D1.1

Final regulatory network inference

Project number:	668858
Project acronym:	**PrECISE**
Project title:	Personalized Engine for Cancer Integrative Study & Evaluation
Project Start Date:	1st January, 2016
Duration:	36 months
Programme:	H2020-PHC-2-2015
Deliverable Type:	Other
Reference Number:	PHC-668858-D1.1
Workpackage:	WP 1
Due Date:	30th June, 2017
Actual Submission Date:	10th November, 2017
Responsible Organisation:	IBM
Editor:	María Rodríguez Martínez
Dissemination Level:	PU
Revision:	2.0
Abstract:	Prostate specific gene regulatory networks were inferred by integration of RNASeq data measurements from TCGA PRAD and ProCOC cohorts using HIPSTER framework.
Keywords:	gene regulatory networks, data integration, network inference, disease-specific regulatory interactions

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 668858.

This work was supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0324-2. The opinions expressed and arguments employed therein do not necessarily reflect the official views of the Swiss Government.

Editor

María Rodríguez Martínez (IBM)

Contributors (ordered according to beneficiary numbers)

Roland Mathis (IBM)
Matteo Manica (IBM)
Ali Oskooei (IBM)
Sara Al-Sayed (TUDA)
Dominik Linzer (TUDA)
Heinz Koeppl (TUDA)
Hua-Sheng Chiu (BCM)
Pavel Sumazin (BCM)

Executive Summary

The goal of this deliverable was to build prostate cancer–specific Gene Regulatory Networks (GRNs) by integrating patient data from different cohorts. The resulting networks constitute graphs listing potential regulatory interactions related to prostate cancer onset and progression. The GRNs were built starting from RNASeq measurements obtained from a heterogeneous set of patient biopsies corresponding to different levels of disease aggressiveness. We applied a consensus strategy to integrate results obtained from different methods using HIPSTER (HIgh Performance SysTEms biology network Recostruction), the framework implemented in the context of D4.1. In order to gain biological insights and better understand the regulatory nature of the interactions, we enriched the consensus networks with information about known transcription factors and identified potential transcription factor candidates.

After a revision request the deliverable was updated integrating more detailed information about the data used, network inference methods considered and results from preliminary work conducted by BCM in T1.1.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent advances in high–throughput biological measurement and sequencing techniques have paved the way for large–scale data analysis and building computational networks that elucidate biological interactions such as gene–gene and protein–protein interactions (PPI), metabolic, signaling and transcription–regulatory networks [Barabasi and Oltvai, 2004]. Among these networks gene regulatory networks (GRNs) are particularly important as they describe how cells regulate expression of genes which in turn control production of proteins that regulate cell function. GRNs can provide us with cues about pathways that affect abnormalities and diseases such as cancer. These pathways can lead to critical information about disease progression and be used for drug target discovery and therapeutic interventions. As such, developing computational methods that can reconstruct gene regulatory networks can provide a wealth of information and tools for cancer diagnosis and treatment [Omranian et al., 2016, Kelemen et al., 2008].

In computational models, gene regulatory networks are often presented as nodes connected with edges, where nodes represent individual genes and the edges provide information about the intensity of the regulatory interaction or even the direction of the regulatory effects. Various network types are used to model GRNs such as: boolean networks, relevance networks, Bayesian networks and differential equation models [Kelemen et al., 2008, Chai et al., 2014]. Boolean networks assume each gene is either “on” or “off” and as such are simple and offer only a qualitative representation of the system. Relevance networks are built based on pairwise distances or similarities between gene pairs. Various measures of similarity or dissimilarity may be used to determine whether a strong enough interaction exists between each pair of genes. Pairwise similarities are often filtered based on a threshold to determine whether an edge between the two nodes should exist. One such network was built by Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe) [Margolin et al., 2006] which determines interaction between genes based on mutual information between gene pairs. Relevance networks are simple to build and unlike Boolean networks provide quantitative measure of gene co–expression. Simplicity renders them suitable for building large networks. Relevance networks, however, cannot explain the dynamic behavior of the network as they do not take into account time variation of parameters. Bayesian networks are built from determining conditional probability distributions of genes in the network. They are suitable for GRN modeling as they take into account the stochastic nature of gene expression naturally. They are however unsuitable for building large networks due to the amount of computation required to determine all conditional probabilities. Bayesian networks are also

static and do not take into account the dynamics of a network. Differential equation networks can be used to obtain a deterministic quantitative and dynamic model of the system. Differential equation inference of networks enables inclusion of chemical reaction kinetics in the model [Kelemen et al., 2008], but at the price of estimating a large amount of parameters and the need of time series data.

Wisdom of crowds, namely network inference based on a community of inference methods rather than individual methods, has been shown to outperform individual inference techniques [Marbach et al., 2012]. In line with the wisdom of crowds paradigm we have developed the HIPSTER (HIgh Performance SysTEms biology network Reconstruction) framework in the context of deliverable D4.1 to estimate PPI networks. HIPSTER constructs a relevance–based regulatory network by means of consensus among a community of previously–established and reputable network inference methods namely: Pearson correlation, Spearman correlation, ARACNe [Margolin et al., 2006], Glasso [Friedman et al., 2008], JRF [Petralia et al., 2016] and FunChisq [Zhang and Song, 2013]. In this work we used HIPSTER to infer networks from transciptomic data where in comparison to the estimation of PPI a much larger number of genes was used. A Detailed description of HIPSTER can be found in Section 3.1.

In the current deliverable, after presenting preliminary results obtained in initial phases of T1.1 (Chapter 2), RNASeq datasets from two different cohorts of prostate cancer patients are analyzed using our previously–devised HIPSTER method, and the corresponding GRNs for each dataset are constructed and compared. Furthermore, using each dataset, prostate cancer–specific pathway GRNs are inferred based on select genes extracted from reported pathways in the literature. The resulting GRNs, obtained using HIPSTER, are described as undirected graphs. To overcome the lack of directionality the networks are contrasted and compared by means of quantitative metrics and studied using prior knowledge of transcription factors from literature focusing on centrality measures, where potential regulatory candidates are selected by examining most central nodes. Details of the aforementioned analyses are presented in Chapters 3 and 4.

Chapter 2

Preliminary results

The work presented in the current deliverable is based on preliminary work developed in T1.1 by BCM. BCM performed an extensive application of different network inference methods to build a prostate–specifc GRN. The data used to build the prostate regulatory networks are two prostrate adenocarcinoma cohorts from TCGA (The Cancer Genome Atlas) and MSKCC (Memorial Sloan Kettering Cancer Center), here reported in Table 2.1.

	TCGA-PRAD	**MSKCC-PRAD**
data-type	RNA profiles	RNA profiles
number of samples	550	218

Table 2.1: Data used for preliminary results in T1.1 and D1.1

Only datasets with over 100 samples accommodate for the prerequisites of some of the analysis methods. Moreover, there was little commonality between predicted targets across datasets, suggesting that the addition of smaller datasets will introduce more noise than benefit.

Name	**Based on**	**Sparsity**
ARACNe	Mutual information	DPI
MARINa	Sequence analysis and delta mutual information	DPI
Hermes	Conditional mutual information	Permutation testing
DME	Sequence analysis	Log–likelihood
OmniMiner	Integration of ARACNe and DME	DPI and permutation testing
LongHorn	Sequence analysis and distance correlation	Permutation testing
Cupid	Conditional mutual information	Brown's method

Table 2.2: Inference methods used in preliminary analysis to reverse engineer prostate-cancer specific regulatory networks.

BCM used multiple methods to reverse engineer prostate–cancer specific regulatory networks (all methods used are reported in Table 2.2). BCM designed Cupid [Chiu et al.,] and LongHorn. Specifically LongHorn was used to improve both transcriptional and post-transcriptional regulatory interactions (more details can be found in the appendix attached to the deliverable *appendix_longhorn.pdf*). All the methodologies generate interaction lists for each analyzed cohort. Not all of the methods were directly applicable in the consequent analysis performed in D1.1. For that reason sequence–analysis–based methods were excluded from the following analysis. Moreover, the combination of dataset analyses with these methods produced spurious results, with little commonality between predicted targets across datasets. Consequently, from the original list of proposed methods contained in T1.1 description, only ARACNe was considered for D1.1 through its integration in HIPSTER (details in Chapter 3). Networks generated in this preliminary analysis phase are available at the following link `https://bcm.box.com/s/evljbwn7voaf25b04lyz42egbw3dp1av`.

Chapter 3

Network inference from RNASeq data

Deciphering regulatory interactions from molecular data plays a key role in understanding cellular processes in complex diseases such as cancer. Thourough analysis of gene expression data, such as RNASeq, is crucial in studying regulatory processes that occur during disease progression since it offers a direct measure of the transcription efficiency for each gene. As previously mentioned, in this work network inference methods are applied to transcripts expression data in prostate cancer. Building GRNs using RNASeq data extracted from prostate cancer patients' tumor samples may reveal novel unknown relationships or regulatory effects that are relevant or critical in cancer growth and development.

3.1 HIPSTER Framework

To infer regulatory interactions from RNASeq data extracted from prostate cancer patient biopsies HIPSTER (HIgh Performance SysTEms biology network Reconstruction) was used. HIPSTER is an inference framework developed in the context of deliverable D4.1. A schematic representation of the framework is given in Figure 3.1.

The HIPSTER framework builds an interaction network using a combination of different data sources:

- Unstructured text from publications or text documents in general using a novel totally unsupervised method developed in D4.1 called INtERAcT
- Public databases using OmniPath[1] [Trei et al., 2016] (a database including interactions from *CancerCellMap, SPIKE, LMPID, DIP,HPRD, PDZBase, dbPTM, Signor, Macrophage, ELM, SignaLink3, NRF2ome, DEPOD, BioGRID, phosphoELM, MPPI, IntAct, PhosphoSite, HPRD-phos, CA1, DeathDomain, ARN*)
- Molecular datasets using a methods consensus approach inspired by DREAM5 [Marbach et al., 2012] (DREAM challenge on network inference for gene regulatory networks). HIPSTER generates an estimate for the network topology from the datasets,

[1] `http://omnipathdb.org/`

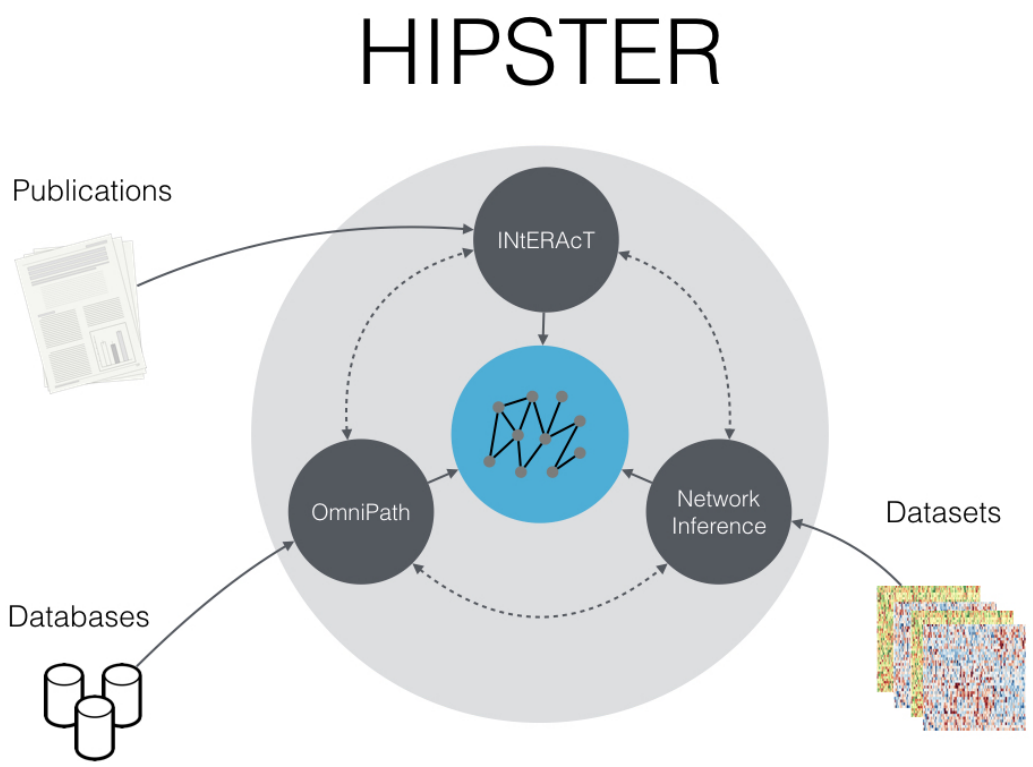

Figure 3.1: Schematic representation of HIPSTER framework - HIPSTER combines different data sources in a robust interaction graph that exploits knowledge at different levels. Specifically it uses data from publications (unstructured text in general), public databases and a consensus approach based on different network inference methods.

and by combining results from single methods, constructing a unique weighted undirected graph with interaction intensities (edge weights) ranging in $[0, 1]$. A list of the considered methods is reported in Table 3.1.

Name	Based on	Sparsity
Pearson	Correlation	Multiple tests correction
Spearman	Ranks correlation	Multiple tests correction
Glasso	Partial correlations	L1 penalization
Aracne	Mutual information	DPI
FunChisq	Functional dependencies	Multiple tests correction

Table 3.1: HIPSTER inference methods

In this deliverable only the consensus inference from the data was used to build a prostate–specific GRN. This was done due to the specific focus of the current deliverable, which is to reverse engineer the regulatory networks from molecular data.

3.2 GRN inference using HIPSTER

HIPSTER was used to infer potential interactions from RNASeq data of two prostate cancer cohorts:

- TCGA PRAD: a well characterized cohort extensively studied from TCGA consortium [Abeshouse et al., 2015]. RNASeq data from biopsies of 498 patients at different stages of the disease were used.
- ProCOC [Umbehr et al., 2008]: a cohort from University Hospital of Zürich. 105 samples at different stages of the disease from 39 patients.

HIPSTER was applied on both cohorts to infer potential regulatory interactions in two different ways. First a high–throughput approach examining all quantified transcripts (i.e., genes) for each cohort was used. The reconstruction of the two cohort–specific transcriptome–wide GRN was the basis for analyzing the estimated topological properties of the estimated networks (see Section 4.1).

Second, in a pathway–specific approach, the focus of analysis was on pathways relevant in cancer. Annotated gene sets from MSigDB (Molecular Signatures Database) [Liberzon et al., 2015] were used to infer pathway–specific GRNs. Pathway networks with similar connectivity patterns in both cohorts were highlighted, enabling a more detailed study of potential regulatory interactions involved in prostate cancer progression (see Section 4.2)

Chapter 4

Results

This chapter encompasses the results and discussion for application of HIPSTER consensus inference to datasets from TCGA PRAD and UZH ProCOC cohorts (reported in Table 4.1). The ProCOC cohort RNASeq data that were used in this work are still unpublished. They were obtained by running RNASeq experiments from multi–area biopsies from 39 patients. A total of 105 samples with different tumor grades were considered. The main reasons to focus on RNASeq and these two specific cohorts were the following:

- RNASeq measurements are suitable to infer gene regulatory networks because they allow to quantify gene expression values and they are the de–facto standard for genome-wide expression analysis.
- The study was limited to these two cohorts because they were extremely comparable. The measurement and processing pipeline adopted in UZH was tailored to obtain results comparable with TCGA PRAD.

First an analysis of the networks constructed from the transcriptomes of the two cohorts is shown (see Section 4.1). The main focus of the section is to compare the two topologies and to analyze common graph properties. In the second part, the emphasis is on a more detailed analysis of pathways relevant to cancer (see Section 4.2). Using the annotated hallmark gene sets ([Liberzon et al., 2015]) it was possible to quantify known transcription factors activity and identify novel potential regulatory elements.

	TCGA-PRAD	**UZH ProCOC**
data-type	RNASeq	RNASeq
number of samples	550	105

Table 4.1: Data considered for prostate-specific GRN reconstruction.

4.1 Complete Transcriptome Inference Results

HIPSTER was used to build prostate–specific GRNs for all the transcripts quantified using RNASeq in the two cohorts studied.

The number of entities/nodes analyzed by HIPSTER for each cohort was significantly higher than previous applications of HIPSTER (in D4.1 interactions were estimated for Protein-Protein Interaction networks consisting of $\sim$3000 nodes). In the current report the baseline for the network inference amounts to 14000 genes in each cohort. In this regime only a subset of the available methods were suitable for performing a robust network estimation with reasonable computational cost. The methods used for network inference (i.e.,estimation) were two correlation–based techniques: Pearson correlations and Spearman correlations (where the significant interactions were selected using 0kBenjamini–Hochberg correction for multiple tests at confidence level p=0.05); and a mutual information–based method: ARACNE [Margolin et al., 2006]. The consensus GRN for each cohort was estimated combining interactions scores from single methods using the hard mean of the scaled ranked interactions .

The hard mean is defined as the sum of the scaled ranked interaction scores from single methods divided by the number of inference methods considered, regardless of the presence of a predicted interaction for each of the single methods (adopted because it outperformed other approaches for PPI estimation in D4.1, compared by looking at the Receiver Operating Characteristic curves on synthetic data generated using GeneNetWeaver [Schaffter et al., 2011]).

In Figure 4.1 the results obtained for the transcriptome reconstruction using molecular data are reported. As demonstrated in the figure, a similar trend can be observed for distributions of edge intensities (interaction scores) for two studied cohorts (see Figure 4.1a and Figure 4.1b). Both estimated networks exhibit a sensible drop in the intensities close to the threshold value t used to filter out low confidence interactions. The threshold value was determined such that it would preserve interactions that were predicted with high confidence by a single method but received low scores from other inference methods (t=0.33), when applying the consensus approach.

To study topological properties of the estimated networks, an analysis of the distribution of node degrees and local clustering coefficient (also known as local transitivity) of the thresholded networks was performed [Wasserman and Faust, 1994].

The degree distribution analysis was performed to determine whether the networks exhibit a scale–free behaviour (as commonly observed in biological networks [Barabasi and Oltvai, 2004]). To check this property a power law fit was used (using package `powerlaw` [Alstott et al., 2014]). Considering a power law for degree (k) distribution:

$$P(k) \sim k^{-\alpha} \tag{4.1}$$

where $2 \leq \alpha \leq 3$ is commonly observed for scale–free networks [Barabasi and Oltvai, 2004]. In Figure 4.1c and Figure 4.1d the degree distributions for the two estimated networks are shown. The power law distribution fitted in both cases doesn't fall in the range expected for scale–free topologies (TCGA PRAD: α=4.47 and ProCOC: α=4.66) but is comparable between the two estimates, exhibiting consistency in inferred network structure across the two studied cohorts.

In addition, an analysis of the local clustering coefficient (also known as local transitivity) was performed to determine whether the two networks demonstrate a hierarchical property.

Since in the current work, weighted undirected networks were considered, the local clustering coefficient was calculated using the method described in [Barrat et al., 2004]. As stated in [Barabasi and Oltvai, 2004], strongly hierarchical networks shows a reciprocal dependency between the clustering coefficient (C) and the degree (k):

$$C(k) \sim k^{-1} \tag{4.2}$$

In Figure 4.1e and Figure 4.1f the log–log scale changes in the local clustering coefficients for both graphs are presented. Neither of the plots demonstrate any indication of a hierarchical network structure. The increase in the clustering coefficient, especially in the network estimated from TCGA data, suggests a network structure that packs highly and strongly connected nodes together, while leaving out loosely connected ones. This behavior leads to the rejection of the hypothesis of a hierarchical structure underlying both estimated networks. While this is not in line with what is commonly observed for biological networks [Barabasi and Oltvai, 2004], it still suggests the existence of highly connected gene sets that can be interpreted as potential regulatory modules.

4.2 Hallmark Gene Sets GRNs analysis

In this section a detailed study of GRNs reconstructed for cancer–relevant pathways annotated in MSigDB was conducted (see Table 4.2 for a list of the studied pathways). HIPSTER was applied to both cohorts focusing on specific subsets of genes reported in MSigDB. In this configuration (pathways containing a number of genes ranging from $\sim$ 30 to $\sim$ 200 genes) all inference methods (see Section 4.2.1) implemented in the framework were able to run and produce an estimate.

4.2.1 Identifying Consistent Pathway GRNs across Cohorts

To understand and identify disease-specific pathways, for each cohort and each pathway (i.e., subset of genes in each pathway) the interactions were inferred using the HIPSTER consensus approach. HIPSTER was able to produce estimates with all the six methods currently implemented in the framework (Pearson correlation, Spearman correlation, ARACNE [Margolin et al., 2006], Glasso [Friedman et al., 2008], JRF [Petralia et al., 2016] and FunChisq [Zhang and Song, 2013]). The consensus approach was applied using the hard mean of the scaled ranks, as described in Section 4.1, using a threshold t=0.16 to preserve high confidence interaction from single methods.

To quantify the similarity of each pathway inferred from two different cohorts, the weighted adjacency similarity between the two inferred pathway networks was computed. The adjacency similarity is the sum of equal entries in the adjacency matrix, given a vertex ordering determined by the vertex labels. It counts the number of edges which have the same source and target labels in both graphs. For undirected weighted graphs, it is defined as:

$$S(\boldsymbol{A}_1, \boldsymbol{A}_2) = E - d(\boldsymbol{A}_1, \boldsymbol{A}_2) \tag{4.3}$$

where:

$$d(\boldsymbol{A}_1, \boldsymbol{A}_2) = \sum_{i<j} |A_{ij}^{(1)} - A_{ij}^{(2)}| \tag{4.4}$$

is the distance between graphs, $\boldsymbol{A}_k$ with $k \in \{1, 2\}$ are the weighted adjacency matrices of the graphs considered, and $E = \sum_{i<j} |A_{ij}^{(1)}| + |A_{ij}^{(2)}|$. The weights were normalized using $S(\boldsymbol{A}_1, \boldsymbol{A}_2)/E$.

Computing for each of the hallmark pathways the distances between the GRNs estimated from TCGA PRAD and ProCOC RNASeq measurements identified stable topologies across the two cohorts. In this way it was possible to find pathways and related GRNs that indicated a disease–specific component independent of the cohorts considered, see Figure 4.2.

From this analysis two pathway GRNs emerged as consistently stable across the two studied cohorts, both closely related to prostate cancer: genes related to TFNα Signaling Via NFKβ ([Srinivasan et al., 2010, Lessard et al., 2003]) and to Epithelial–Mesenchymal Transition ([Grant and Kyprianou, 2013, Imran Khan et al., 2015]).

Table 4.2: Hallmark gene sets from MSigDB. Relevant pathways for cancer reported in MSigDB are ranked according to the weighted adjacency similarity computed between GRN estimates obtained using the HIPSTER consensus approach from TCGA PRAD and ProCOC RNASeq data.

Pathway	Ranked Similarity
HALLMARK_TNFA_SIGNALING_VIA_NFKB	1
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	2
HALLMARK_MYC_TARGETS_V2	3
HALLMARK_INFLAMMATORY_RESPONSE	4
HALLMARK_INTERFERON_ALPHA_RESPONSE	5
HALLMARK_IL6_JAK_STAT3_SIGNALING	6
HALLMARK_MYOGENESIS	7
HALLMARK_ALLOGRAFT_REJECTION	8
HALLMARK_INTERFERON_GAMMA_RESPONSE	9
HALLMARK_KRAS_SIGNALING_UP	10
HALLMARK_TGF_BETA_SIGNALING	11
HALLMARK_ANDROGEN_RESPONSE	12
HALLMARK_MYC_TARGETS_V1	13
HALLMARK_UV_RESPONSE_DN	14
HALLMARK_IL2_STAT5_SIGNALING	15
HALLMARK_APICAL_SURFACE	16
HALLMARK_ESTROGEN_RESPONSE_LATE	17
HALLMARK_ESTROGEN_RESPONSE_EARLY	18
HALLMARK_APICAL_JUNCTION	19
HALLMARK_CHOLESTEROL_HOMEOSTASIS	20
HALLMARK_P53_PATHWAY	21
HALLMARK_ANGIOGENESIS	22
HALLMARK_COMPLEMENT	23
HALLMARK_APOPTOSIS	24
HALLMARK_HYPOXIA	25
HALLMARK_HEDGEHOG_SIGNALING	26
HALLMARK_PI3K_AKT_MTOR_SIGNALING	27
HALLMARK_G2M_CHECKPOINT	28
HALLMARK_MITOTIC_SPINDLE	29
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	30
HALLMARK_WNT_BETA_CATENIN_SIGNALING	31
HALLMARK_E2F_TARGETS	32
HALLMARK_NOTCH_SIGNALING	33
HALLMARK_XENOBIOTIC_METABOLISM	34
HALLMARK_PEROXISOME	35
HALLMARK_ADIPOGENESIS	36
HALLMARK_MTORC1_SIGNALING	37
HALLMARK_FATTY_ACID_METABOLISM	38
HALLMARK_UV_RESPONSE_UP	39
HALLMARK_COAGULATION	40
HALLMARK_GLYCOLYSIS	41
HALLMARK_OXIDATIVE_PHOSPHORYLATION	42
HALLMARK_PROTEIN_SECRETION	43
HALLMARK_DNA_REPAIR	44

Continued on next page

HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	45
HALLMARK_BILE_ACID_METABOLISM	46
HALLMARK_HEME_METABOLISM	47
HALLMARK_PANCREAS_BETA_CELLS	48
HALLMARK_SPERMATOGENESIS	49
HALLMARK_KRAS_SIGNALING_DN	50

4.2.2 Detailed Analysis of Consistent Pathways GRNs

A detailed analysis of the pathways that showed the greatest stability was performed. For each pathway network, the estimates built using TCGA PRAD and ProCOC cohorts were combined using HIPSTER consensus approach applying the hard mean of the scaled rank (see Section 4.1). To selectively preserve interactions that were estimated with high confidence in both cohorts (i.e., the most stable interactions), a threshold t=0.9 was applied to prune low intensity connections.

Since the main goal of the current work consisted of estimation of regulatory interactions, the emphasis was on identifying potential prostate cancer–regulators. To find such candidates a betweenness centrality metric [Freeman, 1977] for all the genes contained in the networks was computed. Most central genes, with respect to this metric, should represent the main actors in regulatory events given their high–intensity interactions and strong connectivity.

Betweenness centrality is defined as:

$$C_B(v) = \sum_{s \neq v \neq t \in V,\, s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{4.5}$$

where σ_{st} is the number of shortest geodesic paths between s and t, and $\sigma_{st}(v)$ is the number of shortest geodesic paths between s and t passing through a v.

Transcripton Factors (TFs) from two different sources were identified and collected to enrich the analysis. An exhaustive list of known TFs was obtained from TFcheckpoint [1] [Chawla et al., 2013], while a list of prostate–specific TFs was extracted from two FANTOM5 [2] [Lizio et al., 2015] tissue–specific gene regulatory networks constructed using adult prostate tissues and prostate cancer cell lines.

In Figures 4.3 and 4.4 the results for the two stable pathways identified in Section 4.2.1 are reported, for TFNα Signaling Via NFKβ and Epithelial–Mesenchymal Transition respectively.

Figures 4.3a and 4.4a report centrality measures for the top 20 most central genes. In both networks it was possible to identify a set of genes that showed a higher centrality measure and that were not reported in the list of the TFs from TFcheckpoint and FANTOM5. In TFNα Signaling Via NFKβ , three genes were identified: DUSP5, PLAUR and SOCS3. While in Epithelial–Mesenchymal Transition two genes emerged as being more central: COL6A3 and

[1] http://www.tfcheckpoint.org/
[2] http://fantom.gsc.riken.jp/5/

GEM. Both sets were investigated in more details in the following paragraphs to understand if the identified genes represent valid prostate cancer–specific regulators candidates.

In Figures 4.3b and 4.4b the high confidence interaction GRNs estimated using HIPSTER are presented in a graph, where it is evident how identified genes exhibit a central role in the network structure.

TFNα Signaling Via NFKβ

Activation of NFKβ signaling has been strongly associated with the progression of prostate cancer; however, the precise underlying mechanisms are not fully understood [Jin et al., 2014, Zhang et al., 2009, Shukla et al., 2004]. NFKβ expression in prostate cancer cells significantly increased AR mRNA and protein levels and AR transactivation activity [Zhang et al., 2009].

Among known regulators in TFNα Signaling Via NFKβ such as CEBPB and ETS2 [Lizio et al., 2015] (see Figure 4.3 yellow bars) other genes were found to be central in the inferred network: DUSP5 (Dual Specificity Phosphatase 5) which has been shown to negatively regulate MAP kinases, which are associated with cellular proliferation [Cai et al., 2015], PLAUR (Plasminogen Activator, Urokinase Receptor) for which elevated mRNA levels have been reported in analysis of androgen independent prostate cancer [Creighton, 2007] and SOCS3 (Suppressor Of Cytokine Signaling 3), which is known to negatively regulate cytokine signaling and is expressed in human prostate cancer [Krebs and Hilton, 2001, Bellezza et al., 2006].

NFKβ and TNF signaling have shown to be of significant importance for prognostic, patient stratification and potentielly therapeutics [Srinivasan et al., 2010, Lessard et al., 2003]

Epithelial–Mesenchymal Transition

While Epithelial-mesenchymal transition (EMT) plays an essential role in regeneration of tissue and development, its activation has also been associated with cancer progression. Especially in the context of prostate cancer EMT activation has been identified to drive the progression to castrate-resistant tumor types [Grant and Kyprianou, 2013, Imran Khan et al., 2015].

Through a molecular mechanism strongly related to EMT, cancer cells can invade neighboring tissues. In our inferred network COL6A3 plays a dominant role. Among other extracellular matrix molecules, collagen has been reported as a candidate that may initiate signals that promote EMT [Shintani et al., 2008]. Collagen type VI α3 (COL6A3) encodes a protein of the extracellular matrix [Zanussi et al., 1992]. In a study using RT-PCR of 18 prostate cancer patients, alternative splicing variants have been detected for nearly half of the metastatic prostate cancer samples compared to normal tissue and localized prostate cancer [Thorsen et al., 2008]. A better understanding of the detailed molecular mechanisms of EMT may help identify new targets for prevention of metastasis [Shintani et al., 2008].

It has been suggested that GTP Binding Protein Overexpressed In Sceletal Muscle (GEM)

could play a role as a regulat gory protein in receptor-mediated signal transduction [Maguire et al., 1994]. In recent work it has been shown that intracellular signaling mechanisms triggered by extracellular hormonal factors acting through (G protein)-coupled receptors can mediate and sustain androgen-independent prostate cancer cell proliferation [Daaka, 2004]. While the signalling pathways are still to be resolved the G protein-dependent activation of the Ras-to-mitogen-activated protein kinase pathway has emerged as a critical regulatory event.

The consensus GRNs inferred for each pathway are attached to the deliverable in an archive (*hipster_hallmark_consensus.zip*).

4.3 Final Remarks

Application of HIPSTER consensus approach to network inference from transcriptomic data proved to be an effective way to identify genes with central and potentially regulatory roles without using any prior knowledge on known transcription factors–targets associations. Using multiple cohorts and known cancer–specific gene sets to identify prostate–specific pathway GRNs showed promising results that may be further improved by incorporating additional cohorts and employing data–driven identification of relevant gene sets.

The analysis presented in this work was focused on generating novel hypotheses on prostate cancer–specific regulation processes and discovery of novel regulatory interactions, for this reason no prior knowledge on existing interactions using HIPSTER framework was added in favor of a fully data–driven approach.

GRN analysis using HIPSTER led to the discovery of candidate regulators that demonstrate relevance in prostate cancer progression. The results reported in this work give rise to new hypotheses for further validation in the context of prostate cancer research.

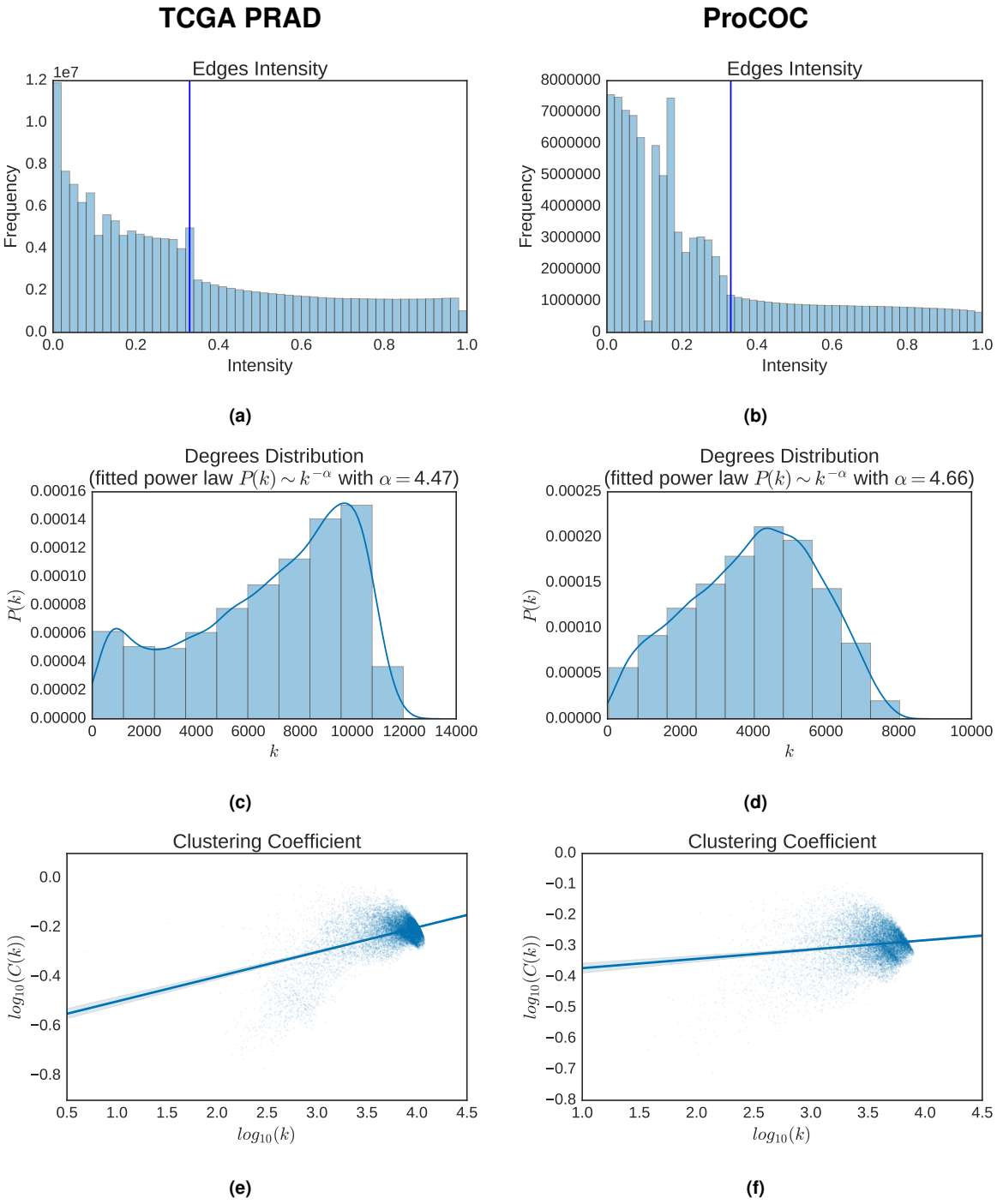

Figure 4.1: HIPSTER inferred GRNs for TCGA PRAD (left) and ProCOC (right). Analysis of GRNs obtained applying the HIPSTER consensus approach on the considered cohorts. Results for both TCGA PRAD and ProCOC cohorts are reported respectively in left and right panels. The histograms in Panel a and b represent distributions of the intensities for the edges predicted using HIPSTER. The blue vertical line corresponds to the threshold imposed on the intensities for the analysis in the following panels. A value for the threshold t=0.33 ensured that we report only interactions that were predicted by at least one method with high confidence. In Panel c and d histograms with kernel density estimations of the degree distributions are reported, together with a fitting of the power law to check connectivity properties. In Panel e and f a scatter plot in log–log scale of the local clustering coefficient versus the degree is shown, the points are in both cases fitted with a robust linear regression with outlier de–weighting and confidence intervals at level 0.95 estimated using bootstrap.

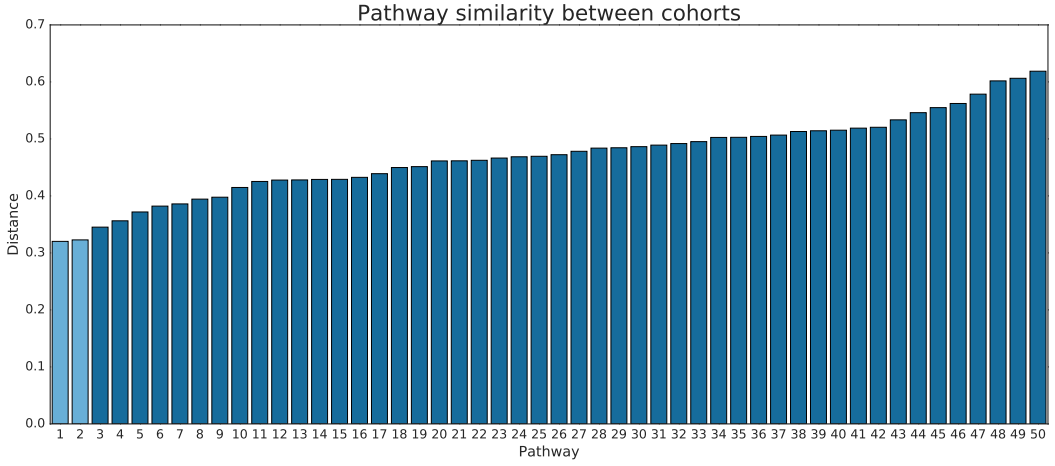

Figure 4.2: Similarity analysis of GRNs estimated with HIPSTER for the hallmark gene sets between the considered cohorts. Pathways with high similarity between cohorts are expected to contain a higher degree of prostate cancer–specific information compared to pathways with low similarity where the cohort effects are influencing the network. The most similar cancer hallmark pathways across cohorts are related to TFNα Signaling Via NFKβ (HALLMARK_TNFA_SIGNALING_VIA_NFKB) and Epithelial–Mesenchymal Transition (HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION). They are highlighted in light blue. A complete list of the pathways can be found in Table 4.2.

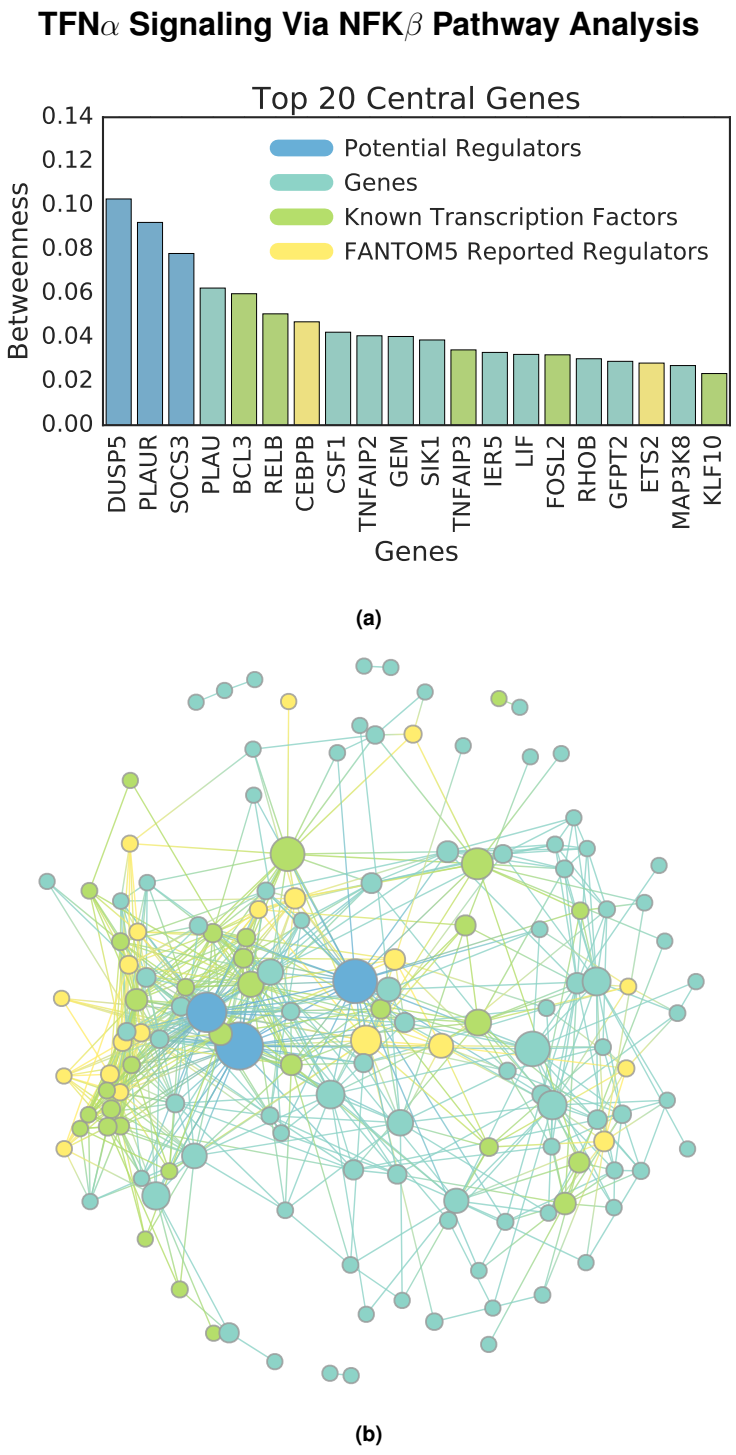

Figure 4.3: HIPSTER inferred consensus GRN for TFNα Signaling Via NFKβ gene set. This figure describes the analysis of high confidence regulatory interactions (pruning edges using a threshold t=0.9) for the most stable hallmark set, TFNα Signaling Via NFKβ . The GRN was obtained using the consensus network estimated after merging the results from both TCGA PRAD and ProCOC cohorts. In Panel a the first top 20 central genes, sorted using betweenness measure, are reported. The legend shows the colors associated with the different genes based on their source. In Panel b a graph reporting all the high confidence interactions is shown. Edge width is a function of the intensity and node size depends on their betweenness. The colors used for the nodes and edges follow the legend in Panel a.

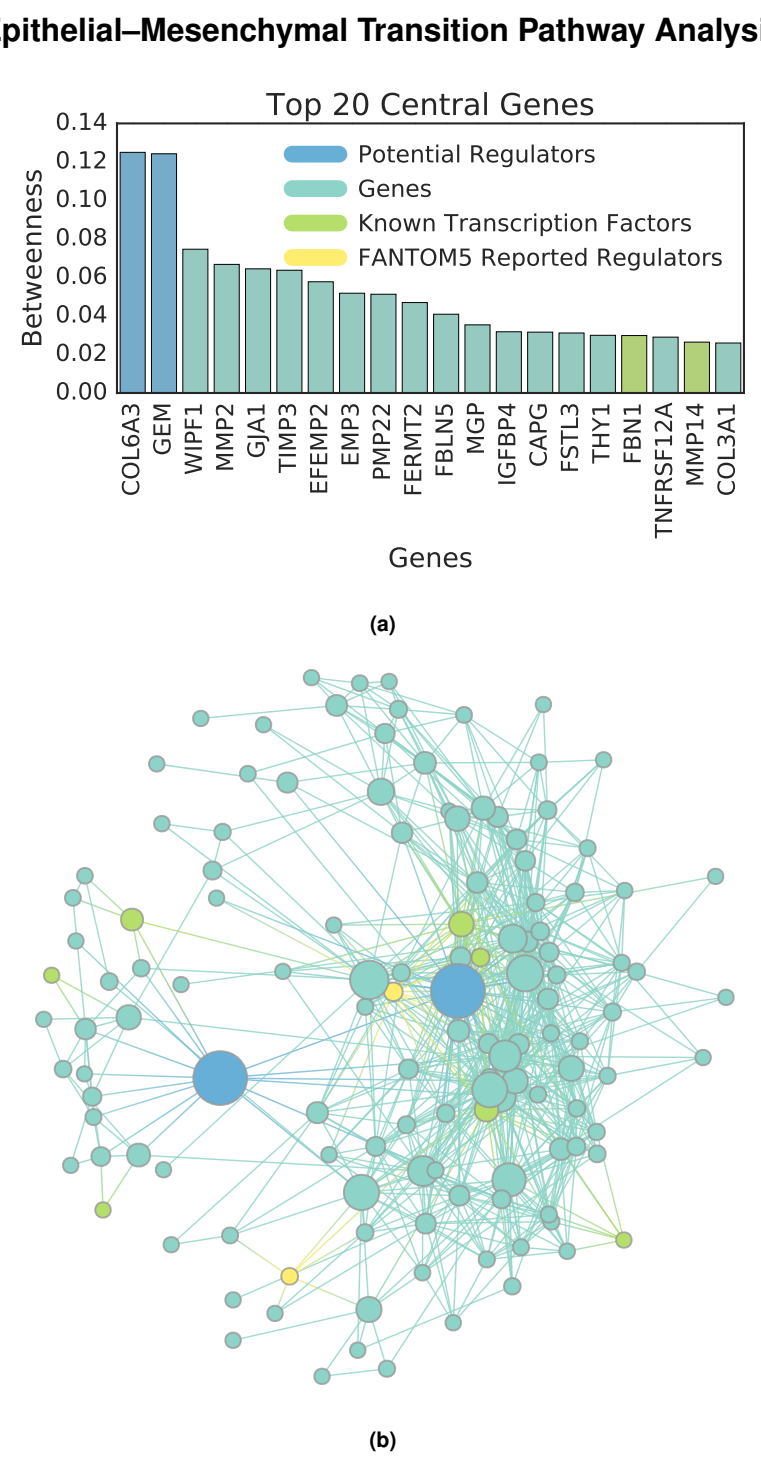

Figure 4.4: HIPSTER inferred consensus GRN for Epithelial–Mesenchymal Transition gene set. This figure describes the analysis of high confidence regulatory interactions (pruning edges using a threshold t=0.9) for the second most stable hallmark set, Epithelial–Mesenchymal Transition . The GRN was obtained using the consensus network estimated after merging the results from both TCGA PRAD and ProCOC cohorts. In Panel a the first top 20 central genes, sorted using betweenness measure, are reported. The legend shows the colors associated with different genes based on their source. No transcription factors reported in prostate–specific regulatory networks from FANTOM5 were detected among the most central nodes. This suggests that disease–induced deregulation is not properly captured by graphs generated from prostate healthy tissues and the cancer cell lines used to build the two FANTOM5 networks considered. In Panel b a graph reporting all the high confidence interactions is shown. Edge width is a function of the intensity and node size depends on their betweenness. The colors used for the nodes and edges follow the legend in Panel a.

Chapter 5

List of Abbreviations

GRN	Gene Regulatory Network
HIPSTER	HIgh Performance SysTEms biology network Reconstruction
TCGA	The Cancer Genome Atlas
MSigDB	Molecular Signatures Database
PPI	Protein–Protein Interaction
TF	Transcripton Factors

Bibliography

[Abeshouse et al., 2015] Abeshouse, A., Ahn, J., Akbani, R., et al. (2015). The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, 163(4):1011–1025.

[Alstott et al., 2014] Alstott, J., Bullmore, E., and Plenz, D. (2014). powerlaw: a python package for analysis of heavy-tailed distributions. *PloS One*, 9(1).

[Barabasi and Oltvai, 2004] Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

[Barrat et al., 2004] Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752.

[Bellezza et al., 2006] Bellezza, I., Neuwirt, H., Nemes, C., et al. (2006). Suppressor of cytokine signaling-3 antagonizes camp effects on proliferation and apoptosis and is expressed in human prostate cancer. *The American Journal of Pathology*, 169(6):2199–2208.

[Cai et al., 2015] Cai, C., Chen, J.-Y., Han, Z.-D., et al. (2015). Down-regulation of dual-specificity phosphatase 5 predicts poor prognosis of patients with prostate cancer. *International Journal of Clinical and Experimental Medicine*, 8(3):4186.

[Chai et al., 2014] Chai, L. E., Loh, S. K., Low, S. T., et al. (2014). A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, 48:55–65.

[Chawla et al., 2013] Chawla, K., Tripathi, S., Thommesen, L., Lægreid, A., and Kuiper, M. (2013). Tfcheckpoint: a curated compendium of specific dna-binding rna polymerase ii transcription factors. *Bioinformatics*.

[Chiu et al.,] Chiu, H.-S., Llobet-Navas, D., Yang, X., et al. Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. 25(2):257–267.

[Creighton, 2007] Creighton, C. J. (2007). A gene transcription signature associated with hormone independence in a subset of both breast and prostate cancers. *BMC Genomics*, 8(1):199.

[Daaka, 2004] Daaka, Y. (2004). G proteins in cancer: the prostate cancer paradigm. *Sci. STKE*, 2004(216).

[Freeman, 1977] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

[Friedman et al., 2008] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

[Grant and Kyprianou, 2013] Grant, C. M. and Kyprianou, N. (2013). Epithelial mesenchymal transition (emt) in prostate growth and tumor progression. *Translational Andrology and Urology*, 2(3):202.

[Imran Khan et al., 2015] Imran Khan, M., Hamid, A., Mustafa Adhami, V., K Lall, R., and Mukhtar, H. (2015). Role of epithelial mesenchymal transition in prostate tumorigenesis. *Current Pharmaceutical Design*, 21(10):1240–1248.

[Jin et al., 2014] Jin, R., Yi, Y., Yull, F. E., et al. (2014). Nf-κb gene signature predicts prostate cancer progression. *Cancer Research*, 74(10):2763–2772.

[Kelemen et al., 2008] Kelemen, A., Abraham, A., and Chen, Y. (2008). *Computational intelligence in bioinformatics*, volume 94. Springer.

[Krebs and Hilton, 2001] Krebs, D. L. and Hilton, D. J. (2001). Socs proteins: negative regulators of cytokine signaling. *Stem Cells*, 19(5):378–387.

[Lessard et al., 2003] Lessard, L., Mes-Masson, A.-M., Lamarre, L., et al. (2003). Nf-κb nuclear localization and its prognostic significance in prostate cancer. *BJU International*, 91(4):417–420.

[Liberzon et al., 2015] Liberzon, A., Birger, C., Thorvaldsdóttir, H., et al. (2015). The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425.

[Lizio et al., 2015] Lizio, M., Harshbarger, J., Shimoji, H., et al. (2015). Gateways to the fantom5 promoter level mammalian expression atlas. *Genome Biology*, 16(1):22.

[Maguire et al., 1994] Maguire, J., Santoro, T., Jensen, P., et al. (1994). Gem: an induced, immediate early protein belonging to the ras family. *Science*, 265(5169):241–245.

[Marbach et al., 2012] Marbach, D., Costello, J. C., Kuffner, R., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat Meth*, 9(8):796–804.

[Margolin et al., 2006] Margolin, A. A., Nemenman, I., Basso, K., et al. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7.

[Omranian et al., 2016] Omranian, N., Eloundou-Mbebi, J. M., Mueller-Roeber, B., and Nikoloski, Z. (2016). Gene regulatory network inference using fused lasso on multiple data sets. *Scientific Reports*, 6.

[Petralia et al., 2016] Petralia, F., Song, W.-M., Tu, Z., and Wang, P. (2016). New method for joint network analysis reveals common and different coexpression patterns among genes and proteins in breast cancer. *Journal of Proteome Research*, 15(3):743–754.

[Schaffter et al., 2011] Schaffter, T., Marbach, D., and Floreano, D. (2011). Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270.

[Shintani et al., 2008] Shintani, Y., Maeda, M., Chaika, N., Johnson, K. R., and Wheelock, M. J. (2008). Collagen i promotes epithelial-to-mesenchymal transition in lung cancer cells via transforming growth factor–β signaling. *American Journal of Respiratory Cell and Molecular Biology*, 38(1):95–104.

[Shukla et al., 2004] Shukla, S., MacLennan, G. T., Fu, P., et al. (2004). Nuclear factor-κb/p65 (rel a) is constitutively activated in human prostate adenocarcinoma and correlates with disease progression. *Neoplasia*, 6(4):390–400.

[Srinivasan et al., 2010] Srinivasan, S., Kumar, R., Koduru, S., Chandramouli, A., and Damodaran, C. (2010). Inhibiting tnf-mediated signaling: a novel therapeutic paradigm for androgen independent prostate cancer. *Apoptosis*, 15(2):153–161.

[Thorsen et al., 2008] Thorsen, K., Sørensen, K. D., Brems-Eskildsen, A. S., et al. (2008). Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Molecular & Cellular Proteomics*, 7(7):1214–1224.

[Trei et al., 2016] Trei, D., Korcsmros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Meth*, 13(12):966–967.

[Umbehr et al., 2008] Umbehr, M., Kessler, T. M., Sulser, T., et al. (2008). Prococ: The prostate cancer outcomes cohort study. *BMC Urology*, 8(1):9.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

[Zanussi et al., 1992] Zanussi, S., Doliana, R., Segat, D., Bonaldo, P., and Colombatti, A. (1992). The human type vi collagen gene. mrna and protein variants of the alpha 3 chain generated by alternative splicing of an additional 5-end exon. *Journal of Biological Chemistry*, 267(33):24082–24089.

[Zhang et al., 2009] Zhang, L., Altuwaijri, S., Deng, F., et al. (2009). Nf-κb regulates androgen receptor expression and prostate cancer growth. *The American Journal of Pathology*, 175(2):489–499.

[Zhang and Song, 2013] Zhang, Y. and Song, M. (2013). Deciphering interactions in causal networks without parametric assumptions. *ArXiv Preprint ArXiv:1311.2707*.