments. Such experiments are often prohibitive with respect to durations and costs, and informative prioritization of experiments is desirable. We therefore consider predicting stable rankings of genes (covariates), according to their total causal effects on a phenotype (response), from observational data. Since causal effects are generally non-identifiable from observational data only, we use a method that can infer lower bounds for the total causal effect under some assumptions. We validated our method, which we call Causal Stability Ranking (CStaR), in two situations. First, we performed knock-out experiments with *Arabidopsis thaliana* according to a predicted ranking based on observational gene expression data, using flowering time as phenotype of interest. Besides several known regulators of flowering time, we found almost half of the tested top ranking mutants to have a significantly changed flowering time. Second, we compared CStaR to established regression-based methods on a gene expression dataset of *Saccharomyces cerevisiae*. We found that CStaR outperforms these established methods. Our method allows for efficient design and prioritization of future intervention experiments, and due to its generality it can be used for a broad spectrum of applications.

Availability: The full table of ranked genes, all raw data and an example R script for CStaR are available from the Bioinformatics website.

Contact: stekhoven@stat.math.ethz.ch

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 22, 2012; revised on July 5, 2012; accepted on August 17, 2012

1 INTRODUCTION

The growing interest in causal inference (e.g. Kruglyak and Storey, 2009) has increased the need not only for methods able to handle this task but also for designed experimental validation. It is of general interest to infer the genotypic causes of a complex phenotypic trait (Glazier *et al*., 2002). The classical approach relies on randomized controlled intervention experiments, e.g. knocking out a gene and observing the effect on the phenotype relative to the wild-type organism. However, such intervention experiments are time consuming and expensive, and a prioritization with respect to most informative new experiments is very desirable. A genetic method to identify loci causing phenotypes or gene expression patterns is based on quantitative trait loci (QTL) and expression QTL (Gilad *et al*., 2008; Kliebenstein, 2009). This can be a very powerful approach but it is limited to loci where genetic variation exists and to situations where segregating progeny of control crosses is available. Often, however, it is desirable to predict causal effects from purely observational data. We therefore consider the problem of predicting total causal effects from data obtained by observing a system without subjecting it to targeted interventions (observational data). This problem is generally ill-posed, but the recently proposed IDA method (Maathuis *et al*., 2009, 2010) provides estimated lower bounds of total causal effects from observational data under some assumptions (Supplementary Section S1). However, these bounds come without a measure of uncertainty. We address this issue by introducing a new method combining IDA and a version of stability selection (Meinshausen and Bühlmann, 2010), which we call Causal Stability Ranking (CStaR; Fig. 1). The addition of stability selection to IDA provides two advantages. First, CStaR leads to a stable ranking of genes (covariates) according to the sizes of lower bounds for their predicted total causal effects, irrespective of the choice of the tuning parameter in stability selection. Second, under some additional assumptions, CStaR allows controlling an error rate of false-positive findings, namely the expected number of false positives and hence also the per-comparison error rate (PCER). CStaR results were confirmed in two biological scenarios using the simple model *Saccharomyces cerevisiae* and the more complex model *Arabidopsis thaliana*. Together, the built-in error measure and the success in finding relevant regulator genes make CStaR an excellent ranking method for the targeted design of experiments based on easily available resources.

2 METHODS

Based on observational training data and a set of required assumptions, CStaR predicts a lower bound for the total causal effect of a covariate on a response of interest, including a PCER for the false-positive selections. This is achieved by combining IDA (Section 2.1) with a version of stability selection (Section 2.2) on a range of different parameters. Predicted

*To whom correspondence should be addressed.

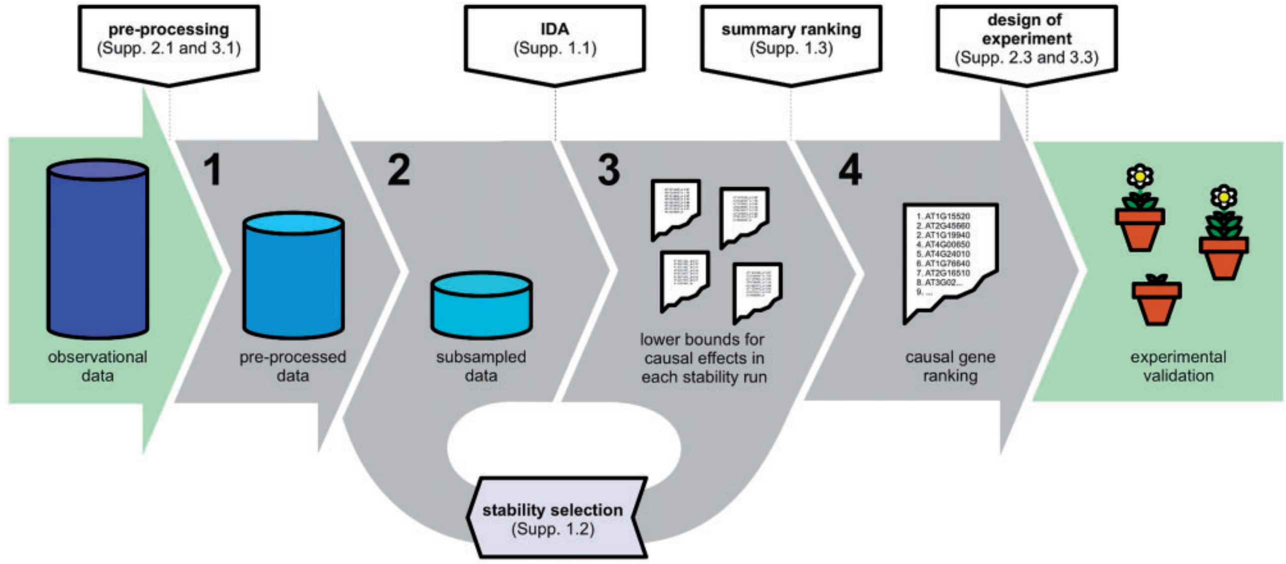

Fig. 1. Schematic overview of the methodological framework used in CStaR. After pre-processing the data (Step 1), lower bounds for the total causal effects are estimated 100 times using stability selection (Meinshausen and Bühlmann, 2010) according to the following procedure. A subsample of size $\lfloor n/2 \rfloor$ is repeatedly drawn from the total of n pre-processed data points (Step 2). On each subsample (or stability run), lower bounds for the total causal effects are estimated using IDA (Maathuis *et al.*, 2009) and used to rank the genes (Step 3, Section 2.1). Next, for a range of different q-values, we record the relative frequencies over the 100 stability runs that each gene appeared in the top q ranks (Section 2.2). The median rank over these different qs is used to generate the final ranking of the genes (Step 4). Furthermore, under additional assumptions, an upper bound for the PCER is estimated for each q-value and its corresponding relative frequency (Section 2.3). Finally, the gene ranking allows for design of new experiments. Thus, a biological validation using intervention experiments can be performed. We tested CStaR in two situations. First, on a publicly available compendium of 31 natural *A. thaliana* accessions consisting of $n = 47$ gene expression measurements, each with 21,326 genes and corresponding flowering time data (Lempe *et al.*, 2005; Supplementary Section S2.1). We performed biological intervention experiments according to the causal gene ranking (Table 1) by focusing on candidates that were not already known to control flowering time and for which mutant seeds were readily available (Supplementary Section S2.3). The biological experiments were analyzed using a two-sample Welch's t-test (Supplementary Section S2.4). The second validation was performed on a publicly available dataset in *S. cerevisiae* containing $n = 63$ observational and 234 interventional full-genome expression profiles, with $p = 5{,}361$ genes (Hughes *et al.*, 2000; Supplementary Section S3). Since this dataset includes both observational and interventional data, the validation was analyzed by comparing estimated total causal effects based on the observational data with inferred effects from the interventional data (Fig. 2)

total causal effects are ranked according to their stability aggregated over this range (Section 2.3).

2.1 Causal inference when the directed acyclic graph (DAG) is absent (IDA)

The IDA procedure (Maathuis *et al.*, 2009) is a statistical method that infers lower bounds for the absolute values of total causal effects on a response of interest from observational data under the assumption that the data come from an unknown DAG without hidden variables.

Suppose we have a dataset with n observations consisting of a response and p explanatory variables. Denoting by θ_j ($j = 1, \ldots, p$), the true total causal effect of gene (covariate) j to the response (the total causal effect θ_j can be interpreted as follows: a change of gene j by one unit (one standard deviation) causes an average change of size θ_j in the response), the output of IDA is the estimated lower bound $\hat{\beta}_j$. It is shown (Maathuis *et al.*, 2009) that under certain assumptions (Supplementary Section 1) and as sample size n tends to infinity:

$$\hat{\beta}_j \overset{n\to\infty}{\to} \beta_j, \beta_j \leq |\theta_j|,$$

justifying the IDA procedure to infer lower bounds. These lower bounds are conservative: for example, if the lower bound is equal to zero, we would not make a statement that there is no causal effect (since the true total causal effect could be indeed equal to zero, or it could be larger than zero but the lower bound would not detect it). Based on the estimated lower bounds, we obtain a ranking of genes (covariates) with j_1 being the index corresponding to the top rank, j_2 for the second best rank and so on:

$$\hat{\beta}_{j_1} \geq \hat{\beta}_{j_2} \geq \ldots \geq \hat{\beta}_{j_p} \tag{1}$$

Under the assumption that the data come from an unknown DAG without hidden variables, the true total causal effect θ_j is generally non-identifiable from observational data, but lower bounds are. The conceptual idea for constructing lower bounds is as follows (Maathuis *et al.*, 2009). We first infer the so-called Markov equivalence class of all the DAGs (see Supplementary Section S1), which are compatible with the observational data. Using intervention calculus (Pearl, 2000), we derive all potential total causal effects based on each DAG G_r in the equivalence class (for every gene (covariate) j)

$$\{\theta_{j;r};\, r = 1, \ldots, m\}(j = 1, \ldots, p),$$

and we define the true lower bounds as

$$\beta_j = \min_{r=1,\ldots,m} |\theta_{j;r}|(j = 1, \ldots, p). \tag{2}$$

Under our assumptions (see Supplementary Section S1), these (true) lower bounds β_j are identifiable from observational data, and the IDA algorithm yields the estimates $\hat{\beta}_j$ ($j = 1, \ldots, p$). The main components of the IDA method are the PC-algorithm for estimating the Markov equivalence class of DAGs (Spirtes *et al.*, 2000) and a local algorithm for calculating the bounds β_j without enumerating all DAG members in the estimated Markov equivalence class (Maathuis *et al.*, 2009). It is

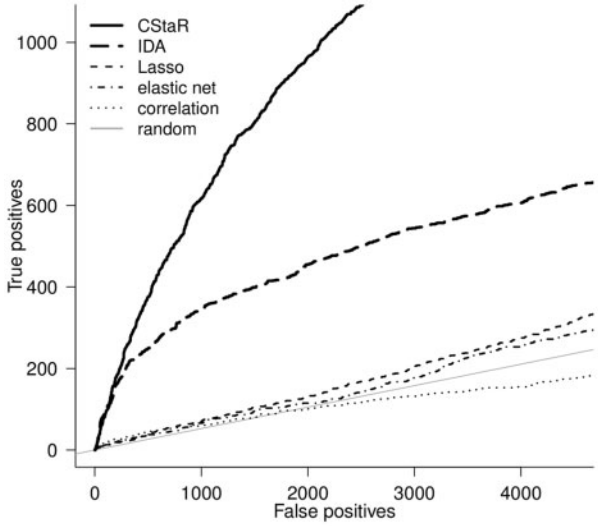

Fig. 2. True-positive selections (y-axis) versus false-positive selections (x-axis) for CStaR (solid) versus plain IDA (Maathuis *et al.*, 2009; long dashed), Lasso (Tibshirani, 1996; short dashed), elastic net (Zou and Hastie, 2005; dash dotted), the latter two using linear models and marginal correlation ranking (dotted) in the *S. cerevisiae* validation (Supplementary Section S3). Random guessing is indicated by the grey line. All methods were trained on the observational data. True positives were defined as the largest 5% of the effects (in absolute value) inferred from the interventional data

this local algorithm that makes the inference of these lower bounds based on thousands of genes (covariates) feasible. IDA is implemented in the R-package `pcalg` (Kalisch *et al.*, 2012).

2.2 Stability selection

CStaR incorporates a stability selection step (Meinshausen and Bühlmann, 2010). We draw 100 independent random subsamples of size $n/2$ and we run IDA on the subsampled data. In each subsampling run, which we also call stability run, we check whether gene (covariate) j has appeared among the top q variables when using the ranking as in equation (1) based on the subsampled data. We can then report the relative selection frequency $\hat{\Pi}_j$, among the 100 stability runs, that gene (covariate) j has appeared (or been selected) among the top q variables. These relative selection frequencies yield a stable list of genes (covariates): the index j_1 corresponds now to the most stably selected variable, and j_p to the least stable variable:

$$\hat{\Pi}_1 \geq \hat{\Pi}_2 \geq \ldots \geq \hat{\Pi}_p. \tag{3}$$

Besides the increased stability in the ranking (3), stability selection is controlling the expected number of false-positive selections. Define the stably selected genes (covariates) as

$$\hat{S}_{\text{stable}} = \{j; \hat{\Pi}_j \geq \pi_{\text{thr}}\},$$

for some threshold $0.5 < \pi_{thr} \leq 1$. Denote the wrongly selected genes (false positives) by $V = |\hat{S}_{stable} \cap S_{false}|$, where S_{false} is the set of (false) genes (covariates) whose true lower bound $\beta_j = 0$, see (2). Then, for a given threshold π_{thr} and a given value of q [which influences (3)] we have, assuming an exchangeability condition (see Supplementary Section S1; Meinshausen and Bühlmann, 2010):

$$E[V] \leq \frac{1}{2\pi_{thr} - 1}\frac{q^2}{p} \tag{4}$$

and this leads to a bound for the PCER ($\text{PCER} = E[V]/p$). If a gene (covariate) j has relative selection frequency $\hat{\Pi}_j$, a bound for the corresponding PCER is given by

$$\frac{1}{2\hat{\Pi}_j - 1}\frac{q^2}{p^2}$$

2.3 Summary ranking

As novelty, we avoid choosing a specific q for the execution of stability selection by assessing the stability and the rank of each gene on a range of different q-values. This constitutes the main modification of the standard stability selection scheme and it also constitutes a useful simplification for the practitioner. This can be summarized graphically (Supplementary Fig. S1 gives an example for a single gene in the *A. thaliana* validation). We found that CStaR is relatively insensitive to the choice of the range of qs. However, down to a certain lower bound, small values of q lead to higher sensitivity and thus better results (see also Supplementary Section S3). If the q-values fall below such a lower bound, the ranking becomes unstable again. Finally, all genes are ranked according to the median rank with respect to the different q-values. Ties in the final ranking are sorted according to median total causal effect size.

2.4 Validation

We validated CStaR in two situations. First, we trained CStaR on a publicly available compendium of *A. thaliana* gene expression data and performed new biological validation experiments (Supplementary Section S2). The compendium contains 47 expression profiles of natural accessions from diverse geographic origins (Lempe *et al.*, 2005). The phenotypic trait of interest is time to flowering, which is robustly measured by the number of days to bolting or the number of rosette leaves formed before bolting (Amasino, 2010). Timing of flowering according to local climatic conditions is a major determinant of the plants' reproductive success and an important agronomical trait that greatly affects yield. Therefore, an improved knowledge about genes controlling flowering time is of substantial economic value (Craufurd and Wheeler, 2009).

As a second validation of the CStaR method, we compared it with the plain IDA method [ranking as in (1)], Lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005) both using linear models (ranking according to absolute values of estimated regression coefficients) and marginal correlation (ranking according to absolute values of marginal correlation to the response) on a publicly available dataset of gene expression profiles in *S. cerevisiae* (Hughes *et al.*, 2000; Supplementary Section S3). This dataset includes both observational and interventional data obtained under similar conditions. Hence, it forms an excellent basis to assess the performance of methods aimed at estimating total causal effects from observational data, as the effects estimated from the observational data can be compared with the effects inferred from the interventional data. These data were used to validate IDA (Maathuis *et al.*, 2010), and we followed the same approach to validate CStaR. In particular, we used the interventional data to infer the total causal effects of the knock-out genes on the remaining genes and defined the top 5% of the effects that were largest in absolute value as the true positives.

3 RESULTS

3.1 Validation for A. thaliana

CStaR scores five known regulators of flowering time (*DWF4*, *FLC*, *FRI*, *RPA2B* and SOC1; Amasino, 2010; Domagalska *et al.*, 2007; Xia *et al.*, 2006) in its top 25 (Table 1). In particular, *SOC1*, *FRI* and *FLC* are curated flowering time genes in

Table 1. Top 25 findings by CStaR for the *A. thaliana* data

	Gene	Summary rank	Median effect	Maximum expression	Error (PCER)	Name/annotation
1	AT2G45660	1	0.60	5.07	0.0032	*SOC1*
2	AT4G24010	2	0.61	5.69	0.0033	*ATCSLG1*
3	**AT1G15520**	**2**	**0.58**	**5.42**	**0.0033**	***PDR12***
4	AT3G02920	5	0.58	7.44	0.0041	*RPA2B*
5	**AT5G43610**	**5**	**0.41**	**4.98**	**0.0069**	***ATSUC6***
6	AT4G00650	7	0.48	5.56	0.0051	*FRI*
7	**AT1G24070**	**8**	**0.57**	**6.13**	**0.0040**	***ATCSLA10***
8	**AT1G19940**	**9**	**0.53**	**5.13**	**0.0045**	***ATGH9B5***
9	**AT3G61170**	**9**	**0.51**	**5.12**	**0.0044**	**PPR protein**
10	**AT1G32375**	**10**	**0.54**	**5.21**	**0.0045**	**F-box protein**
11	**AT2G15320**	**10**	**0.50**	**5.57**	**0.0047**	**LRR protein**
12	**AT2G28120**	**10**	**0.49**	**6.45**	**0.0054**	**Nodulin protein**
13	AT2G16510	13	0.50	10.7	0.0050	*AVAP5*
14	AT3G14630	13	0.48	4.87	0.0056	*CYP72A9*
15	**AT1G11800**	**15**	**0.51**	**6.97**	**0.0053**	**Endonuclease**
16	AT5G44800	16	0.32	6.55	0.0079	*CHR4*
17	AT3G50660	17	0.40	7.60	0.0078	*DWF4*
18	AT5G10140	19	0.30	10.3	0.0085	*FLC*
19	AT1G24110	20	0.49	4.66	0.0071	Peroxidase
20	**AT2G27350**	**20**	**0.48**	**7.06**	**0.0067**	***OTLD1***
21	AT1G27030	20	0.45	10.0	0.0075	Unknown protein
22	**AT2G28680**	**22**	**0.46**	**5.23**	**0.0072**	**Cupin protein**
23	AT3G16370	23	0.43	12.4	0.0099	Lipase/hydrolase
24	AT5G25640	23	0.33	5.59	0.0091	Serine protease
25	AT1G30120	24	0.46	9.97	0.0077	*PDH-E1 BETA*

The genes are ranked by increasing summary rank, where ties are sorted according to the estimated median total causal effect taken over 100 stability runs (third column). The maximum expression is taken over the original $\log_2$ data. The error (PCER) is the median PCER over the range of q values. *SOC1*, *FRI* and *FLC* are 3 of 119 curated flowering time genes in the Arabidopsis Reactome (Tsesmetzis *et al.*, 2008) (http://www.arabidopsisreactome.org). This is a highly significant enrichment of known curated regulators when compared with random guessing ($p<10^{-5}$, hypergeometric test). Although not curated in Arabidopsis Reactome, also *RPA2B* and *DWF4* are known to affect flowering time (Domagalska *et al.*, 2007; Xia *et al.*, 2006). Since the ordering of the genes in the table is given by their summary rank, the values of median total causal effect and PCER are not decreasing monotonously. For instance, *ATSUC6* has a smaller median total causal effect and a larger PCER than the endonuclease, but since its lower bound for the total causal effect is more stable, the former is ranked 10 positions higher than the latter. All genes from this list, for which mutant seeds were readily available and which were not already known to control flowering time, were used in the subsequent intervention experiments (indicated in bold). In total, intervention experiments were performed for 13 of the 25 top genes not previously known to regulate flowering (Supplementary Section S2.3).

Arabidopsis Reactome (Tsesmetzis *et al.*, 2008) containing 119 known regulators of flowering. This is a highly significant enrichment of known curated regulators when compared with random guessing ($p<10^{-5}$ in a hypergeometric test). Interestingly, *FLC* and *FRI* are not only major regulators of flowering time in the model species *A. thaliana* but also in the oil-seed rape crop.

Among the other genes in the top 25, which were not already known to play a role in flowering time, there were 13 genes for which mutant seeds were readily available (Supplementary Table S1). These mutants were used for intervention experiments in order to further validate CStaR and to discover new influential genes for flowering time in *A. thaliana* (Supplementary Section S2.3).

The intervention experiments were performed under two photoperiod conditions, short-day (SD) and long-day (LD) with 8 h and 16 h of light, respectively. As phenotypic responses, the number of days to bolting (DTB, for both SD and LD) as well as the rosette leave number (RLN, only for LD) were recorded. Seed viability varied between different genotypes (Supplementary Tables S2–S4) reducing the number of testable mutants to nine (Supplementary Table S1).

Differences between the knock-out and control group were tested using a two-sided Welch's t-test, because the mutant samples showed different empirical variances compared with the control group. This is most pronounced in the short-day layout. Four new genes were found to have a significant total causal effect on the phenotypic responses at level $\alpha=0.05$ in at least one of the three settings (Table 2). Among the significant genes is *OTLD1*, a gene involved in chromatin modifications, which may potentially regulate *FLC* expression. Another significant gene is *PDH-E1*, which is involved in carbohydrate metabolism, a known regulation point of flowering time. We did not adjust these p-values for multiple testing because we only perform a small number of tests and, in view of small sample sizes, we do not want to sacrifice power. Future studies of the identified novel genes may increase the biological understanding of flowering time control and provide potential targets for breeding strategies in crops. The entire approach from modelling to biological experiments and findings is schematically described in Figure 1.

Table 2. *p*-values from two-sided Welch's *t*-tests in the *A. thaliana* validation

Gene	Welch's *t*-test		
	DTB-SD	DTB-LD	RLN-LD
PDH-E1 BETA	**0.04**	**0.04**	0.91
ATGH9B5	**0.02**	0.15	**0.04**
LRR protein	0.66	**0.03**	0.47
OTLD1	0.43	**0.03**	0.86
PDR12	0.26	0.92	0.77
F-box protein	0.18	–	–
peroxidase	0.18	–	–
PPR protein	–	0.65	0.47
cupin protein	–	0.12	0.93

Only genes are shown for which the insertion was experimentally verified and for which in at least one of the following three settings at least four replicates could be harvested for validation: days to bolting in short days (DTB-SD), days to bolting in long days (DTB-LD) and rosette leave number in long days (RLN-LD). Each mutant was tested versus a control group. p-values < 0.05 are written in bold (for complete results see Supplementary Tables S2–S4). A missing entry indicates insufficient number of replicates for testing, i.e. less than four plants.

3.2 Validation for *S. cerevisiae*

We trained the plain IDA method, Lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005) and marginal correlation ranking on the observational data, and compared their receiver operating characteristic curves on absolute scale (Fig. 2) showing a clear improvement of CStaR over plain IDA. Moreover, CStaR and IDA are clearly superior to high-dimensional regression methods and marginal correlation screening, which is in line with the earlier validation of IDA (Maathuis *et al.*, 2010).

4 DISCUSSION

We propose CStaR as a general method to obtain a stable ranking of genes in terms of the strengths of their total causal effects on a phenotype of interest. An added value of our method is that, under some assumptions, this ranking comes with an error measure controlling false-positive selections. We showed that CStaR exhibits a large increase in sensitivity when compared with plain IDA and modern regression-type methods in *S. cerevisiae* (Fig. 2). Moreover, we demonstrated the success of CStaR for the biologically much more complex multicellular organism *A. thaliana*. However, in view of uncheckable assumptions (Supplementary Section S1), CStaR is not a tool for confirmatory causal inference.

We used insertion mutant lines for experimental validation. This approach can provide very strong evidence for hypotheses about gene function but it often suffers from a high false-negative rate. Genetic networks are characterized by a high degree of functional redundancy, which can buffer effects of single mutations. The *A. thaliana* genome, for instance, underwent a relatively recent duplication causing partial redundancy between many orthologous gene pairs. Thus, often double mutants need to be tested to observe alterations in phenotype. In addition, the function of essential genes cannot be tested with insertion mutants. Therefore, the high proportion of confirmation in the test set of insertion mutants is highly reassuring. This makes it plausible that CStaR is relevant for commercial crops, by pointing to better target genes for marker-assisted breeding and transgenic approaches. In fact since CStaR is mathematically justified under clearly stated assumptions (Maathuis *et al.*, 2009; Meinshausen and Bühlmann, 2010), it has the potential to generalize many other settings in biology, agriculture and other fields where efficient design and prioritization of new intervention experiments is a core aim.

ACKNOWLEDGEMENTS

We thank T. Wey for the technical assistance with plant work.

Funding: The work was partly financed with a grant of the Swiss SystemsX.ch Initiative to the project 'LiverX' of the Competence Centre for Systems Physiology and Metabolic Diseases. The LiverX project was evaluated by the Swiss National Science Foundation.

Conflict of Interest: none declared.

REFERENCES

Amasino,R. (2010) Seasonal and developmental timing of flowering. *Plant J.*, **61**, 1001–1013.

Craufurd,P.Q. and Wheeler,T.R. (2009) Climate change and the flowering time of annual crops. *J. Exp. Bot.*, **60**, 2529–2539.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.

Domagalska,M.A. *et al.* (2007) Attenuation of brassinosteroid signaling enhances FLC expression and delays flowering. *Development*, **134**, 2841–2850.

Gilad,Y. *et al.* (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.

Glazier,A.M. *et al.* (2002) Finding genes that underlie complex traits. *Science*, **298**, 2345–2349.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Kalisch,M. *et al.* (2012) Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.*, **47**, 1–26.

Kliebenstein,D. (2009) Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annu. Rev. Plant Biol.*, **60**, 93–114.

Kruglyak,L. and Storey,J.D. (2009) Cause and express. *Nat. Biotechnol.*, **27**, 544–545.

Lempe,J. *et al.* (2005) Diversity of flowering responses in wild *Arabidopsis thaliana* strains. *PLoS Genet.*, **1**, 109–118.

Maathuis,M.H. *et al.* (2009) Estimating high-dimensional intervention effects from observational data. *Ann. Stat.*, **37**, 3133–3164.

Maathuis,M.H. *et al.* (2010) Predicting causal effects in large-scale systems from observational data. *Nat. Met.*, **7**, 247–248.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. Roy. Stat. Soc. B Met.*, **72**, 417–473.

Pearl,J. (2000) Causality: models, reasoning and inference, Cambridge Univ. Press, 47.

Spirtes,P., Glymour,C.N. and Scheines,R. (2000) Causation, prediction and search, The MIT Press, 81.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. B Met.*, **58**, 267–288.

Tsesmetzis,N. *et al.* (2008) Arabidopsis reactome: a foundation knowledgebase for plant systems biology. *Plant Cell*, **20**, 1426–1436.

Xia,R. *et al.* (2006) ROR1/RPA2A, a putative replication protein A2, functions in epigenetic gene silencing and in regulation of meristem development in Arabidopsis. *Plant Cell*, **18**, 85–103.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B Met.*, **67**, 301–320.