BBA - Gene Regulatory Mechanisms 1863 (2020) 194444

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Gene networks in cancer are biased by aneuploidies and sample impurities

Michael Schubert[a,b,*], Maria Colomé-Tatché[a,b,c], Floris Foijer[a,*]

[a] European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, 9713 AV, Groningen, the Netherlands
[b] Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
[c] TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany

ARTICLE INFO

Keywords:
Gene regulatory networks
Cancer
Method comparison
Aneuploidy

ABSTRACT

Gene regulatory network inference is a standard technique for obtaining structured regulatory information from, for instance, gene expression measurements. Methods performing this task have been extensively evaluated on synthetic, and to a lesser extent real data sets. In contrast to these test evaluations, applications to gene expression data of human cancers are often limited by fewer samples and more potential regulatory links, and are biased by copy number aberrations as well as cell mixtures and sample impurities. Here, we take networks inferred from TCGA cohorts as an example to show that (1) transcription factor annotations are essential to obtain reliable networks, and (2) even for state of the art methods, we expect that between 20 and 80% of edges are caused by copy number changes and cell mixtures rather than transcription factor regulation.

1. Introduction

Gene Regulatory Network (GRN) inference describes the process of identifying regulator-target relationships from experimental molecular data. These data can be protein-protein interactions (often referred to as interaction networks) or protein-DNA binding (obtained by chromatin immunoprecipitation and sequencing, or ChIP-seq), but is most commonly gene expression data (obtained by microarrays or more recently RNA-seq).

For gene expression data, the rationale behind GRN inference is that if a transcription factor (TF) is more highly expressed, it is also likely to be more active and mediate a higher downstream expression of its target genes (TGs). While this ignores potential post-translational modifications that may also influence a transcription factor's activity, as well as epigenetic marks at the enhancer and promoter sites of target genes, GRN inference methods have been shown to be useful in elucidating transcriptional programmes in a variety of contexts [1–7].

There are different kinds of data that one can use to infer networks from. For instance, we can follow a perturbation over time (time-course networks), or take multiple snapshots of the same underlying system in different states. The latter is referred to as observational (meaning comparing different samples) steady-state networks [8], which occur when we for instance measure gene expression in a yeast strain with different growth conditions or in cancer patients across a cohort of the same tumor type.

Each kind of network inference requires different assumptions and hence demands different specialized tools. Here, we focus on steady-state observational networks, where we assume the underlying regulatory structure to be the same or at least its differences small enough so we can ignore them. This is likely true when e.g. a mutation in a signaling molecule activates a certain part of a downstream GRN, but it will not be if a transcription factor loses its affinity to its target genes or a subset thereof. While previous studies have shown this to happen for some genes (reviewed in [9]), observational GRN inference methods assume that this will not change the overall correlation structure across many samples.

Methods that have been developed for observational GRNs can roughly be classified by the theoretical framework they use in order to infer regulatory relationships. The classical approaches come from information theory and employ some kind of mutual information, or correlation and regression-based approaches (classification and theoretical background have been reviewed before [8]). These tools have been continuously developed, but more recently the focus has shifted to machine learning methods such as random forest and neural networks (recent overview of methods reviewed in [10]).

These network inference methods have been extensively evaluated e.g. in the Dialogue of Reverse Engineering and Assessment of Methods (DREAM) competitions [11] and many more comparisons on smaller scale [12–18]. They have provided many biological insights, and have been particularly useful to elucidate mechanisms of pathogenicity in human diseases such as cancer [1,2,19–24]. However, there is a disconnect between evaluation in often relatively simple systems

* Corresponding authors.
E-mail addresses: m.schubert@rug.nl (M. Schubert), f.foijer@umcg.nl (F. Foijer).

https://doi.org/10.1016/j.bbagrm.2019.194444
Received 31 May 2019; Received in revised form 5 September 2019; Accepted 14 October 2019
Available online 23 October 2019

(synthetic networks or GRNs in *E. coli* and yeast) and their application to much more complex mammalian systems.

One application where this disconnect is particularly striking is human cancer, because (a) individual patients harbor different chromosomal aberrations [25] that change the expression of many genes in a coordinated fashion [26], and (b) cancer cells attract different immune and stromal cells that dilute gene expression measurements with their own regulatory programmes [27].

Here, we aim to bridge this gap by investigating how well GRN inference methods perform in the context of cancer, and particularly how much they are influenced by specific confounding factors outside of TF-TG relationships such as aneuploidies and sample impurities. In particular, we show that the inferred links are strongly enriched by these confounding factors, whereas this is not the case for TF binding data.

2. Network inference methods have been extensively evaluated on synthetic data sets

2.1. Network inference methods

For steady-state networks, the basic idea is that the same underlying regulatory structure (the network to be inferred) will be sampled at different states by measuring gene expression of e.g. multiple cancer patients. Genes that are up- or downregulated in a subset of samples compared to the rest will change their expression in accordance to the underlying regulatory network, which can in turn be inferred by looking at this correlation structure: If two genes are correlated across many samples, they are likely to either regulate each other or be regulated by a common third gene (albeit not always directly).

Classic methods that infer networks from gene expression in multiple unperturbed samples can roughly be divided in information-theoretic and correlation-based models. These and more methods have been reviewed in detail [8,10,28,29] and hence we only provide a brief overview. Information-theoretic approaches started out with relevance networks [12], in which the pairwise mutual information (MI) is computed between all pairs of genes. Subsequently, all gene pairs above a certain threshold (that can be estimated from the data itself) are kept. ARACNe (algorithm for the reconstruction of accurate cellular networks) [1,13] added an additional filtering step where the authors eliminate the weakest link in all gene triplets (using the data processing inequality) unless they are protected by a transcription factor link. The recent ARACNe-AP [30] (for adaptive partitioning) implementation adds further performance optimizations. By contrast, the PCIT [31] algorithm only removes edges in triplets if two genes are conditionally independent given the third. Other approaches were taken by CLR [14] (context likelihood of relatedness; using the z-score of the MI distribution) or C3NET [15] (conservative causal core networks; keeping only the strongest MI edge for each gene) and its extension BC3NET [32] (bagging of C3NET results). Yet another approach is taken by MRNET [33], which concurrently maximizes the relevance (MI) while minimizing redundancy (MRMR is a feature selection technique in supervised learning). For practical purposes, it should be noted that MI-based methods are nonparametric, i.e., these methods perform on the ranks of gene expression values rather than the gene expression values themselves. Many of these methods are implemented in the minet R package [34].

Correlation- and regression-based models are another class of gene regulatory network inference methods. In their simplest form, these methods perform a regression or correlation test between two variables. As there are many gene interactions that need to be tested, feature selection is a common feature of these techniques. For instance, Least Angle Regression (LARS) [35,36] starts with the best correlated predictor and then iteratively adds other predictors based on their correlation with the residual. TIGRESS (Trustful Inference of Gene Regulation with Stability Selection) [17] adds the concept of stability selection to LARS. Instead of adding and removing individual predictors, the GeneNet package [37] estimates all predictors simultaneously by inverting the gene expression matrix. Another approach is to combine regression models with decision trees, finding sets of genes that best explain the expression of a target [38–40]. GENIE3 (Gene Network Inference with Ensemble of trees) [16] integrates information of many such trees in order to make regulatory predictions. NIMEFI (Network Inference using Multiple Ensemble Feature Importance algorithms) [18] goes one step further and integrates the results of both TIGRESS and GENIE3 into a combined prediction method.

2.2. Finding a reference set for method evaluation

A challenge in evaluating network inference methods is that in order to score the performance of different methods, we need to compare the edges they infer to edges we know are correct vs. edges we know not to be correct. However, we often do not know the ground truth for real gene regulatory networks. While many interactions may be known, it is likely that only a small fraction of the relevant interactions has been discovered. An alternative approach is to simulate a GRN according to a known network structure and a set of rules about how the different nodes influence each other. Examples of such simulators are SynTren [41] and GeneNetWeaver [42]. The advantage of such a synthetic network is that the ground truth is known, but it may not exhibit all the properties of a real GRN. Method evaluation was often focused on synthetic or synthetic-like datasets. When evaluations on real data were done, they usually were small in scale owing to the limited amount of orthogonal data available.

More recently, the number of available human TF-gene interactions has grown tremendously due to large-scale efforts like the ENCODE project [43], but also curation of individual ChIP binding experiments [44]. These have produced consensus regulons for individual TFs, where binding was observed in a variety of tissues. The latter comprises 100 transcription factors that cover 16,500 target genes in a dataset available from the Enrichr platform [45]. Another option would be the UniBind database with 231 transcription factors [46]. While these consensus interactions are still not proof of actual regulatory interactions, they provide a sufficient number of orthogonally-derived relationships in order to use this set to identify large-scale biases of network inference algorithms with respect to copy number changes or sample impurities.

2.3. Previously published method evaluations

In terms of previous method evaluations, the most comprehensive benchmark studies are the Dialogue of Reverse Engineering and Assessment of Methods (DREAM) challenges [11]. These are community-driven challenges in which a panel of organizers designs competitions about, among other things, network inference. DREAM4 consisted of five synthetic networks with 100 genes and 100 samples. Each of the 100 genes could be a regulator of other genes. These networks were called “multifactorial”, as all nodes were perturbed simultaneously in the simulations. The expression matrices hence contained 100 different steady state realizations of the same underlying network. DREAM5 [11] provided three kinds of networks: a synthetic network, one derived from *E. coli*, and one derived from *S. cerevisiae*. The gene expression matrices were larger and ranged from 1600 to 5900 genes and 536–805 samples, respectively. In contrast to DREAM4, all networks defined a subset of genes that could act as regulators (between 195 and 334; cf. Fig. 1a–b). In addition, newly published methods often perform their own evaluation [12–14,16–18].

Yet, research investigating gene regulatory networks is often applied to more complex systems like human cancers [1,2,19], and not only to simpler and better-defined synthetic networks or networks from microorganisms. In the context of cancer, a single TF was validated using a set of 26 known targets or comparison between 11 known

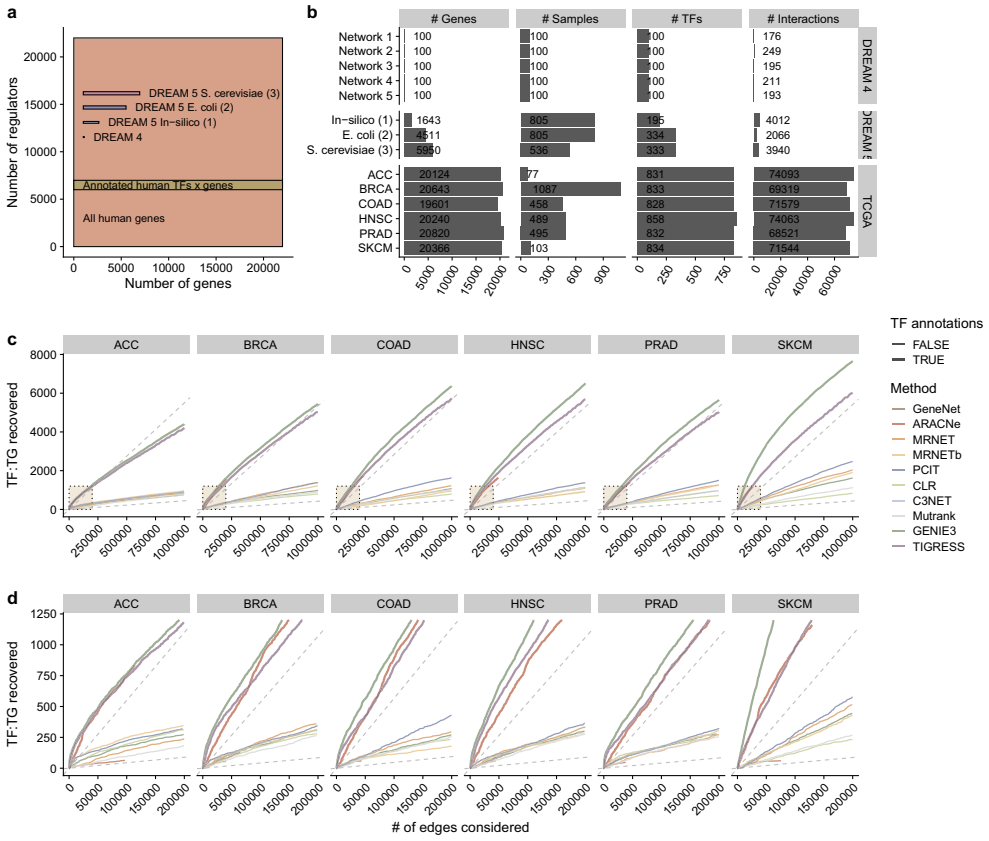


Fig. 1. Translating DREAM challenges to a cancer data set. (a) Comparison of potential interactions in the previous challenges vs. all binary interactions in a human genome with and without known regulators. (b) Number of genes, regulators, samples, and known interactions for the different challenges and TCGA cancer cohorts. (c) Number of known TF-TG pairs recovered by network inference algorithms (y axis) for six selected TCGA cohorts with size cutoff of the inferred network (x axis). Dashed lines indicate expected recovery if randomly sampling from all gene interactions (lower line) or only from known regulators (upper line) (d) zoomed view of c for small sizes of the networks.

targets and 11 negative targets [1]. We are currently not aware of a large-scale study comparing GRN inference methods in the realistic setting of human cancer gene expression [10], which may be explained by the fact that it is difficult to obtain a reference network to compare to.

3. Previous benchmarks do not accurately capture properties of cancer gene expression

3.1. Evaluating inference methods on cancer gene expression data

While the gene expression data sets that were used for evaluation have dramatically increased in size in DREAM 5 compared to DREAM 4, they are still very small in comparison to mammalian organisms (Fig. 1a, b). This starts from the number of genes and regulators present, but is also apparent by the number of confirmed regulatory interactions in the DREAM 5 networks. By contrast, the number of samples available is often not higher than in the much simpler benchmark data set (Fig. 1b). Hence, real cancer gene expression data offers a different kind of challenge for inference methods and may lead to different results compared to the previously published benchmarks. However, as the same methods are commonly employed to infer regulatory programmes in cancer, it is important to gain a better understanding of the opportunities and pitfalls that are specific to this kind of data and may not have been accurately covered in simulation studies or the other DREAM benchmarks.

Here, we are not only interested in the more complex system as defined by the number of genes and regulators, but also in specific biases that cancer gene expression exhibits and that was not covered sufficiently by synthetic or micro-organism networks, like copy number alterations or sample mixtures due to stromal and invading immune cells. As the network simulators discussed above [41,42] do not allow for these kinds of biases, we aim to evaluate GRN inference methods on cancer patient gene expression from The Cancer Genome Atlas (TCGA) [47].

We chose six TCGA cohorts with different numbers of samples available: Adrenocortical carcinoma (ACC), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Head and neck squamous carcinoma (HNSC), Prostate adenocarcinoma (PRAD), and Skin cutaneous melanoma (SKCM). These range from 77 (ACC) to 1087 (BRCA) samples per tumor type (cf. Fig. 1b). We filter all mapped genes to those with 5 or more reads on average per sample, yielding approximately 20,000 genes for all cohorts. We define potential regulators as genes that are annotated with “Transcription factor activity” in Gene Ontology [48] (GO:0003700), leaving approximately 830 regulators per cohort. As a positive set, we used consensus regulons for 101 transcription factors covering 16,500 target genes from ChEA [44] and ENCODE [43] via the Enrichr platform [45]. We did not use a negative set, as those are generally not available for transcriptional regulation.

We ran the network inference algorithms ARACNe-AP [30], GeneNet [37], GENIE3 [16], TIGRESS [17], and other methods available via the NetBenchmark R package [49] on each of our six cohorts using default options. We then evaluated how many known TF-TG pairs were recovered in the top N edges (Fig. 1c and d), prioritized by the score given to each interaction by each method. We found that GENIE3 has a slight edge over TIGRESS, which again performs slightly better than ARACNe-AP. The difference between these three methods and all others is that they take into account TF annotations, which the other methods do not.

3.2. Incorporating prior knowledge is essential for method performance

In terms of network size, the number of potential interactions with knowledge about regulators compared to without is particularly striking: If any gene can also act as a regulator, there are approximately 22,000 genes and 484 million binary interactions. By contrast, using 974 annotated TFs in Gene Ontology [48] and only taking them into account as potential regulators, we are left with only 21 million potential interactions to explore (a 22 fold decrease; cf. Fig. 1a). However, note that all the networks that we infer are undirected, hence these numbers should be halved when considering how well a method recovers known binding interactions. Also, we do not allow self-regulation, i.e. edges of a gene with itself.

This decrease in potential interactions seems to drive a superior performance of methods that are able to incorporate TF annotations in the network they infer. No matter the background or the age of a method, looking naively at the number of links recovered from known ChIP binding, there is a substantial increase (Fig. 1c, d). This result should of course be regarded with respect to the number of potential interactions: sampling randomly from all gene-gene interactions will produce a much worse performance than sampling from known TFs (lower and upper dashed grey lines in Fig. 1c, d, respectively). Nevertheless, if we are interested in recovering true regulatory interactions, it stands to reason to use a method that makes use of TF annotations. If not, no method ignoring these annotations performs anywhere close to randomly sampling from TF-TG interactions (upper dashed line in c, d).

It should be noted that the performance of all methods are close to their respective random lines. This is likely explained both by the fact that our positive set likely contains many non-regulatory binding interactions, and the observation that none of the methods in DREAM 5 performed much better than random for the *S. cerevisiae* network [11]. However, for network sizes up to 200,000 nodes GENIE3, TIGRESS, and ARACNe-AP perform better than random sampling of TF-TG interactions (cf. Fig. 1d).

3.3. Copy number changes and sample impurities are confounding gene expression measurements

As gene regulatory networks are often inferred from gene expression, it is important to consider factors influencing gene expression outside of transcription factor-target gene relationships. This is why care needs to be taken when merging together multiple data sets from e.g. microarrays and RNA-seq, or different processing pipelines that can lead to technical batch effects. These batch effects have been abundantly discussed in literature before (reviewed in [50]), and there are many approaches to correct for them [51–53].

However, cancer cells also harbor biological variability influencing gene expression and hence correlation that has so far not been discussed in depth. For instance, cancer cells often harbor copy number changes ranging from small segments (focal CNAs) up to the level of whole chromosomes (aneuploidies) [54]. Gene expression has been shown to follow these copy number changes [26,55,56], whereas protein expression is often compensated for [57].

Another factor that influences cancer gene expression in particular is that samples obtained from patients will not only consist of a homogeneous population of cancer cells. Instead, samples will also contain stromal cells that have been co-opted in tumorigenesis, as well as immune cells [27] driving inflammation and/or contributing to active clearing of tumor cells. Multiple methods have been developed to estimate cell fractions [58–61], including some that aim to reconstruct the cancer-specific transcriptome from a cell mixture [62,63]. Another level of complexity is that cancer cells themselves often consist of multiple clones and lineages that may exhibit heterogeneous traits not visible in a bulk transcriptomics measurement. While these issues can be overcome with recent single-cell sequencing technologies, it will still take years until these data sets reach a level of comprehensiveness comparable to the TCGA.

Here, we focus both on focal copy number changes and aneuploidies, as well as tumor purity (where the latter is defined as the fraction of cells in a sample estimated to be cancer cells). As focal copy number changes, we take recurrently altered regions (RACS) from the Genomics of Drug Sensitivity in Cancer (GDSC) project [64] as processed by ADMIRE [65]. For different cancer types in the TCGA, we observe a different number as well as different sizes for these regions (Fig. 2a). Glioblastoma multiforme (GBM) and Ovarian serous cystadenocarcinoma (OV) show the highest number of altered regions, with Breast invasive carcinoma (BRCA) and Lung squamous carcinoma (LUSC) showing the highest fraction of their genomes altered due to these local recurrent events. We calculate aneuploidy scores as the average absolute deviation from euploid over whole chromosomes according to copy number segments downloaded via TCGAbiolinks [66] (Fig. 2b), and use consensus purity estimates for different samples from the xCell publication [59].

The cohorts we focus on in subsequent analyses show a heterogeneous level of focal copy number changes and aneuploidies, as well as for sample purity (Fig. 2a–c). Looking at individual samples, we can observe the variability in focal amplifications from an almost euploid cohort (PRAD) up to a very high level (BRCA; cf. Fig. 2d). Similarly, PRAD also shows a low and ACC a high level of aneuploidy (Fig. 2e).

In terms of gene regulatory networks, it is unclear how these factors confounding gene expression influence the inferred edges for different methods. This is why in this study we evaluate the number of edges each of our methods infers that fall within (1) a CNA vs. outside and (2) genes whose expression strongly correlates with sample purity vs. those that do not. As a control, we check for the same enrichment in known transcription factor binding sites.

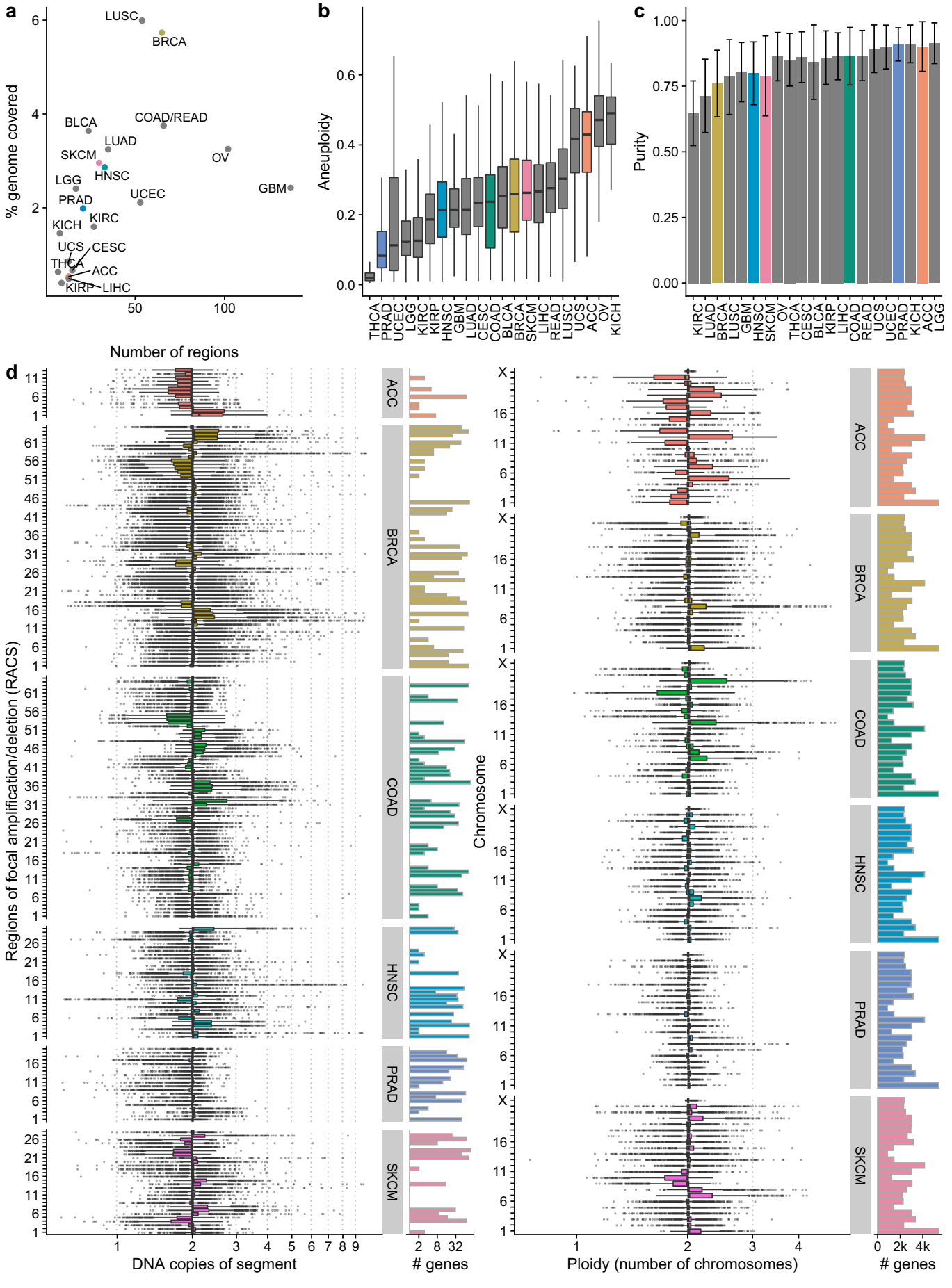
4. Network inference methods are biased towards copy number aberrations and sample purity

4.1. Focal amplifications have strong local but weak genome-wide effect

In order to test for the effect that focal amplifications have on the interactions inferred by different network inference methods, we used our selected methods and cohorts to investigate how many of the inferred edges can likely be explained by the focal amplifications and aneuploidy scores that we previously obtained (Fig. 2). Briefly, we assume that real TF-TG relationships (obtained from ChIP binding information) are equally likely to occur within focal CNAs or aneuploidies as they are between two genes anywhere on the genome. We then compare the edges obtained by the different inference methods to the number of edges theoretically possible within CNAs, and see if this fraction is different to the total number of edges inferred from the total number of possible edges. As a control, we do the same for known transcription factor binding associations to confirm our assumption of equally likely TF-TG relationships within and outside of CNAs.

Because all of the methods we tested provided a score for each inferred edge, we could test different network sizes by setting different cutoffs on the edge scores. We then went on to show the number of expected false positive edges with different network sizes for our six highlighted cohorts (Fig. 3). As our positive set that we test with is likely incomplete, we can only estimate the fraction of these false positive links, and not if any individual link is indeed a false positive.

First, we show the effect of focal amplification on the inferred edges (Fig. 3a). On the top row, we show the sum of observed links within each CNA over the number of possible links within those CNAs. We find that starting from very small networks (1000 edges), almost all of the within-CNA edges inferred by different methods in most cohorts are likely false positives (FPs), as we observe many more edges than we could expect given our null model (that edges within a CNA and outside are equally likely). These false positives, however, have got a relatively minor impact on all the edges inferred across the genome, as the



(caption on next page)

Fig. 2. Abundance of focal amplifications and aneuploidies in the TCGA. (a) Number of recurrently altered focal segments (x axis) vs. the fraction of the genome that they cover (y axis) for different TCGA cohorts. (b) Distribution of aneuploidy scores for different TCGA cohorts with chosen cohorts highlighted in color. (c) Average sample purity by cohort, error bars are standard deviation (d) Distribution of segment (left) and chromosome (right) copy numbers across samples of the six chosen cohorts.

number of genes in the identified recurrent focal amplifications is low (cf. top vs. bottom row in Fig. 3a). GeneNet is the method that shows the strongest enrichment of edges in CNAs, with up to 10% of the total number of edges in a small network (1000 edges). Other methods stay under 5% of genome-wide false positive edges. As the size cutoff gets less stringent (100,000 to 1 million edges), the genome-wide FPR for most of the methods and cohorts drops under 2%. It should, however, be noted that the total number of possible within-CNA edges is low for all cohorts and incorporating all of them in a network will still result in a low genome-wide FPR (cf. Fig. 2d). Hence, a more relaxed definition of recurrent focal CNAs (compared to the one defined by ADMIRE) would likely also yield a higher rate of genome-wide FP edges.

4.2. Aneuploidies have weaker local but strong genome-wide effect

In contrast to the focal amplifications, aneuploidies show a smaller within-segment FPR (Fig. 3, top rows). The effect of the genome-wide FPR, however, is much bigger for aneuploidies than for focal amplifications (Fig. 3, bottom rows). This makes sense intuitively, as the number of genes changed with each aneuploidy is much larger than the number of genes changed with a focal amplification (cf. Fig. 2c–d). Hence, a smaller fraction of incorrectly identified edges within each chromosome already has a large effect on the genome-wide false positive rate. We can observe this in the FP curves between focal amplifications (Fig. 3a) and aneuploidies (Fig. 3b) that reach a much higher level of genome-wide FPR for aneuploidies (up to 85% of the total number of edges inferred, compared to under 10% for focal amplifications).

These results suggest that aneuploidies are likely a major source of bias in the total number of inferred edges for most methods, especially for smaller network sizes. Therefore, caution should be taken when applying these methods to biological samples that may harbor large-scale copy number changes. For the methods that performed well in recovering TF-TG interactions (cf. Fig. 1c–d), we see that TIGRESS is most influenced by aneuploidies, followed by GENIE3 (although the effect is reversed in melanoma). Both methods show a larger FP enrichment with smaller network size, suggesting that they are prone to assigning high scores to genes co-regulated by aneuploidies instead of TF-TG interactions. ARACNe is remarkably stable in the fraction with varying network size, always showing approximately 10–20% of FPs due to aneuploidy. All methods using TF annotations behave equally in the range of 100,000 to 1 million edges. Actual TF binding interactions from ChEA (black line in Fig. 3) are equally likely to be inside and outside of CNAs.

4.3. Networks are biased due to sample composition

Apart from the copy number aberrations, we also investigated the number of false positive edges with sample impurities. The rationale is the same as above: we assume that genes whose expression level is changing with tumor purity are not more likely to be transcription factor and target gene compared to the genes whose expression does not follow that trend. A difference to the analysis before, however, is that there is no clear set of genes that are influenced by sample purity vs. genes that are not. In order to address this, we selected either the top 1000 or top 5000 genes that correlated the most with tumor purity in

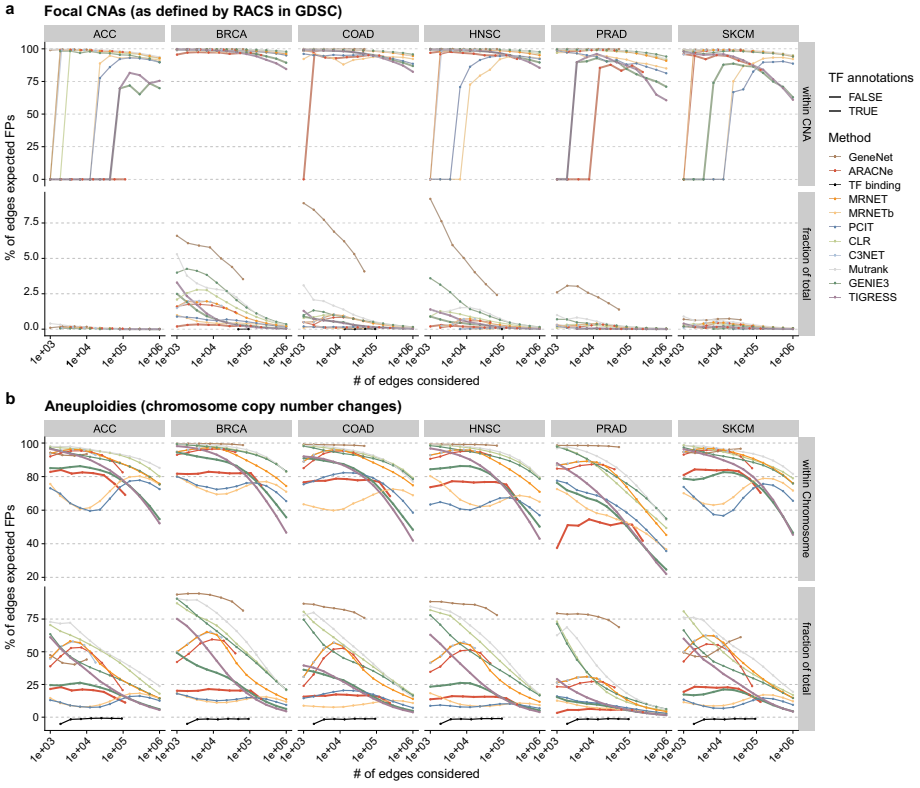


Fig. 3. Effect of (a) focal amplifications and (b) aneuploidies on inferred network edges. False positive rate (y axis) shown for different network size (x axis) either within the CNA (top row) or in comparison to the total number of edges between all genes (bottom row). Dots along lines show where this was quantified.

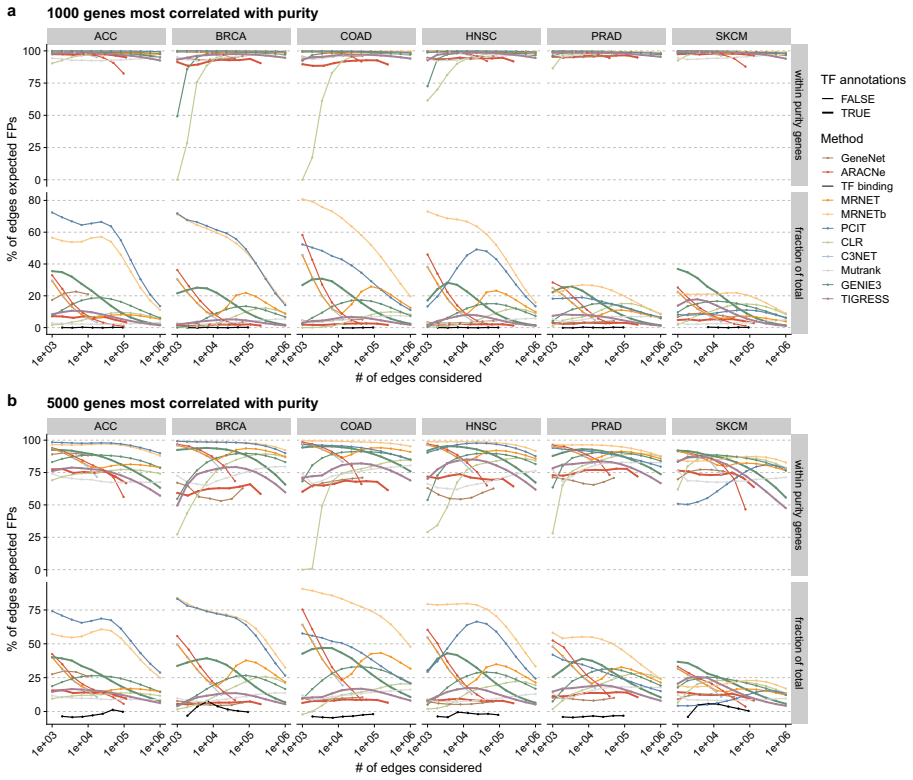


Fig. 4. Effect of sample purity on inferred network edges. For (a) 1000 or (b) 5000 genes most correlated with sample purity, expected false positive rate (y axis) is shown for different network sizes (x axis) in the chosen cohorts.

each cancer type we highlight, and then performed the same enrichment analysis as we did with the focal copy number changes and aneuploidies.

For the methods that performed well in recovering TF-TG interactions, we see that GENIE3 is more susceptible to wrongly inferring links due to samples mixtures than TIGRESS, with more FPs in smaller networks (10–40% vs. 10–20%, respectively). ARACNe shows a stable FPR of 10–15% irrespective of network size. From 100,000 to 1 million links, the methods largely equalize. Again, we do not observe an enrichment of TF binding with purity-correlated genes (black line in Fig. 4). As with the results we find for the CNAs, there is a trade-off between the within-chromosome FPR vs. the genome-wide FPR depending on the cutoff for selecting the purity-correlated genes (albeit less pronounced).

5. Conclusion

When inferring gene networks in the context of cancer, it is important to not only keep in mind the potential technical variability between batches that may induce false positive correlations and hence edges when using network inference methods, but also the biologically intrinsic confounding factors of gene expression, like the one induced by DNA copy number changes [26,55] or a mixture of different proportions of different cell types [60,61].

We have shown that for recovering an accurate network of TF-TG interactions in cancer, methods that incorporate TF annotations should be preferred to those that are not. However, even these methods are largely susceptible to inferring false positive links due to confounding factors. Combined, aneuploidies and sample impurities can be expected to contribute approximately 20–50% of false positive edges for networks with 100,000 to 1 million edges, and up to 80% for smaller networks.

If we are interested in accurately inferring true regulatory interactions, there is a need to correct for these biases. Various methods have been proposed to correct for confounding factors in gene networks in general (like Principal Component Analysis to correct for major axes of variation [67] and linear models to remove unwanted multivariate noise [68]), but the authors have not investigated how well their methods adjust for biological influences like the ones we discussed here. In addition, future method development in this area could address these biases more directly by also modeling aneuploidies and sample impurities explicitly.

6. Methods

6.1. Gene expression and copy number data from the TCGA

We have downloaded the raw read counts for gene expression as well as the inferred continuous regions of the same DNA copy number (copy number segments) from the harmonized TCGA data obtained through the R package TCGAbiolinks [66]. We chose the cohorts (ACC, BRCA, COAD, HNSC, PRAD, and SKCM) because they represented a wide range of sample sizes, ploidy, and purity values (cf. Fig. 2).

We further filtered the samples set to only contain primary tumors (TCGA sample type of “01A”). We filtered the genes to only contain genes on human chromosomes 1 to 21 (excluding X, Y, and MT) and to have more than 20% of the samples with 10 or more reads. We then used the DESeq2 R package [69] to estimate library size factors and get variance stabilized gene expression values.

Finally, we mapped Ensembl IDs to HGNC gene symbols using Ensembl 96 and removed all genes that did not have a valid gene symbol or were duplicated.

6.2. Focal and chromosome copy numbers

For regions with recurrent copy number alterations for both our cohorts, we downloaded Table S2D from https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/ [64].

To get copy numbers of either these segments or whole chromosomes, we calculated the average copy number along the respective regions for each sample in our cohorts.

6.3. Sample purity and purity-correlated genes

We obtained consensus purity estimate from xCell [59] using their "estimate" field in their Additional File 6.

For the primary samples of each of our cohorts, we then calculated differential expression along this estimate (using a likelihood ratio test over the intercept using the DESeq2 package), and selected the top *N* genes by lowest *p*-value.

6.4. Transcription factor annotations and binding data

To generate a list of genes that may act as transcription factors, we downloaded all HGNC symbols associated with GO:0003700 (DNA-binding transcription factor activity) from Ensembl 96 [70].

For our positive set of real transcriptional regulation, we downloaded the "ENCODE_and_ChEA_Consensus_TFs_from_ChIP-X" HGNC symbols from the Enrichr platform [45], encompassing 16,500 different target genes of 100 transcription factors.

6.5. Network inference

For inferring our networks, we used the following methods: ARACNe-AP [30], which is a Java implementation that extends the original ARACNe method; the GeneNet R package [37]; the GENIE3 R package [16], the TIGRESS R package [17], and other methods available in the NetBenchmark R package [49].

We infer one network per cohort per method, and look for enrichment of edges that are likely due to copy number changes or sample mixtures. As not all of these methods provide a significance measure, we instead look at the order of edges inferred, from the highest score to the lowest.

6.6. Quantifying possible TF-TG interactions

To quantify enrichment of real TF-TG interactions (obtained from ChIP binding experiments) within the top *N* genes of a given network, we first need to enumerate the possible number of edges given how many genes and transcription factors we have, and whether a GRN inference method knows the difference between regulators and targets.

In particular, if there are no known regulators we consider the number of possible edges to be:

$$0.5 \times (\mathrm{ng} - 1) \times \mathrm{ng}$$

And if they are known instead:

$$(\mathrm{ng} - \mathrm{ntf}) \times \mathrm{ntf} + 0.5 \times (\mathrm{ntf} - 1) \times \mathrm{ntf}$$

where *ng* is the total number of genes and *ntf* is the number of potential regulators. Note that if every gene can be a regulator, the lower formula simplifies into the upper.

6.7. Enrichment of edges within gene sets

We quantify bias by copy number changes or purity by assuming that the genes in a set (focal regions, chromosomes, or purity-associated genes) are equally likely to form links within the respective set as they are with genes outside the set. We then look for enrichment of edges within a given region over the total number of edges.

In particular, we first compute the odds ratio of a method obtaining links in a segment vs. the overall number of edges:

$$OR = \frac{\textit{inferred edges in segment}}{\textit{inferred network size}} \div \frac{\textit{possible edges in segment}}{\textit{possible edges in genome}}$$

Then, the local false positive rate (FPR) is defined as:

$$FPR_{segment} = 1 - 1/OR$$

While the genome-wide FPR is the number of expected FP links divided by the size of the inferred network:

$$FPR_{genome} = FPR_{segment} \times \frac{\textit{inferred edges in segment}}{\textit{inferred network size}}$$

6.8. Code availability

The analysis code of this manuscript, including code to generate all figures is available at https://github.com/mschubert/GRN-aneup-purity, licensed under GNU GPL version 3 or later.

Funding

FF is funded by Dutch Cancer Society grant 2015-RUG-7833.

MCT acknowledges funding from the Impuls und Vernetzungsfonds of the Helmholtz Gemeinschaft (grant VH-NG-1219).

Transparency document

The Transparency document associated with this article can be found, in online version.

Declaration of competing interest

The authors declare that no conflict of interest exists.

References

[1] K. Basso, et al., Reverse engineering of regulatory networks in human B cells, Nat. Genet. 37 (2005) 382–390.
[2] M.S. Carro, et al., The transcriptional network for mesenchymal transformation of brain tumours, Nature 463 (2009) 318–325.
[3] R. De Smet, K. Marchal, Advantages and limitations of current network inference methods, Nat. Rev. Microbiol. 8 (2010) 717–729.
[4] C. Lefebvre, et al., A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers, Mol. Syst. Biol. 6 (2010).
[5] P.B. Madhamshettiwar, S.R. Maetschke, M.J. Davis, A. Reverter, M.A. Ragan, Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets, Genome Med. 4 (2012) 41.
[6] R.A. Chávez Montes, et al., ARACNe-based inference, using curated microarray data, of Arabidopsis thaliana root transcriptional regulatory networks, BMC Plant Biol. 14 (2014) 97.
[7] M.J. Alvarez, et al., Functional characterization of somatic mutations in cancer using network-based inference of protein activity, Nat. Genet. (2016), https://doi.org/10.1038/ng.3593.
[8] F. Emmert-Streib, G.V. Glazko, G. Altay, R. de Matos Simoes, Statistical inference and reverse engineering of gene regulatory networks from observational expression data, Front. Genet. 3 (2012) 8.
[9] E. Khurana, et al., Role of non-coding sequence variants in cancer, Nat. Rev. Genet. 17 (2016) 93–108.
[10] S. Barbosa, B. Niebel, S. Wolf, K. Mauch, R. Takors, A guide to gene regulatory network inference for obtaining predictive solutions: underlying assumptions and fundamental biological and data constraints, Biosystems. 174 (2018) 37–48.
[11] D. Marbach, et al., Wisdom of crowds for robust gene network inference, Nat. Methods 9 (2012) 796–804.
[12] A.J. Butte, I.S. Kohane, Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements, Pac. Symp. Biocomput. (2000) 418–429.
[13] A. Margolin, et al., ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, BMC Bioinformatics 7 (2006) S7.
[14] J.J. Faith, et al., Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles, PLoS Biol. 5 (2007) e8.
[15] G. Altay, F. Emmert-Streib, Inferring the conservative causal core of gene regulatory networks, BMC Syst. Biol. 4 (2010) 132.
[16] V.A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, Inferring regulatory networks from expression data using tree-based methods, PLoS One 5 (2010).
[17] A.-C. Haury, F. Mordelet, P. Vera-Licona, J.-P. Vert, TIGRESS: trustful inference of gene regulation using stability selection, BMC Syst. Biol. 6 (2012) 145.
[18] J. Ruyssinck, et al., NIMEFI: gene regulatory network inference using multiple ensemble feature importance algorithms, PLoS One 9 (2014) e92709.
[19] L.N. Kwong, et al., Oncogenic NRAS signaling differentially regulates survival and proliferation in melanoma, Nat. Med. 18 (2012) 1503–1510.

[20] J. Zhang, et al., Weighted frequent gene co-expression network mining to identify genes involved in genome stability, PLoS Comput. Biol. 8 (2012) e1002656.
[21] Y. Yang, et al., Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types, Nat. Commun. 5 (2014) 3231.
[22] D.M. Wolf, M.E. Lenburg, C. Yau, A. Boudreau, L.J. van't Veer, Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity, PLoS One 9 (2014) e88309.
[23] N. Zhao, et al., Identification of biomarker and co-regulatory motifs in lung adenocarcinoma based on differential interactions, PLoS One 10 (2015) e0139165.
[24] K. Oros Klein, et al., Gene Coexpression analyses differentiate networks associated with diverse cancers harboring TP53 missense or null mutations, Front. Genet. 7 (2016) 137.
[25] D.J. Gordon, B. Resio, D. Pellman, Causes and consequences of aneuploidy in cancer, Nat. Rev. Genet. 13 (2012) 189–203.
[26] R.S.N. Fehrmann, et al., Gene expression analysis identifies global gene dosage sensitivity in cancer, Nat. Genet. 47 (2015) 115–125.
[27] V. Thorsson, et al., The immune landscape of Cancer, Immunity 48 e14 (2018) 812–830.
[28] F. Markowetz, R. Spang, Inferring cellular networks–a review, BMC Bioinformatics 8 (Suppl. 6) (2007) S5.
[29] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, R. Guthke, Gene regulatory network inference: data integration in dynamic models—a review, Biosystems. 96 (2009) 86–103.
[30] A. Lachmann, F.M. Giorgi, G. Lopez, A. Califano, ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information, Bioinformatics 32 (2016) 2233–2235.
[31] N.S. Watson-Haigh, H.N. Kadarmideen, A. Reverter, PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches, Bioinformatics 26 (2010) 411–413.
[32] R. de Matos Simoes, F. Emmert-Streib, Bagging statistical network inference from large-scale gene expression data, PLoS One 7 (2012) e33624.
[33] P.E. Meyer, D. Marbach, S. Roy, M. Kellis, Information-theoretic inference of gene networks using backward elimination, BIOCOMP, 2010.
[34] P.E. Meyer, F. Lafitte, Bontempi, G. minet: a R/bioconductor package for inferring large transcriptional networks using mutual information, BMC Bioinformatics 9 (2008) 461.
[35] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2004) 407–499.
[36] N. Singh, M. Vidyasagar, bLARS: an algorithm to infer gene regulatory networks, IEEE/ACM Trans. Comput. Biol. Bioinform. 13 (2016) 301–314.
[37] J. Schäfer, R. Opgen-Rhein, K. Strimmer, Reverse engineering genetic networks using the GeneNet package, J. Am. Stat. Assoc. 96 (2001) 1151–1160.
[38] J. Sławek, T. Arodź, ENNET: inferring large gene regulatory networks from expression data using gradient boosting, BMC Syst. Biol. 7 (2013) 106.
[39] J. Wu, X. Zhao, Z. Lin, Z. Shao, Large scale gene regulatory network inference with a multi-level strategy, Mol. BioSyst. 12 (2016) 588–597.
[40] S. Guo, Q. Jiang, L. Chen, D. Guo, Gene regulatory network inference using PLS-based methods, BMC Bioinformatics 17 (2016) 545.
[41] T. Van den Bulcke, et al., SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms, BMC Bioinformatics 7 (2006) 43.
[42] T. Schaffter, D. Marbach, D. Floreano, GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods, Bioinformatics 27 (2011) 2263–2270.
[43] The ENCODE, Project consortium, An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 489 (2012) 57–74.
[44] A. Lachmann, et al., ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments, Bioinformatics 26 (2010) 2438–2444.
[45] E.Y. Chen, et al., Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinformatics 14 (2013) 128.
[46] M. Gheorghe, et al., A map of direct TF-DNA interactions in the human genome, Nucleic Acids Res. e21 (2019) 47.
[47] The Cancer Genome Atlas Research Network et al, The cancer genome atlas pan-cancer analysis project, Nat. Genet. 45 (2013) 1113–1120.
[48] Gene Ontology Consortium, The gene ontology (GO) database and informatics resource, Nucleic Acids Res. 32 (2004) D258–D261.
[49] P. Bellot, C. Olsen, P. Salembier, A. Oliveras-Vergés, P.E. Meyer, NetBenchmark: A bioconductor package for reproducible benchmarks of gene regulatory network inference, BMC Bioinformatics 16 (2015) 312.
[50] J.T. Leek, et al., Tackling the widespread and critical impact of batch effects in high-throughput data, Nat. Rev. Genet. 11 (2010) 733–739.
[51] H.S. Parker, H.C. Bravo, J.T. Leek, Removing batch effects for prediction problems with frozen surrogate variable analysis, PeerJ 2 (2014) e561.
[52] M.N. McCall, B.M. Bolstad, R.A. Irizarry, Frozen robust multiarray analysis (fRMA), Biostatistics 11 (2010) 242–253.
[53] V. Nygaard, E.A. Rødland, E. Hovig, Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses, Biostatistics 17 (2016) 29–39.
[54] A.M. Taylor, et al., Genomic and functional approaches to understanding cancer aneuploidy, Cancer Cell 0 (2018).
[55] N. Pavelka, et al., Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast, Nature 468 (2010) 321–325.
[56] J. Roszik, et al., Somatic copy number alterations at oncogenic loci show diverse correlations with gene expression, Sci. Rep. 6 (2016) 19649.
[57] E. Gonçalves, et al., Widespread post-transcriptional attenuation of genomic copy-number variation in cancer, Cell Syst 5 e4 (2017) 386–398.
[58] A.M. Newman, et al., Robust enumeration of cell subsets from tissue expression profiles, Nat. Methods 12 (2015) 453–457.
[59] D. Aran, Z. Hu, A.J. Butte, xCell: Digitally portraying the tissue cellular heterogeneity landscape, Genome Biol. 18 (2017) 220.
[60] F. Finotello, et al., Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data, *bioRxiv* 223180, 2018, https://doi.org/10.1101/223180.
[61] G. Sturm, et al., Comprehensive evaluation of cell-type quantification methods for immuno-oncology, *bioRxiv* 463828, 2018, https://doi.org/10.1101/463828.
[62] J. Ahn, et al., DeMix: deconvolution for mixed cancer transcriptomes using raw measured data, Bioinformatics 29 (2013) 1865–1871.
[63] Z. Wang, et al., Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration, *bioRxiv* 146795, 2017, https://doi.org/10.1101/146795.
[64] F. Iorio, et al., A landscape of pharmacogenomic interactions in cancer, Cell 166 (2016) 740–754.
[65] E. van Dyk, M.J.T. Reinders, L.F.A. Wessels, A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control, Nucleic Acids Res. 41 (2013) e100.
[66] A. Colaprico, et al., TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data, Nucleic Acids Res. 44 (2016) e71.
[67] P. Parsana, et al., Addressing confounding artifacts in reconstruction of gene co-expression networks, Genome Biol. 20 (2019) 94.
[68] S. Freytag, J. Gagnon-Bartsch, T.P. Speed, M. Bahlo, Systematic noise degrades gene co-expression signals but can be corrected, BMC Bioinformatics 16 (2015) 309.
[69] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biology 15 (2014) 550, https://doi.org/10.1186/s13059-014-0550-8.
[70] Hubbard, T. *et al.* The Ensembl genome database project. Nucleic Acids Res. 30, 38–41 (2002).