bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL

Anika Liu[1,2,*], Panuwat Trairatphisan[1,*], Enio Gjerga[1,2,*], Athanasios Didangelos[3], Jonathan Barratt[3], Julio Saez-Rodriguez[1,2,#]

[1] Heidelberg University, Faculty of Medicine, Institute of Computational Biomedicine, 69120 Heidelberg, Germany
[2] RWTH Aachen University, Faculty of Medicine, Joint Research Centre for Computational Biomedicine (JRC-COMBINE), 52074, Aachen, Germany
[3] Department of Infection, Immunity and Inflammation, University of Leicester, Leicester, United Kingdom

* co-first authors
[#] Corresponding author: julio.saez@bioquant.uni-heidelberg.de

Abstract

While gene expression profiling is commonly used to gain an overview of cellular processes, the identification of upstream processes that drive expression changes remains a challenge. To address this issue, we introduce CARNIVAL, a causal network building tool which derives network architectures from gene expression footprints.

CARNIVAL (CAusal Reasoning pipeline for Network identification using Integer VALue programming) integrates different sources of prior knowledge, including signed and directed protein-protein interactions, transcription factor targets, and pathway signatures. The use of prior knowledge in CARNIVAL allows the capture of a broad set of upstream cellular processes and regulators, which in turn delivered results with higher accuracy when benchmarked against related tools. Implementation as an integer linear programming (ILP) problem also guarantees efficient computation. As a case study, we applied CARNIVAL to contextualize signaling networks from gene expression data in IgA nephropathy, a chronic kidney disease. CARNIVAL identified specific signaling pathways and associated mediators with important bioactivity in IgAN including WNT and TGF-β, that we subsequently validated experimentally.

In summary, we demonstrated how CARNIVAL generates hypotheses on potential upstream alterations that propagate through signaling networks, providing valuable insights into disease processes. CARNIVAL, freely available as an R-package at https://saezlab.github.io/CARNIVAL, can be applied to any field of biomedical research to contextualize signaling networks and identify the causal relationships between downstream gene expression and upstream regulators.

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Introduction

Cells possess a sophisticated and finely tuned signaling architecture, and the de-regulation of signaling processes can mislead cellular behaviour in many diseases. A better understanding of signaling, therefore, allows to gain insights into disease processes and to prioritize potential targets for drug development.

The computational inference of signaling networks' can be performed based on phosphoproteomics data that directly measure key signaling players such as receptors and kinases (Giudice and Petsalaki, 2017; Invergo and Beltrao, 2018), preferably in combination with prior knowledge (Hill *et al.*, 2016). However, the availability of phosphoproteomics data is often limited in many research fields while gene expression data is more abundant. The inference of signaling networks based on gene expression is, therefore, an attractive approach to uncover the organisation of cellular signal transduction.

There are multiple computational tools which allow the inference of regulatory signaling networks based on gene expression data. Many of these methods assume gene expression levels as a proxy for signaling protein activities and use them to construct networks (Chen *et al.*, 2014). For instance, (Huang and Fraenkel, 2009) mapped transcriptomics data onto signaling pathways then applied a Steiner's tree algorithm for network contextualization. Such methods can provide valuable insight, but are limited by the fact that the abundance and activities of signaling proteins only partially correlate with gene expression (Vogel and Marcotte, 2012).

To overcome such limitation, an alternative way is to identify upstream signaling regulators from the profiles of downstream gene targets. One approach is to analyse gene expression footprints of signalling pathways obtained from perturbation experiments (Schubert *et al.*, 2018; Parikh *et al.*, 2010; Bild *et al.*, 2005). Another one to predict transcription factor (TF) activities based on their regulons (Alvarez *et al.*, 2016; Garcia-Alonso *et al.*, 2018). However, these approaches do not provide information on signaling pathways' topology. These insights can be obtained by applying network-based approaches which allow the incorporation of network structures as prior information.

In recent years, a causal network inference approach, which integrates a causal reasoning principle into network-based modeling, was introduced. Given a starting prior knowledge network (PKN), upstream signaling regulators can be inferred from downstream signaling targets in the form of a sub-network that connects not only just direct connections but also as a cascade of signaling events. This causal network approach was implemented e.g. in (Melas *et al.*, 2015) and in CausalR (Bradley and Barrett, 2017; Chindelevitch *et al.*, 2012). These tools, however, only take the PKN as prior knowledge. X2K, in contrast, uses expression footprint as prior knowledge to link gene expression to upstream regulatory kinases using TF and kinase enrichment, but without considering the causality of the cascades (Chen *et al.*, 2012).

We set out to integrate the causal network approach with expression footprints to infer the whole signalling cascade**.** For this, we developed the causal reasoning tool CARNIVAL (CAusal Reasoning pipeline for Network identification using Integer VALue programming). CARNIVAL expands an

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ILP implementation for causal reasoning (Melas *et al.*, 2015) to integrate information from TF and signaling pathway activity scoring. In addition, it can be applied not only to perturbation experiments as in the original implementation but also generally to compare between two or more conditions. CARNIVAL uses a comprehensive collection of pathway resources available in OmniPath as PKN (Türei *et al.*, 2016), though other sources can be used. We performed a benchmarking study using the SBVimprover Species Translation Challenge dataset (Poussin *et al.*, 2014) and compared its performance to an alternative causal reasoning network building tool CausalR (Bradley and Barrett, 2017). As a case study, we apply CARNIVAL to glomerular gene expression data on IgA nephropathy (IgAN) to gain insights on the cellular processes that regulate its pathophysiology. These were confirmed by independent experimental validation.

2 Results

2.1 CARNIVAL pipeline. The ILP-based causal reasoning pipeline by Melas *et al.* requires a prior knowledge network, differential gene expression, as well as potential or known target(s) of perturbation for which input values are discretized. CARNIVAL incorporates several modifications: First, the objective function was customized to incorporate TF and pathway activity levels in a continuous scale. Second, deregulated TFs are first derived with DoRothEA (Garcia-Alonso *et al.*, 2018; Schubert *et al.*, 2018) summarizing potentially noisy gene expression data into TF activities to be used as input. Last, CARNIVAL overcomes the need for known targets of perturbations which restricted the original method's applicability (see Methods). Two CARNIVAL pipelines are introduced here which will be referred henceforth as Standard CARNIVAL 'StdCARNIVAL' (with known perturbation targets as an input) and Inverse CARNIVAL 'InvCARNIVAL' (without information on targets of perturbation), see Figure 1.

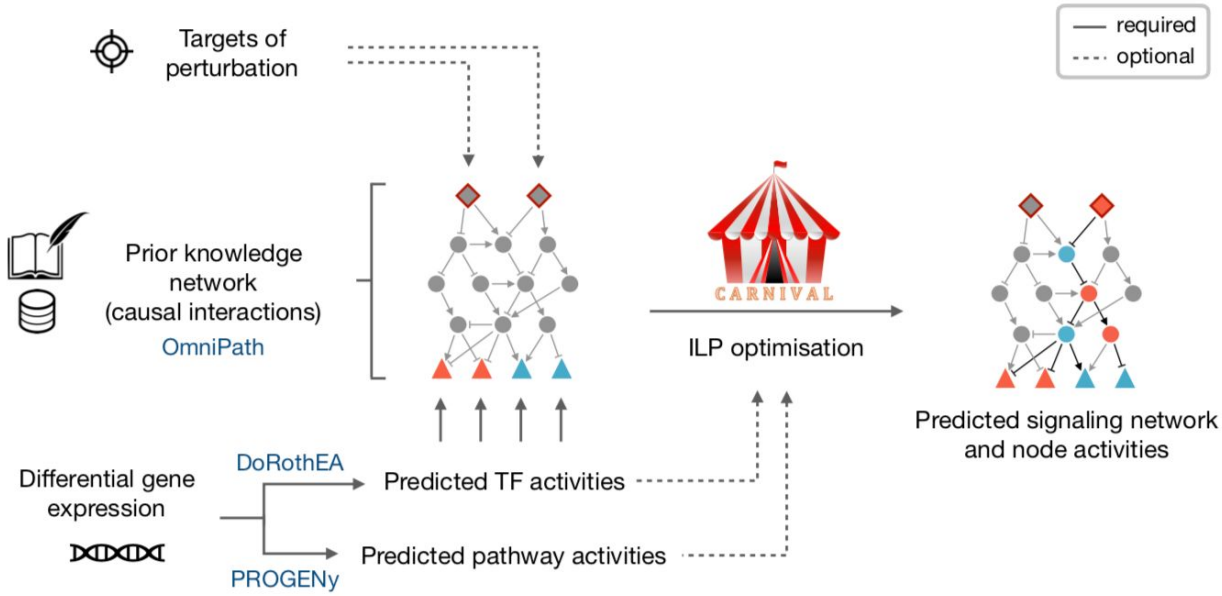

Figure 1: CARNIVAL pipeline. The CARNIVAL pipeline requires as input a prior knowledge network and differential gene expression. The information on perturbed targets and their effects can be assigned (StdCARNIVAL) or omitted (InvCARNIVAL). The differential gene expression is used to infer transcription factor (TF) activities with DoRothEA, which are subsequently discretized in

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

order to formulate ILP constraints. As a result, CARNIVAL derives a family of highest scoring networks which best explain the inferred TF activities. Continuous pathway and TF activities can be additionally considered in the objective function.

2.2 Benchmarking on the SBV improver dataset. To evaluate the performance of the CARNIVAL pipeline, we applied it to the SBVimprover Species Translation Challenge dataset which provides phosphorylation and gene expression data for multiple perturbations (Poussin *et al.*, 2014), see Suppl. Figure S1A. We applied both the StdCARNIVAL and InvCARNIVAL pipeline to evaluate the effect of information on perturbation targets on the resulting networks. The results from both pipelines were then compared to the ones generated by CausalR.

In this study, the phosphoprotein data could not be sufficiently mapped to CARNIVAL node activities (see Methods). We therefore first determined which pathways are known to be linked to the perturbations by different molecules according to the KEGG database (Kanehisa *et al.*, 2017); referred to perturbation-attributed pathways henceforth). We then performed an enrichment analysis to define whether they are more up- or down-regulated (see Methods and Suppl. Figure S1B). This gives insights into how well expected pathways and their regulatory direction are captured by CARNIVAL.

2.2.1 Incorporation of predicted TF and pathway activities. The normalized enrichment scores (NES) from DoRothEA were used as an estimate for the degree of dysregulation (see Methods). For StdCARNIVAL, significant enrichment of the perturbation-attributed pathway set in up-regulated pathways is only achieved for IL1-β (IL1B) and TGF-α (TGFA) with the introduction of TF weights (Suppl. Figure S2). In InvCARNIVAL, where the targets of perturbations are not known, the results with pathway weights showed a significant enrichment of perturbation-attributed pathways in up-regulated pathways for PDGF-β (PDGFB), IL1-β, EGF, and TGF-α while only TGF-α was significant without pathway weights (Suppl. Figure S3). TGF-α was more enriched in activated than in inhibited pathways with pathway weights but this trend was inverse for FSL1 and IGF2. While being imperfect, PROGENy weight still provides an overall improvement in detecting more deregulated pathways (4 versus 1).

2.2.2 Comparison of StdCARNIVAL, InvCARNIVAL, GSEA and CausalR. The perturbations found in InvCARNIVAL (PDGF-β, IL1-β, EGF, and TGF-α) were also identified in StdCARNIVAL (Figure 2; Suppl. Figure S4). In contrast, NTF3 and FSL1 only showed a significant enrichment in StdCARNIVAL.

The same number of enrichment of activated pathways in the perturbation-attributed pathways were captured with InvCARNIVAL and with pathway inference from differential gene expression directly. TGF-α and IL1-β were captured with both approaches, while IFN-γ (IFNG) and TNF-α (TNFA) were only significantly enriched in activated pathways inferred by GSEA, and PDGF-β and EGF by InvCARNIVAL. However, the trend of directionality was more often correct in pathway activities inferred with InvCARNIVAL. CausalR only shows a significant enrichment of activated pathways in the perturbation-attributed pathway set for IFNG, and therefore does not perform as well as CARNIVAL and GSEA (Figure 2).

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

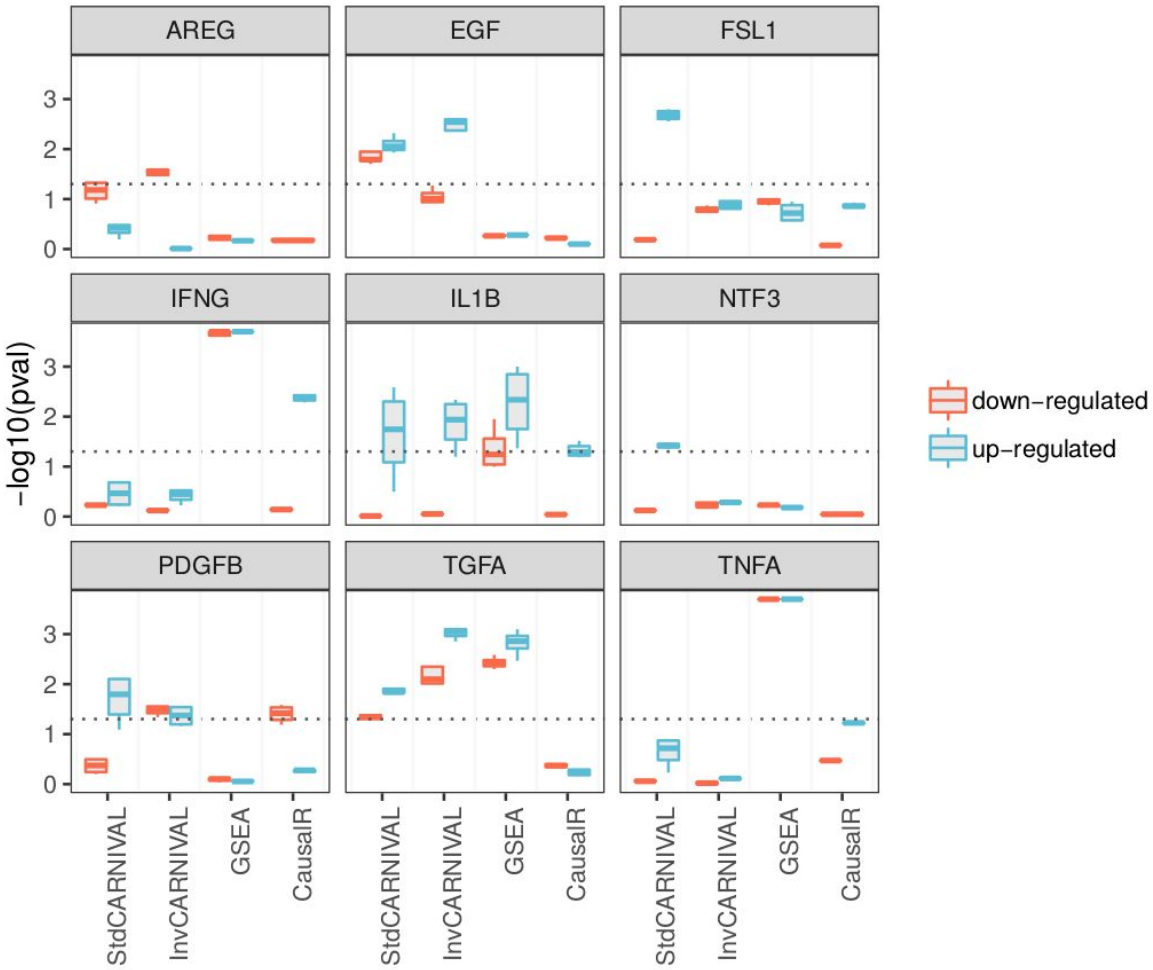

Figure 2. Comparison of the enrichment results of the perturbation-attributed pathway set in dysregulated pathways inferred with different tools. An enrichment of the perturbation-attributed pathway set among the significant pathways was determined. The significance level of 0.05 is indicated by the dotted lines.

On another perspective, a bias towards hub nodes was observed for CausalR while this was not true for CARNIVAL. The degree distribution of edges in CARNIVAL was highly similar to the one of the prior knowledge network (PKN), see Suppl. Figure S5.

2.3 Inferring glomerular signaling in IgAN. Immunoglobulin A nephropathy (IgAN) is a common chronic kidney disease (CKD) accounting for 35% of all renal transplantations in adults (Wyatt and Julian, 2013) and is the most frequent form of glomerulonephritis. It is characterized by the deposition of aberrant IgA in the glomeruli of the kidney. The pathogenic IgA-containing immune complexes trigger the activation of inflammation and fibrosis (Yeo *et al.*, 2018). Further improvement in early diagnosis and treatment of IgAN are needed and can only be achieved by a better understanding of disease mechanisms.

2.3.1 InvCARNIVAL results. In this study, we generated the causal networks from the differential glomerular gene expression between groups of healthy subject versus IgAN patients. Given that the

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

node penalty did not affect the performance but might result in minor fluctuations, this analysis was performed with different node penalties to achieve more robust results (β in (0.03; 0.1; 0.3; 0.5; 0.8), see Methods). To give an example of a solution network from CARNIVAL, the combined solutions are shown in Figure 3. This network consists of 43 TFs, 37 input nodes and 62 associated nodes which are connected through 231 edges. Given that the adherens junction set is the most strongly dysregulated one, the set members are highlighted in the network. Thereby, only one transcription factor (TCF7) inferred by DoRothEA is represented in this gene set, while 13 associated nodes and four input nodes are solely inferred by causal reasoning with CARNIVAL.

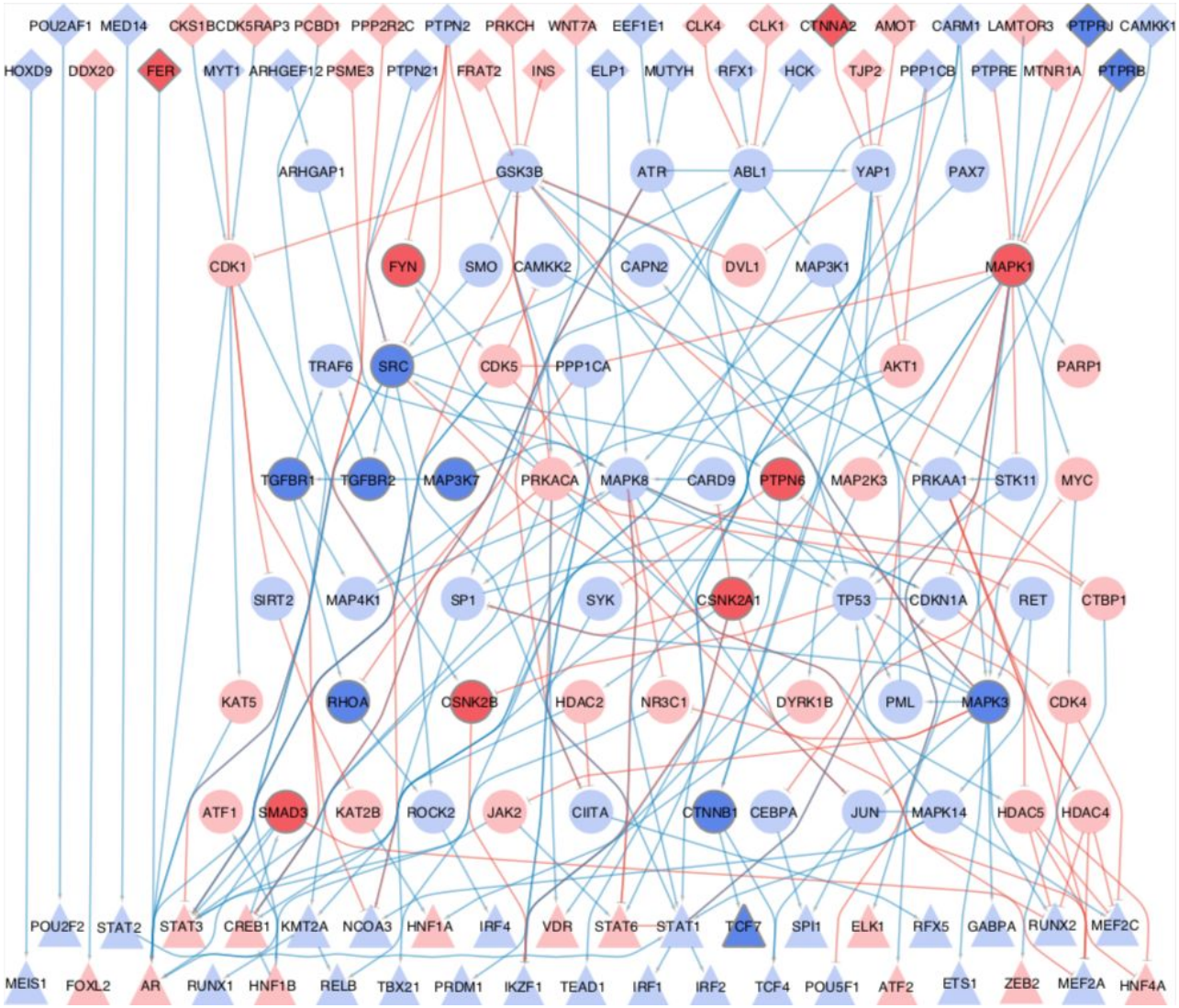

Figure 3. IgAN-contextualized network from CARNIVAL. The network summarizes the CARNIVAL results for node penalty β=0.5. Up-regulated nodes and activatory reactions are indicated in blue while down-regulated nodes and inhibitory edges are colored in red. Triangles correspond to transcription factors, squares represent input nodes and circles correspond to purely inferred nodes. Members of the most dysregulated gene set, i.e. adherens junctions, are labeled by more intense background colors.

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2.3.2 Inferred dysregulated cellular processes by CARNIVAL. Dysregulated cellular processes were inferred by over-representation analysis of the CARNIVAL nodes in the KEGG pathway maps. The most significantly dysregulated pathways were identified by median p-value over different node penalties (Figure 4). Among these, known drivers of renal fibrosis including TGF-β, WNT, and EGFR/ERbB signaling stand out (Tang *et al.*, 2018; Zhou and Liu, 2015; Rayego-Mateos *et al.*, 2018). These processes are significantly over-represented in CARNIVAL, but are not significant in GSEA. Additionally, focal adhesions were reported as an activated process in CARNIVAL and GSEA. Interactions between extracellular matrix and the cytoskeleton are particularly important in matrix-producing cells like fibroblasts and mesangial cells in the kidney (Rustad *et al.*, 2013).

Wnt/β-catenin signaling has a central role in mediating podocyte dysfunction and β-catenin is activated in podocytes in various proteinuric kidney diseases, including IgAN (Zhou and Liu, 2015). Consistent with clinical data in IgA nephropathy reporting podocyte injury and podocyturia as important prognostic features, we identified adhesion junctions as the most significantly dysregulated gene set and tight junctions as the most significantly down-regulated one (Bellur *et al.*, 2017).

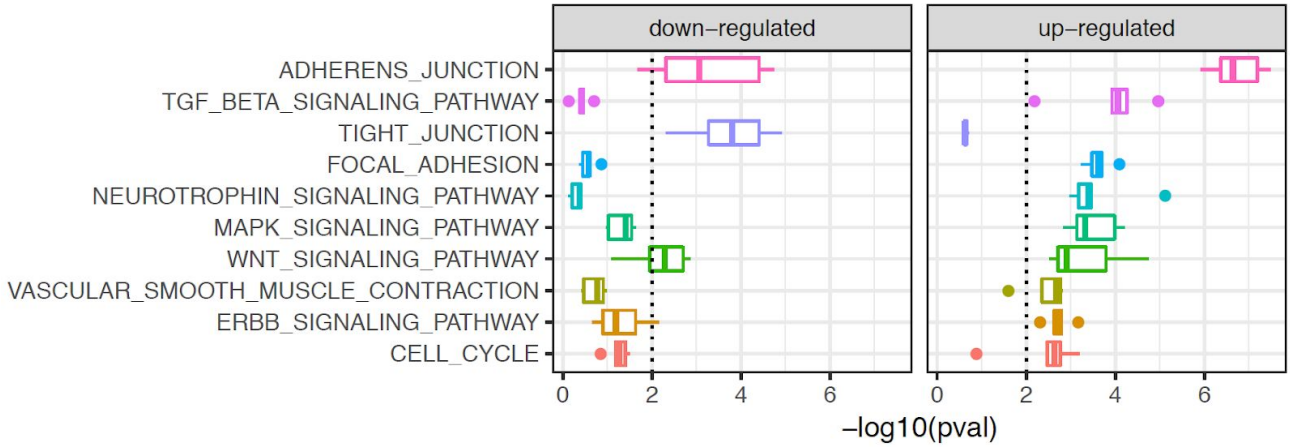

Figure 4. Dysregulated cellular processes in IgAN. Up- and down-regulated pathways are shown with decreasing median significance from top to bottom. The significance level is 0.01. Among others, these point to podocyte injury and the disruption of the slit diaphragms, as well as fibrosis.

2.3.3 Fluorescence immunohistology detection of RhoA and β-catenin. To validate our findings, we examined the protein levels of components of the TGF-β and Wnt signaling pathways which contribute to the pathophysiology of IgAN. Here, we chose RhoA and β-catenin as the representatives to perform fluorescence immunostaining on human renal biopsies from healthy pre-transplantation (controls) and IgAN diagnostic biopsies (see Suppl. Figure S6 and Suppl. Text S1). As expected, mesangial IgA staining was present in IgAN and absent in control specimens (Suppl. Figure S6D-F and Suppl. Figure S6A-C, respectively). In contrast, RhoA and β-catenin were present, albeit differentially expressed, in both control and IgAN biopsies. RhoA was expressed in glomeruli and renal tubular epithelial cells (proximal and distal), where it displayed a luminal preference. β-catenin was expressed in tubular epithelial cells at varying intensity. Notably, we observed stronger β-catenin staining in tubular epithelial cells in IgAN and an increase in β-catenin staining in IgAN glomeruli which was most likely in mesangial cells (Suppl. Figure S6D-F).

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3 Discussion

We present a novel causal network building tool, CARNIVAL. CARNIVAL seamingly integrates gene expression data with various types of prior biological knowledge (signaling networks, TF-targets, and pathway-footprints) to identify processes driving changes in gene expression. Importantly, the perturbation targets do not need to be known to produce the contextualized networks.

According to the benchmarking study, the introduction of TF weights from DoRothEA improves the performance of StdCARNIVAL (Suppl. Figure S2). For InvCARNIVAL, a performance was improved with the additional introduction of pathway weights inferred from PROGENy (Suppl. Figure S3), although the same advantage was not observed in StdCARNIVAL (results not shown). This shows that the pathway weights only guide the network search if a direction is not provided through a known perturbed target node.

The benchmarking results show that the InvCARNIVAL implementation with TF and pathway weights can obtain perturbation-attributed pathways with a comparable accuracy to StdCARNIVAL (Figure 2). Therefore, we recommend to include the TF and pathway implementations as the default setting. Comparing to GSEA, perturbation-attributed pathways were more frequently identified with the correct direction. Additionally, CausalR did not perform as well as InvCARNIVAL or GSEA and was biased towards hub nodes while CARNIVAL was not (Suppl. Figure S5).

In the application study in IgAN, WNT signaling was reported as a dysregulated process in CARNIVAL and is known to be involved in podocyte injury and renal fibrosis (Zhou and Liu, 2015). The IgAN network included representative mediators of the classical WNT signaling pathway from the messenger WNT7A to the TFs TCF4 and TCF7. However, it should be noted that not all of these are linked in the expected ways nor do all members show the expected activity.

TGF-β signaling is a main driver of fibrosis (Tang *et al.*, 2018). CARNIVAL's IgAN network captured all members of the TGF-β/RhoA pathway as up-regulated and linked through the biologically expected interactions. This includes the TGF-β receptors (TGFBR1 and TGFBR2), the ras homolog gene family member A (RHOA) and the Rho-associated protein kinase 2 (ROCK2). This is consistent with the previously reported up-regulation of protein levels of TGF-β receptors and RhoA in IgAN (Ebefors *et al.*, 2016; Mattii *et al.*, 2005) and illustrates how CARNIVAL can identify highly relevant and specific processes and regulators from gene expression data. Our validation experiment shows both β-catenin and RhoA are present in glomerular and tubular epithelial cells in IgAN, consistent with the involvement of these cells in the pathophysiology of the disease (Suppl. Figure S6; (Yeo *et al.*, 2018).

Overall, we demonstrate the superior performance of CARNIVAL over existing methods in the benchmarking study and also its applicability to biomedical data in our IgAN case study. However, it should be noted that the benchmarking is performed at the cellular process level due to the limited information on protein activities (see Methods). Moreover, the two-step inference approach also has a few limitations: 1) Not all attributed pathways are represented in MSigDB nor equally relevant, 2) Gene sets do not account for directionality and 3) Gene sets for the same process are inconsistent in

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

different databases. All of these factors could affect the benchmarking results, and further analyses should be performed to determine the generality of our findings.

Although we demonstrated that incorporating prior knowledge into the network inference can lead to a higher accuracy, the drawback is the inherent bias towards known biology. Even if CARNIVAL only uses the known interactions as a scaffold, the construction of the network is data-driven. Hence, it can the predict the status of proteins and their connections for specific contexts. Since CARNIVAL can not propose *de novo* connections between signaling molecules, it could be combined with pure data-driven network inference approaches such as nested effect model (NEM) (Markowetz *et al.*, 2007) in the future.

4 Conclusion

We present CARNIVAL as an open-source causal reasoning tool to infer upstream signaling networks from downstream gene expression. The network inference process is swiftly performed with an ILP formulation of causal reasoning. CARNIVAL can integrate prior information on TF and pathway activities from DoRothEA and PROGENy in a quantitative manner, demonstrating a better performance in our benchmark. As a case study, we applied CARNIVAL to IgA nephropathy and identified regulatory molecules that govern the disease which were experimentally validated. We believe that, given the flexibility of the CARNIVAL pipeline, it can be a useful tool to infer context-specific signaling network architectures from gene expression in many studies.

5 Methods

5.1 TF and pathway activity predictions with DoRothEA and PROGENy. DoRothEA version 2 (Garcia-Alonso *et al.*, 2018) provides a framework to estimate TF activity from the gene expression of its direct target genes. The provided regulon set, was filtered to include only the 289 TF-regulons with at least ten TF-target gene interactions with medium to high confidence (A, B, C). Subsequently, the differential gene expression t-values processed by the *limma* R-package (Ritchie *et al.*, 2015) and the filtered DoRothEA regulon were passed to the viper function in the *VIPER* package (Alvarez *et al.*, 2016) to perform an analytic Rank-based Enrichment Analysis (aREA). The activities of each transcription factor in the form of NES were then derived from the rank of the genes and the top 50 TF scores were used as the input in CARNIVAL.

The 14 PROGENy pathway signatures were obtained from https://github.com/saezlab/progeny ((Schubert *et al.*, 2018); Holland *et al.*, submitted) and applied to differential expression t-values from limma (Ritchie *et al.*, 2015). Based on an empirical null distribution generated through 10,000 times gene-wise permutation and the percentile corresponding to the observed value, the significance score (termed 'score' henceforth, Equation 1) was derived.

$$score(x) = 2 \cdot (percentile(x) - 0.5) \qquad \ldots \ (Eq\ 1)$$

5.2 ILP implementation and objective function. We re-implemented the causal reasoning ILP formulation of Melas *et al.* (Suppl. Text S2) in R. Here, we present the additional features on the objective function that we applied to CARNIVAL where the parameter α refers to the mismatch

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

penalty, β to the node penalty, and the newly introduced γ to the node penalty adjustment (Equation 2).

$$min\ (\sum|\alpha||x_j - m_j| + \sum\beta(1-\gamma_j)\cdot x_j^+ + \sum\beta(1+\gamma_j)\cdot x_j^-) \qquad \ldots\ (Eq\ 2)$$

In the previous work of Melas *et al.*, the objective function prioritizes the models which can explain the observed discretized measurements while minimizing the number of nodes in the network. In CARNIVAL, we introduce the effect of inferred TF and pathway activities to fine tune this tradeoff and model selection (Equation 2). For TF scores, we applied a TF-specific mismatch penalty α corresponding to the magnitude of the NES derived from DoRothEA. For pathway scores, a minimal set of representative downstream nodes was chosen for each PROGENy pathway to capture all known signal transduction routes involved while avoiding overlapping information between pathways, and with TF predictions (Suppl. Table S1). The node penalty is sign-adjusted through the γ weights which means that the anticipated direction is penalized less in the expected direction while more in the counterpart. This is implemented using the PROGENy significance scores (Equation 1), which range from -1 to 1.

Additional information regarding parameter settings and the ILP problem formulation for InvCARNIVAL can be found in Suppl. Text S3. Summarized results from the study of multiple α-to-β ratios to assess parameters' robustness can be found in Suppl. Text S4.

5.3 Benchmark dataset. The benchmark dataset was taken from the SBVimprover project which contains perturbations on normal human bronchial epithelial cells (Poussin *et al.*, 2014). Gene expression was measured at 6 hours after perturbation (E-MTAB-2091) in a processed form (log2 expression after GC robust multiarray averaging). Probe IDs were mapped to HGNC gene symbols and multiple entries were summarized by the mean value. Batch effects were removed using the combat function of the *sva* R-package (Leek *et al.*, 2012). Differential gene expression was then computed with the *limma* R package (Ritchie *et al.*, 2015).

The measurements with 19 phosphoprotein-binding antibodies were mapped to 14 differential protein activities using the curated regulatory sites in the PhosphositePlus knowledgebase (Hornbeck *et al.*, 2015). Given that only a small fraction of the PKN nodes is reported as dysregulated in CARNIVAL, the overlap between dysregulated nodes and measured protein activities was low and not suited for statistical testing.

5.4 Kidney datasets. Microarray data on glomerular gene expression in IgAN patients and healthy living donors (HLD) were obtained from 5 publicly accessible studies (Berthier *et al.*, 2012; Hodgin *et al.*, 2014; Liu *et al.*, 2017; Woroniecka *et al.*, 2011), see details in Suppl. Table S2. Study- and platform-dependent batch effects were mitigated using the combat function from the *sva* R-package (Leek *et al.*, 2012), and differential gene expression is determined with the limma R-package (Ritchie *et al.*, 2015).

5.5 CausalR. The CausalR package identifies dysregulated nodes and networks by scanning for nodes with sign-consistent shortest paths to the observations (Bradley and Barrett, 2017). With the SCAN (Sequential Causal Analysis of Networks) method, path lengths from one to five edges are scanned and potentially dysregulated nodes are identified which constantly score among the top 150 based on the number of explained observations. Matched observations increase the score (+1),

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mismatched ones decrease it (-1), and unmatched or ambiguously matched nodes are not included in the scoring.

5.6 Two-step inference approach to KEGG pathways attribution. In our study, we assume that the inferred node activities from CARNIVAL should represent upstream signaling and should hence map well with the attributed KEGG pathways. Hence, a two-step inference approach was developed to serve the validation propose (Suppl. Figure S1B). In the first step, up- and down-regulated pathways were predicted by an over-representation analysis with a hypergeometric test from the *Category* package in R on the deregulated nodes inferred by CARNIVAL. It is assumed that an over-representation of up-regulated CARNIVAL nodes indicates higher activity pathway and vice versa. The universe in this regard was set to all nodes present in the PKN and the curated KEGG pathway sets were obtained from MSigDB (Subramanian *et al.*, 2005). A significance test was only performed if at least one set member of the pathway was present in the given CARNIVAL node set.

In the second step, with the assumption that the pathways attributed to the corresponding perturbation are expected to be up-regulated upon perturbation, the inferred pathway activities were divided according to whether they are attributed to the respective perturbation target protein in the corresponding entry or not. In the case of good performance, it is expected that those pathways are up-regulated while others are not. To evaluate this hypothesis and control for unspecific results, up- and down-regulated pathways were enriched in the set of attributed pathways using the *piano* R-package (Väremo *et al.*, 2013). An enrichment analysis was applied to gene expression directly to identify a baseline performance.

6 Acknowledgements

We thank funding of the European Union's H2020 program (675585 Marie-Curie ITN ''SymBioSys'') and the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement no. 116030 (TransQST). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

We also thank Aurélien Dugourd for sharing the *runPROGENy* script and for the inputs for benchmarking using enrichment analysis, Bence Szalai, Luz Garcia-Alonso and Luis Tobalina for feedbacks during the early development of the CARNIVAL pipeline and Attila Gábor for the support on compiling the CARNIVAL package.

7 Authors contributions

AL performed benchmarking and IgAN studies. PT compiled CARNIVAL workflow as an R-package. EG implemented ILP formulation in R. AL, PT and EG designed the CARNIVAL pipeline, analysed results and wrote manuscript. AD and JB analysed IgAN results and performed the validation experiment. JSR conceived the project. PT and JSR supervised the project. All authors read and revised the manuscript.

8 References

Alvarez,M.J. *et al.* (2016) Functional characterization of somatic mutations in cancer using

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

network-based inference of protein activity. *Nat. Genet.*, **48**, 838–847.

Bellur,S.S. *et al.* (2017) Evidence from the Oxford Classification cohort supports the clinical value of subclassification of focal segmental glomerulosclerosis in IgA nephropathy. *Kidney Int.*, **91**, 235–243.

Berthier,C.C. *et al.* (2012) Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis. *J. Immunol.*, **189**, 988–1001.

Bild,A.H. *et al.* (2005) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Bradley,G. and Barrett,S.J. (2017) CausalR: extracting mechanistic sense from genome scale data. *Bioinformatics*, **33**, 3670–3672.

Chen,B. *et al.* (2014) Identifying protein complexes and functional modules--from static PPI networks to dynamic PPI networks. *Brief. Bioinform.*, **15**, 177–194.

Chen,E.Y. *et al.* (2012) Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers. *Bioinformatics*, **28**, 105–111.

Chindelevitch,L. *et al.* (2012) Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics*, **28**, 1114–1121.

Ebefors,K. *et al.* (2016) Mesangial cells from patients with IgA nephropathy have increased susceptibility to galactose-deficient IgA1. *BMC Nephrol.*, **17**, 40.

Garcia-Alonso,L. *et al.* (2018) Benchmark and integration of resources for the estimation of human transcription factor activities.

Giudice,G. and Petsalaki,E. (2017) Proteomics and phosphoproteomics in precision medicine: applications and challenges. *Brief. Bioinform.*

Hill,S.M. *et al.* (2016) Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods*, **13**, 310.

Hodgin,J.B. *et al.* (2014) The molecular phenotype of endocapillary proliferation: novel therapeutic targets for IgA nephropathy. *PLoS One*, **9**, e103413.

Hornbeck,P.V. *et al.* (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43**, D512–20.

Huang,S.-S.C. and Fraenkel,E. (2009) Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci. Signal.*, **2**, ra40.

Invergo,B.M. and Beltrao,P. (2018) Reconstructing phosphorylation signalling networks from quantitative phosphoproteomic data. *Essays Biochem.*, **62**, 525–534.

Kanehisa,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

Leek,J.T. *et al.* (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.

Liu,P. *et al.* (2017) Transcriptomic and Proteomic Profiling Provides Insight into Mesangial Cell Function in IgA Nephropathy. *J. Am. Soc. Nephrol.*, **28**, 2961–2972.

Markowetz,F. *et al.* (2007) Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, **23**, i305–12.

Mattii,L. *et al.* (2005) Kidney Expression of RhoA, TGF-β1, and Fibronectin in Human IgA Nephropathy. *Nephron Exp. Nephrol.*, **101**, e16–e23.

Melas,I.N. *et al.* (2015) Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. *Integr. Biol.* , **7**, 904–920.

Parikh,J.R. *et al.* (2010) Discovering causal signaling pathways through gene-expression patterns. *Nucleic Acids Res.*, **38**, W109–17.

bioRxiv preprint doi: https://doi.org/10.1101/541888; this version posted February 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Poussin,C. *et al.* (2014) The species translation challenge—A systems biology perspective on human and rat bronchial epithelial cells. *Scientific Data*, **1**.

Rayego-Mateos,S. *et al.* (2018) Connective tissue growth factor induces renal fibrosis via epidermal growth factor receptor activation. *J. Pathol.*, **244**, 227–241.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Rustad,K.C. *et al.* (2013) The role of focal adhesion complexes in fibroblast mechanotransduction during scar formation. *Differentiation*, **86**, 87–91.

Schubert,M. *et al.* (2018) Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.*, **9**, 20.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.

Tang,P.M.-K. *et al.* (2018) Transforming growth factor-β signalling in renal fibrosis: from Smads to non-coding RNAs. *J. Physiol.*, **596**, 3493–3503.

Türei,D. *et al.* (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods*, **13**, 966–967.

Väremo,L. *et al.* (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.*, **41**, 4378–4391.

Vogel,C. and Marcotte,E.M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.*, **13**, 227–232.

Woroniecka,K.I. *et al.* (2011) Transcriptome Analysis of Human Diabetic Kidney Disease. *Diabetes*, **60**, 2354–2369.

Wyatt,R.J. and Julian,B.A. (2013) IgA nephropathy. *N. Engl. J. Med.*, **368**, 2402–2414.

Yeo,S.C. *et al.* (2018) New insights into the pathogenesis of IgA nephropathy. *Pediatr. Nephrol.*, **33**, 763–777.

Zhou,L. and Liu,Y. (2015) Wnt/β-catenin signalling and podocyte dysfunction in proteinuric kidney disease. *Nat. Rev. Nephrol.*, **11**, 535–545.