Molecular interactomes

Network-guided cancer prognosis prediction & multi-way chromatin interaction analysis

Allahyar, Amin

DOI
10.4233/uuid:3d65f306-0e41-4d5c-b53d-a536b845851b

Publication date
2018

Document Version
Final published version

Citation (APA)
Allahyar, A. (2018). *Molecular interactomes: Network-guided cancer prognosis prediction & multi-way chromatin interaction analysis*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:3d65f306-0e41-4d5c-b53d-a536b845851b

Molecular interactomes

network-guided cancer prognosis prediction & multi-way chromatin interaction analysis

Molecular interactomes

NETWORK-GUIDED CANCER PROGNOSIS PREDICTION & MULTI-WAY CHROMATIN INTERACTION ANALYSIS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, prof.dr.ir. T.H.J.J. van der Hagen,
Chair of the Board for Doctorates,
to be defended publicly on Monday 12 November 2018 at 15:00 o'clock

by

Amin Allahyar

Master of Science in Computer Engineering, Artificial Intelligence,
Ferdowsi University of Mashhad (Iran),
born in Shiraz, Iran.

This dissertation has been approved by the

promotor: Prof. dr. ir. M.J.T. Reinders and
copromotor: Dr. ir. J. de Ridder

Composition of the doctoral committee:

Rector Magnificus,	chairman
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology
Dr. ir. J. de Ridder,	University Medical Center Utrecht
Prof. dr. W. de Laat	University Medical Center Utrecht
Independent members:	
Prof. dr. R. van Ham,	Delft University of Technology
Prof. dr. ir. D. de Ridder,	Wageningen University
Dr. ir. P. D. Moerland,	Amsterdam University Medical Center, University of Amsterdam
Prof. dr. T. Lenaerts,	Université Libre de Bruxelles
Prof. dr. ir. R. L. L. Lagendijk,	Delft University of Technology, reserve member

TU Delft Delft University of Technology

NWO Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Keywords: Bioinformatics, breast cancer outcome prediction, 3D organization of the genome

Printed by: IPSKAMP printing, ipskampprinting.nl

Front & Back: Amin Allahyar

ISBN 000-00-0000-000-0

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

To whom I live for,
my wife, my parents and my sister.

CONTENTS

1

INTRODUCTION

To apply local interventions that may cure a particular disease,
we cannot avoid understanding the cells' global organization.

A. L. Barabási [1]

1.1. BRIEF HISTORY OF CANCER RESEARCH

In contrast to popular belief, cancer is not a modern disease. Evidence of cancer cells is found in dinosaur fossils living 70-80 million years ago [2–5]. Malignant tumor cells were also found in Homo erectus, an extinct ancestor of modern humans that lived about 4 million years ago in Kenya [6]. At the same time, cancer therapy of patients has primeval origin. The first evidence for cancer treatment is found in the Edwin Smith papyrus, containing an ancient Egyptian medical text estimated to be from 3000 BCE. According to this document, ancient physicians like Hippocrates (460 BC - 370 BC) believed that cancer originates from the excess of black bile, one of the four "humors" thought to be the basic substances of the human body. At the end of 18th century, pathologist Rudolph Virchow (1821-1902) revealed that cancer cells originate from normal and healthy cells [7]. In the past century, cancer research reached the consensus that this disease is likely caused by damage to *DNA*. This long molecule in the cell's nucleus contains instructions necessary for diverse functions in the cell. This understanding explains why DNA damaging factors such as exposure to radiations (e.g. ultraviolet or gamma rays) and chemical substances (e.g. those encountered in cigarette smoke) are common causes of cancer.

It turned out that characterizing the origin of cancer is only the proverbial tip of the iceberg. While in simple genetic diseases, such as cystic fibrosis or muscular dystrophy, alteration of a sole base in the DNA was found to be associated with the phenotype (i.e. *monogenic* diseases), comparison between normal and tumor cells demonstrated independent mutations in several genes, suggesting that cancer is a *polygenic* disease. This collusion between the so called *oncogenes* was first exemplified in embryonic fibroblast cells harboring RAS mutations where their tumorigenic potential were conditioned on impairment of a second oncogene MYC [8]. Meanwhile presence of *tumor suppressors* in mouse and later in human cells were confirmed [9–11]. Collectively, these pieces of evidence fueled the theory of *multistage carcinogenesis* which postulates that healthy cells require several independent aberrations before they can become neoplastic cells [12, 13].

Clonal evolution is the modern equivalent of the multistep carcinogenesis theory [14]. In this model, few cells with acquired "advantageous" mutations overpower nearby cells by taking over their resources and grow out into so called *benign* tumors. As their name suggests, these benign tumors are often harmless. The malignant step occurs when cells in these tumors acquire additional mutations allowing them to metastasize (i.e. traveling to other organs) which accounts for as much as 90% of cancer mortality [15]. Breast cancer presents an infamous example of this event where the majority of deaths from this disease are not due to the primary (benign) tumor but from metastasis [16]. Breast cancer is the most common type of cancer in women worldwide and is the prime cause of death among them [17, 18]. Substantial efforts have been made to discern the complex aberrations that are frequently observed in patients diagnosed with this cancer. In Chapter 2 and 4, we will focus on breast cancer and investigate its abnormalities.

It is believed that proliferating cells originated from the primary tumor in breast "intravasate" to blood or lymphatic vessels and later "extravasate" into the target organ [19]. These cells require ample resources due to their need to maintain high levels of proliferation, thus necessitating higher blood supply (achieved through angiogenesis) [20]. At the same time, these cells must stay hidden from the immune system and ignore apoptotic signals in order to sustain their expansion [21]. In a seminal paper, Hanahan and Weinberg consolidated this theory into six cellular hallmarks including evading apoptosis, self-sufficiency in growth signals, insensitivity to anti-growth signals, sustained angiogenesis, limitless replicative potential, tissue invasion and metastasis [22]. In their subsequent influential paper, Hanahan and Weinberg complemented this theory with the additional hallmarks such as reprogramming of energy metabolism and immune destruction evasion [23].

1.2. GLOBAL VIEW OF CELL STATE

As Hanahan and Weinberg argued, the formation and spread of neoplastic cells rely on disruption of diverse processes (i.e. hallmarks of cancer). Furthermore, each individual mis-regulation is known to be essential for cancer development in these cells [21, 24]. It should be noted that these perturbed processes might be result of abnormal expression of many genes. Consequently, assessing aberrations in only few candidate genes may not be sufficient to describe and further understand the underlying mechanisms driving the progression and metastasis in cancerous cells [25]. Therefore, it is nowadays accepted in cancer research that piecing together the cancer “puzzle” is nearly impossible without considering the entire set of genes operating in the cell [26].

1.2.1. MEASURING THE TRANSCRIPTOME

Completion of the human genome project promoted several measurement techniques with unprecedented capabilities. *Microarrays* [27, 28] in particular, have proved to provide a global view of transcribed genes at the level of messenger RNAs (mRNA). Using this exciting technology, for the first time, investigation of aberrated cellular processes and regulatory mechanisms in cancer cells at a genome-wide scale became possible. This notable step forward, coupled with a relatively cheap and convenient laboratory protocol, led to widespread application of microarrays in a variety of biological problems and revolutionized cancer research [29]. Microarrays exploit the hybridization property of DNA in which two complementary strands of DNA bind to each other to form a double stranded molecule. A microarray chip contains *probes*, which are spots of single stranded DNA representing all genes in a host of interest. These probes hybridize specifically to their complementary mRNA originating from the host genes. Next, a laser beam excites fluorescent dyes mounted on the mRNA molecules during library preparation. The fluorescence emissions are then captured by a high resolution camera to provide a genome-wide picture of expression for that particular sample.

In the past few years, next generation sequencing and in particular sequencing of RNA-derived molecules (RNA-seq) gained popularity. This technology is specially interesting as it is not limited to known sequences which enables exploration of organisms with unknown transcripts or splice-variants. However in breast cancer research (which is the focus of this thesis), the available datasets based on RNA-seq are still limited in number of samples. Most notably, The Cancer Genome Atlas (TCGA) [30] encompasses 1092 survival labeled breast cancer RNA-seq samples (at the moment of writing this thesis) which is still far fewer than available samples in microarray datasets such as METABRIC [31, 32]. Nevertheless, number of publicly available RNA-seq samples are increasing at a fast pace while subtle caveats and biases in this technology [33] are being identified and resolved with computational methodologies [34, 35]. Therefore, It is expected that RNA-seq technology will replace microarrays in transcriptome profiling if these datasets grow in size and robust methods to process and normalize its data are introduced [36]. Considering recent reports, one could argue that this transition has already happened [37]. It is worth noting that this shift will make a huge pile of microarray samples obsolete. Maybe the best approach in machine learning applications would be to develop and utilize cross-platform normalization strategies to combine microarray and RNA-seq datasets (see section 1.2.4 and [38]).

1.2.2. MACHINE LEARNING

Advances in sequencing platforms and microarrays provided affordable genome-wide measurements for many laboratories. It was believed that all necessary pieces of the cancer puzzle have finally become available [39]. However, the sheer amount of data produced by these technologies and extracting the relevant information soon gave rise to a series of unique challenges [40]. The complex nature of cancer combined with the high dimensionality of genomic data required an automated approach to assemble the puzzle pieces. This coincided with an explosion of computation power in personal computers and a drop in their price. As a result, techniques like machine learning flourished in genomic research. Soon after the introduction of microarrays, a flow of papers utilizing machine learning methodologies to tackle various problems in cancer research appeared in top journals.

Perou *et al.* pioneered one of the early applications of pattern recognition in genomic research by demonstrating that these models can be used to segregate breast cancer patients into clinical groups (i.e. *subtypes*) with homogeneous patterns of expression in each group [41]. This application of machine learning is commonly known as “unsupervised” learning as no prior categorization (e.g. subtype) for patients is considered to determine parameters of the model. Later, van ’t Veer *et al.* introduced a *supervised* application of machine learning called *outcome prediction*. In contradiction to unsupervised methods, a supervised method “learns” the relationship between expression patterns of patients for which a phenotype of interest (in this case their outcome) is known and aims to predict the phenotype for new patients. van ’t Veer *et al.*’s tool (called *MammaPrint*) could classify patients into "good" (survival more than 5 years) or "poor" (survival less than 5 years) prognosis by analyzing expression levels of 70 pre-defined genes. This allowed breast cancer patients with expected good prognosis to be excluded from treatments with drastic side effects (e.g. chemotherapy). Introduction of MammaPrint triggered a great excitement in cancer research community because it aimed to address the main limitation in clinical practices where each physician would rely on his or her own criteria for determining chemotherapy administration, introducing inconsistencies among prognosis [42].

As exemplified by MammaPrint, exploitation of genomic data has profound implications toward more personalized treatments for breast cancer patients. Within this thesis we will study several classifiers (e.g. Lasso [43]) that were previously employed as outcome predictors in more detail and further utilize their extensions (e.g. Sparse Group Lasso [44]) as models that can exploit existing knowledge about the cellular processes and functions represented in gene-gene interaction networks. Classical outcome predictors (such as MammaPrint) are often linear models. In these models, for each patient, gene expression levels are combined (with different weights) into a single value that represents patient's membership to good/poor prognosis group. In the next section, we will explore linear outcome predictors in more detail.

1.2.3. Linear regression

Consider a problem in which a researcher aims to explain a phenotype of interest $y_i \in \mathbb{R}$ from a set of d observations (genes) that are collected from n samples (patients) $\mathbf{x}_i = \left[x_{i_1}, x_{i_2}, \dots, x_{i_d}\right]$ $i \in \{1, 2, \dots, n\}$ [1]. Let β_j represents the contribution (weight) of a gene j to the patient's outcome:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix}, \quad p_i = \sum_{j=1}^{d} \beta_j x_{ij} \tag{1.1}$$

where p_i is the predicted outcome by this linear model for sample $\mathbf{x}_i$. Optimal weights for this problem minimizes the *Mean Square Error* (MSE) between prediction p_i and the observed phenotype y_i, across all patients:

$$\min L(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \beta_j x_{ij} \right)^2 = \frac{1}{2} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 \tag{1.2}$$

where $\|.\|_2$ is the L_2 norm i.e. $\left\|\boldsymbol{\beta}\right\|_2 = \sqrt[2]{\beta_1^2 + \beta_2^2 + \cdots + \beta_d^2}$. Furthermore, $\mathbf{X} \in \mathbb{R}^{n \times d}$ [2] is an expression matrix containing n patients and d genes while $\mathbf{y} \in \mathbb{R}^{n \times 1}$ is a vector of n phenotypes of interest for all patients. cost function (1.2) is known as *Ordinary Least Square* (OLS) problem. Owing to its simplicity, optimal coefficients β_j in this minimization problem can be found by a few linear algebra operations:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{\beta} \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} \end{aligned} \tag{1.3}$$

Differentiating equation (1.3) with respect to $\boldsymbol{\beta}$ and equating the result to zero gives:

$$-\boldsymbol{X}^T \boldsymbol{y} + (\boldsymbol{X}^T \boldsymbol{X}) \boldsymbol{\beta}^* = 0 \tag{1.4}$$

where $\boldsymbol{\beta}^*$ indicates the optimal vector of values that minimizes cost function (1.2). Further reordering of these terms yields:

$$\boldsymbol{\beta}^* = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{1.5}$$

[1] In this thesis, boldface letters are used to represent vectors.

[2] In this thesis, capital boldface letters represent matrices.

Precise calculation of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ requires $\boldsymbol{X}$ to have full *rank*. Meaning that this matrix has to satisfy several prerequisites (see [45] for details). Most notably, the number of rows (patients) should be at least as large as the number of columns (genes). This is why having many samples to train an outcome predictor is a major factor in the accuracy of the trained model. At the same time, having many samples assists the model to disregard the technical noise that may be introduced during the preparation of samples or expression measurements [46]. Unfortunately, despite substantial progress [47], it is still expensive to produce such massive datasets for outcome prediction and consequently most datasets for this problem are limited to few thousand patients at best while encompassing several tens of thousands of genes [48].

1.2.4. BATCH EFFECTS

One basic solution to acquire more samples is to *pool* data from different studies [49]. However, this approach brings its own challenges. The primal difficulty in sample pooling is that technical variations in expression profiles are often study-specific [50]. Many sources of such variations can be attributed to e.g. a difference in library preparation, microarray (or RNA-seq) platform or image acquisition [51]. These disparities often result in study-specific alterations of expression levels. This variation can be observed even in early microarrays datasets with an ordinary visualization method like t-SNE [52]. For example Figure 1.1.a represents a t-SNE visualization of a dataset formed by pooling the original expression data measured by Perou *et al.* as well as van 't Veer *et al.*. In this figure, one can see that patients in the Perou *et al.* dataset are more similar to each other compared to patients in the van 't Veer *et al.* dataset. In this example, outcome prediction of a linear regression model that is trained using the van 't Veer *et al.* data would not be better than a random guess when applied to patients in the Perou *et al.* dataset. To overcome these study-specific effects several pre-processing methods have been developed. COMBAT [53] is one of the commonly used methods to remove batch effects from microarray data. Figure 1.1.b demonstrates the same dataset after correction of expression levels (using COMBAT) showing that study-specific clusters of patients do not exist anymore.

In Chapter 4, we show that although apparent batch effects can be removed (e.g. using COMBAT) from an expression dataset that is formed by pooling samples from independent studies, more subtle batch effects remain in the dataset. More critically, recent studies reported evidence for new batch effects that are introduced by batch effect removal methods themselves [54]. In fact, dealing with these batch effects is expected to be the next major challenge in the large-scale analysis of biological datasets [55].

1.2.5. MODERATING COMPLEXITY OF THE MODEL

It should be noted that, finding the optimal parameters for an outcome predictor is not the final goal. This is because the optimal $\boldsymbol{\beta}^*$ coefficients only guarantee precise prediction of y_i across all "seen" patients (*training* set). Yet, such an optimization does not warrant accurate prediction of survival for "unseen" patients (*test* set). This concept is often known as the *generalization* capability of a classifier. Many factors can have negative impact on the generalization competence of a classifier. For example, the intrinsic batch effects discussed previously can hamper generalization of an outcome predictor

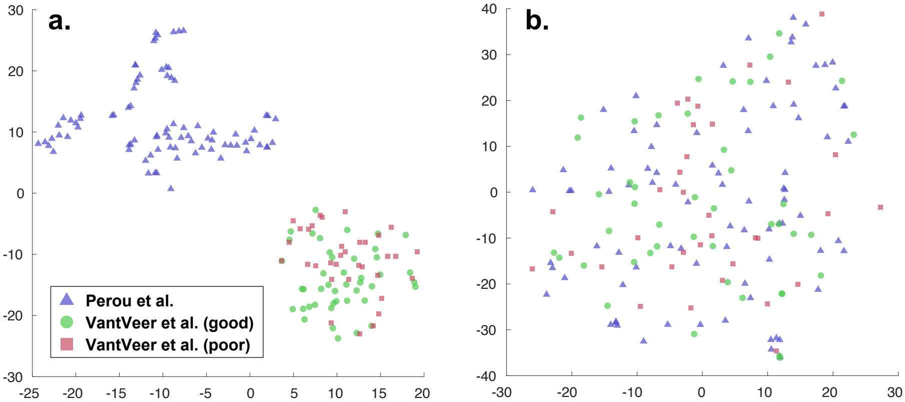


Figure 1.1: Due to variations in library preparation and data pre-processing across studies, pooling datasets is challenging. Methods like COMBAT can potentially mitigate the cross-study variation. To visualize this variation, t-SNE can be used to represent the patients in a 2-dimensional space in a way that patients with similar expression profile reside closer in 2D space while disparate patients end up far away from each other. **a.** t-SNE visualization of gene expression data pooled from Perou *et al.* and van 't Veer *et al.*. **b.** Visualization of the same datasets after reducing batch effects using COMBAT.

to a large degree. In those cases, the classifier usually *overfits* to the training set meaning that the prediction accuracy of the training samples are noticeably higher than samples in the test set.

Overfitting is also prevalent when the utilized dataset has many features (genes) and few samples (patients) (*curse of dimensionality*) which is a typical property of biological datasets [44].

Additionally in medical applications, it is crucial to identify genes whose expression levels are mostly associated to the phenotype of interest (i.e. *interpretation* of the trained model). However, the sheer number of measurements that loosely correlate to the phenotype of interest make it challenging to discern the involved genes. Accordingly in such applications, it is commonly sought to determine a smaller subset of elements that exhibits the strongest effect at the expense of accuracy of the model (in the training samples). Empirical evidence showed that this procedure improves the generalization of the model [56].

To mitigate these issues, Tibshirani devised LASSO (Least Absolute Shrinkage and Selection Operator) which aims to reduce the number and influence of the parameters (genes) in the linear model by "shrinking" the β_j while preserving the prediction power (minimal MSE). The resulting cost function is designed to be a mixture of the OLS problem with a term that regularizes the non-essential (or expendable) β_j to zero. Tibshirani proposed the following cost function [43]:

$$L(\boldsymbol{\beta}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}\right\|_1, \quad \lambda \geq 0 \tag{1.6}$$

where $\|.\|_1$ is the L_1 norm e.g. $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{d} |\beta_j|$. The added penalty term effectively encourages sparsity in the weights of the solution vector $\boldsymbol{\beta}$ which leads to feature selection. At the same time, λ provides a balance between the MSE and the shrinkage.

1.2.6. CROSS STUDY GENERALIZATION

As described, pooling datasets (which increases the number of samples) [49] and regularization of coefficients (which reduces the number and influence of genes) [56, 57] is reported to improve the generalization in outcome predictors. But model generalization in these reports is often measured in a single study (i.e. *within-study* generalization). Meaning that cross-study variation which is a typical property in real world applications of outcome predictors is not considered in the evaluation procedure. In real world applications, predictors trained on expression profiles obtained in one hospital, by a certain preparation protocol or measurement technology, are expected to perform well when applied to data from a different hospital, protocol or technology. Unfortunately, empirical evidence showed that such variations substantially impede performance of outcome predictors [58]. In fact, lack of performance in cross-study validation greatly hampered the clinical application of outcome predictors [59, 60]. Thus, generalization of an outcome predictor should be assessed in a *cross-study* validation procedure to closely simulate real world application of these models [61, 62].

One may argue that once suitable batch effect removal methods are developed, classical outcome predictors could be readily used with high generalization [58]. However, if batch effect is the only limiting factor, then independent analysis of each cohort should have found similar sets of genes associated with the outcome of patients. Yet, several studies reported lack of overlap between survival markers identified by independent analysis of different datasets [63]. Notably, from 70 markers identified by Wang *et al.* only three genes were in common with markers identified by van 't Veer *et al.* in her independent dataset [65]. Even more striking, it was shown that many random gene sets can be predictive as long as this set contained sufficient (i.e. >100) genes [66]. Together these findings suggest that irrespective of intrinsic batch effects, the characterized markers are not describing the primary *driver* mechanisms of the disease and are limited to secondary *passenger* manifestations which may differ substantially from patient to patient [67].

1.3. CELL WIRING DIAGRAM: A VIABLE RESOURCE FOR OUTCOME PREDICTION

During the past few decades, cancer research has unraveled various pathways that are often mis-regulated in carcinogenic cells. Perturbations in such diverse driver processes manifest in extensive expression profile heterogeneity of breast cancer tumors [68]. Similarly, deregulation of multiple pathways can have impact on the expression of an individual gene. A notable example is TP53 (responsible for apoptosis) which is shown to be inactivated by many different pathways [69]. Therefore, pathway membership of genes could be potentially informative for the role a gene can have in the risk of developing cancer.

1.3.1. EXTENDING THE CLASSICAL OUTCOME PREDICTORS

However, classical outcome predictors do not enforce prior constraints on these relationships. Considering the number of genes, this flexibility in the classical models could identify many spurious markers as long as their mixture is predictive. For example, genes that are related to positive feelings in humans or genes active in localization of skin fibroblast in mice were shown to be viable markers in outcome prediction of breast cancer patients [66]. These findings encouraged a new type of predictive model that promotes the predictive variables to be formed from sets of genes with priory known relationships [67].

1.3.2. GENE INTERACTOMES TO GOVERN PREDICTOR MODELS

One way to present gene relationships is to conceptualize genes (as well as proteins or other metabolites in the cell) as nodes and their interactions as links in a network, giving rise to many different biological networks [70, 71]. For example in a metabolic network, directed edges can be used to connect reaction substrates to products [72]. Alternatively, the physical binding of two proteins can be depicted using an undirected edge which collectively form a physical *Protein–Protein Interaction* (PPI) network [73]. Networks can also depict relationships between genes and their regulators (such as other genes, transcription factors, RNA or other small molecules) [74] or organize them into sets of overlapping modules commonly known as *pathway* networks [75]. Generally, sources of interaction evidence could be experimental [76], literature mined [77, 78], extracted from expression analysis [79, 80], or even combination of these methodologies [81]. Such network representations are conceptually appealing in computational biology as many well-established concepts in network theory can be directly applied on these representations. Notably, it has been reported that many biological networks are *scale-free* (i.e. enclosing few *hubs* which are highly connected nodes) [82, 83]. Additionally, hubs are shown to perform well to predict survival of patients [84].

1.3.3. NETWORK BASED OUTCOME PREDICTORS

An intracellular *interactome* could be a valuable source of information for an outcome predictor to identify groups of genes that once perturbed could give rise to breast cancer and its metastasis [67]. This was the dawn of Network-based Outcome Predictor (NOP) models [85]. These models often incorporate network information in two steps: gene set formation (selections) and expression aggregation (integration) [86]. The initial step utilizes a network and outputs gene sets each of which representing (part of) a cellular process or pathway [87]. In the integration step, the expressions of genes in each set are combined (often by averaging) to produce a single "meta-gene" [88]. These meta-genes are then considered as typical features and (similar to a classical model) are used to train an outcome prediction model [89]. Figure 1.2 depicts a schematic overview of this procedure.

1.3.4. MODELING EXPRESSIONS IN GENE SETS

While the NOP concept is promising, devising a network-aware model has proven to be difficult. This is mainly because selection and integration steps in a NOP are interdependent. Specifically, it is difficult to group genes without knowing 1) how these genes are

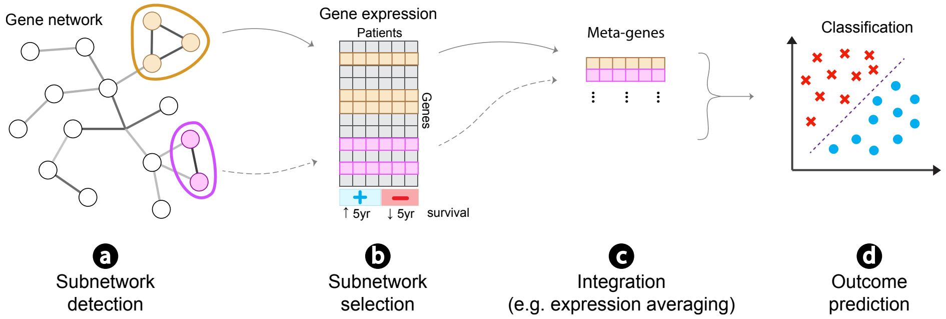


Figure 1.2: Schematic overview of NOPs. NOPs are usually trained in multiple steps. **a.** Groups of genes are identified (often through clustering). **b.** Performance of each gene set is measured. **c.** Meta-genes are formed by aggregating the expressions of genes (usually by averaging). **d.** Final model is trained using produced meta-genes.

integrated into meta-gene and 2) how the produced meta-gene would perform in collaboration with other meta-genes in the final model. In Chapter 2, we propose FERAL that exploits a derivative of lasso called Sparse Group Lasso (SGL) to simultaneously pick the most suitable meta-genes from each gene set while aggregating the chosen meta-genes to form appropriate markers for predicting breast cancer outcome. In the next section, we will briefly describe the cost functions and properties of these lasso derivatives.

1.3.5. LASSO DERIVATIVES TO SUPPORT GENE GROUPS

Suppose the d genes are divided into G groups and m_k where $k = \{1,2,\ldots,G\}$ denotes the number of genes in k^{th} group. To simplify the notation, we utilize $\boldsymbol{X}_k$ to represent the expression matrix of genes residing in the k^{th} group while $\boldsymbol{\beta}_k$ corresponds to the coefficient vector for this particular group. For clarity, we assume that the patient's outcome ($\boldsymbol{y}$) and the gene expression matrix ($\boldsymbol{X}$) are centered (i.e. zero column mean). The Group Lasso (GL) proposed by Yuan and Lin solves the following convex cost function to identify the optimal coefficients for each group [90]:

$$L(\boldsymbol{\beta}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \lambda \sum_{k=1}^{G} \sqrt{m_k}\left\|\boldsymbol{\beta}_k\right\|_2, \ \ \lambda \geq 0 \tag{1.7}$$

GL is structurally similar to lasso (which regularize features) but applies regularization at the group level. That is, an entire group of predictors may drop out of the model. The group lasso is a generalization of the standard lasso because if the group sizes are all equal to one, cost function (1.7) reduces to the classical lasso cost function shown in equation (1.6).

One limitation in the GL model is its inability to select relevant genes within each group. That is, if group j is active (i.e. its corresponding coefficients in $\boldsymbol{\beta}_j$ are non-zero), individual coefficients in this group are free to have any arbitrary value (i.e. it becomes similar to the OLS in equation (1.3)). To mitigate this issue, Friedman *et al.* proposed Sparse Group Lasso (SGL), which is formed by coupling the penalty terms of lasso and

GL, yielding sparsity at both individual feature (gene) and group (i.e. pathway) level. This cost function is defined as follows [91]:

$$L(\boldsymbol{\beta}) = \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \lambda_1\left\|\boldsymbol{\beta}\right\|_1 + \lambda_2\sum_{k=1}^{G}\sqrt{m_k}\left\|\boldsymbol{\beta}_k\right\|_2, \quad \lambda_1,\lambda_2 \geq 0 \tag{1.8}$$

SGL is capable of achieving a simultaneous selection of genes and groups (or meta-genes). In Chapter 2, we investigate such a model (i.e. FERAL) and show its superior performance compared to several existing NOPs.

1.4. REFINING NETWORKS TO THE PROBLEM OF INTEREST

In Chapter 2, we reproduce two previously reported observations. Initially, we (among others) note that most (if not all) models trained using meta-genes do not outperform classical models trained using individual features [88, 92, 93]. This result on the one hand may suggest that meta-genes do not add to predictive power of trained models. On the other hand and even more surprising is the observation where a shuffled network does not reduce NOPs performance. This is clearly in contrast with the promise of NOPs (i.e. exploiting network information to guide the model) and calls for a fundamental reevaluation of NOPs structure and how these models are usually trained.

In Chapter 4, we take a critical look at NOPs structure to provide an explanation for these observations. We point out that biological networks capture only a partial picture of the cell's multifaceted system. For example, such networks describe gene expression correlations or known signaling pathways, but not both at the same time. This incomplete perspective may not be sufficient to link the wide range of aberrations that may occur in a complex and heterogeneous disease such as breast cancer [94, 95]. In addition, many links in these networks are experimentally obtained from model organisms such as yeast and therefore not specific for humans [96–98]. Finally, it should be recognized that many links in these networks are unreliable [99, 100], missing [101] or redundant [102]. For this reason, considerable efforts have been made to refine these networks [103]. Additionally, interactions are often biased towards well-studied genes while many other genes are rarely connected to the rest of the network. Taken together, the employed networks may have little (or insufficient) relevance to outcome prediction potentially explaining why a shuffled network provides a comparable performance to biological networks.

To address this issue, we will effectively turn the problem around in Chapter 4. Instead of using a generic biological network to improve outcome prediction, we use the expression data to identify a network of genes that truly improves outcome prediction. To this end, we search for synergistic gene pairs, i.e. genes whose joint prediction power is beyond what is attainable by both genes individually [104]. The resulting network, called SyNet, is specific to the phenotype under study and will be used to govern a NOP model. In this chapter, we show that integrating genes according to SyNet provides superior accuracy and stability (in terms of performance and marker consistency) and we also demonstrate that shuffling nodes in SyNet results in a substantial performance drop which confirms relevance of SyNet links to outcome prediction. Further, while SyNet is inferred without use of prior biological knowledge, we show that its genes are markedly enriched for well-known factors in survival of breast cancer. These findings suggest that compared to general purpose gene networks, phenotype-specific networks provide valuable mechanistic insights into the aetiology of breast cancer that is missed when restricting towards well-studied genes.

1.5. Multi-way vs. pairwise interactions

Due to computational burden, we limited SyNet to pairwise interactions. Downstream analysis of SyNet revealed that highly connected (hub) genes in this network are in fact well-known driver genes in breast cancer. This observation corroborates previous findings that these driver genes are involved in multiple fundamental mechanism in this disease [105]. Based on this observation, it would be interesting to investigate if synergy could become stronger in triplets of genes. To this end, we selected a limited set of the top 1000 highly variable genes in a collected cohort of more than 4000 patients (1 billion gene triplets). Next, we searched for synergistic triplets that did not show predictivity (i.e. average AUC across 5 repeats of cross-study validation) when constitute genes were analyzed separately or in pairs. Intriguingly, we found many triples to have such a property. Figure 1.3 represents the performance of the top 100 triplets with highest synergy. Most notably, the top triplet consists of RPL5, SORBS and DDX5 genes, well-known for their role in invasive capacity of tumor cells in breast cancer patients and their response to chemotherapy treatment [106–108]. This preliminary evidence suggests that the pairwise representation of gene interactions (which are used in most if not all biological networks), might be insufficient to truly depict gene relationships. Specially in a complex disease such as cancer, complete characterization of the cell wiring diagram may require a more complex representation of the interactions between genes. Nonetheless, representation and integration of these higher order interactions are only trivial parts of this problem. The primal challenge is to experimentally identify and validate these complex multi-way interactions [109, 110]. In fact, inefficiency of measurement techniques and their low throughput is currently the limiting factor in multi-way interaction assessments [111–114].

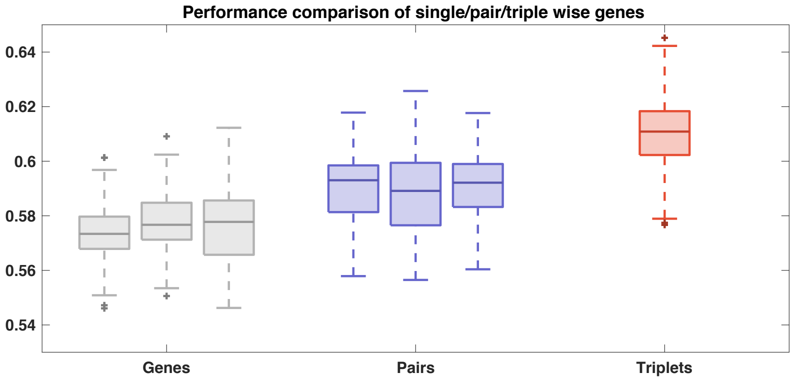


Figure 1.3: Gene triplets may reveal performance that cannot be captured by individual or pairs of genes. Red bars represents cross-validated performance of top 100 gene triplets. Gray bars represent individual performance of the same set of genes. Blue bars represent gene pair performance of selected genes.

1.6. SPATIAL CONFORMATION OF THE GENOME AS A CELLULAR NETWORK

In the previous sections, we discussed how biological networks represent a myriad of intertwined regulatory mechanisms by which gene expression in the mammalian genome is regulated. One important process by which a gene's expression is regulated is through promoter enhancer loops. An enhancer is a short (50-1500 bp) piece of DNA that attracts transcription factors and thereby increases the expression of genes that are brought into its 3D vicinity through the looping of DNA. A genome wide sketch of such relationships can be represented as a network of interconnections between enhancers and their target genes. It has been reported that a similar network made for 3D proximity of genes resembles co-expression network of genes [115], which we showed to be a suitable candidate for guiding network based outcome predictors [93]. Further, it has been shown that perturbation in 3D conformation of the genome could promote neoplasm in cells [116–118]. Consequently, employing a 3D proximity interactome could potentially guide existing network based predictors to identify abnormal activity in expression profiles of carcinogenic cells.

CHROMATIN FOLDING

It is widely established that packing of DNA in the nucleus is not just a compaction mechanism [119]. In fact, this "conformation" is known to be responsible for fine-tuning activity of many genes in mammalian cells [120]. This important function entails careful organization of functional elements in the nucleus even at chromosome level. As depicted in Figure 1.4.a, each chromosome preferentially occupies a *territory* in the nucleus [121]. Active and gene dense chromosomes tend to be positioned in the center while other chromosomes are mostly found close to the nuclear periphery [122]. Zooming-in to chromosome territories, one can observe Topologically Associating Do-

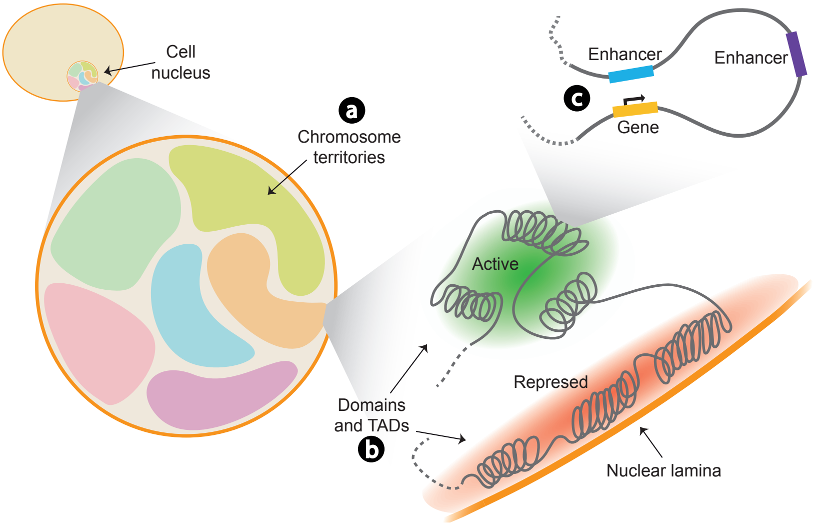


Figure 1.4: Hierarchical organization of the DNA within the nucleus of the mammalian cells. **a.** Chromosome territories. **b.** Chromosome domains and topologically associated domains or TADs. **c.** Enhancer-promoter loops

mains (TADs) that are 200kb - 1mb regions within which regulatory DNA elements (i.e. enhancers and promoters) are often stationed close together in 3D space and form chromatin "interactions" [123]. At the boundaries of these TADs, architectural proteins like the CCCTC-binding factor (CTCF) are located that focus chromatin interactions to intra TADs and reduce inter-TAD interactions. TAD configurations are stabilized by a ring-shaped protein called *Cohesin*, which is believed to hold distantly bound CTCF sites together [124]. CTCF sites are known to have directionality preferences where chromatin loops are often found to be formed between convergent CTCF sites.

Proper formation and dynamics of such a complex and hierarchical organization is known to be essential for appropriate gene activity, and perturbation of these regulatory mechanisms has been shown to promote cancer development [116, 125]. For breast cancer, the role of these chromatin interactions in the deregulation of pathways is subject of research [126].

Taken together, the current understanding of genome organization states that these deleterious factors may reside far away from the location of gene sequence. Such a distal associations between genes (and other functional elements) can be represented in a genome wide network of elements that collectively govern the expression profile of a cell in its nucleus. Therefore, such a network may provide another view of the cell wiring diagram that can be readily used in NOPs. It should be noted that such a comprehensive understanding of this regulatory system is acquired by at least two decades of intensive world-wide research. In the next section, we will give a brief overview of these efforts.

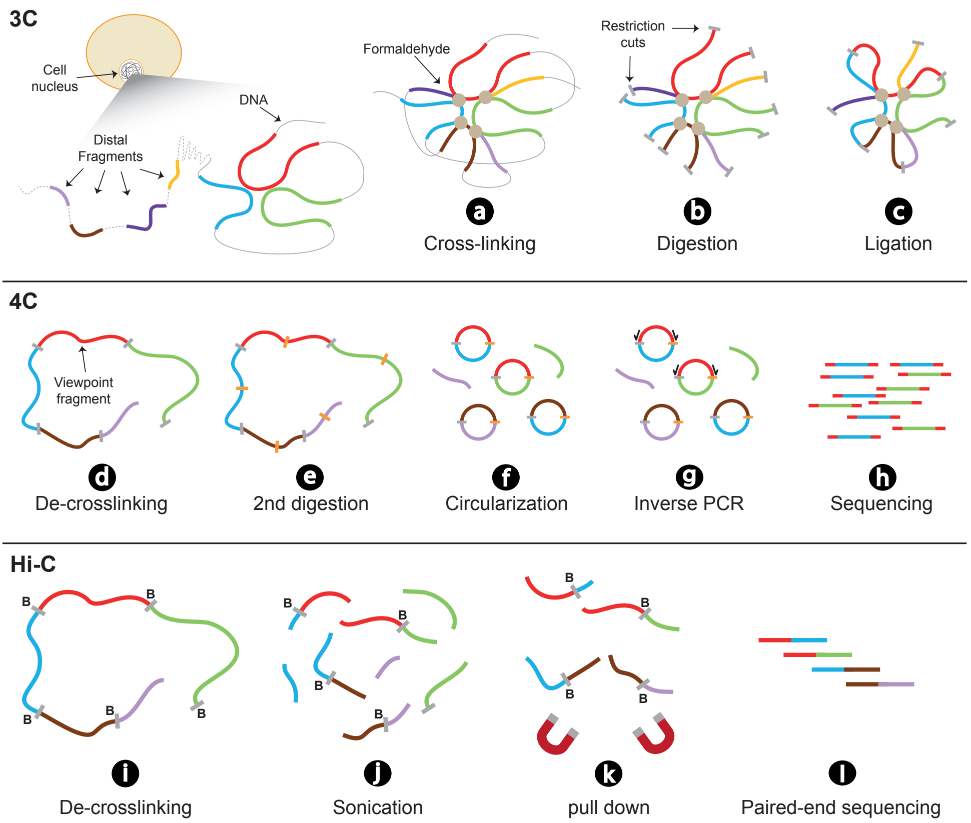


Figure 1.5: Schematic overview of Chromatin Conformation Capture (3C) methodologies. **a.** Native 3D conformation of DNA within nucleus of cell is fixed by Formaldehyde (beige circles). **b.** DNA are digested using restriction enzymes (grey rectangles). Type of restriction site used determines the cut sites. **c.** Cut sites are ligated back to random fragments in their vicinity to form long (~20kb) stretch of DNA which are often called *concatemer*. 3C derivatives often use this construct as their base material. **d.-h.** 4C method is able to uncover all 3D DNA contacts made with specific region of the genome (i.e. the *view point*). To this end, **e.** 3C templates are again cut using another restriction site to shrink their size. **f.** The cut templates are then circularized and then **g.** circles that contain view point fragment are amplified using inverse PCR which produces linear sequences. **h.** Sequencing adapters are then added to concatemer ends to prepare them for sequencing. **i.-l.** Instead of focusing on the genome contacts made with the view point region (*one vs. all*), Hi-C is designed to reveal all genome wide contacts (i.e. *all vs. all*). **i.** Hi-C uses special restriction enzymes that mark the cut sites in DNA with a specific magnetic molecules (biotin). **j.** 3C templates are then shattered to smaller concatemers using sonication. **k.** The marked concatemers are pooled down using magnets and **l.** prepared for sequencing.

1.6.1. 3C TECHNOLOGIES

Exploring the cell nucleus and its content using the microscope has been subject of interest since 1873 [127, 128]. However, due to limitations of light microscopy, these findings were mostly limited to large events such as chromosome separation. The study of chromatin conformation entered a new era after the introduction of Chromatin Conformation Capture (3C) based technologies (Figure 1.5a-c), which allowed probing the relative interaction frequency between a pair of DNA elements within a population of cells [129]. To measure this, 3C initially fixates the DNA fiber (using formaldehyde cross-linking) so that its conformation would not be disrupted during later steps. Next, the chromatin is digested into fragments using *restriction enzymes* that cut the DNA at particular enzyme-specific recognition sites. Further, by catalyzing the DNA ligation (via DNA ligase), fragments in close spatial proximity fuse together and form a *concatemer* (i.e. a collection of fragments linked together) (Figure 1.5c). Removing the cross-links from the concatemer produces the so-called *3C template* [130]. Several million nuclei can be simultaneously treated this way to obtain genome-wide spatially linked DNA concatemers in a population of cells. These concatemers are later analyzed using sequencing platforms (or PCR in classical 3C) to reveal enclosed fragments that were in close spatial distance at the moment of fixation. The premise in 3C technology (and other proximity ligation based methods) is that the observed number of fragment ligations are a proxy for the 3D interaction frequency of corresponding elements and their preferential looping in the genome. This technology formed the basis of several fundamental discoveries in genome organization including experimental confirmation of chromatin loops in transcription regulation [131].

1.6.2. 4C, THE NEXT STEP: CHROMATIN CONFORMATION CAPTURE ON CHIP

3C can only examine contacts formed by few pre-selected regions of the genome. To alleviate this limitation, 4C (Chromatin Conformation Capture on Chip) was introduced which is capable of interrogating interactions between a restriction fragment of interest (often called *viewpoint*) and any other restriction fragment in the genome [132] (Figure 1.5; 4C). To this end, 4C uses two primers that are designed to bind to each end of the viewpoint. Next, inverse PCR is employed to amplify and enrich for reads that carry the viewpoint. Inverse PCR necessitates circularization and shortening of reads which is achieved by a secondary digestion and a subsequent re-ligation of each read. The PCR products (or concatemers) then need to be prepared for microarray quantification. With the introduction of next-gen sequencing, this technology was adopted by 4C to simplify the protocol and enhance its throughput [133]. Nowadays, 4C is widely used to identify promoter-enhancer [134] or architectural loops [135]. Inspired by 3C and 4C approaches, many other methods were developed. Notably, *Hi-C* was proposed to investigate a genome-wide view of DNA interactions (at the expense of resolution) [136]. To achieve this, Hi-C exploits particular restriction enzymes that are capable of incorporating magnetic molecules in the cut sites (Figure 1.5; Hi-C) and then employ magnets to enrich for concatemers that contain concatemers with cut sites.

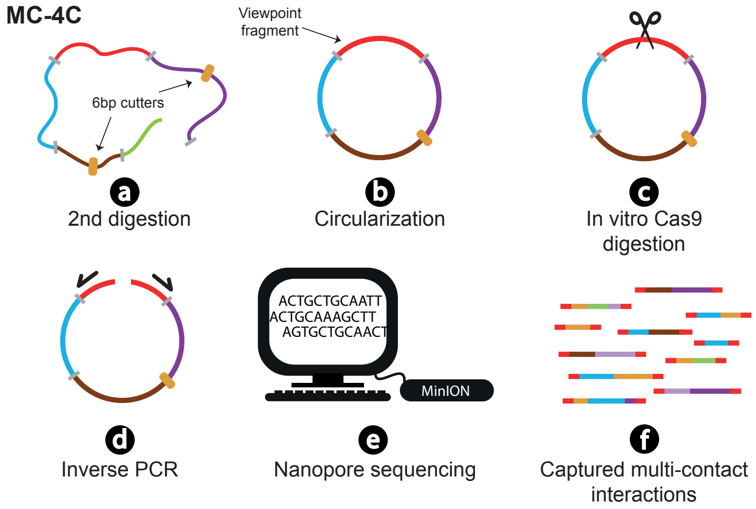


Figure 1.6: Schematic overview of Multi-Contact 4C (MC-4C) method. **a.** After de-crosslinking of 3C templates, a restriction enzyme with 6 base pairs recognition site is used to cut the concatemers and shrink their size to approximately 2-5kb. **b.** The cut templates are then circularized. **c.** Cas9-mediated in vitro digestion of the viewpoint fragment (and its neighbors) is used to block continues rolling circle amplification of products and reduce abundance of undigested circles **d.** Circles containing the view point fragment are amplified using inverse PCR. **e.** Libraries are then prepared and sequenced in MinION device. **f.** Sequenced reads enclose fragments that were in close proximity at the moment of fixation.

1.6.3. UNRAVELING MULTI-WAY INTERACTIONS

In contrast to PPI research which is still focused on pairwise interactions between proteins, the value of higher order 3D interactions (i.e. more than pairwise; see section 1.5) in the genome was well recognized in the genome conformation community [129]. There are two fundamental challenges for multi-way interaction appraisal. Primarily, higher order assessment requires exponentially higher throughput (i.e. number of sequenced reads). Using the latest advances in sequencing technology, the throughput is still insufficient, even to characterize the full genome wide pairwise interactions [137–139]. At the same time, multi-way interactions require a complex preparation protocol [140, 141]. Therefore, research in 3D conformation was focused on pairwise interactions [142]. In the wake of the 3rd generation sequencing revolution it has become possible to start interrogating the multi-way interaction landscape of the genome. In Chapter 6, we layout the first steps in revealing this higher level DNA interaction by exploiting the long-read sequencing platform MinION. Our approach, called Multi-Contact 4C (MC4C) (Figure 1.6), targets a specific region of the genome and unravels the multi-way interactions of functional elements in this locus [143].

Specifically, we focus on β-globin and PCDHa locus in mice where multi-way interactions between its genes and *enhancers* was speculated but never experimentally validated [144–148]. Using MC4C, we provide the first experimental validation that the individual enhancers of the β-globin locus in liver cells can cooperatively interact to form a spatial enhancer hub (i.e. commonly known as *LCR*). Additionally, we confirmed that the collection of enhancers in this LCR can physically accommodate two genes at a time.

It should be noted that 3D interactions between elements in the genome is speculated to be governed by numerous factors (e.g. CTCF, cohesin, etc.) many of which are still unknown [130, 149]. Another component in this complex regulatory system is *Wings APart-Like* (WAPL) protein, which is cohesin's DNA release factor. Without WAPL, cohesin remains bound to chromatin for longer periods of time [150]. Therefore, it was speculated that absence of WAPL would enable a given CTCF to engage with new CTCF partners over much larger distances (i.e. *loop extension*) [151]. However, experimental confirmation of this hypothesis required the assessment of multi-way interactions between multiple CTCF sites which, until now, was impossible due to the pairwise nature of the state of the art methods (interaction between A and B in addition to A and C do not imply interaction between all three elements). To address this question, we applied MC4C on WAPL deficient Hap1 cells to ascertain the validity of the loop extension hypothesis. Our experiments suggest that in the absence of WAPL, the reeled in CTCF sites are immobilized in the Cohesin loops. Ultimately, this “trapping” of CTCF sites in the Cohesin loops brings together distal CTCF sites and form a CTCF “traffic jam”.

1.6.4. CONTRIBUTIONS OF THIS THESIS

The contributions of this thesis can be summarized as follows. In Chapter 2 and 3 we describe several limitations in current Network-based Outcome Prediction (NOP) models and propose a novel method called FERAL that exploits various aggregation operators to represent diverse aberrations that may occur in tumors. In Chapter 4 and 5, we introduce SyNet which initially infers a gene network and then builds a NOP from the same data, exploiting synergistic effects between pairs of genes. We demonstrate how such a network not only improves performance beyond individual genes but also stabilizes the performance across independent datasets. We further show that SyNet corroborates well with existing biological networks which suggests that it can be used to discover new pathways that were missed in generic interactions networks.

In Chapter 6, we focus on DNA-DNA interactions and take the first steps in expansion from the pairwise to multi-way view of these networks. We demonstrate how these multi-way interactions can reveal higher order relationships between elements that were missed when assessing pairwise interactions. In Chapter 7, we focus on the computational aspect of multi-way 3D interactions analysis and explore prospective avenues to augment its efficiency and fidelity.

Taken together this thesis provides further insights into how networks can be inferred and used to improve breast cancer outcome prediction and delineates the starting point for further multi-way interaction assessments that could bring our understanding of complex diseases one step closer to elucidation.

REFERENCES

[1] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease,* Nature reviews genetics **12**, 56 (2011).

[2] *Early history of cancer | american cancer society,* https://www.cancer.org/cancer/cancer-basics/history-of-cancer/what-is-cancer.html (2018), accessed: 2018-1-23.

[3] R. L. Moodie, *Studies in Paleopathology: General Consideration of the Evidences of Pathological Conditions Found Among Fossil Animals. I* (Paul B. Hoeber, 1918).

[4] B. M. Rothschild, B. J. Witzke, and I. Hershkovitz, *Metastatic cancer in the jurassic,* Lancet **354**, 398 (1999).

[5] B. M. Rothschild, D. H. Tanke, M. Helbling, 2nd, and L. D. Martin, *Epidemiologic study of tumors in dinosaurs,* Naturwissenschaften **90**, 495 (2003).

[6] L. S. B. Leakey, *The stone age races of Kenya* (Oxford University Press, 1935).

[7] M. Avery and S. Mccarty, *Anecdotal, historical and critical commentaries on genetics,* Genetics **117**, 1 (1987).

[8] H. Land, L. F. Parada, and R. A. Weinberg, *Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes,* Nature **304**, 596 (1983).

[9] A. G. Knudson, Jr, *Mutation and cancer: statistical study of retinoblastoma,* Proc. Natl. Acad. Sci. U. S. A. **68**, 820 (1971).

[10] D. E. Comings, *A general theory of carcinogenesis,* Proc. Natl. Acad. Sci. U. S. A. **70**, 3324 (1973).

[11] H. Harris, O. J. Miller, G. Klein, P. Worst, and T. Tachibana, *Suppression of malignancy by cell fusion,* Nature **223**, 363 (1969).

[12] P. Armitage and R. Doll, *The age distribution of cancer and a multi-stage theory of carcinogenesis,* Br. J. Cancer **8**, 1 (1954).

[13] R. A. Weiss, *Multistage carcinogenesis,* Br. J. Cancer **91**, 1981 (2004).

[14] M. Greaves and C. C. Maley, *Clonal evolution in cancer,* Nature **481**, 306 (2012).

[15] C. L. Chaffer and R. A. Weinberg, *A perspective on cancer cell metastasis,* Science **331**, 1559 (2011).

[16] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, *Gene expression profiling predicts clinical outcome of breast cancer,* Nature **415**, 530 (2002).

[17] R. Siegel, E. Ward, O. Brawley, and others, *Cancer statistics,* CA Cancer J. Clin. (2011).

[18] S. R. L., M. K. D., and J. Ahmedin, *Cancer statistics, 2018,* CA: A Cancer Journal for Clinicians **68**, 7 (2018), https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21442 .

[19] K. W. Hunter, N. P. S. Crawford, and J. Alsarraj, *Mechanisms of metastasis,* Breast Cancer Res. **10 Suppl 1**, S2 (2008).

[20] I. J. Fidler, D. M. Gersten, and I. R. Hart, *The biology of cancer invasion and metastasis,* Adv. Cancer Res. **28**, 149 (1978).

[21] G. Poste and I. J. Fidler, *The pathogenesis of cancer metastasis,* Nature **283**, 139 (1980).

[22] D. Hanahan and R. A. Weinberg, *The hallmarks of cancer,* Cell **100**, 57 (2000).

[23] D. Hanahan and R. A. Weinberg, *Hallmarks of cancer: The next generation,* Cell **144**, 646 (2011).

[24] I. J. Fidler, *The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited,* Nat. Rev. Cancer **3**, 453 (2003).

[25] P. J. Stephens, P. S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D. C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, G. R. Bignell, L. R. Yates, E. Papaemmanuil, D. Beare, A. Butler, A. Cheverton, J. Gamble, J. Hinton, M. Jia, A. Jayakumar, D. Jones, C. Latimer, K. W. Lau, S. McLaren, D. J. McBride, A. Menzies, L. Mudie, K. Raine, R. Rad, M. S. Chapman, J. Teague, D. Easton, A. Langerød, Oslo Breast Cancer Consortium (OSBREAC), M. T. M. Lee, C.-Y. Shen, B. T. K. Tee, B. W. Huimin, A. Broeks, A. C. Vargas, G. Turashvili, J. Martens, A. Fatima, P. Miron, S.-F. Chin, G. Thomas, S. Boyault, O. Mariani, S. R. Lakhani, M. van de Vijver, L. van 't Veer, J. Foekens, C. Desmedt, C. Sotiriou, A. Tutt, C. Caldas, J. S. Reis-Filho, S. A. J. R. Aparicio, A. V. Salomon, A.-L. Børresen-Dale, A. L. Richardson, P. J. Campbell, P. A. Futreal, and M. R. Stratton, *The landscape of cancer genes and mutational processes in breast cancer,* Nature **486**, 400 (2012).

[26] I. Aksan and J. A. Stinson, *Piecing together the cancer puzzle,* Trends Biochem. Sci. **27**, 387 (2002).

[27] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, *Parallel human genome analysis: microarray-based expression monitoring of 1000 genes,* Proc. Natl. Acad. Sci. U. S. A. **93**, 10614 (1996).

[28] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton, and E. L. Brown, *Expression monitoring by hybridization to high-density oligonucleotide arrays,* Nat. Biotechnol. **14**, 1675 (1996).

[29] M. Habeck, *DNA microarray technology to revolutionise cancer treatment,* Lancet Oncol. **2**, 5 (2001).

[30] C. X. Ma and M. J. Ellis, *The cancer genome atlas: clinical applications for breast cancer,* Oncology **27**, 1263 (2013).

[31] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, METABRIC Group, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio, *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups,* Nature **486**, 346 (2012).

[32] B. Pereira, S.-F. Chin, O. M. Rueda, H.-K. M. Vollan, E. Provenzano, H. A. Bardwell, M. Pugh, L. Jones, R. Russell, S.-J. Sammut, *et al.*, *The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes,* Nature communications **7**, 11479 (2016).

[33] M. Teng, M. I. Love, C. A. Davis, S. Djebali, A. Dobin, B. R. Graveley, S. Li, C. E. Mason, S. Olson, D. Pervouchine, *et al.*, *A benchmark for rna-seq quantification pipelines,* Genome biology **17**, 74 (2016).

[34] M. Griffith, J. R. Walker, N. C. Spies, B. J. Ainscough, and O. L. Griffith, *Informatics for RNA sequencing: A web resource for analysis on the cloud,* PLoS Comput. Biol. **11**, e1004393 (2015).

[35] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, *A survey of best practices for RNA-seq data analysis,* Genome Biol. **17**, 13 (2016).

[36] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, *Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells,* PLoS One **9**, e78644 (2014).

[37] W. Zhang, Y. Yu, F. Hertwig, J. Thierry-Mieg, W. Zhang, D. Thierry-Mieg, J. Wang, C. Furlanello, V. Devanarayan, J. Cheng, *et al.*, *Comparison of rna-seq and microarray-based models for clinical endpoint prediction,* Genome biology **16**, 133 (2015).

[38] J. A. Thompson, J. Tan, and C. S. Greene, *Cross-platform normalization of microarray and rna-seq data for machine learning applications,* PeerJ **4**, e1621 (2016).

[39] P.-E. Colombo, F. Milanezi, B. Weigelt, and J. S. Reis-Filho, *Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction,* Breast Cancer Res. **13**, 212 (2011).

[40] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane, *Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification,* J. Natl. Cancer Inst. **95**, 14 (2003).

[41] C. M. Perou, S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. Lee, D. Lashkari, D. Shalon, P. O. Brown, and D. Botstein, *Distinctive gene expression patterns in human mammary epithelial cells and breast cancers,* Proc. Natl. Acad. Sci. U. S. A. **96**, 9212 (1999).

[42] R. H. Carlson, *Mammaprint assay & adjuvant chemotherapy use in early brca,* Oncology Times **38** (2016).

[43] R. Tibshirani, *Regression shrinkage and selection via the lasso,* Journal of the Royal Statistical Society. Series B (Methodological) **58**, 267 (1996).

[44] R. Tibshirani, *Regression shrinkage and selection via the lasso: a retrospective,* J. R. Stat. Soc. Series B Stat. Methodol. **73**, 273 (2011).

[45] N. Matloff, *Statistical Regression and Classification: From Linear Models to Machine Learning* (CRC Press, 2017).

[46] S. O. Zakharkin, K. Kim, T. Mehta, L. Chen, S. Barnes, K. E. Scheirer, R. S. Parrish, D. B. Allison, and G. P. Page, *Sources of variation in affymetrix microarray experiments,* BMC Bioinformatics **6**, 214 (2005).

[47] S. E. Clare and P. L. Shaw, *"big data" for breast cancer: where to look and what you will find,* Npj Breast Cancer **2**, 16031 EP (2016), review Article.

[48] J. Yli-Hietanen, A. Ylipää, and O. Yli-Harja, *Cancer research in the era of next-generation sequencing and big data calls for intelligent modeling,* Chinese Journal of Cancer **34**, 12 (2015).

[49] M. H. van Vliet, F. Reyal, H. M. Horlings, M. J. van de Vijver, M. J. T. Reinders, and L. F. A. Wessels, *Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability,* BMC Genomics **9**, 375 (2008).

[50] S. M. Gibbons, C. Duvallet, and E. J. Alm, *Correcting for batch effects in case-control microbiome studies,* PLOS Computational Biology **14**, 1 (2018).

[51] D. M. Leigh, H. E. L. Lischer, C. Grossen, and L. F. Keller, *Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths,* Molecular Ecology Resources **18**, 778 (2018), https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12779 .

[52] L. v. d. Maaten and G. Hinton, *Visualizing data using t-SNE,* J. Mach. Learn. Res. **9**, 2579 (2008).

[53] W. E. Johnson, C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical bayes methods,* Biostatistics **8**, 118 (2007).

[54] V. Nygaard, E. A. Rødland, and E. Hovig, *Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses,* Biostatistics **17**, 29 (2016).

[55] W. W. B. Goh, W. Wang, and L. Wong, *Why batch effects matter in omics data, and how to avoid them,* Trends Biotechnol. **35**, 498 (2017).

[56] M. Y. Park, T. Hastie, and R. Tibshirani, *Averaged gene expressions for regression,* Biostatistics **8**, 212 (2007).

[57] J. Das, K. M. Gayvert, F. Bunea, M. H. Wegkamp, and H. Yu, *ENCAPP: elastic-net-based prognosis prediction and biomarker discovery for human cancers,* BMC Genomics **16**, 263 (2015).

[58] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, *Tackling the widespread and critical impact of batch effects in high-throughput data,* Nat. Rev. Genet. **11**, 733 (2010).

[59] S. Michiels, S. Koscielny, and C. Hill, *Prediction of cancer outcome with microarrays: a multiple random validation strategy,* Lancet **365**, 488 (2005).

[60] C. Sotiriou and M. J. Piccart, *Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?* Nat. Rev. Cancer **7**, 545 (2007).

[61] P. J. Castaldi, I. J. Dahabreh, and J. P. A. Ioannidis, *An empirical assessment of validation practices for molecular classifiers,* Brief. Bioinform. **12**, 189 (2011).

[62] C. Bernau, M. Riester, A.-L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa, *Cross-study validation for the assessment of prediction algorithms,* Bioinformatics **30**, i105 (2014).

[63] G. Abraham, A. Kowalczyk, S. Loi, I. Haviv, and J. Zobel, *Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context,* BMC Bioinformatics **11**, 277 (2010).

[64] Y. Wang, J. G. M. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. J. J. Berns, D. Atkins, and J. A. Foekens, *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,* Lancet **365**, 671 (2005).

[65] L. Ein-Dor, O. Zuk, and E. Domany, *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer,* Proc. Natl. Acad. Sci. U. S. A. **103**, 5923 (2006).

[66] D. Venet, J. E. Dumont, and V. Detours, *Most random gene expression signatures are significantly associated with breast cancer outcome,* PLoS computational biology **7**, e1002240 (2011).

[67] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, *Network-based classification of breast cancer metastasis,* Molecular systems biology **3**, 140 (2007).

[68] K. Polyak, *Heterogeneity in breast cancer,* J. Clin. Invest. **121**, 3786 (2011).

[69] J. D. Amaral, J. M. Xavier, C. J. Steer, and C. M. Rodrigues, *The role of p53 in apoptosis,* Discov. Med. **9**, 145 (2010).

[70] F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains, *Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks,* Frontiers in cell and developmental biology **2**, 38 (2014).

[71] Y. R. Wang and H. Huang, *Review on statistical methods for gene network reconstruction using expression data,* Journal of Theoretical Biology **362**, 53 (2014), network-based biomarkers for complex diseases.

[72] N. Auslander, A. Wagner, M. Oberhardt, and E. Ruppin, *Data-driven metabolic pathway compositions enhance cancer survival prediction,* PLoS computational biology **12**, e1005125 (2016).

[73] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, *Protein-protein interaction networks (ppi) and complex diseases,* Gastroenterology and Hepatology from bed to bench **7**, 17 (2014).

[74] E. Davidson and M. Levin, *Gene regulatory networks,* Proceedings of the National Academy of Sciences **102**, 4935 (2005), http://www.pnas.org/content/102/14/4935.full.pdf .

[75] M. Kanehisa and S. Goto, *Kegg: kyoto encyclopedia of genes and genomes,* Nucleic acids research **28**, 27 (2000).

[76] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, *Human protein reference database—2009 update,* Nucleic Acids Research **37**, D767 (2009).

[77] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, *Diseases: Text mining and data integration of disease–gene associations,* Methods **74**, 83 (2015), text mining of biomedical literature.

[78] T.-K. Jenssen, A. Lægreid, J. Komorowski, and E. Hovig, *A literature network of human genes for high-throughput analysis of gene expression,* Nature genetics **28**, 21 (2001).

[79] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. Di Bernardo, *How to infer gene networks from expression profiles,* Molecular systems biology **3**, 78 (2007).

[80] W.-P. Lee and W.-S. Tzou, *Computational methods for discovering gene networks from expression data,* Briefings in Bioinformatics **10**, 408 (2009).

[81] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. von Mering, *The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible,* Nucleic Acids Research **45**, D362 (2017).

[82] A.-L. Barabasi and Z. N. Oltvai, *Network biology: understanding the cell's functional organization,* Nature reviews genetics **5**, 101 (2004).

[83] F. Kepes, *Biological Networks*, Complex systems and interdisciplinary science (World Scientific, 2007).

[84] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, *Dynamic modularity in protein interaction networks predicts breast cancer outcome,* Nature biotechnology **27**, 199 (2009).

[85] W. W. B. Goh and L. Wong, *Integrating networks and proteomics: Moving forward,* Trends Biotechnol. **34**, 951 (2016).

[86] C. Staiger, S. Cadot, B. Györffy, L. F. A. Wessels, and G. W. Klau, *Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis,* Front. Genet. **4**, 289 (2013).

[87] N. Alcaraz, M. List, R. Batra, F. Vandin, H. J. Ditzel, and J. Baumbach, *De novo pathway-based biomarker identification,* Nucleic Acids Res. **45**, e151 (2017).

[88] C. Staiger, S. Cadot, R. Kooter, M. Dittrich, T. Müller, G. W. Klau, and L. F. A. Wessels, *A critical evaluation of network and Pathway-Based classifiers for outcome prediction in breast cancer,* PLOS ONE **7**, e34796 (2012).

[89] S. Strunz, O. Wolkenhauer, and A. de la Fuente, *Network-Assisted disease classification and biomarker discovery,* Methods Mol. Biol. **1386**, 353 (2016).

[90] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables,* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 49 (2006).

[91] J. Friedman, T. Hastie, and R. Tibshirani, *A note on the group lasso and a sparse group lasso,* arXiv preprint arXiv:1001.0736 (2010), 1001.0736 .

[92] D. Hou and M. Koyutürk, *Comprehensive evaluation of composite gene features in cancer outcome prediction,* Cancer Inform. **13**, 93 (2014).

[93] A. Allahyar and J. de Ridder, *Feral: network-based classifier with application to breast cancer outcome prediction,* Bioinformatics **31**, i311 (2015).

[94] M. Kotlyar, C. Pastrello, N. Sheahan, and I. Jurisica, *Integrated interactions database: tissue-specific view of the human and model organism interactomes,* Nucleic Acids Res. **44**, D536 (2016).

[95] G. de Anda-Jáuregui, T. E. Velázquez-Caldelas, J. Espinal-Enríquez, and E. Hernández-Lemus, *Transcriptional network architecture of breast cancer molecular subtypes,* Front. Physiol. **7**, 568 (2016).

[96] C. S. Greene and O. G. Troyanskaya, *Chapter 2: Data-driven view of disease biology,* PLoS Comput. Biol. **8**, e1002816 (2012).

[97] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, D. I. Chasman, G. A. FitzGerald, K. Dolinski, T. Grosser, and O. G. Troyanskaya, *Understanding multicellular function and disease with human tissue-specific networks,* Nat. Genet. **47**, 569 (2015).

[98] E. Yeger-Lotem and R. Sharan, *Human protein interaction networks across tissues and diseases,* Front. Genet. **6**, 257 (2015).

[99] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. Carvunis, N. Simonis, J.-F. Rual, H. Borick, P. Braun, M. Dreze, J. Vandenhaute, M. Galli, J. Yazaki, D. E. Hill, J. R. Ecker, F. P. Roth, and M. Vidal, *Literature-curated protein interaction datasets,* Nature methods **6**, 39 (2009).

[100] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, *Comparative assessment of large-scale data sets of protein-protein interactions,* Nature **417**, 399 (2002).

[101] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, *Towards a proteome-scale map of the human protein-protein interaction network,* Nature **437**, 1173 (2005).

[102] M. A. Mahdavi and Y.-H. Lin, *False positive reduction in protein-protein interaction predictions using gene ontology annotations,* BMC Bioinformatics **8**, 262 (2007).

[103] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal, *A proteome-scale map of the human interactome network,* Cell **159**, 1212 (2014).

[104] J. Watkinson, X. Wang, T. Zheng, and D. Anastassiou, *Identification of gene interactions associated with disease from gene expression data using synergy networks,* BMC Syst. Biol. **2**, 10 (2008).

[105] B. K. Rajendran and C.-X. Deng, *Characterization of potential driver mutations involved in human breast cancer by computational approaches,* Oncotarget **8**, 50252 (2017).

[106] N. Wortham, E. Ahamed, S. Nicol, R. Thomas, M. Periyasamy, J. Jiang, A. Ochocka, S. Shousha, L. Huson, S. Bray, *et al.*, *The dead-box protein p72 regulates erα-/oestrogen-dependent transcription and cell growth, and is associated with improved survival in erα-positive breast cancer,* Oncogene **28**, 4053 (2009).

[107] C. Ploeger, N. Waldburger, A. Fraas, B. Goeppert, S. Pusch, K. Breuhahn, X. W. Wang, P. Schirmacher, and S. Roessler, *Chromosome 8p tumor suppressor genes sh2d4a and sorbs3 cooperate to inhibit interleukin-6 signaling in hepatocellular carcinoma,* Hepatology **64**, 828 (2016).

[108] L. Song, R. Chang, C. Dai, Y. Wu, J. Guo, M. Qi, W. Zhou, and L. Zhan, *Sorbs1 suppresses tumor metastasis and improves the sensitivity of cancer to chemotherapy drug,* Oncotarget **8**, 9108 (2017).

[109] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. A. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. de Santiago, L.-M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. W. Edwards, M. Nicodemi, and A. Pombo, *Complex multi-enhancer contacts captured by genome architecture mapping,* Nature **543**, 519 (2017).

[110] S. A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. M. Lai, A. A. Shishkin, P. Bhat, Y. Takei, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, L. Cai, P. McDonel, M. Garber, and M. Guttman, *Higher-order inter-chromosomal hubs shape 3d genome organization in the nucleus,* Cell (XXXX), 10.1016/j.cell.2018.05.024.

[111] S. Schoenfelder, T. Sexton, L. Chakalova, N. F. Cope, A. Horton, S. Andrews, S. Kurukuti, J. A. Mitchell, D. Umlauf, D. S. Dimitrova, C. H. Eskiw, Y. Luo, C.-L. Wei, Y. Ruan, J. J. Bieker, and P. Fraser, *Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells,* Nature Genetics **42**, 53 EP (2009), article.

[112] N. Gheldof, E. M. Smith, T. M. Tabuchi, C. M. Koch, I. Dunham, J. A. Stamatoyannopoulos, and J. Dekker, *Cell-type-specific long-range looping interactions identify distant regulatory elements of the cftr gene,* Nucleic Acids Research **38**, 4325 (2010).

[113] E. Apostolou, F. Ferrari, R. Walsh, O. Bar-Nur, M. Stadtfeld, S. Cheloufi, H. Stuart, J. Polo, T. Ohsumi, M. Borowsky, P. Kharchenko, P. Park, and K. Hochedlinger, *Genome-wide chromatin interactions of the <em>nanog</em> locus in pluripotency, differentiation, and reprogramming,* Cell Stem Cell **12**, 699 (2013).

[114] C. Sinoquet, *Probabilistic graphical models for genetics, genomics, and postgenomics* (OUP Oxford, 2014).

[115] S. Babaei, W. Akhtar, J. de Jong, M. Reinders, and J. de Ridder, *3D hotspots of recurrent retroviral insertions reveal long-range interactions with cancer genes,* Nat. Commun. **6**, 6381 (2015).

[116] D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young, *Activation of proto-oncogenes by disruption of chromosome neighborhoods,* Science **351**, 1454 (2016).

[117] P. C. Taberlay, J. Achinger-Kawecka, A. T. Lun, F. A. Buske, K. Sabir, C. M. Gould, E. Zotenko, S. A. Bert, K. A. Giles, D. C. Bauer, G. K. Smyth, C. Stirzaker, S. I. O'Donoghue, and S. J. Clark, *Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations,* Genome Research **26**, 719 (2016), http://genome.cshlp.org/content/26/6/719.full.pdf+html .

[118] M. J. Zeitz, F. Ay, J. D. Heidmann, P. L. Lerner, W. S. Noble, B. N. Steelman, and A. R. Hoffman, *Genomic interaction profiles in breast cancer reveal altered chromatin architecture,* PLOS ONE **8**, 1 (2013).

[119] C. L. Woodcock and S. Dimitrov, *Higher-order structure of chromatin and chromosomes,* Curr. Opin. Genet. Dev. **11**, 130 (2001).

[120] S. V. Razin and S. V. Ulianov, *Gene functioning and storage within a folded genome,* Cell. Mol. Biol. Lett. **22**, 18 (2017).

[121] T. Cremer and M. Cremer, *Chromosome territories,* Cold Spring Harb. Perspect. Biol. **2**, a003889 (2010).

[122] A. J. Fritz, A. R. Barutcu, L. Martin-Buley, A. J. van Wijnen, S. K. Zaidi, A. N. Imbalzano, J. B. Lian, J. L. Stein, and G. S. Stein, *Chromosomes at work: Organization of chromosome territories in the interphase nucleus,* J. Cell. Biochem. **117**, 9 (2016).

[123] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, *Topological domains in mammalian genomes identified by analysis of chromatin interactions,* Nature **485**, 376 (2012).

[124] A. Gonzalez-Sandoval and S. M. Gasser, *On TADs and LADs: Spatial control over gene expression,* Trends Genet. **32**, 485 (2016).

[125] A.-L. Valton and J. Dekker, *TAD disruption as oncogenic driver,* Curr. Opin. Genet. Dev. **36**, 34 (2016).

[126] N. H. Dryden, L. R. Broome, F. Dudbridge, N. Johnson, N. Orr, S. Schoenfelder, T. Nagano, S. Andrews, S. Wingett, I. Kozarewa, I. Assiotis, K. Fenwick, S. L. Maguire, J. Campbell, R. Natrajan, M. Lambros, E. Perrakis, A. Ashworth, P. Fraser, and O. Fletcher, *Unbiased analysis of potential targets of breast cancer susceptibility loci by capture Hi-C,* Genome Res. **24**, 1854 (2014).

[127] A. F. Schneider, *Untersuchungen über Plathelminthen / von Anton Schneider* (J. Ricker,, Giessen :, 1873).

[128] J. R. McIntosh and T. Hays, *A brief history of research on mitotic mechanisms,* Biology **5** (2016).

[129] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, *Capturing chromosome conformation,* Science **295**, 1306 (2002).

[130] A. Denker and W. de Laat, *The second decade of 3C technologies: detailed insights into nuclear organization,* Genes Dev. **30**, 1357 (2016).

[131] O. I. Kulaeva, E. V. Nizovtseva, Y. S. Polikanov, S. V. Ulianov, and V. M. Studitsky, *Distant activation of transcription: mechanisms of enhancer action,* Mol. Cell. Biol. **32**, 4892 (2012).

[132] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat, *Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c),* Nat. Genet. **38**, 1348 (2006).

[133] H. J. G. van de Werken, G. Landan, S. J. B. Holwerda, M. Hoichman, P. Klous, R. Chachik, E. Splinter, C. Valdes-Quezada, Y. Oz, B. A. M. Bouwman, M. J. A. M. Verstegen, E. de Wit, A. Tanay, and W. de Laat, *Robust 4c-seq data analysis to screen for regulatory DNA interactions,* Nat. Methods **9**, 969 (2012).

[134] E. de Wit, B. A. M. Bouwman, Y. Zhu, P. Klous, E. Splinter, M. J. A. M. Verstegen, P. H. L. Krijger, N. Festuccia, E. P. Nora, M. Welling, E. Heard, N. Geijsen, R. A. Poot, I. Chambers, and W. de Laat, *The pluripotent genome in three dimensions is shaped around pluripotency factors,* Nature **501**, 227 (2013).

[135] E. de Wit, E. S. M. Vos, S. J. B. Holwerda, C. Valdes-Quezada, M. J. A. M. Verstegen, H. Teunissen, E. Splinter, P. J. Wijchers, P. H. L. Krijger, and W. de Laat, *CTCF binding polarity determines chromatin looping,* Mol. Cell **60**, 676 (2015).

[136] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, *Comprehensive mapping of long-range interactions reveals folding principles of the human genome,* Science **326**, 289 (2009).

[137] S. Schoenfelder, M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B.-M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, E. Dimitrova, A. Dimond, L. B. Edelman, S. Elderkin, K. Tabbada, E. Darbo, S. Andrews, B. Herman, A. Higgs, E. LeProust, C. S. Osborne, J. A. Mitchell, N. M. Luscombe, and P. Fraser, *The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements,* Genome Research **25**, 582 (2015), http://genome.cshlp.org/content/25/4/582.full.pdf+html .

[138] T.-H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando, *Mapping nucleosome resolution chromosome folding in yeast by micro-c,* Cell **162**, 108 (2015).

[139] S. Sati and G. Cavalli, *Chromosome conformation capture technologies and their impact in understanding genome function,* Chromosoma **126**, 33 (2017).

[140] L. Yao, B. P. Berman, and P. J. Farnham, *Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes,* Crit. Rev. Biochem. Mol. Biol. **50**, 550 (2015).

[141] P. Olivares-Chauvet, Z. Mukamel, A. Lifshitz, O. Schwartzman, N. O. Elkayam, Y. Lubling, G. Deikus, R. P. Sebra, and A. Tanay, *Capturing pairwise and multi-way chromosomal conformations using chromosomal walks,* Nature **540**, 296 (2016).

[142] W. de Laat and F. Grosveld, *Spatial organization of gene expression: the active chromatin hub,* Chromosome Res. **11**, 447 (2003).

[143] A. Allahyar, C. Vermeulen, B. Bouwman, P. Krijger, M. Verstegen, G. Geeven, M. van Kranenburg, M. Pieterse, R. Straver, J. Haarhuis, H. Teunissen, I. Renkens, W. Kloosterman, B. Rowland, E. de Wit, J. de Ridder, and W. de Laat, *Locus-specific enhancer hubs and architectural loop collisions uncovered from single allele dna topologies,* bioRxiv (2017), 10.1101/206094.

[144] M. A. Bender, M. Bulger, J. Close, and M. Groudine, *Beta-globin gene switching and DNase I sensitivity of the endogenous beta-globin locus in mice do not require the locus control region,* Mol. Cell **5**, 387 (2000).

[145] X. Hu, S. Eszterhas, N. Pallazzi, E. E. Bouhassira, J. Fields, O. Tanabe, S. A. Gerber, M. Bulger, J. D. Engel, M. Groudine, and S. Fiering, *Transcriptional interference among the murine beta-like globin genes,* Blood **109**, 2210 (2007).

[146] S. Esumi, N. Kakazu, Y. Taguchi, T. Hirayama, A. Sasaki, T. Hirabayashi, T. Koide, T. Kitsukawa, S. Hamada, and T. Yagi, *Monoallelic yet combinatorial expression of variable exons of the protocadherin-α gene cluster in single neurons,* Nat. Genet. **37**, 171 (2005).

[147] P. Kehayova, K. Monahan, W. Chen, and T. Maniatis, *Regulatory elements required for the activation and repression of the protocadherin-alpha gene cluster,* Proc. Natl. Acad. Sci. U. S. A. **108**, 17195 (2011).

[148] S. Yokota, T. Hirayama, K. Hirano, R. Kaneko, S. Toyoda, Y. Kawamura, M. Hirabayashi, T. Hirabayashi, and T. Yagi, *Identification of the cluster control region for the protocadherin-beta genes located beyond the protocadherin-gamma cluster,* J. Biol. Chem. **286**, 31885 (2011).

[149] E. de Wit and W. de Laat, *A decade of 3C technologies: insights into nuclear organization,* Genes Dev. **26**, 11 (2012).

[150] H. Yu, *Chromosome biology: Wapl spreads its wings,* Curr. Biol. **23**, R923 (2013).

[151] J. H. I. Haarhuis, R. H. van der Weide, V. A. Blomen, J. O. Yáñez-Cuna, M. Amendola, M. S. van Ruiten, P. H. L. Krijger, H. Teunissen, R. H. Medema, B. van Steensel, T. R. Brummelkamp, E. de Wit, and B. D. Rowland, *The cohesin release factor WAPL restricts chromatin loop extension,* Cell **169**, 693 (2017).

2

FERAL: A NETWORK BASED CLASSIFIER

Amin Allahyar
Jeroen de Ridder

This chapter is published in Bioinformatics (2015) 31 (12): i311-i319 [1].

FERAL: Network Based Classifier with Application to Breast Cancer Outcome Prediction

Amin Allahyar and Jeroen de Ridder

2.1. Abstract

Breast cancer outcome prediction based on gene expression profiles is an important strategy for personalized patient care. To improve performance and consistency of discovered markers of the intial molecular classifiers, Network based Outcome Prediction methods (NOPs) have been proposed. In spite of the initial claims, recent studies revealed that neither performance nor consistency can be improved using these methods. NOPs typically rely on the construction of meta-genes by averaging the expression of several genes connected in a network that encodes protein interactions or pathway information. In this paper, we expose several fundamental issues in NOPs that impede the prediction power, consistency of discovered markers and obscure biological interpretation. To overcome these issues, we propose FERAL, a network-based classifier that hinges upon Sparse Group Lasso which performs simultaneous selection of marker genes and training of the prediction model. An important feature of FERAL, and a significant departure from existing NOPs, is that it uses multiple operators to summarize genes into meta-genes. This gives the classifier the opportunity to select the most relevant meta-gene for each gene set. Extensive evaluation revealed that the discovered markers are markedly more stable across independent datasets. Moreover, interpretation of the marker genes detected by FERAL reveals valuable mechanistic insight into the aetiology of breast cancer.
All scripts used in this manuscript are available for download at:
http://homepage.tudelft.nl/53a60/resources/FERAL/FERAL.zip.

2.2. Introduction

Breast cancer is the most frequently diagnosed type of cancer and one of the leading causes of death in women [2]. The main cause of death in these patients is, however, not the primary tumor, but its metastases at distant sites (e.g. in bone, lung, liver and brain) [3]. Typical risk factors such as lymph node status and tumor size are insufficient to accurately predict the risk of metastasis in patients [3, 4]. Over the last few years, substantial efforts have been made in deriving molecular classifiers to predict clinical outcome based on gene expression profiles obtained from the primary tumor [3, 5, 6].

A fundamental limitation of breast cancer outcome prediction is that it proved very difficult to obtain a robust classifier performance across different datasets. It was found that, despite properly cross-validated classifier training, prediction performance decreases dramatically when a classifier trained on one dataset is applied to another one [7, 8]. Moreover, the prognostic gene signatures identified using these classifiers have poor concordance across different studies [9, 10]. This points to a lack of a unified mechanism through which clinical outcome can be explained from gene expression profiles, which is still a major hurdle in clinical cancer biology.

Several studies ascribe the lack of classification robustness to insufficient patient sample size [11]. Other causes may be the inherent measurement noise in microarray experiments or heterogeneity in the samples [9, 12]. To mitigate these issues breast cancer datasets are often pooled in order to capture the information of as many samples as possible in the predictor [10, 13]. It remains, however, an open question how many samples are sufficient to account for all the noise and heterogeneity.

One of the hallmarks of cancer is that it is caused by deregulation of several processes or cellular pathways through multiple somatic mutations [14, 15]. More recent efforts of outcome prediction aim to exploit this hallmark by taking existing knowledge on relations between genes and pathways into account in the classifier. A common approach is to aggregate several functionally related genes to produce discriminative meta-genes or subnetworks [16–20]. Often, functional relationships between genes are determined based on the topology of a pre-defined biological network such as a co-expression network [21], cellular pathway map [22] or Protein-Protein Interaction (PPI) network [23]. Therefore we refer to such approaches as Network-based Outcome Prediction methods (NOPs).

The approach proposed by Park *et al.* is among the first NOPs [21]. Initially, the co-expression network is partitioned into gene sets using a linkage algorithm. Next, meta-genes are formed by taking the average expression of the genes in each gene set. Consequently, highly correlated genes will be aggregated which reduces the number of features as well as co-linearity among genes. The appropriate number of clusters, which determines the scale at which meta-genes are assembled, is determined by cross-validation.

Chuang *et al.* exploit the PPI network to identify predictive gene sets (called subnetworks in their work) [23]. Gene sets are constructed by a greedy procedure which starts with a gene (i.e. seed gene) and extends iteratively by adding the neighboring gene that provides the highest mutual information between corresponding average meta-gene and target label.

Taylor *et al.* exploit the topology of the PPI network [19]. In this method, predictive hub genes (i.e. genes with more than five connections) are ranked based on the absolute difference in within-class correlation between the hub and its neighbors. The corresponding meta-genes are constructed by taking the difference of expression between the hub and its neighbors.

Unfortunately, contrary to previous claims, recent studies reported that many NOPs do not outperform a model trained over single gene features [24–26]. Notably, in the analysis carried out by Staiger *et al.*, no significant improvement of classification performance nor an improvement of gene signature stability was observed, despite the fact that these authors examined many different methods and experimented with several biological networks [26]. Perhaps even more striking is the finding that utilizing random networks [25] or integrating random genes as markers [27] performs on par with complex NOPs. Taken together, it appears that current NOPs have produced very limited progress on solving the issue of robust classification performance and robust prognostic gene signature selection. This also casts doubt on the potential to extract useful insights from the derived prognostic gene signatures into the mechanisms underlying the disease.

The main goal of this paper is to identify and alleviate several fundamental issues in current NOPs that impede on reaching robust prediction performance and identify a stable prognostic gene signature. We find that the main bottleneck in current NOPs is that the frequently used average operator is a poor choice to integrate the expression of

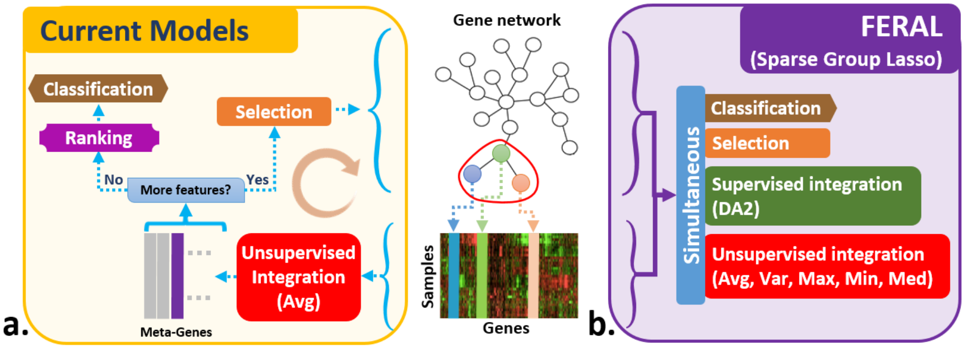


Figure 2.1: Overview of the proposed model (FERAL). **a.** Current models follow a similar path in which several nearby genes (according to a given network) are selected and then integrated using an average operator resulting in a meta-gene. These meta-genes are then ranked based on a pre-defined scoring function and top candidates are presented to the final classifier. **b.** Instead of being limited to average-based meta-genes, FERAL computes several meta-genes using different operators and employs the sparse group lasso to select the most appropriate meta-gene for each specific gene set while simultaneously performing selection, integration and classification.

functionally related genes. Moreover, the use of a single operator may not be sufficient to capture and summarize the aberration of higher level functions in cell. In addition, we conclude that decoupling the training of the classifier from the selection of genes to be used in meta-genes or the selection of the meta-genes themselves hampers the stability of gene signature identification.

To address these issues, we propose FERAL (Del**F**t n**E**two**R**k b**A**sed c**L**assifier), a new NOP method that is based on the Sparse Group Lasso (SGL) [28, 29]. SGL exploits groups of features (i.e. gene sets) and yields sparsity at both group (i.e. gene set) and feature (i.e. gene / meta-genes) levels [30]. In this way, simultaneous selection of features and training of the prediction model is achieved. Furthermore, instead of using a single operator to integrate gene-expression into meta-genes, FERAL exploits a wide range of such operators, including a previously unexplored supervised integration strategy.

We present extensive experiments using a compendium dataset called ACES (Amsterdam Classification Evaluation Suite), which was recently used for NOP model evaluation [26]. FERAL achieves statistically significant performance improvement, owing to the regularization of the SGL and inclusion of multiple integration operators. We moreover find substantially improved stability of the selected prognostic gene sets. Taken together, these feats enable biological interpretation of the trained classifier which, we find, results in highly relevant mechanistic insights.

2.3. METHOD

To motivate the design choices of FERAL we start by outlining the basic properties of existing NOPs. We focus on three well-known models proposed for network-based outcome prediction. Nonetheless, there are numerous network based methods which we do not take into consideration. A closer look at these methods reveals that in fact they all take two main steps to incorporate network information: gene set selection and integra-

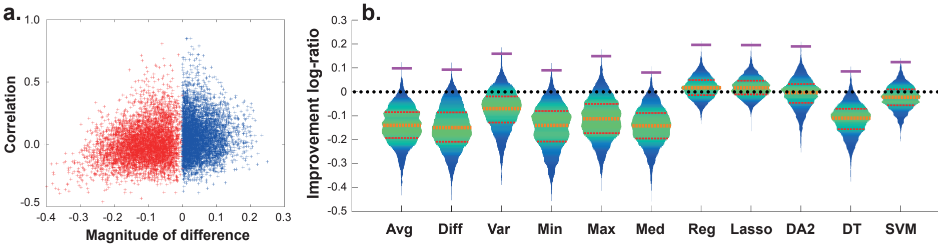


Figure 2.2: Evaluation of different integration operators. **a.** Visualization of the consistency in the direction of association with the target label for connected gene pairs in the I2D network. The x-axis represents the magnitude of difference, defined as $abs(C_a - C_b) \times \text{Sgn}(C_a \times C_b)$, where C_x denotes the correlation between gene x and the target label and Sgn is sign function. The y-axis is the correlation between two genes. **b.** Performance comparison between 11 operators including (from left to right): Average, Average of differences between seed gene and its interactors (implemented in Taylor), Variance, Minimum, Maximum, Median, Regression, Lasso, Direction Aware Average (DA2), Decision tree (DT) and Support Vector Machine (SVM) with a RBF kernel. To generate each violin plot, 5000 randomly selected gene sets were integrated into a meta-gene using one of the operators, and the predictive performance (AUC) is determined. The y-axis represents the improvement log-ratio of the AUC obtained with the meta-gene with the highest AUC of the individual genes. Purple lines indicate maximum ratio obtained in each distribution. This comparison shows that other operators are able to provide similar or even better performance compared to average operator. Interestingly, adjusting the direction of genes before taking the average can improve the performance considerably.

tion (Figure 2.1.a). The selection step should result in gene sets that represent (part of) a cellular process or pathway that collectively exhibit aberrant behavior. In the integration step the selected genes are summarized to produce a meta-gene capable of representing the aberrant behavior in the corresponding cellular process. Typically, this is followed by an additional round of selection and integration in which meta-genes are selected and integrated to produce a final prediction.

2.3.1. INTEGRATION OF GENE SETS INTO META-GENES

Most NOPs use the average operator to summarize gene expression into meta-gene expression. However, other biologically inspired operations, such as the max/min (to model AND/OR relations) or the variance (to capture variability of expression levels among genes close in the network) might also be suitable for representing higher level functions in cell. The assumption in many NOPs is that the directionality of the aberrant activity is the same (i.e. over/under expression) for nearby genes in the network. This may be inappropriate, for instance when genes exhibit opposite association w.r.t. the class label. In such cases the average operator can even cancel out their predictive contribution. By assessing the expression correlation of protein-protein interactions we established that this is a frequent event (Figure 2.2a and S3).

This problem arises because the aforementioned operators are unsupervised, i.e. an identical meta-gene would be produced using shuffled sample labels. This can be resolved by using a linear or non-linear regressor that considers the labels for achieving the best performance. In spite of their superior performance (See Figure 2.2b; S4), supervised integration operators may promote overfitting. This issue is apparent when linear operators are compared to non-linear ones (e.g. Decision Tree and SVM). Hence, in the integration procedure a trade-off exists between performance and complexity.

To alleviate this issue, we propose the Direction Aware Average (DA2) operator which adjusts the direction of genes before taking the average. DA2 for each gene g is defined as:

$$\mathrm{DA2_g} = \frac{1}{\left|\Psi_g\right|} \sum_{j\in\psi_g} \operatorname{sgn}\left(C_j\right) \times E_j,$$

where Ψ_g is the gene set of seed gene g and E_j and C_j contain the expression and correlation values with the class label of gene j, respectively. Just like all supervised meta-gene constructors, DA2 only uses training samples for calculating C_j. The DA2 provides a balance between stability of unsupervised operators (owing to its simplicity) and performance of supervised operators. It suffers less from overfiting due to the fact that labels are only employed to detect the direction of genes which is more stable compared to their individual predictive power. This is also apparent from our experiment (Figure 2.2b), as the DA2 provided a comparable performance to top integrating operators (e.g. regression and the Lasso).

It is worth noting that different integration operators offer different representations of higher level cellular functions. The proper operator for each gene set is not known a priori. It might be beneficial to use multiple of such operators, and allow the classifier to select the appropriate operator to describe a gene set, or allow a single gene set to be described using multiple operators. In addition to potentially achieving better performance, it provides insights into the underling aberrant behavior of each gene set. To the best of our knowledge, there are no NOPs that use multiple integration operators.

In FERAL, gene sets are formed by the individual gene expression profiles extended with several meta-genes produced by aggregating gene expression of these genes. We included the following unsupervised aggregations. The average operator, to model the overall expression level of the gene set in a fully unsupervised way. The median operator, similar to average but with reduced sensitivity to outliers. The variance operator, to measure the fluctuation in expression of interacting genes as this may point to a loss of regulation due to rewiring. Min and max, to model the AND/OR relationship between genes. In addition to these unsupervised operators, also supervised operators were included. The linear integration is implicitly provided by the SGL. DA2, as described above was also included. The non-linear integration methods, which are included in the analysis presented in Figure 2.2b were not included, since it was observed that they were prone to overfitting (data not shown).

2.3.2. SELECTION OF GENES IN GENE SETS

To determine which genes will be summarized in a meta-gene, Park selects all genes in a correlation cluster whereas Taylor uses all genes that are connected to the same hub gene in the PPI network. Both of these methods are likely to produce a highly skewed cluster size distribution, with a few very large clusters and many smaller ones [31, 32]. These large clusters will contain a substantial number of irrelevant genes that may not only hamper the performance, but also limit the interpretability of the meta-gene as it is difficult to identify the driver genes amongst all genes in the gene set [33]. Moreover, in

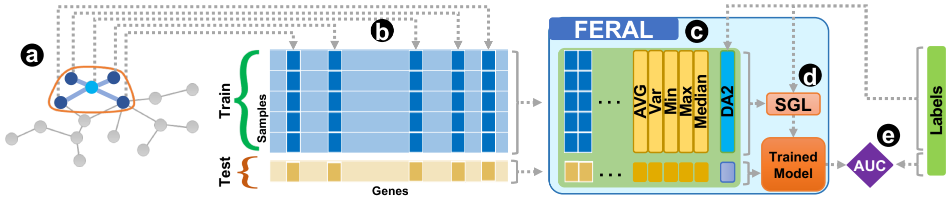


Figure 2.3: **Schematic of the training and testing procedures of FERAL. a.** In the first step, 10 genes are selected using given network. **b.** Corresponding genes in expression dataset are selected and normalized using z-score. **c.** Meta-genes are computed using the expression profiles of the gene set and target label (in case of a supervised integration). The expression of the individual genes are retained within the gene set. **d.** The Sparse Group Lasso is trained using training samples. **e.** Test samples are used to assess the prediction performance (in terms of AUC) in the current fold.

case of Taylor, only genes connected to hub genes can appear in a meta-gene, which a priori greatly limits the repertoire of genes that can be used in the final predictor.

Instead, in FERAL the gene set size is kept constant. This is achieved by defining gene sets as groups of ten genes - a seed gene with nine of its closest neighbors. Moreover, all genes were considered as seed genes, resulting in a total of N gene sets and ensuring each gene is included in at least one gene set. In case a seed gene has more than nine neighbors, the gene set is reduced to a total of ten genes by randomly removing genes. In case a seed gene has less than nine neighbors, the neighbors of the neighbors are considered in a similar fashion. When a weighted network is used, the edge weights are taken into account while determining the closest neighbors.

Chuang employs a greedy search to define subnetworks. This is done by iteratively extending the network from a seed gene guided by a supervised performance criterion. Because label information is used to guide the network growing, this increases the risk of overfitting and thereby reduces the stability of selected gene sets. Moreover, this procedure also critically depends on the accuracy of gene-gene interactions, which may be problematic as concerns exist about the reliability of individual interactions in these networks [34, 35].

Instead of including all genes in a group (Park and Taylor) or using a greedy search in a noisy network (Chuang), FERAL leverages the fact that the SGL performs embedded feature selection. This is realized because SGL provides regularization both at the level of the individual genes as well as the gene set level. As a result, selection of the most relevant genes will be performed if sufficiently large gene sets are provided. Because feature selection and classifier training are performed simultaneously, classifiers that offer embedded feature selection often provide improved performance and select more relevant features [36]. This approach also eliminates the need for additional cross-validation round that is often incorporated when a feature selection procedure is employed to reduce overfitting.

2.3.3. PRE-RANKING AND INTEGRATION OF META-GENES

After producing the meta-genes, most NOPs employ a ranking step. This step can be considered as a second selection step at the meta-gene level. Typically, each meta-gene is assessed based on a pre-defined ranking function (e.g. mutual information, t-test or permutation test) and the top candidates will be used in the final prediction step (akin to so-called individual feature selection). Evaluation of meta-genes in the methods of Chuang and Taylor is performed one at a time. Hence, the ranking procedure cannot identify multiple synergistic meta-genes when they have poor individual performance nor can it determine if several meta-genes contain the same information and are therefore redundant (see Figure S2.2 for an example of such cases in Chuang's method).

As FERAL employs the SGL, which performs embedded feature selection at the gene set level, the need of meta-gene selection is circumvented altogether. This greatly improves gene set stability.

2.3.4. Improvements on standard NOPs

To compare against, we use the methods from Park, Chuang and Taylor, henceforth referred to as standard methods. Based on our discussion so far it seems reasonable to change a few parts of these standard methods that evidently impede their performance. The original version of each method (prefixed by "o") is implemented by strictly following the procedure described in the author's paper. Additionally, we implemented an improved version (prefixed by "i") which includes obvious improvements beneficial for their performance and stability (See S2 for details). More specifically for Park's method, instead of training individual Lasso over the meta-genes produced in each level of hierarchical tree, single Lasso was trained over all meta-genes collected from levels of hierarchical tree. For Taylor's method, similar to Staiger *et al.*, we took the average of differences between hub and its interactor for corresponding meta-gene. Finally, we removed the ranking procedure in Taylor and Chuang methods and, similar to Park, used the Lasso to achieve a simultaneous selection and integration of the meta-genes. To assess the utility of biological networks in the outcome prediction problem we also included a Lasso trained on the individual genes, i.e. without exploiting network information.

2.3.5. Ranking and scoring of marker genes

One of the main objectives in NOPs is to detect marker genes that play a role in driving this complex disease. This can be achieved by ranking them on a pre-defined score that captures the contribution of the genes on the final prediction performance. In the Chuang method, gene sets (i.e. sub-networks) are ranked based on p-value that is obtained using a permutation test. In Taylor, the average difference of the correlation coefficient between classes is used. Finally in Park, the coefficients provided by lasso are used as gene sets score, which are subsequently propagated to the genes in the cluster. In FERAL, genes are scored based on the coefficients of the SGL. In addition, to take into account the contribution of the meta-genes in each feature group the largest meta-gene coefficient value is added to the score of the genes in the gene set. If a gene receives multiple scores, which is possible due to overlapping gene sets, the scores are averaged (See S5 for more details on the ranking of methods).

2.3.6. Implementation of FERAL

The implementation of the Sparse Group Lasso in this work is based on SLEP [37]. We further added a wrapper around this package to implement sample weighting to mitigate unbalanced classes along with a search for estimating the optimal parameters using an inner cross-validation. The following steps are taken to train FERAL (Figure 2.3). Initially, for all genes, nine of its closest neighbors are selected based on a gene network. After z-score normalization of the expression data, meta-genes are computed. Next, the SGL classifier is trained using the training samples. The parameters λ_1 and λ_2, which control the sparsity at the group level and within the groups, respectively, are determined by an inner cross-validation. Finally, the performance of the current fold is determined using the AUC measure.

2.4. Results and discussion

For evaluation of FERAL we use the Amsterdam Classification Evaluation Suite (ACES) [25], a cohort of 1606 breast cancer samples collected from 12 studies in NCBI's Gene Expression Omnibus (GEO) (see S7 for details). The label for each patient corresponds to recurrence free survival time with respect to a 5-year threshold ("good" vs. "poor" outcome). Three different networks are used in the evaluation: I2D, a PPI network also employed in Staiger *et al.*, a co-expression network and a random network. The co-expression network was defined on training data only, and thresholded at a correlation of 0.6. To produce the random network, we shuffle the nodes in the I2D network to destroy any biological knowledge while keeping its structure.

We used Area Under Curve (AUC) as the main measure of performance throughout the paper. Two types of cross-validation are considered. In the first type (sub-type stratified CV), the ratio of breast cancer sub-types is kept constant in the training and test set. In the second type (sampled leave-one-study-out CV), half of samples in each study is randomly selected (with replacement) while all samples from one study is excluded from selection and kept hidden as a test set. This configuration forms 12 folds, equal to the number of studies available in ACES. For both cross-validations, the indices of training and testing samples in each fold are kept identical across all methods.

2.4.1. Performance comparison

Figure 2.4a shows the obtained average AUCs for 10 repeats of the subtype stratified CV. As a first observation we note that the improved versions of the standard methods offer better performance, demonstrating the importance of simultaneous selection and integration of meta-genes. Interestingly, this improvement is most notable for Park's method, which achieves this improved performance despite the fact that the clearly sub-optimal average operation was used to construct meta-genes. A likely explanation for this improvement is that iPrk includes meta-genes at several different levels (i.e. meta-gene scales) in the hierarchical clustering tree. Apparently, it is important that the Lasso predictor can choose the best scale at which meta-genes are defined, suggesting that scale is another key factor in the performance of NOPs.

We moreover observe that FERAL offers superior performance across all three networks considered. This performance improvement is very significant (p-value $< 7 \cdot 10^{-8}$; paired t-test), obtained for the comparison with the best other method, iPrk using the co-expression network. This demonstrates that the SGL approach applied to gene sets containing several meta-gene definitions are beneficial in terms of predictive performance.

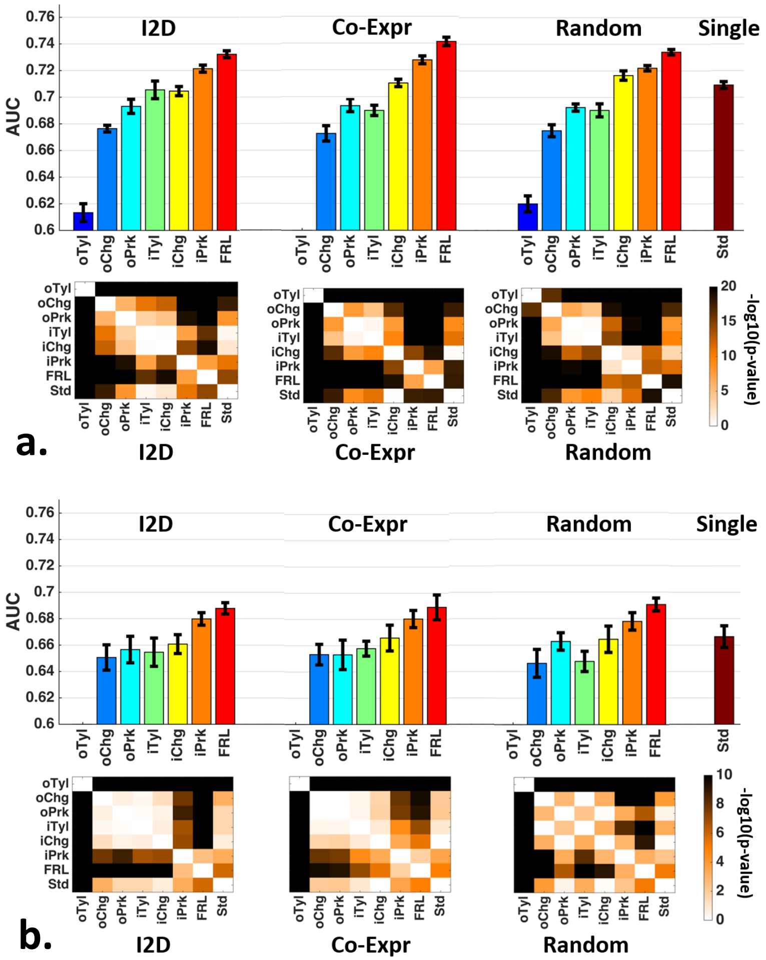


Figure 2.4: **Performance evaluation (AUC).** Performance of the methods under study for the protein-protein interaction network (I2D), a co-expression network (Co-Expr) and a random network (Random). We also added the result when a classical Lasso is employed (Single). Error bars denote the 95% confidence interval. The heatmaps indicate the p-value of the paired t-test between pairwise comparison of the AUCs of the individual CV folds. **a.** Sub-type stratified CV. **b.** Sampled leave-one-study-out CV.

Figure 2.4b shows the results for 10 repeats of the sampled leave-one-study-out CV. As expected, all classifiers showed performance reduction, but the general trends remain the same, that is, the standard methods performed poorly compared to their improved counterparts and FERAL significantly outperforms all other classifiers. It should be noted that, although FERAL achieves a better overall performance, the overall classification performance is relatively modest. It is likely that there is a limit on the maximum performance that can be achieved for the problem at hand (70% AUC). This is in line with previous observations [10, 26, 27].

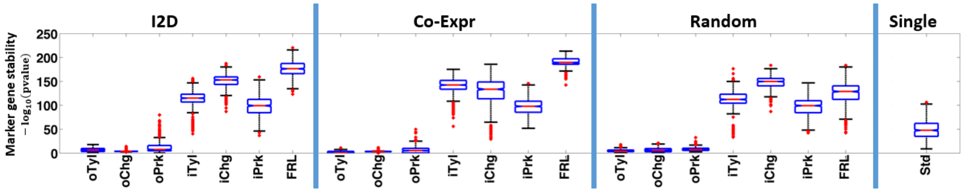


Figure 2.5: **Stability measurement (using Fisher's exact test) for three different networks including I2D, Co-Expr and random network.** The original version of the standard methods produced a much a lower overlap between folds due to pre-ranking of meta-genes. Similarly, Lasso produced a low overlap due to random selection of correlated features. FERAL obtained a higher gene set stability across folds for the I2D and Co-Expr network.

An interesting observation can be made from the performance of methods when the random network is utilized. As a general trend, all methods produced a comparable performance when networks that contain some biological information are used. The only exception is the oPrk method, which performs better when random network is used. Further investigation showed that genes with higher prognostic power often had higher degree in I2D network. For this reason, these genes would show up in large clusters diluting their predictive power after average integration. On the other hand, in the random network, they will mostly appear in the smaller clusters and can therefore contribute to the prediction of the Lasso [9, 25, 26]. The lack of a positive contribution on predictive performance of NOPs that use a biological network has been previously observed. The most likely explanation for this is the presence of large number of genes that are correlated with the target label which, in turn, makes it possible to construct many alternative features with comparable performance [9, 27].

2.4.2. STABILITY OF MARKER GENES

Finding robust marker genes is one of the key challenges in breast cancer research as prognostic gene signatures identified in independent datasets often show little to no overlap. To assess how FERAL and the (improved) standard methods perform in terms of signature stability we follow Staiger *et al.* and assess the stability of selected gene across folds by means of a Fisher's exact test. To this end, we measured the overlap between the top 100 genes selected by each of the methods in every fold (see S5 for details on these score functions). The leave-one-study-out CV was used without subsampling, resulting in a 12-fold cross-validation.

Figure 2.5 shows box plots of the marker gene stability for all pairwise comparisons between the 12 folds. It is striking to see that FERAL as well as the improved standard methods clearly have better marker gene stability compared to the standard methods (least significant p-value: $1.7 \cdot 10^{-52}$), which perform poorly, irrespective of the network employed. For the oChg and oTyl methods this can be explained by the fact that only very few meta-genes are used in the classifier, which apparently vary substantially between folds. The poor consistency for the oPrk method is caused by a combination of variability of the linkage tree and unstable regression coefficients resulting from the Lasso.

The concordance is highest for FERAL, which even has significantly improved marker gene stability compared to the improved standard methods (least significant p-value:

$1.8{\cdot}10^{-10}$). This demonstrates that FERAL's approach to refrain from a pre-filtering of top genes or gene sets and providing the embedded feature selection of SGL with all genes and many meta-genes using different operators is beneficial for marker gene stability.

Marker gene stability is also improved compared to the single gene classifier. This method performs a Lasso using all genes as predictors and therefore also no pre-filtering is applied in this method. Nevertheless, the overlap of marker genes between folds is still much lower than that obtained with FERAL (p-value: $5.3\cdot10^{-53}$). One explanation is that Lasso randomly selects features if they are highly correlated [38]. Another reason is that in different samples different - yet functionally related - genes play a role. As a result, in any subset of the data different marker genes will be selected. FERAL (and to some extent also the improved standard methods) are able to mitigate this by exploiting network information and summarize functionally related or interacting genes into meta-genes. This is supported by the observation that marker gene stability is significantly reduced when the random network is used (p-value: $1.6\cdot10^{-29}$). For the improved standard measures there is no significance different in case the random network is used. Thus, while utilizing network information does not improve performance, it is helpful in producing more stable sets of marker genes.

2.4.3. Functional enrichment of marker genes

If a NOP attains reasonable and robust performance and the marker genes selected across the folds are stable, the selected genes may be amenable to interpretation. This facilitates improved understanding of the underlying aberrant processes that play a role in this complex disease. One way to assess whether the methods under study are capable of detecting relevant genes is to compare the identified gene sets to already known cancer-related genes. To accomplish this, we collected nine cancer-related gene sets, including six cancer related GO terms. To measure the enrichment of cancer-related genes in a ranked list of genes produced by each method, we use an AUC measure. We also included a rankings based on a gene's individual AUC (indicated by Ind*) and one random ranking (indicated by Rnd*).

The observed enrichments obtained using the I2D network are depicted in Figure 2.7a. The results show that all methods have very modest enrichments not exceeding 0.6 for all but one cancer-related gene set. The notable exception is the enrichment obtained with FERAL, which is vastly superior and close to 0.7 for most cancer-related gene sets and 0.75 for two of them. The enrichment obtained using the Ind* ranking is generally poor, which confirms that differential expression analysis is unsuitable for finding genes involved in the disease. Surprisingly, we observed a severe reduction of gene enrichment using the Co-Expression network for all methods (See S6). This corroborates previous findings that protein-protein interaction networks capture regulatory interaction and functional relations [39].

Taken together, these observations support those made in Section 2.4.1 and 2.4.2, that is, incorporating network information does not improve performance, but it does contribute to stabilizing the marker gene sets and finding the biologically relevant genes.

Finally, we used BiNGO [40] to determine enrichment across all available gene sets. The hypergeometric test with a Benjamini Hochberg false discovery rate of 5% is performed for detecting over representation of the top 400 genes in the GO_Biological_Process

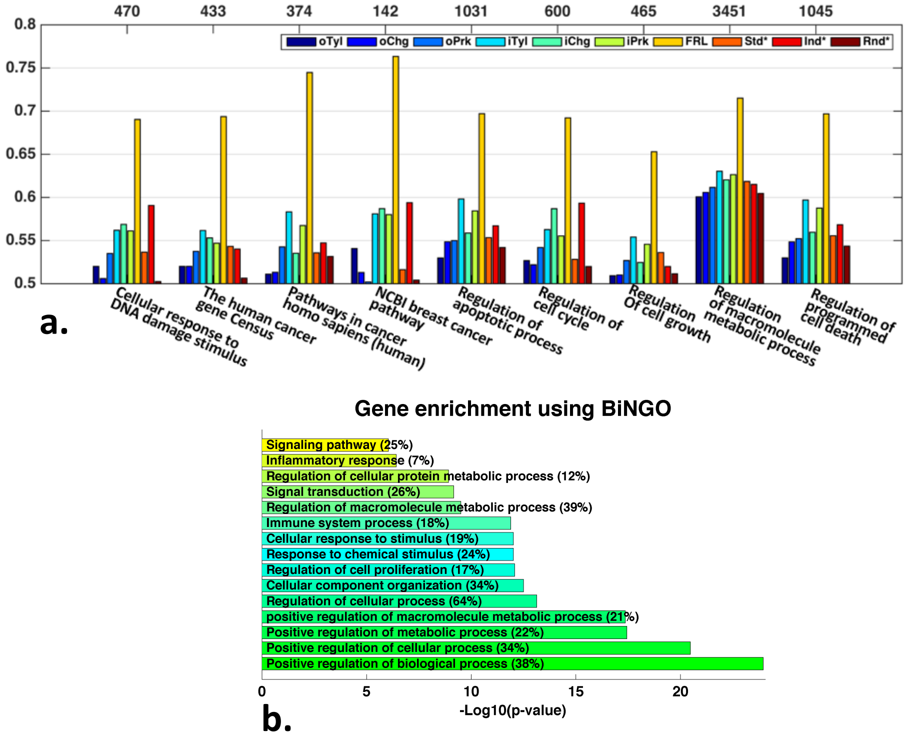


Figure 2.6: **Gene enrichment**. **a.** Gene enrichment of top genes for each method when the I2D network is employed. The values on top of each group represent the number of genes in each gene set. A notably increased enrichment is obtained using the gene sets produced by FERAL **b.** Result of top 15 gene enrichments by BiNGO applied to top 400 genes provided by FERAL.

category. The top 15 most enrichment GO categories are summarized in Figure 2.7b. Very significant enrichments are observed in various functional categories related to regulation, signaling and proliferation. This finding suggests that FERAL is able to uncover a wide diversity of genes that may play a role in the processes underlying breast cancer metastasis.

2.4.4. INTERPRETATION OF META-GENES IN FREQUENTLY SELECTED NETWORKS

Next, we investigated the selected gene sets and meta-genes by FERAL and determine whether they provide new insights into the mechanisms of breast cancer metastasis. To this end, we trained FERAL using the leave-one-study-out CV and obtained optimized λ_1 and λ_2. In this model, still about 1000 gene sets received non-zero coefficients. In an effort to reduce this further, while retaining the most essential ones, λ_2 was increased until the number of selected gene sets was less than 100 in each fold. The majority (66) of the selected gene sets were selected in at least 10 of the 12 folds, demonstrating that the selected gene sets were stable across folds. These 66 gene sets were then investigated

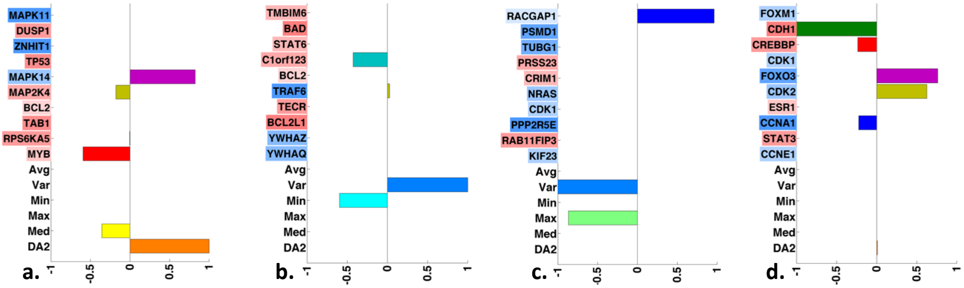


Figure 2.7: **Frequently identified gene-sets by FERAL.** The bars represent the median coefficient across folds, normalized to the range $\{-1, 1\}$ and represented along the x-axis. Y-axis represents four selected gene sets (10 genes) along calculated operators (e.g. Avg, std, etc.). Text background colors along y-axis indicate the expression correlation of corresponding gene with patient prognosis ranging from positive correlation (+1, full blue) to negative correlation (-1, full red).

for relevance to breast cancer in general and metastasis in particular.

We performed gene set enrichment for all 66 gene sets using BiNGO. The vast majority of gene sets (94%) were enriched (hypergeometric test with a Benjamini Hochberg false discovery rate of 5%) for key processes involved in cancer development, such as signaling of cell growth and survival, (regulation of) cell cycle, cell division, proliferation and apoptosis. This shows that FERAL is able to retrieve coherent sets of genes that are involved in cancer.

Figure 2.7 displays four of the selected gene sets, along with their median coefficient across the folds (horizontal bars) and association of the individual genes with the survival label (shading behind the gene names). We observe that for all gene sets, at least one of the genes was selected as a predictor in the final model. In the complete set of 66 gene sets there were 11 that exclusively used expression of individual genes. This corroborates the finding that it is important to supply the classifier with the actual expression profiles of the genes [16, 18]. Interestingly, in all four gene sets the direction of association with the class label was inconsistent between genes in the set. The average operator is therefore a poor choice to construct a meta-gene, and is consequently not used by the classifier.

In all four gene sets (and in 83% of all gene sets) a meta-gene obtained a non-zero coefficient. In three cases (and in 62% of all gene sets) even more than one meta-gene was selected. This demonstrates the importance of including multiple summarizations of the gene expression in addition to expression profiles of the genes. Finally, we note that the simple, yet effective, DA2 operator was selected in gene set (a). This was the case in 33% of all gene sets. Taken together, we observe that the final predictor was able to exploit both the raw gene expression profiles as well as a number of carefully constructed meta-genes.

Next we investigated each of the gene sets in Figure 2.7 using Ingenuity Pathway Analysis (IPA). Gene set (a) is strongly enriched for p38 MAPK signaling (p-value=$1.4 \cdot 10^{-14}$). There is ample evidence to suggest that MAPK signaling plays an important role in breast cancer, specifically through Notch regulation [41]. Interestingly, among the genes in this gene set is P53, which typically is not detected through differential expression analysis [23]. In this gene set P53 is also not directly selected, but is included in the final prediction model through the meta genes that are constructed using the DA2 and Median constructors. IPA also suggested a strong involvement of these genes in proliferation of T-lymphocytes (p-value= $1.5 \cdot 10^{-12}$). This is of particular interest as tumor-infiltrating lymphocytes may be a good biomarker and have recently been implicated in predicting response to neoadjuvant chemotherapy in breast cancer [42].

Gene set (b) was most enriched for PI3K / AKT signaling (p-value=$8.4 \cdot 10^{-8}$), which is one of the major pathways directly related to proliferation and cancer, and for which there exist promising therapeutic intervention possibilities [43]. For the genes in gene set (c), IPA revealed a strong enrichment for breast cancer regulation by stathmin-1, a downstream target of CDK1, which is included in gene set (c) (p-value=$1.4 \cdot 10^{-6}$). This gene set also included RCGAP1, which was recently shown to have prognostic significance in high-risk early breast cancer [44]. Finally, the gene set (d) was significantly enriched for estrogen-mediated S-phase entry (p-value=$2.9 \cdot 10^{-13}$). Estrogen is strongly implicated in breast cancer risk due to its role in promoting division of breast cells [45].

2.5. Conclusion

In this work, we proposed a network based outcome prediction method FERAL, that exploits network information in molecular classification of breast cancer outcome. Our method deviates from traditional NOPs in two important aspects. First of all, FERAL includes several different integration strategies to construct meta-genes, including a novel supervised integration strategy. Our results indicate that the final classification model frequently uses meta-genes produced by these constructors, often even multiple meta-genes based on the same gene set. This underscores the importance of extending traditional meta-genes based on a simple average. The second important improvement is that FERAL performs simultaneous selection and training of the classifier by employing the SGL. This mitigates the need for pre-ranking of genes and/or meta-genes, which is likely to severely reduce the stability of selected genes.

FERAL reached a significant performance increase compared to all standard NOP methods, including those that contained significant improvements made by us. This improvement was also obtained using a random network, leading to the conclusion that the biological knowledge encoded in the network is not used to obtain these improvements. The stability of marker genes improves substantially as a result of the procedure implemented in FERAL. This improvement was not observed when the random network was used, indicating that the biological knowledge contributes to the stability of the gene signatures.

Because FERAL attains robust performance and stable marker gene selection, the selected genes and gene sets might reveal insight into the underlying aberrant processes that play a role in this complex disease. We find that almost all of the gene sets used in the final model were enriched for cancer related processes. The four gene sets that were studied in more detail revealed very strong suggestive evidence for their involvement in breast cancer, with clear links to MAPK, PI3K and AKT signaling and regulation by stathmin-1. In summary, while classification performance of breast cancer outcome obtained with NOPs is unlikely to improve beyond ~70% AUC, we have shown that FERAL achieves much more stable marker gene selection that enables valuable mechanistic insight into the aetiology of breast cancer.

2.6. ACKNOWLEDGMENTS

Funding: This work was carried out on the Dutch national e-infrastructure with the support of the SURF Foundation. J. de Ridder is supported by the Netherlands Organization for Scientific Research (NWO-Veni: 639.021.233)

REFERENCES

[1] A. Allahyar and J. de Ridder, *Feral: network-based classifier with application to breast cancer outcome prediction,* Bioinformatics **31**, i311 (2015).

[2] A. Fantozzi and G. Christofori, *Mouse models of breast cancer metastasis,* Breast Cancer Research **8**, 212 (2006).

[3] B. Weigelt, J. L. Peterse, and L. J. van't Veer, *Breast cancer metastasis: markers and models,* Nature reviews cancer **5**, 591 (2005).

[4] C. L. Shapiro and A. Recht, *Side effects of adjuvant treatment of breast cancer,* New England Journal of Medicine **344**, 1997 (2001), pMID: 11430330, http://dx.doi.org/10.1056/NEJM200106283442607 .

[5] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, *Gene expression profiling predicts clinical outcome of breast cancer,* Nature **415**, 530 (2002).

[6] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, *A gene-expression signature as a predictor of survival in breast cancer,* New England Journal of Medicine **347**, 1999 (2002).

[7] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solís, R. Duque, H. Bersini, and A. Nowé, *Batch effect removal methods for microarray gene expression data integration: a survey,* Briefings in Bioinformatics (2012), 10.1093/bib/bbs037.

[8] C. Soneson, S. Gerster, and M. Delorenzi, *Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation,* PLoS ONE **9**, e100335 (2014).

[9] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, *Outcome signature genes in breast cancer: is there a unique set?* Bioinformatics **21**, 171 (2005), http://bioinformatics.oxfordjournals.org/content/21/2/171.full.pdf+html .

[10] M. H. van Vliet, F. Reyal, H. M. Horlings, M. J. van de Vijver, M. J. T. Reinders, and L. F. A. Wessels, *Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability,* BMC Genomics **9**, 375 (2008).

[11] J. Hua, W. D. Tembe, and E. R. Dougherty, *Performance of feature-selection methods in the classification of high-dimension data,* Pattern Recognition **42**, 409 (2009).

[12] W. F. Symmans, J. Liu, D. M. Knowles, and G. Inghirami, *Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions,* Human pathology **26**, 210 (1995).

[13] R. Shen, D. Ghosh, and A. M. Chinnaiyan, *Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data,* Bmc Genomics **5**, 94 (2004).

[14] D. Hanahan and R. A. Weinberg, *The hallmarks of cancer,* Cell **100**, 57 (2000).

[15] D. Hanahan and R. A. Weinberg, *Hallmarks of cancer: The next generation,* Cell **144**, 646 (2011).

[16] S. Babaei, E. Van Den Akker, J. De Ridder, and M. Reinders, *Integrating protein family sequence similarities with gene expression to find signature gene networks in breast cancer metastasis,* in *Pattern Recognition in Bioinformatics* (Springer, 2011) pp. 247–259.

[17] M. A. Pujana, J.-D. J. Han, L. M. Starita, K. N. Stevens, M. Tewari, J. S. Ahn, G. Rennert, V. Moreno, T. Kirchhoff, B. Gold, *et al.*, *Network modeling links breast cancer susceptibility and centrosome dysfunction,* Nature genetics **39**, 1338 (2007).

[18] E. B. Van den Akker, B. Verbruggen, B. T. Heijmans, M. Beekman, J. N. Kok, P. E. Slagboom, and M. Reinders, *Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis,* J Integr Bioinform **8**, 188 (2011).

[19] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, *Dynamic modularity in protein interaction networks predicts breast cancer outcome,* Nature biotechnology **27**, 199 (2009).

[20] P. Dao, R. Colak, R. Salari, F. Moser, E. Davicioni, A. Schönhuth, and M. Ester, *Inferring cancer subnetwork markers using density-constrained biclustering,* Bioinformatics **26**, i625 (2010), http://bioinformatics.oxfordjournals.org/content/26/18/i625.full.pdf+html .

[21] M. Y. Park, T. Hastie, and R. Tibshirani, *Averaged gene expressions for regression,* Biostatistics **8**, 212 (2007).

[22] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, *Inferring pathway activity toward precise disease classification,* PLoS Comput Biol **4**, e1000217 (2008).

[23] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, *Network-based classification of breast cancer metastasis,* Molecular systems biology **3**, 140 (2007).

[24] Y. Cun and H. Frohlich, *Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions,* BMC Bioinformatics **13**, 69 (2012).

[25] C. Staiger, S. Cadot, R. Kooter, M. Dittrich, T. Müller, G. W. Klau, and L. F. A. Wessels, *A critical evaluation of network and Pathway-Based classifiers for outcome prediction in breast cancer,* PLOS ONE **7**, e34796 (2012).

[26] C. Staiger, S. Cadot, B. Györffy, L. F. A. Wessels, and G. W. Klau, *Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis,* Front. Genet. **4**, 289 (2013).

[27] D. Venet, J. E. Dumont, and V. Detours, *Most random gene expression signatures are significantly associated with breast cancer outcome,* PLoS computational biology **7**, e1002240 (2011).

[28] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables,* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 49 (2006).

[29] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, *A sparse-group lasso,* Journal of Computational and Graphical Statistics **22**, 231 (2013), http://dx.doi.org/10.1080/10618600.2012.681250 .

[30] J. Friedman, T. Hastie, and R. Tibshirani, *A note on the group lasso and a sparse group lasso,* arXiv preprint arXiv:1001.0736 (2010), 1001.0736 .

[31] G. Chen, S. A. Jaradat, N. Banerjee, T. S. Tanaka, M. S. Ko, and M. Q. Zhang, *Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data,* Statistica Sinica **12**, 241 (2002).

[32] R. Albert, *Scale-free networks in cell biology,* Journal of cell science **118**, 4947 (2005).

[33] W. Cheng, X. Zhang, Z. Guo, Y. Shi, and W. Wang, *Graph-regularized dual lasso for robust eqtl mapping,* Bioinformatics **30**, i139 (2014).

[34] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, *Comparative assessment of large-scale data sets of protein-protein interactions,* Nature **417**, 399 (2002).

[35] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. Carvunis, N. Simonis, J.-F. Rual, H. Borick, P. Braun, M. Dreze, J. Vandenhaute, M. Galli, J. Yazaki, D. E. Hill, J. R. Ecker, F. P. Roth, and M. Vidal, *Literature-curated protein interaction datasets,* Nature methods **6**, 39 (2009).

[36] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)* (Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006).

[37] J. Liu, S. Ji, J. Ye, and Others, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University (2009).

[38] E. Grave, G. R. Obozinski, and F. R. Bach, *Trace lasso: a trace norm regularization for correlated designs,* in *Advances in Neural Information Processing Systems* (2011) pp. 2187–2195.

[39] R. Kelley and T. Ideker, *Systematic interpretation of genetic interactions using protein networks,* Nature biotechnology **23**, 561 (2005).

[40] S. Maere, K. Heymans, and M. Kuiper, *Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks,* Bioinformatics **21**, 3448 (2005).

[41] J. Izrailit, H. K. Berman, A. Datti, J. L. Wrana, and M. Reedijk, *High throughput kinase inhibitor screens reveal trb3 and mapk-erk/tgfβ pathways as fundamental notch regulators in breast cancer,* Proc Natl Acad Sci U S A **110**, 1714 (2013).

[42] Y. Mao, Q. Qu, Y. Zhang, J. Liu, X. Chen, and K. Shen, *The value of tumor infiltrating lymphocytes (tils) for predicting response to neoadjuvant chemotherapy in breast cancer: A systematic review and meta-analysis,* PLoS One **9**, e115103 (2014).

[43] N. M. Davis, M. Sokolosky, K. Stadelman, S. L. Abrams, M. Libra, S. Candido, F. Nicoletti, J. Polesel, R. Maestro, A. D'Assoro, L. Drobot, D. Rakus, A. Gizak, P. Laidler, J. Dulińska-Litewka, J. Basecke, S. Mijatovic, D. Maksimovic-Ivanic, G. Montalto, M. Cervello, T. L. Fitzgerald, Z. Demidenko, A. M. Martelli, L. Cocco, L. S. Steelman, and J. A. McCubrey, *Deregulation of the egfr/pi3k/pten/akt/mtorc1 pathway in breast cancer: possibilities for therapeutic intervention,* Oncotarget **5**, 4603 (2014).

[44] K. Pliarchopoulou, K. T. Kalogeras, R. Kronenwett, R. M. Wirtz, A. G. Eleftheraki, A. Batistatou, M. Bobos, N. Soupos, G. Polychronidou, H. Gogas, E. Samantas, C. Christodoulou, T. Makatsoris, N. Pavlidis, D. Pectasides, and G. Fountzilas, *Prognostic significance of racgap1 mrna expression in high-risk early breast cancer: a study in primary tumors of breast cancer patients participating in a randomized hellenic cooperative oncology group trial,* Cancer Chemother Pharmacol **71**, 245 (2013).

[45] J. S. Foster, D. C. Henley, A. Bukovsky, P. Seth, and J. Wimalasena, *Multifaceted regulation of cell cycle progression by estrogen: regulation of cdk inhibitors and cdc25a independent of cyclin d1-cdk4 function,* Mol Cell Biol **21**, 794 (2001).

3

FERAL: Supplementary Material

3.1. Lasso and its variants

To avoid redundancy, this section is omitted in the thesis. Please refer to supplementary material in the original publication [1], or to the introduction chapter of this thesis (Chapter 1) for details about Linear Regression (section 1.2.3), Lasso (section 1.2.5) and its derivatives Group Lasso and Sparse Group Lasso (section 1.3.5).

3.2. A brief overview of previous network based outcome prediction models

In this section we will provide a brief overview of previous network based methods proposed for breast cancer outcome prediction problem. The aim is to explain their procedure with focus on providing an insight on how they are essentially similar. We will also mention their strong and weak points.

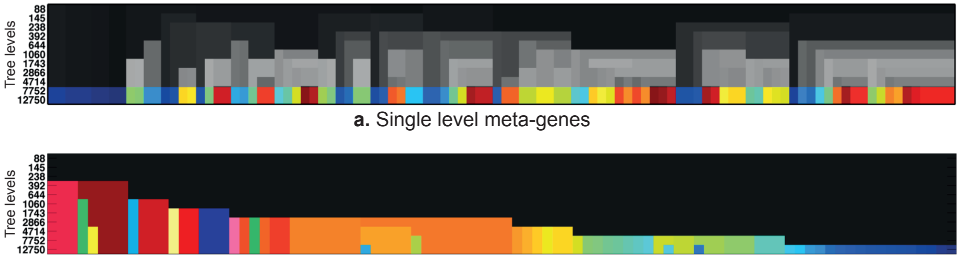
3.2.1. Park

The main goal in Park's method is to reduce the collinearity among the genes, which results in large variance of the estimates and inaccurate prediction [2]. They proposed a simple yet efficient method which utilizes the noise reduction property of the average operator to solve this issue. In the first step, they applied hierarchical clustering with correlation as similarity measure. This will produce dendrogram that exposes the nested correlation structure. At each level, a meta-genes will be constructed per set by computing the average expression of genes in that particular set. In other words, they simply aggregated the highly correlated genes, which not only eliminate the co-linearity among genes but also reduces the number of features. As the next step, they trained a Lasso regression over each level of dendrogram and selected the best level using cross-validation.

Although the collinearity among genes can be reduced using Park's method, the standard average linkage will provide a highly skewed distribution of cluster sizes with few large clusters along with thousands of small ones many of which even contain only a single gene [3]. In this situation, the constructed meta-gene might potentially lose its performance if enclosing genes have a different sign of correlation with the target label (See 3.3). On top of that, a single Lasso is trained on the meta-genes obtained for one level of the hierarchical clustering tree (Figure 3.1.a). Therefore, this model cannot exploit synergies from two meta-genes that arise at different levels of the clustering tree. This is not in line with biology, as cellular functions arise at different scales [4, 5] (Figure 3.1.b).

To validate the multi-scale property of the problem at hand, we constructed all possible meta-genes using every levels of the given linkage hierarchical tree. In the next step, Lasso is trained over all of these meta-genes. It should be noted that in the new method, the scale parameter is eliminated from model. Therefore, only the Lambda parameter should be determined automatically using an inner cross-validation step. The result of this experiment is represented in Figure 3.1. It can be observed that useful meta-genes might be found at different levels. It is interesting to observe that Lasso identified few parent meta-genes in addition to their child. In other words, a predictive meta-gene can further produce new information in higher levels of hierarchical tree when aggregated with other meta-genes. This shows that outcome prediction is not only multi-scale but also hierarchical.

Based on the observed improvement, we considered this strategy to add the multi-scale support to the improved version of Park method.



a. Single level meta-genes

b. Meta-genes from all levels

Figure 3.1: Multi-scale property of breast cancer outcome prediction. **a.** Demonstration of 79 selected meta-genes (in color) and their corresponding parents (in black and white) in a level (7752 clusters) of hierarchical tree in original version of Park's method. **b.** Result of employing all possible meta-genes for training single Lasso. The meta-genes with non-zero coefficient is demonstrated in color. The height of each cluster represents the level in which it is selected. It is clear that predictive meta-genes are identified in different levels which signify the multi-scale property of problem at hand.

3.2.2. CHUANG

Chuang et. al used sub-networks of Protein-Protein Interaction (PPI) network to identify several predictive sub-networks (a set of functionally related genes) [6]. A meta-gene is constructed by taking the average expression of genes in the corresponding sub-network. Each sub-network has a score which is defined as the mutual information between labels and the corresponding meta-gene. The sub-network selection procedure is a greedy method. In each step, the current sub-network will be expanded by adding the nearest gene in the corresponding PPI network. The expansion will end when its score (mutual information) stops increasing. This is a powerful method to quickly find discriminative sub-networks and it also supports overlapping gene sets. At end, the top meta-genes are sequentially added to a logistic regression model until no further improvement in the performance is observed.

Apart from being limited to average operator, assessing gene sets and using top sub-networks solely based on their discriminative power in the training set might cause the final model to over fit. In addition, greedy search for predictive sub-network over a noisy network might not provide an ideal solution. Finally, sequential selection of top sub-networks for final model might select sub-networks with identical enclosed information.

In order to evaluate the effect of pre-ranking meta-genes, we considered the top 500 meta-genes produced in original method of Chuang. Following the corresponding definition we sequentially added these genes and its corresponding performance is measured using AUC (Figure 3.2; blue curve). On the other hand, we trained a lasso over all of these meta-genes and measured its performance (Figure 3.2; red line). Based on this experiment, Lasso achieved higher performance compared to logistic regression. It is interesting to observe that the predictive meta-genes are selected irrespective of their ranking. This demonstrates that ranking is inadequate in this method. For the improved version of this method, we followed a similar path and applied Lasso to all identified meta-genes instead of the top ones.

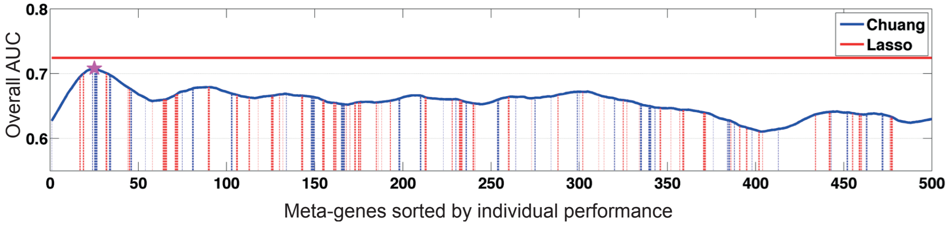


Figure 3.2: Performance comparison for sequential vs. global selection of meta-genes. Blue curve represents the step by step performance of logistic regression when i-th top meta-gene is added to the model. In the original version of this method, the addition terminates when performance stops increasing (marked with yellow star). However, the performance can be improved if all these meta-genes are utilized using Lasso (~71%; represented with red horizontal line). The vertical lines represent the identified meta-genes. The wideness and color of these lines represent the prominence and sign of corresponding coefficient (red and blue represent negative and positive respectively). Interestingly, it can be observed that predictive features are not identified from top meta-genes and many important feature are positioned after breaking point.

3.2.3. TAYLOR

Taylor et al. looked for predictive hub genes (i.e. genes with more than five connections). Each hub is scored based on the absolute difference of within-class correlation between the hub and its neighbors:

$$S_h = \frac{\sum_{i \in N(h)} Cr^{-1}(E_h, E_i) - Cr^{+1}(E_h, E_i)}{|N(h)|}$$

Where E_h indicates the hub's expressions and $Cr^k(E_h, E_i)$ indicates the Pearson's correlation of h and i genes calculated from samples of class k. In addition, $N(i)$ produces a set of genes that have a direct link to gene i. In addition |.| specify the number of items in a set. The corresponding meta-genes are constructed by taking the average difference of expression between the hub and its neighbors:

$$F_h = \frac{\sum_{i \in N(h)} E_h - E_i}{|N(h)|}$$

The Taylor method can detect the sub-networks that change in the correlation of their enclosing genes is prognostic of metastasis. While this is in line with biology, the top ranked sub-networks using this procedure might not be predictive. This is because a strong change of correlation between classes does not indicate the separability of classes. On the other hand, two completely separated classes might show a similar correlation between features.

Apart from that, the hub genes regularly have high degrees and most probably not all of these interactions are predictive. In order to investigate this issue, we applied Lasso to the BRCA1 and SP1 interactors that was identified in the author's paper [7]. Surprisingly, we observed that a model using 111 genes (Figure 3.3.a) deliver a similar performance compared to a model which uses only two genes (i.e. CCNB1 and ESR1) (Figure 3.3.b). These genes are known to be important in breast cancer sub-typing.

In addition, the meta-gene integration operator is in fact a special case of linear integration where hub and its interactors are averaged with −1 and +1 weights respectively. Finally, in the Taylor method, similar to Park and Chuang a pre-ranking and incremental addition of top meta-genes are considered for meta-genes before the final training step which might impede the performance and stability of identified genes. In the improved version of this method (i.e. iTyl), we excluded the ranking procedure and used all meta-genes to train Lasso classifier.

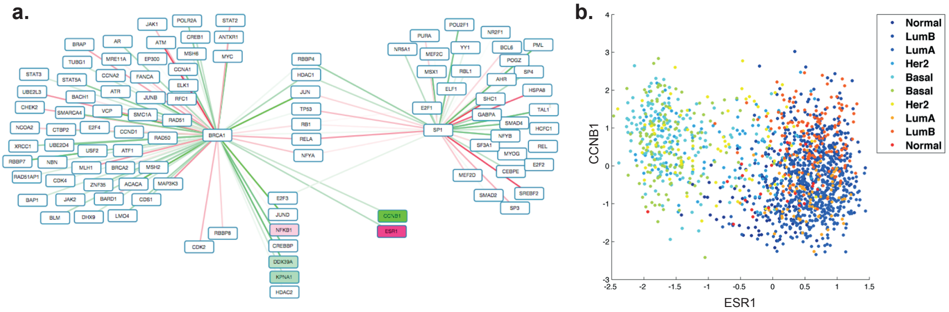


Figure 3.3: Distribution selection (Taylor) vs. Predictive selection (Lasso). **a.** Demonstration of a predictive sub-network identified by Taylor's method and corresponding predictive genes found by Lasso. While Taylor achieved 0.682 AUC with 111 genes, Lasso identified two genes which produces a similar performance (AUC=0.679). **b.**Visualization of two gene expressions (i.e. ESR1 and CCNB1) which was identified by Lasso. These genes are known to be important in sub-type breast cancer identification.

3.3. DIRECTION OF ASSOCIATION OF NEARBY GENES

In NOPs, meta-genes are often constructed by average operator. However, this operator would lose its prediction power if included genes in the gene set have different sign of correlation with the target label. In this experiment we would like to assess the probability of observing this situation in the nearby genes. Let c_i be the correlation between expression of gene i and target label. It is easy to check that the average meta-gene produced from gene i and j lose its predictive power if $c_i \times c_j < 0$. In order to assess the frequency of this event in I2D network, we selected all non-singleton genes available in ACE as seed ($n = 9871$) along with their closest neighbor. For each pair, we computed two values as follows:

$$x_{axis} = abs\left(c_i - c_j\right) \times sign\left(c_i \times c_j\right)$$
$$y_{axis} = corr\left(e_i, e_j\right)$$

Where e_i represents the expression of gene i and c_i denote the correlation between e_i and target label. In addition, $sign$ and $corr$ signify the sign and correlation functions respectively. It is worth noting that the first term in x_{axis} provides a magnitude for this issue. Result of plotting these measures is represented in Figure 3.4.a. We observed a different direction between gene pairs in roughly 49.3% of the cases (4866 out of 9871). It is important to note that this high frequency is observed in gene pairs only. Generally, gene sets consist of more than two genes and hence chance of observing different sign of correlation increases accordingly. Based on this result, it is evident that average operator would frequently lose performance in gene sets created using I2D. To remove the effect of this network, we plotted the same measures for all combination of genes included in ACE dataset in Figure 3.4.b. We observed a similar frequency in this experiment (49.9% vs. 50.1% for negative vs. positive sign respectively). As a final note, it can be observed from two experiments that an average meta-gene constructed from correlated genes suffers less from this issue compared to uncorrelated ones. This suggests the utility of Co-Expression networks in NOPs which utilize average based meta-gene.

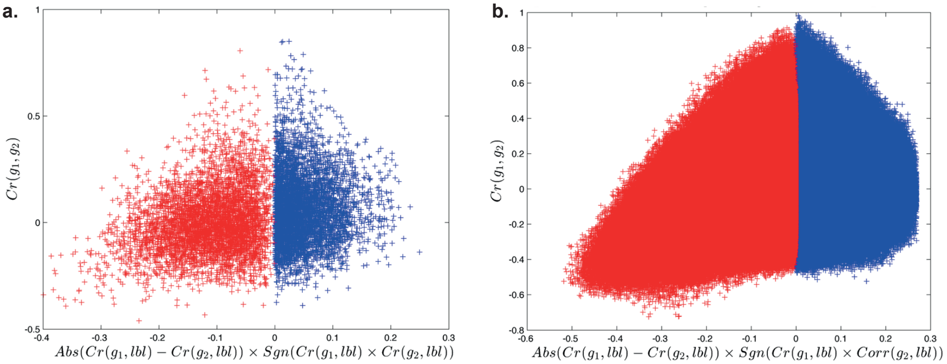


Figure 3.4: Direction of nearby genes. **a.** Distribution of direction for 9871 seed genes and their closest neighbor. Nearly half of the nearby genes (4866 vs. 5005) showed a different sign of correlation with target label compared to the seed gene. **b.** Distribution of genes for all combination of genes ($n = 81274875$) included in ACE. Similar frequency (~50%) for observing a different sign of correlation with target label for gene pairs is observed.

3.4. Performance comparison of different operators

In order to compare the average-based aggregation with other aggregation operators and methods, we created 5000 gene sets based on one randomly selected gene (i.e. a seed gene) and its 10 closest neighbors according to the PPI network. For each gene set, we computed meta-genes using 11 different aggregation operators. To determine the performance of these meta-genes, we calculated the ratio of AUC (Area Under Curve) obtained using the produced meta-gene and the gene with highest individual AUC in the gene set. Hence, values larger than one indicate improvement of the meta-gene over the best individual gene in the gene set. Result of such experiment is demonstrated in Figure 2.3.b.

We repeated the same experiment for Co-Expression and random network. Results of such experiment are demonstrated in 3.5. We also find no improvement if we utilize a random network for finding neighbor genes.

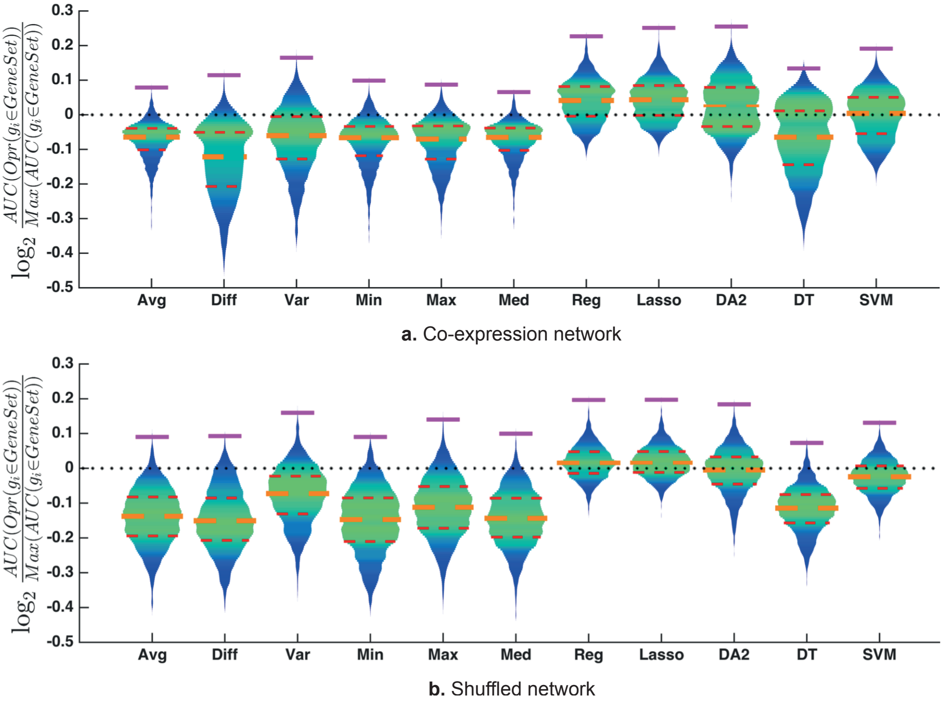


Figure 3.5: Comparison between operators for 5000 gene sets. **a.** Co-expression network produces a better gene sets compared to I2D and random network. **b.** A comparable performance could be observed when random network is utilized.

3.5. SCORING FUNCTIONS

Each method under study has its own ranking strategy. The following procedures are considered for identifying top genes:

- oPrk/iPrk: After training Lasso over generated meta-gene and determining the optimal Lambda, each meta-gene has a corresponding coefficient in the optimal model. The corresponding coefficients are assigned to enclosing genes within each meta-gene.

- oChg/iChg: The p-values obtained from the final permutation test in Chuang's method is considered for score of each sub-network. Next, these scores are propagated to the enclosing genes inside each sub-network. For original version of this method we only considered the p-values of sub-networks that are included in the final model (i.e. incremental addition of sub-networks) are considered.

- oTyl/iTyl: For each sub-network, the difference of correlation between hub and its interactors is considered as it's score. Next, these scores are propagated to the enclosing genes inside each sub-network. For original version, we only considered the p-values of sub-networks that are included in the final model (i.e. incremental addition of sub-networks) are considered.

- Std: After training Lasso and determining its optimal Lambda, each gene has a dedicated coefficient which can be considered as its score for current fold.

- FERAL: After training the SGL for current fold and determining its corresponding optimal λ_1 and λ_2, the λ_1 in SGL increased while the optimal λ_2 kept constant until less than 100 groups have at least one non-zero coefficient. Then for each gene set, the maximum coefficient value of its meta-genes are added to the coefficient of encompassing genes. In other words, each gene gets two scores for each gene set: its individual coefficient and the maximum value of coefficient of meta-genes.

After collecting the scores, if multiple scores are assigned to a gene (resulted from overlapping gene sets), average of these scores are considered for the final score of this gene in current fold. On the other hand, if no score is assigned to a gene, a random value that is smaller than smallest score in the set is assigned for this gene.

3.6. RESULTS OF GENE ENRICHMENT USING CO-EXPRESSION AND SHUFFLED VERSION

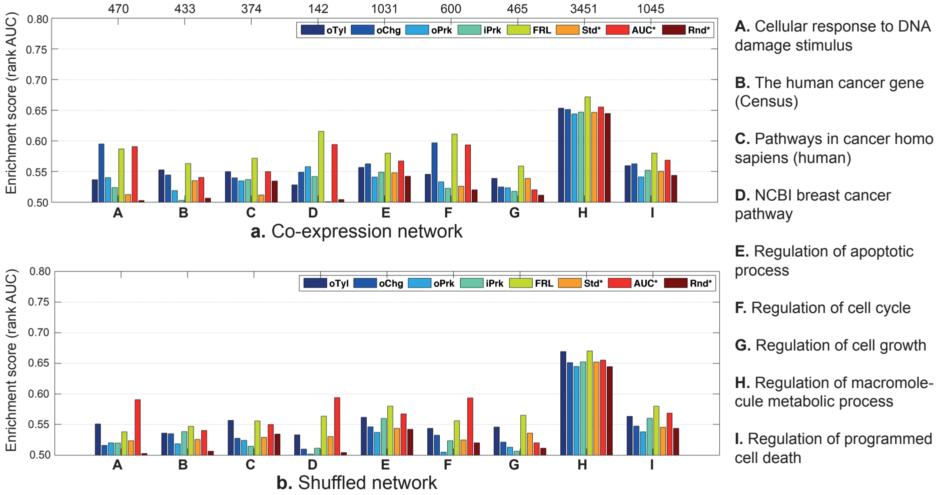


Figure 3.6: Gene enrichment for Co-Expression and a shuffled version. **a.**Enrichment of markers identified by FERAL when co-expression network is employed. **b.** Similar enrichment is obtained when a shuffled version of network is utilized.

3.7. DETAILS OF ACES STUDIES

Table 3.1: Collected studies in ACES and their specifications

Dataset	Geo accession (GSE)	No. of poor	No. of good
Ivshina	4922	30	72
Hatzis-Pusztai	25066	102	48
Desmedt-June07	7390	56	127
Minn	2603	21	44
Miller	3494	21	68
WangY-ErasmusMC	2034	88	169
Schmidt	11121	24	145
Pawitan	1456	33	114
Symmans	17705	37	187
Loi	6532	24	33
Zhang	12093	9	112
WangY	5327	10	42
Total		**445**	**1161**

REFERENCES

[1] A. Allahyar and J. de Ridder, *Feral: network-based classifier with application to breast cancer outcome prediction,* Bioinformatics **31**, i311 (2015).

[2] M. Y. Park, T. Hastie, and R. Tibshirani, *Averaged gene expressions for regression,* Biostatistics **8**, 212 (2007).

[3] G. Chen, S. A. Jaradat, N. Banerjee, T. S. Tanaka, M. S. Ko, and M. Q. Zhang, *Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data,* Statistica Sinica **12**, 241 (2002).

[4] J. de Ridder, A. Uren, J. Kool, M. Reinders, and L. Wessels, *Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens,* PLOS Computational Biology **2**, 1 (2006).

[5] J. de Ridder, J. Kool, A. Uren, J. Bot, L. Wessels, and M. Reinders, *Co-occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes,* Bioinformatics **23**, i133 (2007).

[6] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, *Network-based classification of breast cancer metastasis,* Molecular systems biology **3**, 140 (2007).

[7] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, *Dynamic modularity in protein interaction networks predicts breast cancer outcome,* Nature biotechnology **27**, 199 (2009).

4

SyNet: Synergistic gene pairs

Amin Allahyar
Joske Ubels
Jeroen de Ridder

This chapter is published in bioRxiv [1], and is under revision.

A data-driven interactome of synergistic genes improves network-based cancer outcome prediction

Amin Allahyar, Joske Ubels and Jeroen de Ridder

4.1. Abstract

Robustly predicting outcome for cancer patients from gene expression is an important challenge on the road to better personalized treatment. Network-based outcome predictors (NOPs), which considers the cellular wiring diagram in the classification, hold much promise to improve performance, stability and interpretability of identified marker genes. Problematically, reports on the efficacy of NOPs are conflicting and for instance suggest that utilizing random networks performs on par to networks that describe biologically relevant interactions. In this paper we turn the prediction problem around: instead of using a given biological network in the NOP, we aim to identify the network of genes that truly improves outcome prediction. To this end, we propose SyNet, a gene network constructed ab initio from synergistic gene pairs derived from survival-labelled gene expression data. To obtain SyNet, we evaluate synergy for all 69 million pairwise combinations of genes resulting in a network that is specific to the dataset and phenotype under study and can be used to in a NOP model. We evaluated SyNet and 11 other networks on a compendium dataset of >4000 survival-labelled breast cancer samples. For this purpose, we used cross-study validation which more closely emulates real world application of these outcome predictors. We find that SyNet is the only network that truly improves performance, stability and interpretability in several existing NOPs. We show that SyNet overlaps significantly with existing gene networks, and can be confidently predicted (~85% AUC) from graph-topological descriptions of these networks, in particular the breast tissue-specific network. Due to its data-driven nature, SyNet is not biased to well-studied genes and thus facilitates post-hoc interpretation. We find that SyNet is highly enriched for known breast cancer genes and genes related to e.g. histological grade and tamoxifen resistance, suggestive of a role in determining breast cancer outcome. All corresponding scripts are publicly available through github: https://github.com/UMCUGenetics/SyNet.

4.2. Introduction

Metastases at distant sites (e.g. in bone, lung, liver and brain) is the major cause of death in breast cancer patients [2]. However, it is currently difficult to assess tumor progression in these patients using common clinical variables (e.g. tumor size, lymph-node status, etc.) [3]. Therefore, for 80% of these patients, chemotherapy is prescribed [4]. Meanwhile, randomized clinical trials showed that at least 40% of these patients survive without chemotherapy and thus unnecessarily suffer from the toxic side effect of this treatment [4, 5]. For this reason, substantial efforts have been made to derive molecular classifiers that can predict clinical outcome based on gene expression profiles obtained from the primary tumor at the time of diagnosis [6, 7].

An important shortcoming in molecular classification is that 'cross-study' generalization is often poor [8, 9]. This means that prediction performance decreases dramatically when a classifier trained on one patient cohort is applied to another one [9]. Moreover, the gene signatures found by these classifiers vary greatly, often sharing only few or no genes at all [10–12]. This lack of consistency casts doubt on whether the identified signatures capture true 'driver' mechanisms of the disease or rather subsidiary 'passenger' effects [13].

Several reasons for this lack of consistency have been proposed, including small sample size [12, 14, 15], inherent measurement noise [16] and batch effects [17, 18]. Apart from these technical explanations, it is recognized that traditional models ignore the fact that genes are organized in pathways [19]. One important cancer hallmark is that perturbation of these pathways may be caused by deregulation of disparate sets of genes which in turn complicates marker gene discovery [20, 21].

To alleviate these limitations, the classical models (i.e. outcome predictors that use ordinary classifiers) are superseded by Network-based Outcome Predictors (NOP) which incorporate gene interactions in the prediction model [22]. NOPs have two fundamental components: aggregation and prediction. In the aggregation step, genes that interact, belong to the same pathway or otherwise share functional relation are aggregated (typically by averaging expressions) into so called "meta-genes" [23]. This step is guided by a supporting data source describing gene-gene interactions such as cellular pathway maps or protein-protein interaction networks. In the consequent prediction step, meta-genes are selected and combined into a trained classifier, similar to a traditional classification approach. Several NOPs have been reported to exhibit improved discriminative power, enhanced stability of the classification performance and signature and better representation of underlying driving mechanisms of the disease [19, 24–26].

In recent years, a range of improvements to the original NOP formulation has been proposed. In the prediction step, various linear and nonlinear classifiers have been evaluated [27, 28]. Problematically, the reported accuracies are often an overestimation as many studies neglected to use cross-study evaluation scheme which more closely resembles the real-world application of these models [8]. Also for the aggregation step, which is responsible for forming meta-genes from gene sets, several distinct approaches are proposed such as clustering [24] and greedy expansion of seed genes into subnetworks [19]. Moreover, in addition to simple averaging, alternative means by which genes can be aggregated, such as linear or nonlinear embeddings, have been proposed [18, 29]. Most recent work combines these steps into a unified model [9, 30]. Meanwhile, efforts that extend these concepts to sequencing data by exploiting the concept of cancer hallmark networks have also been proposed [31].

Despite these efforts and initial positive findings, there is still much debate over the utility of NOPs compared to classical methods, with several studies showing no performance improvement [22, 32, 33]. Perhaps even more striking is the finding that utilizing a permuted network [33] or aggregating random genes [11] performs on par with networks describing true biological relationships. Several meta-analyses attempting to establish the utility of NOPs have appeared with contradicting conclusions. Notably, Staiger et al. compared performance of nearest mean classifier [34] in this setting and concluded that network derived meta-genes are not more predictive than individual genes [22, 33]. This is in contradiction to Roy et al. who achieved improvements in outcome prediction when genes were ranked according to their t-test statistics compared to their page rank property [35] in PPI network [29, 36]. It is thus still an open question

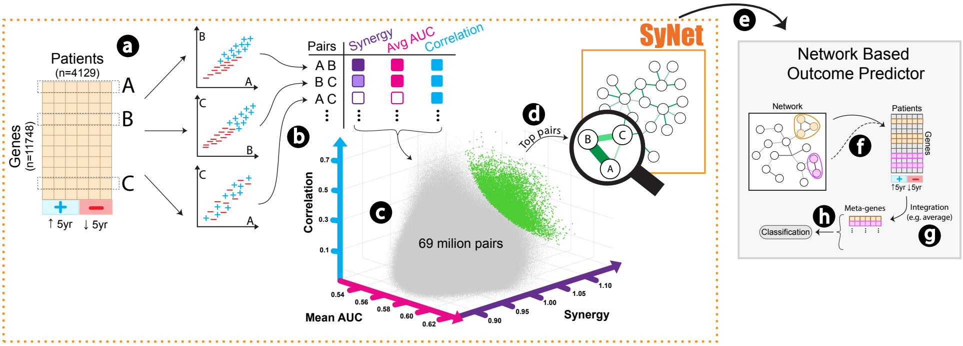


Figure 4.1: Schematic overview of SyNet inference and NOP training. For every 69 million combinations of gene pairs **(a)** we compute three criteria including synergy (S_{ij}, purple), average AUC (M_{ij}, pink), and correlation (C_{ij}, blue) **(b)**. These three criteria form a three-dimensional space **(c)** from which Fitness (F_{ij}) can be calculated for each pair. Top pairs (green dots) in this space are considered as SyNet **(d)**. SyNet is subsequently used in a NOP **(e)**, in which the links in SyNet guide the construction of "meta-genes". Within a NOP, groups of genes are formed **(f)** and then integrated into meta-genes (typically using averaging) **(g)**. The constructed meta-genes are then used as regular features to train standard classifiers **(h)**. The phenotype of interest is patient outcome (i.e. 5-year survival).

whether NOPs truly improve outcome prediction in terms of predictive performance, cross-study robustness or interpretability of the gene signatures.

A critical - yet often neglected - aspect in the successful application of NOPs is the contribution of the biological network. In this regard, it should be recognized that many network links are unreliable [37, 38], missing [39] or redundant [40] and considerable efforts are being made to refine these networks [39, 41–43]. In addition, many links in these networks are experimentally obtained from model organisms and therefore may not be functional in human cells [44–46]. Finally, most biological networks capture only a part of a cell's multifaceted system [47]. This incomplete perspective may not be sufficient to link the wide range of aberrations that may occur in a complex and heterogeneous disease such as breast cancer[48, 49]. Taken together, these issues raise concerns regarding the extent to which the outcome predictors may benefit from inclusion of common biological networks in their models.

In this work, we propose to construct a network ab initio that is specifically designed to improve outcome prediction in terms of cross-study generalization and performance stability. To achieve this, we will effectively turn the problem around: instead of using a given biological network, we aim to use the labelled gene expression datasets to identify the network of genes that truly improves outcome prediction (see Figure 4.1 for a schematic overview).

Our approach relies on the identification of synergistic gene pairs, i.e. genes whose joint prediction power is beyond what is attainable by both genes individually [50]. To identify these pairs, we employed grid computing to evaluate all 69 million pairwise combinations of genes (see section 4.3.5 for details). The resulting network, called SyNet, is specific to the dataset and phenotype under study and can be used to infer a NOP model with improved performance.

To obtain SyNet, and allow for rigorous cross-study validation, a dataset of substantial size is required. For this reason, we combined 14 publicly available datasets to form a compendium encompassing 4129 survival labeled samples. To the best of our knowledge, the data combined in this study represents the largest breast cancer gene expression compendium to date. Further, to ensure unbiased evaluation, sample assignments in the inner as well as the outer cross-validations folds are kept equal across all assessments throughout the paper.

In the remainder of this paper, we will demonstrate that integrating genes based on SyNet provides superior performance and stability of predictions when these models are tested on independent cohorts. In contrast to previous reports, where shuffled versions of networks also performed well, we show that the performance drops substantially when SyNet links are shuffled (while containing the same set of genes), suggesting that SyNet connections are truly informative. We further evaluate the content and structure of SyNet by overlaying it with known gene sets and existing networks, revealing marked enrichment for known breast cancer prognostic markers. While overlap with existing networks is highly significant, the majority of direct links in SyNet is absent from these networks explaining the observed lack of performance when NOPs are guided by the phenotype-unaware networks. Interestingly, SyNet links can be reliably predicted from existing networks when more complex topological descriptors are employed. Taken together, our findings suggest that compared to generic gene networks, phenotype-specific networks, which are derived directly from labeled data, can provide superior performance while at the same time revealing valuable insight into etiology of breast cancer.

4.3. MATERIALS AND METHODS

4.3.1. INFERRING A SYNERGISTIC NETWORK (SYNET)

We hypothesized that, in order to improve outcome prediction by network-based classification, interconnections in the network should correspond to gene pairs for which integration yields a performance beyond what is attainable by either of the individual genes (i.e. synergy). Accordingly, we formulated the synergy S_{ij} between gene i and gene j as

$$S_{ij} = \frac{A_{ij}}{Max(A_i, A_j)}$$

where A_i, A_j and A_{ij} respectively represent the Area Under Curve (AUC) of gene i, the AUC of gene j and the AUC of meta-gene ij formed by aggregation of gene i and gene j. Meta-gene formation is carried out by a linear regression model which demonstrated superior performance in our experiments (see Chapter 5, supplementary figure 5.1 for details). Cross-validation performance of the linear regression (see section on Cross validation design for details) is obtained and the median of 65 AUCs (13 folds and 5 repeats) is used as the final score A_{ij} for each pair. The AUC of the individual genes (i.e. A_i and A_j) is obtained in a similar fashion.

Defining the synergy as a function of AUC yields a phenotype-specific (i.e. label-specific) measure which effectively ignores extraneous relationships between gene pairs that are not relevant in outcome prediction. The synergy measure S_{ij} depends on the performance of individual genes where poorly performing genes tend to achieve higher degree of synergy compared to two predictive genes (see Chapter 5, supplementary figure 5.2 for corresponding analysis). In order to account for this effect, the average AUC of individual genes is included as a second criterion. Furthermore, our preliminary tests confirmed previous findings [9, 23, 50], that integrating highly correlated genes (which reduces meta-gene noise) may improve survival prediction. For this reason, we added correlation of pairs as a third criterion. To combine these three measures, each measure is normalized independently between [0, 1] and then combined into an overall fitness score F_{ij} for gene pair ij:

$$F_{ij} = -\sqrt[2]{(1-\overline{S_{ij}})^2 + (1-\overline{M_{ij}})^2 + (1-\overline{C_{ij}})^2}$$

Here, M_{ij} and M_{ij} represent mean AUC and absolute spearman correlation of gene i and j respectively. Bars above letters indicate that the corresponding values are normalized to the [0 , 1] interval. Employing the Dutch grid infrastructure, we quantified the fitness for all 69 million possible pairs of genes (n=11748). Figure 4.1.c visualizes the fitness of all pairs in a three-dimensional space. Finally, the top 50,000 pairs with highest fitness are considered as SyNet.

4.3.2. Expression Data

Accurately estimating survival risk and identifying markers relevant for progression of a complex disease such as breast cancer requires a large number of samples [12]. To this end, samples from METABRIC [51] (n=1981) are combined with 12 studies collected in ACES [22] (n=1606) as well as samples from the TCGA breast invasive carcinoma dataset [52] (n=532) (see supplementary text for details). Collectively, these datasets, spanning 14 distinct studies, form a compendium encompassing 4129 samples. To the best of our knowledge, the data combined in this paper represents the largest breast cancer gene expression compendium to date. As a result, our compendium should capture a large portion of the biological heterogeneity among breast cancer patients, as well as technical biases originating from the variability in platforms and study-specific sample preparations [53]. This variability will assist the trained models to achieve better generalization which is crucial in real world application of the final classification model [10, 14, 54]. To correct for technical variations that may arise during the library preparation, initially the expression data within each study is quantile normalized and then batch-effect corrected using Combat [55] where the outcome of patients was modeled as an additional covariate to maintain the variance associated with the prognostics. This procedure was shown to perform well among many batch effect removal methods [56, 57]. Successful removal of batch effects was confirmed using t-SNE visualization [58] (see Chapter 5, supplementary figure 5.4 for details). The label for each patient corresponds to overall survival time (or recurrence free survival if available) with respect to a 5-year threshold (good vs. poor outcome).

4.3.3. Regular classifiers and Network based prediction models

Ascertaining the relevance of networks in outcome prediction should be performed using a robust predictor capable of providing adequate performance in prognostic prediction. Previous assessments in this regard have been limited to only few classifiers [22, 24, 29, 35]. To identify the optimal predictor, we have compared performance of wide range of linear and nonlinear classifiers (see Chapter 5, supplementary figure 5.3 for details). Supporting our previous findings [9], this evaluation demonstrates that simple linear classifiers outperform the more complex ones, with the regularized linear classifier (Lasso) reaching the highest AUC. This classifier supports both classical as well as network-based prediction by its derivative called Group Lasso (GL) [59]. The GL is structurally analogous to standard Lasso with the exception of the way in which the regularization is performed; Lasso applies regularization to genes while GL enforces selection of groups of genes (See supplementary text for details). In order to incorporate network information in the GL, similar to our previous work [9], each gene in the corresponding network is considered as seed gene and together with its K neighbors the group structure provided to the GL. Priority of neighbor selection is determined by edge weights between each neighbor and corresponding seed gene. The hyperparameters for each classifier (e.g. K in the GL) are determined by means of a grid search in the inner cross validation loop (see Chapter 5, supplementary figure 5.5 for schematic overview).

For comparison, we include three well-known NOPs in our analysis. Park *et al.* utilized hierarchical clustering to group highly correlated genes [24]. Each group is summarized into a meta-gene by averaging the expression profile of the genes in that group. These meta-genes are then employed as regular features to train a Lasso classifier. The optimal cluster size for hierarchical clustering is identified by iterative application of Lasso in an inner cross-validation. Chuang *et al.* employs a greedy search to define subnetworks [19]. This is done by iteratively expanding a sub-network initiated from a seed gene guided by a supervised performance criterion which halts when performance no longer increases (in the training set). After groups are formed, the meta-genes are constructed by averaging expression of each gene within each group similar to Park et al. Finally, Taylor *et al.* focus on hubs (i.e. highly connected genes, degree>5) in a network [25]. To identify dysregulated subnetworks, the change in correlation between each hub and its direct neighbors across two classes of outcome (poor vs. good) is assessed. Meta-genes are formed from candidate subnetworks similar to the procedure employed by Park et al.

4.3.4. Networks

In addition to SyNet, we considered a range of publicly available networks, including generic networks (HumanInt, BioPlex, BioGRID, IntAct and STRING) as well as a correlation network (Corr) which was previously shown to be an effective network in outcome prediction [9, 24]. Additionally, we assessed five tissue-specific networks (including brain, kidney, ovary, breast, lymph node) that are recently introduced by Greene et al. [45]. These tissue-specific networks are inferred by integrating protein-protein interactions collected from Human Protein Reference Database [60] and tissue-specific information from BRENDA tissue ontology [61] and then filtered using expert-selected Gene Ontology (GO) terms. The tissue-specificity of each network is then validated by a comprehensive collection of expression and interaction datasets encompassing about 38000 conditions collected from approximately 14000 publications. To the best of our knowledge, our study is the first to evaluate tissue-specific networks in the context of NOPs. To maintain a reasonable network size, we utilized only the top 50,000 links (based on the link weight) in each network (similar to number of links in SyNet). For the only unweighted network, HumanInt [39], all interactions (n=~14k) were included and links were weighted according to the average degree of the two interacting genes. Moreover, a randomized version of each network is constructed by shuffling nodes in the network which destroys the biological information of the links while preserving the overall network structure (see supplementary text for full details on preparation of networks).

4.3.5. Cross validation design

In order to ascertain if network information truly aids outcome prediction, the evaluation should be based on a rigorous cross-validation that closely resembles the real-world application of these models. To this end, we perform cross-study validation in order to mimic a realistic situation in which a classifier is applied to data from a different hospital than it was trained on [8]. Briefly, one study is taken out for validation of the final performance (outer loop test set). SyNet inference and NOP training are carried out on the 13 remaining studies (outer loop training set). Within each fold of the outer loop training set, again one study is left out to obtain the inner loop test set and the rest of studies for inner loop training set. The inner loop training set is sub sampled (with replacement) to 70% and regression is performed for every gene as well as gene pairs (identical set of samples are used across all genes and pairs). The AUC scores (A_i, A_j and A_{ij}) are calculated on the inner loop test set. This is repeated 5 times. To train a NOP for this fold, a new inner loop training set is formed by redrawing 70% of the samples from the outer loop training set. This set is also used to infer correlation network. To assess the final performance of the NOP the outer loop test set is used (see Chapter 5, supplementary figure 5.5 for a detailed schematic). Our initial experiments showed a large variation of performance across studies (see Chapter 5, supplementary figure 5.6 for details). To prevent this variation from influencing our comparisons, assignment of samples to folds in both inner and outer cross-validation loops are kept identical across all comparisons throughout the paper. We used Area Under the ROC Curve (AUC) as the main measure of performance in this paper.

4.4. Results

4.4.1. SyNet improves NOP performance

We first evaluated NOP performance for three existing methods (Park, Chuang and Taylor) and the Group Lasso (GL) when supplied with a range of networks, including generic networks, tissue-specific networks and SyNet. As a baseline model, we used a Lasso classifier trained using all genes in our expression dataset (n=11748) without network guidance. The Lasso exhibits superior performance among many linear and non-linear classifiers evaluated on our expression dataset (see Chapter 5, supplementary figure 5.3 for details).

The AUC of the four NOPs, presented in Figure 4.2, clearly demonstrates that SyNet

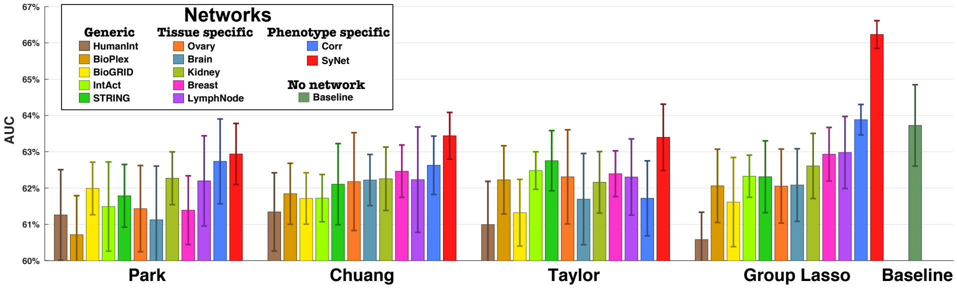


Figure 4.2: Performance comparison of NOPs for 4 methods and 12 networks including SyNet. Bars represent the averaged performance in terms of the AUC and error bars represent the standard deviation of performances across 10 repeats. The rightmost bar represents the performance of standard Lasso which considers all individual genes as features (i.e. no network is used in this model).

improves the performance of all NOPs, except for the Park method in which it performs on par to the Correlation (Corr) network. Notably, SyNet is inferred using training samples only, which prevents 'selection bias' in our assessments [62]. Furthermore, comparison of baseline model performance (i.e. Figure 4.2, rightmost bar) and other NOPs supports previous findings that many existing NOPs do not outperform regular classifiers that do not use networks [9, 22, 33].

The GL clearly outperforms all other methods, in particular when it exploits the information contained in SyNet. This corroborates our previous finding [9] that existing methods which construct meta-genes by averaging are suboptimal (see Chapter 5, supplementary figure 5.1 for a more extensive analysis). The GL using the Corr network also outperforms the baseline model, albeit non-significantly (p≃0.6), which is in line with previous reports [24]. It should be noted that across all these experiments an identical set of samples is used to train the models so that any performance deviation must be due to differences in (i) the set of utilized genes or (ii) the integration of the genes into meta-genes. In the next two sections, we will investigate these factors in more details.

4.4.2. SYNET PROVIDES FEATURE SELECTION CAPABILITIES

Networks only include genes that are linked to at least one other gene. As a result, networks can provide a way of ranking genes based on the number and weight of their connections. One explanation for why NOPs can outperform regular classifiers is that networks provide an a priori gene (feature) selection [33]. To test this hypothesis and determine the feature selection capabilities of SyNet, we compare classification performances obtained using the baseline classifier (i.e. Lasso) that is trained using enclosed genes in each network. While this classifier performs well compared to other standard classifiers that we investigated (see Chapter 5, supplementary figure 5.3 for details), it cannot exploit information contained in the links of given network. So, any performance difference must be due to the genes in the network. The number of genes in each network under study is optimized independently by varying the threshold on the weighted edges in the network and removing unconnected genes (see 4.3 for network size optimization details). The edge weight threshold and the Lasso regularization parameter

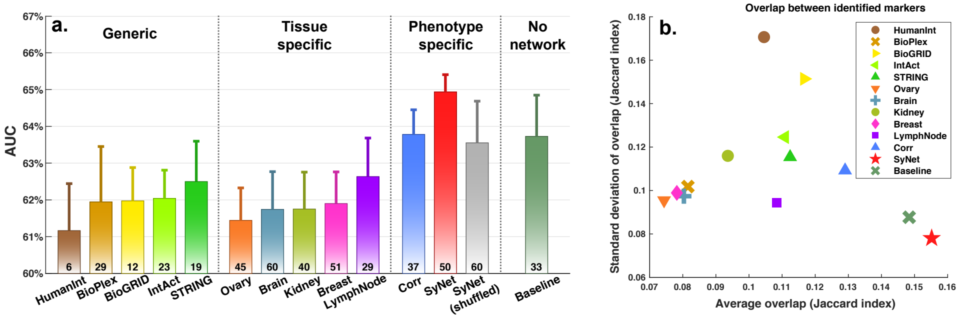


Figure 4.3: Performance comparison between networks when interconnections are ignored and genes contained in each network are utilized to train Lasso. **a.** Performance (AUC) of Lasso classification using individual genes in 12 networks. Numbers below each bar represent the median number of non-zero coefficients after training Lasso across 10 repeats and 14 folds. **b.** Stability of identified signatures measured by overlap between identified gene sets using Jaccard index. X and y-axis represent average and standard deviation of the Jaccard index measured across 10 repeats and 14 folds.

were determined simultaneously using a grid search cross-validation scheme (see Chapter 5, supplementary figure 5.5 for details). Figure 4.3 provides the optimal performances for 12 distinct networks along with number of genes used in the final model (i.e. genes with non-zero Lasso coefficients). We also included the baseline model where all genes (n=11748) are utilized to train Lasso classifier (rightmost bar).

The results presented in Figure 4.3.a demonstrate that SyNet is the only network that performs markedly better than the baseline model which is trained on all genes. Interestingly, we observe that SyNet is the top performing network while utilizing a comparable number of genes to other networks. The second-best network is the Corr network. We argue that superior performance of SyNet over the Corr network stems from the disease specificity of genes in SyNet which helps the predictor to focus on the relevant genes only. It should be noted that the data on which SyNet and the Corr network are constructed are completely independent from the validation data on which the performance is based due to our multi-layer cross-validation scheme (see Methods and Chapter 5, supplementary figure 5.5) which avoids selection bias [62]. We conclude that dataset-specific networks, in particular SyNet which also exploits label information, provides a meaningful feature selection that is beneficial for classification performance.

Our result show that none of the tissue-specific networks outperform the baseline. Despite the modest performance, it is interesting to observe that performance for these networks increases as more relevant tissues (e.g. breast and lymph node networks) are utilized in the classification. Additionally, we observe that tissue-specific networks do not outperform the generic networks. This may be the result of the fact that generic networks predominantly contain broadly expressed genes with fundamental roles in cell function which may still be relevant to survival prediction. A similar observation was made for GWAS where SNPs in these widely-expressed genes can explain a substantial amount of missed heritability [63].

In addition to classifier performance, an important motivation for employing NOPs is to identify stable gene signatures, that is, the same genes are selected irrespective of the study used to train the models. Gene signature stability is necessary to confirm that the identified genes are independent of dataset specific variations and therefore are true biological drivers of the disease under study. To measure the signature consistency, we assessed the overlap of selected genes across all repeats and folds using the Jaccard Index. Figure 4.3.b shows that a Lasso trained using genes preselected by SyNet, identifies more similar genes across folds and studies compared to other networks. Surprisingly, despite the fact that the expression data from which SyNet is inferred changes in each classification fold, the signature stability for SyNet is markedly better than for generic or tissue-specific networks that use a fixed set of genes across folds. Therefore, our results demonstrate that synergistic genes in SyNet truly aid the classifier to robustly select signatures across independent studies.

4.4.3. SyNet connections are beneficial for NOP

The ultimate goal of employing NOPs compared to classical models that do not use network information is to improve prognosis prediction by harnessing the information contained in the links of the given network. Therefore, we next aimed to assess to what extent also connections between the genes, as captured in SyNet and other networks, can help NOPs to improve their performance beyond what is achievable using individual genes. As before, we utilized identical datasets (in terms of genes, training and test samples) in inner and outer cross-validation loops to train all four NOPs as well as the baseline model which uses Lasso trained using all genes (n=11748). Our results presented in Figure 4.4.a, clearly demonstrate that compared to other NOPs under study, GL guided by SyNet achieves superior prognostic prediction for unseen patients selected from an independent cohort. To confirm that NOP performance using SyNet is the result of the network structure, we also applied the GL to a shuffled version of SyNet (Figure 4.4.a). We observe a substantial deterioration of the AUC, supporting the conclusion that not only the genes, but also links contained in SyNet are important to achieve good prediction. Moreover, this observation rules out that the GL by itself is able to provide enhanced performance compared to standard Lasso. The result of a similar assessment for the Corr network is given in Chapter 5, supplementary figure 5.12. Additionally, we found that SyNet remains predictive even when the dataset is down sampled to 25% of samples (see Chapter 5, supplementary figure 5.13 for details). We also evaluated a recently developed set of subtype-specific networks for breast cancer [31] and found that SyNet markedly outperforms these networks in predictive performance (see Chapter 5, supplementary figure 5.18 for details). We next assessed the performance gain of the network-guided model compared to a Lasso model that cannot exploit network information. To this end, the GL was trained based on each network whereas the Lasso is was trained based on the genes present in the network. Figure 4.4.b demonstrates the results of this analysis. We find that the largest gain in GL performance is achieved when using SyNet (Figure 4.4.b, x-axis), indicating that the links between genes in SyNet truly aid classification performance beyond what is obtained as a result of the feature selection capabilities of Lasso. Figure 4.4.c provides the Kaplan-Meier plot when each patient is assigned to a good or poor prognostic class according to frequency of predicted prognosis across 10 repeats (ties are broken by random assignment to one of the classes) for Lasso as well as Group Lasso. Result of this analysis suggests that superior performance of the GL compared to

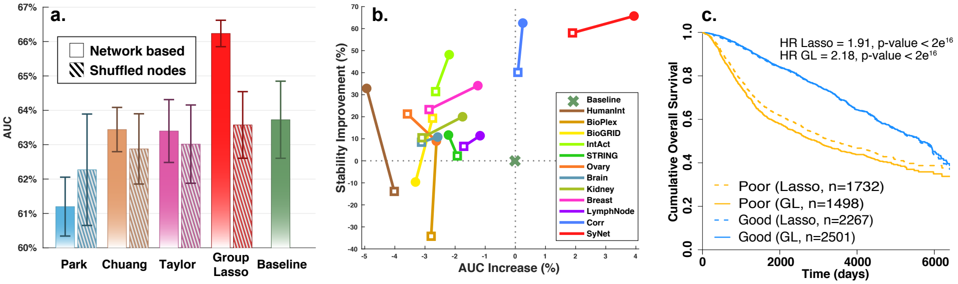


Figure 4.4: Performance of NOP models trained using SyNet compared to a shuffled version of this network (i.e. the same genes are present but randomly connected while keeping their degree intact). **a.** Bars indicate average performance of models across repeats and error bars denote the corresponding standard deviation. Solid bars represent average performance of models trained using SyNet. Dashed bars denote performance of the same model using shuffled SyNet. **b.** Improvement of performance (x-axis) and stability (in terms of the standard deviation of the AUC; y-axis) compared to the baseline model. Square and circle markers represent performance obtained using genes only (i.e. Lasso) and the network (i.e. GL), respectively. **c.** Kaplan-Meier plot for patients predicted to have good or poor prognosis. Dashed lines represent the Lasso prediction and solid lines the Group Lasso (GL) prediction.

the Lasso is mostly stemming from GLs ability to better discern the patients with poor prognosis.

An important property of an outcome predictor is to exhibit constant performance irrespective of the dataset used for training the model (i.e. performance stability). This is a highly desirable quality, as concerns have been raised regarding the highly variable performances of breast-cancer classifiers applied to different cohorts [8, 64]. To measure performance stability, we calculated the standard deviation of the AUC for Lasso and GL. The y-axis in Figure 4.4.b represents the average difference of standard deviation for Lasso and GL across all evaluated folds and repeats (14 folds and 10 repeats). Based on this figure, we conclude that a NOP model guided by SyNet not only provides superior overall performance, it also offers improved stability of the classification performance.

Finally, we investigated the importance of hub genes in SyNet (genes with >4 neighbors) and observe that a comparable performance can be obtained with a network consisting of hub genes exclusively at the cost of reduced performance stability (see Chapter 5, supplementary figure 5.14 for details). Moreover, we did not observe performance gain for a model that is governed by combined links from multiple networks (either by intersection or unification, see Chapter 5, supplementary figure 5.15 for details). We further confirmed that the performance gain of the network-guided GL is preserved when networks are restricted to have equal number of links (see Chapter 5, supplementary figure 5.7 for details), or when links with lower confidence are included in the network (see Chapter 5, supplementary figure 5.16 for details). We also considered the more complex Sparse Group Lasso (SGL), which offers an additional level of regularization (see Supplementary text for details). No substantial difference between GL and SGL performance was found (see Chapter 5, supplementary figure 5.8 for details). Likewise, we did not observe substantial performance differences when the number of genes, group size and regularization parameters were simultaneously optimized in a grid search (see Chapter

5, supplementary figure 5.9 for details). Together, these findings can be considered as the first unbiased evidence of true classification performance improvement in terms of average AUC and classification stability by a NOP.

4.4.4. GENE ENRICHMENT ANALYSIS FOR SYNET

Many curated biological networks suffer from an intrinsic bias since genes with well-known roles are the subject of more experiments and thus get more extensively and accurately annotated [65]. Post-hoc interpretation of the features used by NOPs, often by means of an enrichment analysis, will therefore be affected by the same bias. SyNet does not suffer from such bias, as its inference is purely data driven. Moreover, since SyNet is built based on gene pairs that contribute to the prediction of clinical outcome, we expect that the genes included in SyNet not only relate to breast cancer; they should play a role in determining how aggressively the tumor behaves, how advanced the disease is or how well it responds to treatment.

To investigate the relevance of genes contained in SyNet in the development of breast cancer and, more importantly, clinical outcome, we ranked all pairs according to their median Fitness (F_{ij}) across 14 studies and selected the top 300 genes (encompassing 3544 links). This cutoff was frequently chosen by the GL as the optimal number of genes in SyNet (see section 3.1). Figure 4.5 visualizes this network revealing three main subnetworks and a few isolated gene pairs. We performed functional enrichment for all genes as well as for the subcomponents of the three large subnetworks in SyNet using Ingenuity Pathway Analysis (IPA) [66].

IPA reveals that out of 300 genes in SyNet, 287 genes have a known relation to cancer (2e-06<p<1e-34) of which 222 are related to reproductive system disease (2e-06<p<1e-34). Furthermore, according to IPA analysis, the top five upstream regulators of genes in SyNet (orange box, Figure 4.5) are CDKN1A, E2F4, RABL6, TP53 and ERBB2, all of which are well known players in the development of breast cancer [67–71]. The mean degree of the 300 genes in SyNet is 24, but there are 12 genes which have a degree of 100 or above: ASPM [72], BUB1 [73], CCNB2 [74], CDKN3 [75], CENPA [76], DLGAP5 [77], KIF23 [78], MCM10 [79], MELK [80], RACGAP1 [81], TTK [82] and UBE2C [83]. All these genes play a vital role in progression through the cell cycle and mitosis, by ensuring proper DNA replication, correct formation of the mitotic spindle and proper attachment to the centromere.

In addition to a clear involvement of genes linked to breast cancer generically, IPA also finds clear indications that the genes in SyNet are relevant to clinical outcome and prognosis of the disease. For instance, the most highly enriched cluster (Figure 4.5; green cluster) is found by IPA to be associated to histological grade of the tumor (p=6e-201). The histological grade, which is based on the morphological characteristics of the tumor, has been shown to be informative for the clinical behavior of the tumor and is one of the best-established prognostic markers [84]. Notably, the blue cluster is enriched for genes involved in tamoxifen resistance (p<2e-3), one of the important treatments of ER-positive breast cancer.

Two other sub-clusters (yellow and purple in Figure 4.5), contain genes from distinctly different biological processes than the main cluster. In these clusters we also observe clear hub genes: SLC7A7 and CD74 in the yellow and ACKR1 and MFAP4 in the

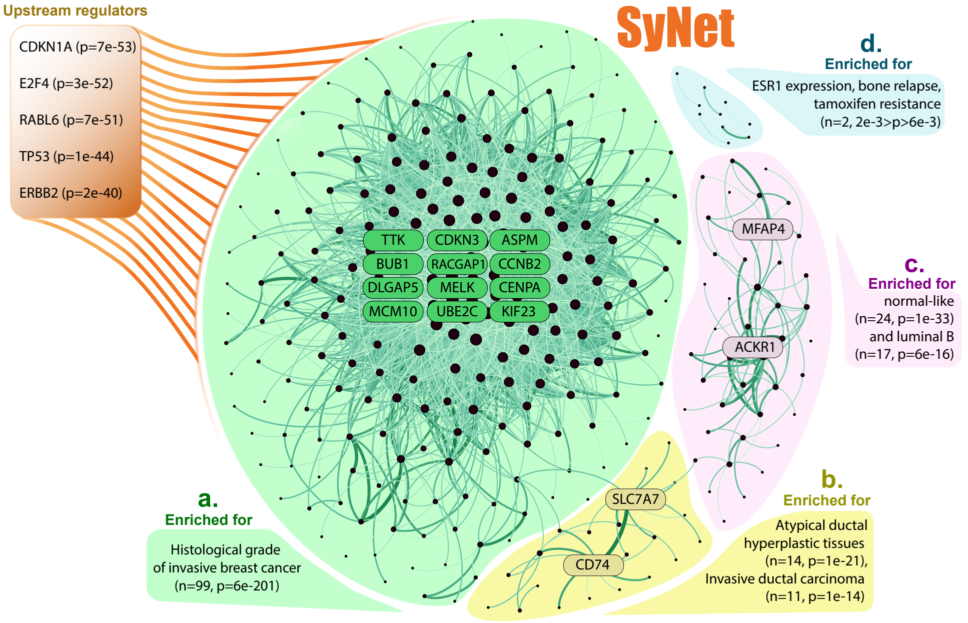

Upstream regulators
CDKN1A (p=7e-53)
E2F4 (p=3e-52)
RABL6 (p=7e-51)
TP53 (p=1e-44)
ERBB2 (p=2e-40)
SyNet
d.
Enriched for
ESR1 expression, bone relapse, tamoxifen resistance (n=2, 2e-3>p>6e-3)
TTK
CDKN3
ASPM
BUB1
RACGAP1
CCNB2
DLGAP5
MELK
CENPA
MCM10
UBE2C
KIF23
MFAP4
ACKR1
c.
Enriched for
normal-like (n=24, p=1e-33) and luminal B (n=17, p=6e-16)
SLC7A7
CD74
a.
Enriched for
Histological grade of invasive breast cancer (n=99, p=6e-201)
b.
Enriched for
Atypical ductal hyperplastic tissues (n=14, p=1e-21), Invasive ductal carcinoma (n=11, p=1e-14)

Figure 4.5: Visualization of SyNet. SyNet consists of three main subnetworks (a, b and c) and five separated gene pairs (d). Node size represents degree of node and link thickness indicates fitness of the corresponding pair. **a.** The largest subnetwork encompassing 223 genes is enriched for histologic grade of invasive breast cancer tumors. **b.** The second subnetwork is directly connected to the first cluster and contains risk factors for developing breast cancer. **c.** The third cluster is enriched for genes upregulated in normal-like subtype of breast cancer. **d.** Out of five pairs, only TFF3 and TFF1 pair is enriched for genes up-regulated in early primary breast tumors.

purple cluster. ACKR1 is a chemokine receptor involved in the regulation of the bioavailability of chemokine levels and MFAP4 is involved in regulating cell-cell adhesion. The recruitment of cells, as regulated by chemokines, and reducing cell-cell adhesion both play an important role in the process of metastasis. CD74 has also been linked to metastasis in triple negative breast cancer [85]. Metastasis, and not the primary tumor, is the main cause of death in breast cancer [4].

IPA highly significantly identifies the SyNet genes as upstream regulators of canonical pathways implicated in breast cancer (Figure 4.5), such as Cell Cycle Control of Chromosomal Replication (8e-18), Mitotic Roles of Polo-Like Kinase (4e-15), Role of CHK Proteins in Cell Cycle Checkpoint Control (6e-12), Estrogen-mediated S-phase Entry (2e-11), and Cell Cycle: G2/M DNA Damage Checkpoint Regulation (5e-10). Although all cancer cells deregulate cell cycle control, the degree of dysregulation may contribute to a more aggressive phenotype. For instance, it is recognized that the downregulation of certain checkpoint regulators is related to a worse prognosis in breast cancer [86, 87]. In summary, SyNet predominantly appears to contain genes relevant to two main processes in the progression of breast cancer: increased cell proliferation and the process of metastasis. Although many genes have not previously been specifically linked to breast cancer

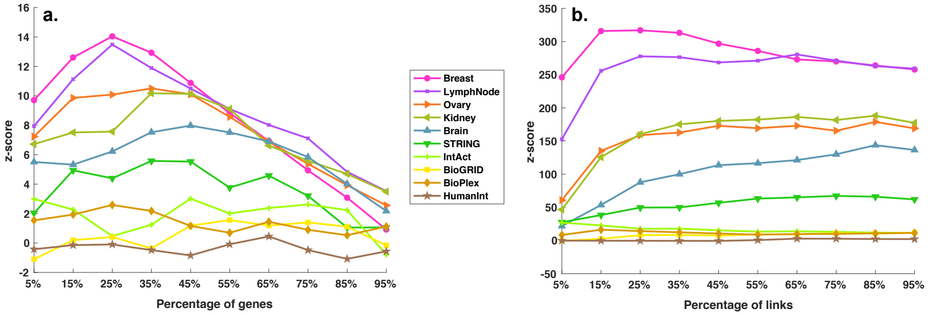


Figure 4.6: Similarity of existing biological networks to SyNet in terms of **a.** genes and **b.** links. The x-axis represents the percentage of top gene/links used, the y-axis the z-score of observed vs. expected number of gene/links. The z-score is calculated by relating the observed number of SyNet gene/links that are present in existing biological networks to the expected distribution. To calculate the expected distribution, genes in biological networks are shuffled.

prognosis, their role in regulating different stages of replication and mitosis points to a genuine biological role in the progression and prognosis of breast cancer.

4.4.5. SIMILARITY OF SYNET TO EXISTING BIOLOGICAL NETWORKS

We next sought to investigate the similarity between SyNet and existing biological networks that directly or indirectly capture biological interactions. To enable a comparison with networks of different sizes, we compare the observed overlap (both in terms of genes as well as links) to the distribution of expected overlap obtained by shuffling each network 1000 times (while keeping the degree distribution intact). Overlap is determined for varying network sizes by thresholding the link weights such that a certain percentage of genes or links remains. Results are reported in terms of a z-score in Figure 4.6.

Figure 4.6.a shows that for the majority of networks a significantly higher than expected number of SyNet genes is contained in the top of each network. The overlap is especially pronounced for the tissue-specific networks, in particular the Breast-specific and Lymph node-specific networks, supporting our observation that SyNet contains links that are relevant for breast cancer. The enrichment becomes even more significant when considering the overlap between the links (Figure 4.6.b). In this respect, SyNet is also clearly most similar to the Breast-specific and Lymph node-specific networks. We confirmed that these enrichments are not only driven by the correlation component of SyNet by repeating this analysis with a variant of the SyNet network without the correlation component (i.e. only average and synergy of gene pairs are used for pair-ranking; see Chapter 5, supplementary figure 5.10 for details). It should moreover be noted that, although a highly significant overlap is observed, the vast majority of SyNet genes and links are not present in the existing networks, explaining the improved performance obtained with NOPs using SyNet. Specifically, out of the 300 genes in SyNet, only 142 are contained within the top 25% of genes (n=1005) in the Breast-specific network, and 151 in the top 25% of genes (n=1290) in the Lymph node-specific network. Similarly, out of the

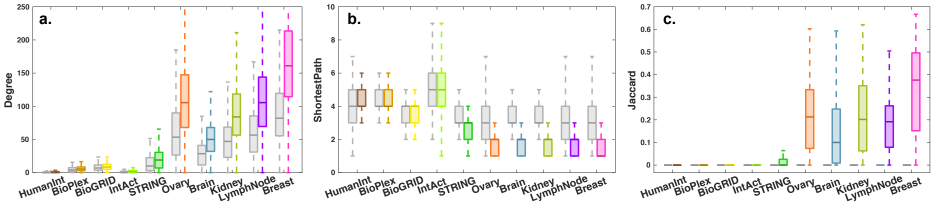


Figure 4.7: Comparison of three topological measures calculated over biological networks. Each color represents a network. Gray boxes represent the same topological measures calculated on the shuffled network.

3544 links in SyNet, only 1182 are contained within the top 25% of links (n=12500) in the Breast-specific network, and 617 in the top 25% (n=12500) of the Lymph node-specific network (see Chapter 5, supplementary figure 5.11 for details). We further confirmed that the overall trend in observed overlaps between SyNet and other networks does not change when the size of these networks (in terms of the number of links) are increased or reduced (see Chapter 5, supplementary figure 5.17 for details).

4.4.6. Higher order structural similarity of SyNet and existing biological networks

In addition to direct overlap, we also aimed to investigate if genes directly connected in SyNet may be indirectly connected in existing networks. To assess this for each pair of genes in SyNet, we computed several topological measures characterizing their (indirect) connection in the biological networks. We included degree (Figure 4.7.a), shortest path (Figure 4.7.b) and Jaccard (Figure 4.7.c) (see Supplementary text for details). To produce an edge measure from degree and page rank (which are node based), we computed the average degree and page rank of genes in a pair respectively. Furthermore, we produced an expected distribution for each pair by computing the same topological measures for one of the genes and another randomly selected gene. The results from this analysis supports our previous observation that the information contained in the links of SyNet is markedly - yet only partially - overlapping with the information in the existing networks. Notably, the similarity increases for networks of increased relevance to the tissue in which the gene expression data is measured (i.e. breast tissue).

4.4.7. Predicting SyNet links from biological networks

Encouraged by the overlap with existing biological networks, we next asked whether links in SyNet can be predicted from the complete collection of topological measures calculated based on existing networks. To this end, we characterized each gene-pair by a set of 12 graph-topological measures that describe local and global network structure around each gene-pair. In addition to the degree, shortest path and Jaccard, we included several additional graph-topological measures including direct link, page rank (with four betas), closeness centrality, clustering coefficient and eigenvector centrality (see Supplementary text for details). While converting node-based measures to edge based measures, in addition to using the average, we also used the difference between the score for each gene in the pair, similar to our previous work [88]. We applied these measures

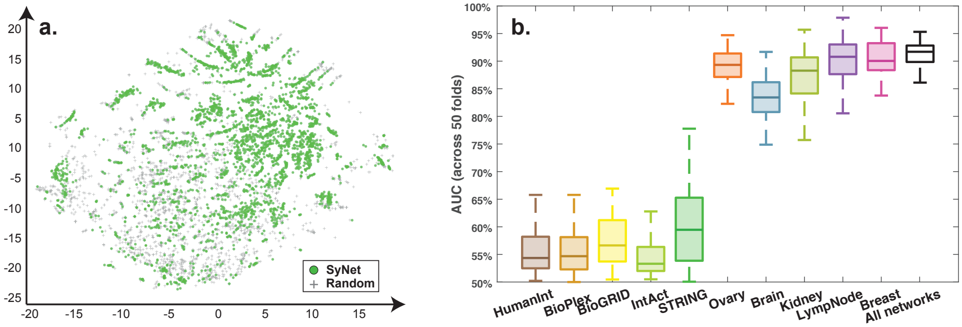


Figure 4.8: Characterizing SyNet links by a range of graph topological measures. **a.** t-SNE (unsupervised) visualization of the combined 180 topological measures. Each dot represents one gene pair. Green dots indicate SyNet links while gray markers represent an equal number of random pairs. **b.** Performance of Lasso model trained over all topological measure for different networks and all networks combined (rightmost bar).

to all 10 networks in our collection yielding a total of 210 features. The gene-pairs are labeled according to their presence or absence in SyNet. Inspection of this dataset using the t-SNE [58] reveals that the links in SyNet occupy a distinct part of the 2D embedding obtained (Figure 4.8.a).

We trained a Lasso and assessed classification performance in a 50-fold cross validation scheme where in each fold 1/50 of pairs in SyNet is kept hidden and the rest of pairs is utilized to train the classifier. To avoid information leakage in this assessment, we removed gene pairs from the training set in case one of the genes is present in the test set. Based on this analysis we find that a simple linear classifier can reach ~85% accuracy in predicting the synergistic gene relationships from SyNet (Figure 4.8.b, rightmost bar). The contribution from generic networks is notably smaller than for the tissue-specific networks. In particular the networks relevant to breast cancer are highly informative, to the extent that combining multiple networks no longer improves prediction performance. Further investigation of feature importance revealed that the page rank topological measure was commonly used as a predictive marker across folds. Apparently, while direct overlap between SyNet and existing networks is modest, the topology of the relevant networks (i.e. breast-specific and lymph node-specific networks) are highly informative for the links contained in SyNet. This corroborates findings from Winter et al. in which the page rank topological measure was proposed to identify relevant genes in outcome prediction [35, 36, 89].

4.5. Discussion and Future work

Although the principle of using existing knowledge of the cellular wiring diagram to improve performance, robustness and interpretability of gene expression classifiers appears attractive, contrasting reports on the efficacy of such approach have appeared in literature [22, 29, 33, 36]. Consensus in this field has particularly been frustrated by an evaluation of a limited set of sub-optimal classifiers [22, 24, 29, 36], small sample size [19, 25, 27], or the use of standard K-fold cross-validation instead of cross-study evaluation schemes, which results in inflated performance estimates [25, 27]. For this reason, it remained unclear if network-based classification, and in particular network-based outcome prediction, is beneficial. Here, we present a rigorously cross-validated procedure to train and evaluate Group Lasso-based NOPs using a variety of networks, including tissue-specific networks in particular, which have not been evaluated in the context of NOPs before.

Based on our analyses, we conclude that none of the existing networks achieve improved performance compared to using properly regularized classifiers trained on all genes. In this work we therefore present a novel gene network, called SyNet, which is computationally derived directly from the survival-labeled samples. The links in SyNet connect synergistic gene pairs. We followed a cross-validation procedure in which the inference of SyNet and validation of its utility in a NOP is strictly separated. We find that SyNet-based NOPs yields superior performance with higher stability across the folds compared to both the baseline model trained on all genes as well as models that use other existing gene networks. We therefore conclude that at least in outcome prediction problem, network guidance can improve model performance, but only if this network is phenotype-specific. Supporting this conclusion, we also show that a correlation network, which is dataset-specific but not phenotype specific, also improved performance but much less compared to SyNet.

A major benefit of SyNet over manually curated gene networks is that its inference is purely data driven, and therefore not biased to well-studied genes. Post-hoc interpretation of the genes selected by a NOP that utilized SyNet is therefore expected to provide a more unbiased interpretation of the important molecular players underlying breast cancer and patient survival. Analysis of the genes contained in SyNet shows strong enrichment for genes with known relevance to breast cancer. More importantly, the largest subcomponent of SyNet is strongly linked to patient prognosis as it includes many genes with a known relation to the histological grade of the tumor.

To investigate if SyNet captures known biological gene interactions, we extensively compared SyNet with existing networks. We find highly significant overlaps between links, indicating that SyNet connects genes that also have a known biological interaction. Despite this significant overlap, the majority of the SyNet links are not recapitulated by direct links in the existing networks. However, we find that accurate prediction of links in SyNet are possible if more complex graph topological descriptions of the indirect connections in the existing networks are employed. Interestingly, accurate predictions are only obtained when using the breast specific networks. Apparently, although the information contained in SyNet is similar to other gene interaction networks, the wiring of SyNet much better supports GL-based classification. This might explain why using existing biological networks in NOPs directly is unsuccessful and why graph topological measures have been successful in identifying relevant genes in outcome prediction [35, 36, 89]. Taken together, our results underline that network-based outcome prediction is a promising approach to improving patient prognosis prediction and therefore can provide an important contribution towards more personalized healthcare. At the same time, the SyNet approach provides an unbiased interactome which makes the NOP more amenable for model interpretation, thus providing important insights into the etiology of the disease under study.

REFERENCES

[1] A. Allahyar, J. de Ridder, and J. Ubels, *A data-driven interactome of synergistic genes improves network based cancer outcome prediction,* bioRxiv (2018), 10.1101/349688.

[2] A. Fantozzi and G. Christofori, *Mouse models of breast cancer metastasis,* Breast Cancer Research **8**, 212 (2006).

[3] C. L. Shapiro and A. Recht, *Side effects of adjuvant treatment of breast cancer,* New England Journal of Medicine **344**, 1997 (2001), pMID: 11430330, http://dx.doi.org/10.1056/NEJM200106283442607 .

[4] B. Weigelt, J. L. Peterse, and L. J. van't Veer, *Breast cancer metastasis: markers and models,* Nature reviews cancer **5**, 591 (2005).

[5] F. Cardoso, L. J. van't Veer, J. Bogaerts, L. Slaets, G. Viale, S. Delaloge, J.-Y. Pierga, E. Brain, S. Causeret, M. DeLorenzi, A. M. Glas, V. Golfinopoulos, T. Goulioti, S. Knox, E. Matos, B. Meulemans, P. A. Neijenhuis, U. Nitz, R. Passalacqua, P. Ravdin, I. T. Rubio, M. Saghatchian, T. J. Smilde, C. Sotiriou, L. Stork, C. Straehle, G. Thomas, A. M. Thompson, J. M. van der Hoeven, P. Vuylsteke, R. Bernards, K. Tryfonidis, E. Rutgers, M. Piccart, and MINDACT Investigators, *70-gene signature as an aid to treatment decisions in Early-Stage breast cancer,* N. Engl. J. Med. **375**, 717 (2016).

[6] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, *Gene expression profiling predicts clinical outcome of breast cancer,* Nature **415**, 530 (2002).

[7] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, *A gene-expression signature as a predictor of survival in breast cancer,* New England Journal of Medicine **347**, 1999 (2002).

[8] C. Bernau, M. Riester, A.-L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa, *Cross-study validation for the assessment of prediction algorithms,* Bioinformatics **30**, i105 (2014).

[9] A. Allahyar and J. de Ridder, *Feral: network-based classifier with application to breast cancer outcome prediction,* Bioinformatics **31**, i311 (2015).

[10] D. F. Ransohoff, *Opinion: Bias as a threat to the validity of cancer molecular-marker research,* Nat. Rev. Cancer **5**, 142 (2005).

[11] D. Venet, J. E. Dumont, and V. Detours, *Most random gene expression signatures are significantly associated with breast cancer outcome,* PLoS computational biology **7**, e1002240 (2011).

[12] L. Ein-Dor, O. Zuk, and E. Domany, *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer,* Proc. Natl. Acad. Sci. U. S. A. **103**, 5923 (2006).

[13] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway, and D. Pe'er, *An integrated approach to uncover drivers of cancer,* Cell **143**, 1005 (2010).

[14] S.-Y. Kim, *Effects of sample size on robustness and prediction accuracy of a prognostic gene signature,* BMC Bioinformatics **10**, 147 (2009).

[15] J. Hua, W. D. Tembe, and E. R. Dougherty, *Performance of feature-selection methods in the classification of high-dimension data,* Pattern Recognition **42**, 409 (2009).

[16] P. A. Bryant, G. K. Smyth, R. Robins-Browne, and N. Curtis, *Technical variability is greater than biological variability in a microarray experiment but both are outweighed by changes induced by stimulation,* PLoS One **6**, e19556 (2011).

[17] H. S. Parker and J. T. Leek, *The practical effect of batch on genomic prediction,* Stat. Appl. Genet. Mol. Biol. **11**, Article 10 (2012).

[18] N. Alcaraz, M. List, R. Batra, F. Vandin, H. J. Ditzel, and J. Baumbach, *De novo pathway-based biomarker identification,* Nucleic Acids Res. **45**, e151 (2017).

[19] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, *Network-based classification of breast cancer metastasis,* Molecular systems biology **3**, 140 (2007).

[20] D. Hanahan and R. A. Weinberg, *Hallmarks of cancer: The next generation,* Cell **144**, 646 (2011).

[21] D. Hanahan and R. A. Weinberg, *The hallmarks of cancer,* Cell **100**, 57 (2000).

[22] C. Staiger, S. Cadot, B. Györffy, L. F. A. Wessels, and G. W. Klau, *Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis,* Front. Genet. **4**, 289 (2013).

[23] W.-Y. Cheng, T.-H. Ou Yang, and D. Anastassiou, *Biomolecular events in cancer revealed by attractor metagenes,* PLoS Comput. Biol. **9**, e1002920 (2013).

[24] M. Y. Park, T. Hastie, and R. Tibshirani, *Averaged gene expressions for regression,* Biostatistics **8**, 212 (2007).

[25] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, *Dynamic modularity in protein interaction networks predicts breast cancer outcome,* Nature biotechnology **27**, 199 (2009).

[26] W. Zhang, J. Chien, J. Yong, and R. Kuang, *Network-based machine learning and graph theory algorithms for precision oncology,* npj Precision Oncology **1** (2017).

[27] V. Popovici, W. Chen, B. G. Gallas, C. Hatzis, W. Shi, F. W. Samuelson, Y. Nikolsky, M. Tsyganova, A. Ishkin, T. Nikolskaya, K. R. Hess, V. Valero, D. Booser, M. Delorenzi, G. N. Hortobagyi, L. Shi, W. F. Symmans, and L. Pusztai, *Effect of training-sample size and classification difficulty on the accuracy of genomic predictors,* Breast Cancer Res. **12**, R5 (2010).

[28] L. F. A. Wessels, M. J. T. Reinders, A. A. M. Hart, C. J. Veenman, H. Dai, Y. D. He, and L. J. van't Veer, *A protocol for building and evaluating predictors of disease state based on microarray data,* Bioinformatics **21**, 3755 (2005).

[29] J. Roy, C. Winter, and M. Schroeder, *Meta-analysis of cancer gene profiling data,* Methods in Molecular Biology , 211 (2016).

[30] J. Dutkowski and T. Ideker, *Protein networks as logic functions in development and cancer,* PLoS Comput. Biol. **7**, e1002180 (2011).

[31] N. Zaman, L. Li, M. L. Jaramillo, Z. Sun, C. Tibiche, M. Banville, C. Collins, M. Trifiro, M. Paliouras, A. Nantel, *et al.*, *Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets,* Cell reports **5**, 216 (2013).

[32] Y. Cun and H. Fröhlich, *Network and data integration for biomarker signature discovery via network smoothed t-statistics,* PLoS One **8**, e73074 (2013).

[33] C. Staiger, S. Cadot, R. Kooter, M. Dittrich, T. Müller, G. W. Klau, and L. F. A. Wessels, *A critical evaluation of network and Pathway-Based classifiers for outcome prediction in breast cancer,* PLOS ONE **7**, e34796 (2012).

[34] E. Alpaydin, *Introduction to Machine Learning* (MIT Press, 2014).

[35] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, V. Hentrich, F. Rückert, M. Niedergethmann, W. Weichert, M. Bahra, H. J. Schlitt, U. Settmacher, H. Friess, M. Büchler, H.-D. Saeger, M. Schroeder, C. Pilarsky, and R. Grützmann, *Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes,* PLoS Comput. Biol. **8**, e1002511 (2012).

[36] J. Roy, C. Winter, Z. Isik, and M. Schroeder, *Network information improves cancer outcome prediction,* Briefings in Bioinformatics **15**, 612 (2014).

[37] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. Carvunis, N. Simonis, J.-F. Rual, H. Borick, P. Braun, M. Dreze, J. Vandenhaute, M. Galli, J. Yazaki, D. E. Hill, J. R. Ecker, F. P. Roth, and M. Vidal, *Literature-curated protein interaction datasets,* Nature methods **6**, 39 (2009).

[38] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, *Comparative assessment of large-scale data sets of protein-protein interactions,* Nature **417**, 399 (2002).

[39] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, *Towards a proteome-scale map of the human protein-protein interaction network,* Nature **437**, 1173 (2005).

[40] M. A. Mahdavi and Y.-H. Lin, *False positive reduction in protein-protein interaction predictions using gene ontology annotations,* BMC Bioinformatics **8**, 262 (2007).

[41] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal, *A proteome-scale map of the human interactome network,* Cell **159**, 1212 (2014).

[42] E. L. Huttlin, L. Ting, R. J. Bruckner, F. Gebreab, M. P. Gygi, J. Szpyt, S. Tam, G. Zarraga, G. Colby, K. Baltier, R. Dong, V. Guarani, L. P. Vaites, A. Ordureau, R. Rad, B. K. Erickson, M. Wühr, J. Chick, B. Zhai, D. Kolippakkam, J. Mintseris, R. A. Obar, T. Harris, S. Artavanis-Tsakonas, M. E. Sowa, P. De Camilli, J. A. Paulo, J. W. Harper, and S. P. Gygi, *The BioPlex network: A systematic exploration of the human interactome,* Cell **162**, 425 (2015).

[43] E. L. Huttlin, R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, J. Szpyt, S. Tam, G. Zarraga, L. Pontano-Vaites, S. Swarup, A. E. White, D. K. Schweppe, R. Rad, B. K. Erickson, R. A. Obar, K. G. Guruharsha, K. Li, S. Artavanis-Tsakonas, S. P. Gygi, and J. W. Harper, *Architecture of the human interactome defines protein communities and disease networks,* Nature **545**, 505 (2017).

[44] C. S. Greene and O. G. Troyanskaya, *Chapter 2: Data-driven view of disease biology,* PLoS Comput. Biol. **8**, e1002816 (2012).

[45] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, D. I. Chasman, G. A. FitzGerald, K. Dolinski, T. Grosser, and O. G. Troyanskaya, *Understanding multicellular function and disease with human tissue-specific networks,* Nat. Genet. **47**, 569 (2015).

[46] E. Yeger-Lotem and R. Sharan, *Human protein interaction networks across tissues and diseases,* Front. Genet. **6**, 257 (2015).

[47] S. Zhang, G. Jin, X.-S. Zhang, and L. Chen, *Discovering functions and revealing mechanisms at molecular level from biological networks,* Proteomics **7**, 2856 (2007).

[48] M. Kotlyar, C. Pastrello, N. Sheahan, and I. Jurisica, *Integrated interactions database: tissue-specific view of the human and model organism interactomes,* Nucleic Acids Res. **44**, D536 (2016).

[49] G. de Anda-Jáuregui, T. E. Velázquez-Caldelas, J. Espinal-Enríquez, and E. Hernández-Lemus, *Transcriptional network architecture of breast cancer molecular subtypes,* Frontiers in physiology **7**, 568 (2016).

[50] J. Watkinson, X. Wang, T. Zheng, and D. Anastassiou, *Identification of gene interactions associated with disease from gene expression data using synergy networks,* BMC Syst. Biol. **2**, 10 (2008).

[51] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, METABRIC Group, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio, *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups,* Nature **486**, 346 (2012).

[52] Cancer Genome Atlas Network, *Comprehensive molecular portraits of human breast tumours,* Nature **490**, 61 (2012).

[53] E. H. Allott, J. Geradts, X. Sun, S. M. Cohen, G. R. Zirpoli, T. Khoury, W. Bshara, M. Chen, M. E. Sherman, J. R. Palmer, C. B. Ambrosone, A. F. Olshan, and M. A. Troester, *Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification,* Breast Cancer Res. **18**, 68 (2016).

[54] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, *Machine learning applications in cancer prognosis and prediction,* Comput. Struct. Biotechnol. J. **13**, 8 (2015).

[55] W. E. Johnson, C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical bayes methods,* Biostatistics **8**, 118 (2007).

[56] C. Müller, A. Schillert, C. Röthemeier, D.-A. Trégouët, C. Proust, H. Binder, N. Pfeiffer, M. Beutel, K. J. Lackner, R. B. Schnabel, L. Tiret, P. S. Wild, S. Blankenberg, T. Zeller, and A. Ziegler, *Removing batch effects from longitudinal gene expression - quantile normalization plus ComBat as best approach for microarray transcriptome data,* PLoS One **11**, e0156594 (2016).

[57] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu, *Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods,* PLoS One **6**, e17238 (2011).

[58] L. van der Maaten, L. van der Maaten, and G. Hinton, *Visualizing non-metric similarities in multiple maps,* Mach. Learn. **87**, 33 (2011).

[59] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables,* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 49 (2006).

[60] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, *Human protein reference database—2009 update,* Nucleic Acids Research **37**, D767 (2009).

[61] M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg, *The brenda tissue ontology (bto): the first all-integrating ontology of all organisms for enzyme sources,* Nucleic Acids Research **39**, D507 (2011).

[62] C. Ambroise and G. J. McLachlan, *Selection bias in gene extraction on the basis of microarray gene-expression data,* Proceedings of the National Academy of Sciences **99**, 6562 (2002), http://www.pnas.org/content/99/10/6562.full.pdf .

[63] E. A. Boyle, Y. I. Li, and J. K. Pritchard, *An expanded view of complex traits: From polygenic to omnigenic,* Cell **169**, 1177 (2017).

[64] P. J. Castaldi, I. J. Dahabreh, and J. P. A. Ioannidis, *An empirical assessment of validation practices for molecular classifiers,* Brief. Bioinform. **12**, 189 (2011).

[65] P. Khatri, M. Sirota, and A. J. Butte, *Ten years of pathway analysis: current approaches and outstanding challenges,* PLoS Comput. Biol. **8**, e1002375 (2012).

[66] A. Krämer, J. Green, J. Pollard, Jr, and S. Tugendreich, *Causal analysis approaches in ingenuity pathway analysis,* Bioinformatics **30**, 523 (2014).

[67] S. S. Khaleel, E. H. Andrews, M. Ung, J. DiRenzo, and C. Cheng, *E2F4 regulatory program predicts patient survival prognosis in breast cancer,* Breast Cancer Res. **16**, 486 (2014).

[68] M. Gasco, S. Shami, and T. Crook, *The p53 pathway in breast cancer,* Breast Cancer Res. **4** (2002).

[69] J.-X. Wei, L.-H. Lv, Y.-L. Wan, Y. Cao, G.-L. Li, H.-M. Lin, R. Zhou, C.-Z. Shang, J. Cao, H. He, Q.-F. Han, P.-Q. Liu, G. Zhou, and J. Min, *Vps4A functions as a tumor suppressor by regulating the secretion and uptake of exosomal microRNAs in human hepatoma cells,* Hepatology **61**, 1284 (2015).

[70] M. Tan and D. Yu, *Molecular mechanisms of ErbB2-Mediated breast cancer chemoresistance,* Advances in Experimental Medicine and Biology , 119 (2007).

[71] J. Montalbano, W. Jin, M. S. Sheikh, and Y. Huang, *RBEL1 is a novel gene that encodes a nucleocytoplasmic ras superfamily GTP-binding protein and is overexpressed in breast cancer,* J. Biol. Chem. **282**, 37640 (2007), accessed: 2017-11-12.

[72] J. L. Fish, Y. Kosodo, W. Enard, S. Pääbo, and W. B. Huttner, *Aspm specifically maintains symmetric proliferative divisions of neuroepithelial cells,* Proc. Natl. Acad. Sci. U. S. A. **103**, 10438 (2006).

[73] D. A. Skoufias, P. R. Andreassen, F. B. Lacroix, L. Wilson, and R. L. Margolis, *Mammalian mad2 and bub1/bubr1 recognize distinct spindle-attachment and kinetochore-tension checkpoints,* Proc. Natl. Acad. Sci. U. S. A. **98**, 4492 (2001).

[74] G. Draetta, F. Luca, J. Westendorf, L. Brizuela, J. Ruderman, and D. Beach, *Cdc2 protein kinase is complexed with both cyclin a and b: Evidence for proteolytic inactivation of MPF,* Cell **56**, 829 (1989).

[75] G. Nalepa, J. Barnholtz-Sloan, R. Enzor, D. Dey, Y. He, J. R. Gehlhausen, A. S. Lehmann, S.-J. Park, Y. Yang, X. Yang, S. Chen, X. Guan, Y. Chen, J. Renbarger, F.-C. Yang, L. F. Parada, and W. Clapp, *The tumor suppressor CDKN3 controls mitosis,* J. Cell Biol. , jcb.201205125 (2013).

[76] D. R. Foltz, L. E. T. Jansen, B. E. Black, A. O. Bailey, J. R. Yates, 3rd, and D. W. Cleveland, *The human CENP-A centromeric nucleosome-associated complex,* Nat. Cell Biol. **8**, 458 (2006).

[77] A.-P. Tsou, C.-W. Yang, C.-Y. F. Huang, R. C.-T. Yu, Y.-C. G. Lee, C.-W. Chang, B.-R. Chen, Y.-F. Chung, M.-J. Fann, C.-W. Chi, J.-H. Chiu, and C.-K. Chou, *Identification of a novel cell cycle regulated gene, HURP, overexpressed in human hepatocellular carcinoma,* Oncogene **22**, 298 (2003).

[78] V. Pavicic-Kaltenbrunner, M. Mishima, and M. Glotzer, *Cooperative assembly of CYK-4/MgcRacGAP and ZEN-4/MKLP1 to form the centralspindlin complex,* Mol. Biol. Cell **18**, 4992 (2007).

[79] R. M. Ricke and A.-K. Bielinsky, *Mcm10 regulates the stability and chromatin association of DNA Polymerase-α,* Mol. Cell **16**, 173 (2004).

[80] I. Nakano, A. A. Paucar, R. Bajpai, J. D. Dougherty, A. Zewail, T. K. Kelly, K. J. Kim, J. Ou, M. Groszer, T. Imura, W. A. Freije, S. F. Nelson, M. V. Sofroniew, H. Wu, X. Liu, A. V. Terskikh, D. H. Geschwind, and H. I. Kornblum, *Maternal embryonic leucine zipper kinase (MELK) regulates multipotent neural progenitor proliferation,* J. Cell Biol. **170**, 413 (2005).

[81] K.-Y. Lee, B. Esmaeili, B. Zealley, and M. Mishima, *Direct interaction between centralspindlin and PRC1 reinforces mechanical resilience of the central spindle,* Nat. Commun. **6**, 7290 (2015).

[82] H. A. Fisk, C. P. Mattison, and M. Winey, *Human mps1 protein kinase is required for centrosome duplication and normal mitotic progression,* Proc. Natl. Acad. Sci. U. S. A. **100**, 14875 (2003).

[83] Z. Hao, H. Zhang, and J. Cowell, *Ubiquitin-conjugating enzyme UBE2C: molecular biology, role in tumorigenesis, and potential as a biomarker,* Tumour Biol. **33**, 723 (2012).

[84] E. A. Rakha, J. S. Reis-Filho, F. Baehner, D. J. Dabbs, T. Decker, V. Eusebi, S. B. Fox, S. Ichihara, J. Jacquemier, S. R. Lakhani, J. Palacios, A. L. Richardson, S. J. Schnitt, F. C. Schmitt, P.-H. Tan, G. M. Tse, S. Badve, and I. O. Ellis, *Breast cancer prognostic classification in the molecular era: the role of histological grade,* Breast Cancer Res. **12**, 207 (2010).

[85] C. Greenwood, G. Metodieva, K. Al-Janabi, B. Lausen, L. Alldridge, L. Leng, R. Bucala, N. Fernandez, and M. V. Metodiev, *Stat1 and CD74 overexpression is co-dependent and linked to increased invasion and lymph node metastasis in triple-negative breast cancer,* J. Proteomics **75**, 3031 (2012).

[86] C. Catzavelos, N. Bhattacharya, Y. C. Ung, J. A. Wilson, L. Roncari, C. Sandhu, P. Shaw, H. Yeger, I. Morava-Protzner, L. Kapusta, E. Franssen, K. I. Pritchard, and J. M. Slingerland, *Decreased levels of the cell-cycle inhibitor p27kip1 protein: Prognostic implications in primary breast cancer,* Nat. Med. **3**, 227 (1997).

[87] C. Craig, R. Wersto, M. Kim, E. Ohri, Z. Li, D. Katayose, S. J. Lee, J. Trepel, K. Cowan, and P. Seth, *A recombinant adenovirus expressing p27kip1 induces cell cycle arrest and loss of cyclin-cdk activity in human breast cancer cells,* Oncogene **14**, 2283 (1997).

[88] M. Hulsman, C. Dimitrakopoulos, and J. de Ridder, *Scale-space measures for graph topology link protein network architecture to function,* Bioinformatics **30**, i237 (2014).

[89] M. E. J. Newman, *Analysis of weighted networks,* Phys. Rev. E **70**, 056131 (2004).

5

SyNet: Supplementary Material

Gene expression preprocessing

For the METABRIC dataset, clinical data was collected from the Synapse Commons archive (syn2133322; www.synapse.org) and normalized gene expression profiles were retrieved from the European genome-phenome archive (EGAS00000000083). For this study, gene expression was measured using Illumina HT-12 v3 platform. For the ACES dataset (see Supplementary table 5.1 for accession number of individual studies in ACES), apart from quantile normalization and batch effect removal, no preprocessing was performed. TCGA breast invasive carcinoma (BRCA) gene expression profiles were retrieved from UCSC Xena Browser [1]. These data were obtained using Agilent 244K custom gene expression ($G4502A_{0}7_{3}$) microarrays.

Table 5.1: GEO accession and number of samples per study in ACES

Study	# Patients	Geo accession
Ivshina	102	4922
Hatzis-Pusztai	150	25066
Desmedt-June07	183	7390
Minn	65	2603
Miller	89	3494
WangY-ErasmusMC	257	2034
Schmidt	169	11121
Pawitan	147	1456
Symmans	224	17705
Loi	57	6532
Zhang	121	12093
WangY	52	5327

Network preprocessing

The Human Interactome (HumanInt, vII-14) network [2] is collected from interactome.baderlab.org. This network does not have weighted links and hence all interactions are utilized (n=14057). BioPlex v2.0 [3] is obtained from bioplex.hms.harvard.edu. The weights for each pair of genes is collected from a column with header of "p(Interaction)" which reflects the likelihood of an interaction to be a true positive. The organism specific version of BioGRID (Homo sapiens, v3.4.155) [4] was obtained from thebiogrid.org and the “score” column is used for link weights. The Homo Sapiens version of the STRING network (9606, v10) [5] is collected from string-db.org and the “combined score” is utilized for link weights. The tissue specific networks are downloaded from the HumanBase [6] website (hb.flatironinstitute.org). Each link in these networks has a weight which reflects the tissue specificity of that interaction. UniProt (used by IntAct) and Ensembl gene IDs (used by STRING) are converted to HGNC IDs using Ensembl Biomart [7]. Entrez IDs (used by BioPlex, HumanBase and BioGRID) are mapped to HGNC using the Hugo server (genenames.org). HumanInt uses HGNC IDs to refer to genes and hence no further conversion is needed for this network.

Lasso and Lasso derivatives

To avoid redundancy, this section is omitted in the thesis. Please refer to supplementary material in the original publication [8], or to the introduction chapter of this thesis (Chapter 1) for details about Linear Regression (section 1.2.3), Lasso (section 1.2.5) and its derivatives Group Lasso and Sparse Group Lasso (section 1.3.5).

Topological measures

In our work, a range of graph topological measures are calculated that describe local graph structure around a node or a edge. The degree is defined as the number of edges connected to the node. The shortest path between two nodes is defined as smallest number of edges from one node that need to be traversed to reach the other node. Pairs of nodes that are not connected have a shortest path equal to infinity. Betweenness of a node is the number of times that node resides on a shortest path of any other pair of nodes in the network normalized by total number of possible pairs in the network. Closeness of a node measures the inverse sum of distances from that node to all other nodes in the network. The Jaccard index between two nodes is defined as number of shared neighbors between two nodes normalized by the total number of unique neighbors of those nodes. The clustering coefficient of a node computes the number of links between direct neighbors of that node normalized by total number of possible links between those direct neighbors. The eigenvector centrality of a node is equal to its component of the related eigenvector of the network. The page rank of a node corresponds to probability of a random surfer to visit that particular node [9]. At each step, this surfer visits a direct neighbor of current node (with probability of β) or restarts its walk from a randomly chosen node in the network (with probability of $1-\beta$). Pagerank is known to be statistically similar to node degree [10].

Gene enrichment analysis for SyNet clusters

Gene enrichment analysis is performed using GSEA [11] for genes in each individual cluster of SyNet. Each row refers to a study in which given genes are enriched. Columns respectively represent study name (Gene set name), number of genes in the study (# Genes in Gene Set (K)), a brief explanation of the enriched set (description), number of genes in the given SyNet cluster that overlap with get set in the study (# Genes in Overlap (k)), chi-square statistics for significance of enrichment (k/K), p-value of corresponding chi-square statistics (p-value) and corrected q-value after false discovery rate control (FDR q-value).

Table 5.2: Gene enrichment analysis for SyNet clusters

Gene set source	# total genes (K)	# overlap (k)	k/K	p-value	FDR q-value
Sotiriou *et al.*	151	99	0.6556	1.38E-204	6.55E-201
Kobayashi *et al.*	251	103	0.4104	8.78E-184	2.08E-180
Shedden *et al.*	456	115	0.2522	4.88E-178	7.71E-175
Fischer *et al.*	929	131	0.141	6.31E-170	7.47E-167
Pescini Gobert *et al.*	570	116	0.2035	1.28E-167	1.21E-164
Rosty *et al.*	140	82	0.5857	2.01E-161	1.59E-158
Dutertre *et al.*	324	91	0.2809	1.09E-142	7.39E-140
Dodd *et al.*	1375	124	0.0902	3.33E-134	1.97E-131
Kinsey *et al.*	1278	121	0.0947	5.25E-133	2.77E-130
Graham *et al.*	181	72	0.3978	1.83E-124	8.65E-122

5.1. Supplementary Figures

5.1.1. Determining the optimal operator for meta-features formation

In this section, we will investigate how integration of gene expressions using different operators could influence the yielded performance. Result of this analysis is represented in Figure 5.1.

5.1.2. Poor genes tend to yield more synergy compared to predictive genes

In this analysis, we will show that an anti-correlated relationship exists between highly predictive genes and degree of synergy that can be observed when combined to best synergistic gene. Result of this experiment is represented in Figure 5.2.

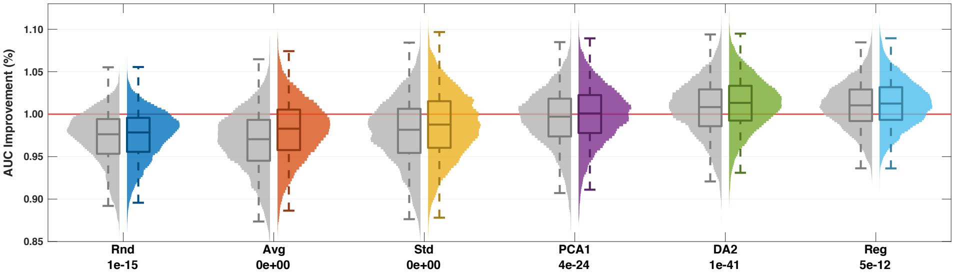


Figure 5.1: Performance gained from meta-genes assembled with different operators. To form a gene set, each gene is selected as "seed" along with its closest 20 neighbor genes according to STRING network. A meta-gene is constructed by averaging the genes (Avg distribution), taking the standard deviation (Std), taking the largest principal component (PCA1), negating the expression of genes that are anti-correlated with outcome before averaging (DA2 [22]) and finally for the Reg meta-gene a linear regression is trained on 50% of patients (randomly selected) and applied to the test set (other 50%) to form the meta-gene. In this figure, the AUC of each meta-gene is compared to the AUC of the best gene in the set (determined in the training set). As a control, we also included a meta-gene where a random gene in the gene set is chosen as the best gene (Rnd). The gray distributions indicate performance gained by a shuffled version of the STRING network (as described in the Methods). The number below each pair of distributions, represents the p-value of a one-sided t-test comparing the distributions. This result shows that the average operator performs on par with the random operator although it is widely used in NOPs. Meanwhile, the DA2 operator, which simply adjust gene directions according to outcome, performs substantially better in improving performance of meta-genes. The top performing operator is regression.

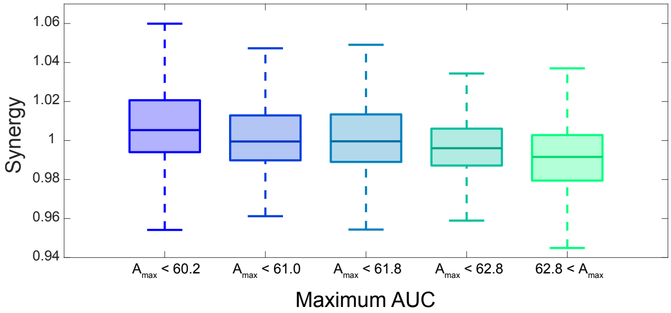


Figure 5.2: Poorly performing genes tend to yield more synergy compared to predictive genes. The top $10k$ pair in SyNet is selected and grouped into 5 non-overlapping bins according to $max(A_i, A_j)$ where A_i and A_j are the AUC of gene i and gene j respectively. For each group, bars represent distribution of synergy (S_{ij}). This result show that for higher individual AUCs, synergy is reduced.

5.1.3. PERFORMANCE OF CLASSICAL PREDICTORS

In order to identify a baseline model, we investigated performance of 12 common regular classifiers for predicting outcome of patients in our collected dataset. Result of this analysis is represented in Figure 5.3.

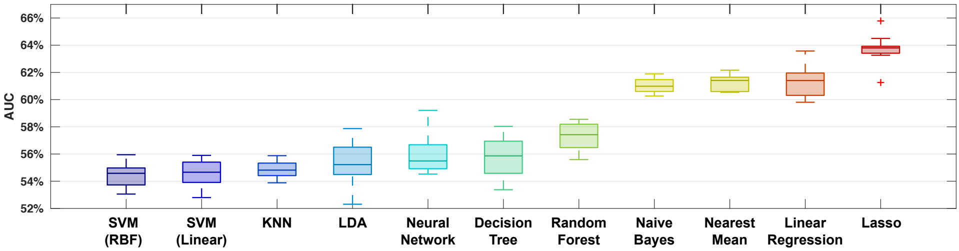


Figure 5.3: Performance of standard classifiers trained using individual features (no network information is utilized). The cross-study validation scheme is employed to evaluate range of linear and nonlinear classifiers including Naive Bayes (NB) , Nearest Mean Classifier (NMC), Linear Discriminant Analysis (LDA), Linear Regression (LR), Lasso, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) using linear and Radial Basis Functions (RBF), Decision-Tree (DT), Random Forest (RF) and Neural Networks (NN) (Mitchell 1997). The hyperparameters of each model (e.g. the gamma parameter in SVM) is optimized by means of inner fold cross-study evaluations. Samples used to train and validate these classifiers are identical to the samples used for the performance evaluations in the main manuscript. According to these results, linear classifiers offer improved performance compared to non linear (and more complex) classifiers (with exception of LDA). The best performing classifier is the Lasso which regularises gene weights to perform a simultaneous selection and integration of gene expressions. Therefore, performance is improved by marginal increase of complexity.

5.1.4. BATCH EFFECT REMOVAL USING COMBAT

We employed COMBAT to remove the expected batch effects from our collected dataset. To confirm reduction of batch effects, we visualized expression profile of patients using t-SNE [23]. Result of this analysis is represented in Figure 5.4.

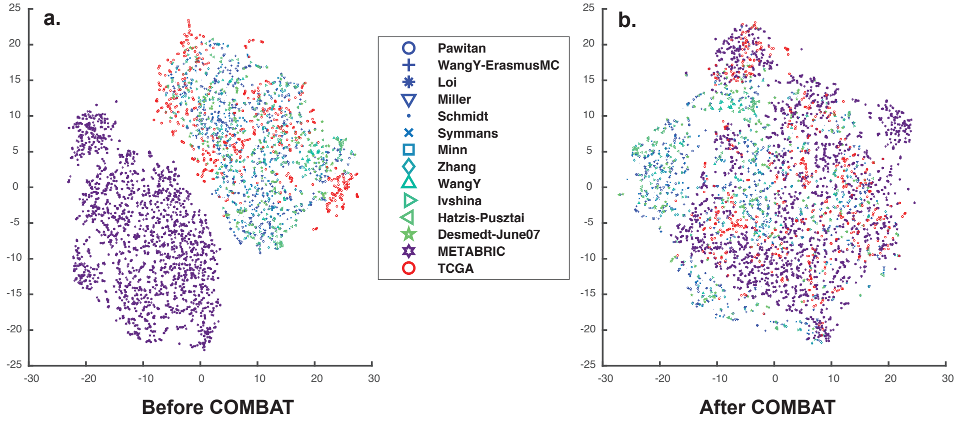


Figure 5.4: COMBAT is successfully employed to remove batch effects between studies. **a.** Expression data (n=4109) is first quantile normalized within each study and then visualized using t-SNE (perplexity=20) in two-dimensional space. Based to this visualization, METABRIC data is clearly occupying a different part of gene expression space compared to other studies. **b.** Batch effect removed data after applying COMBAT. According to t-SNE visualization, data from different studies show homogenous patterns in the expression space.

5.1.5. CROSS-VALIDATION SCHEME

Figure 5.5 demonstrates an schematic overview of cross-validation procedure we considered to evaluate performance of models under study.

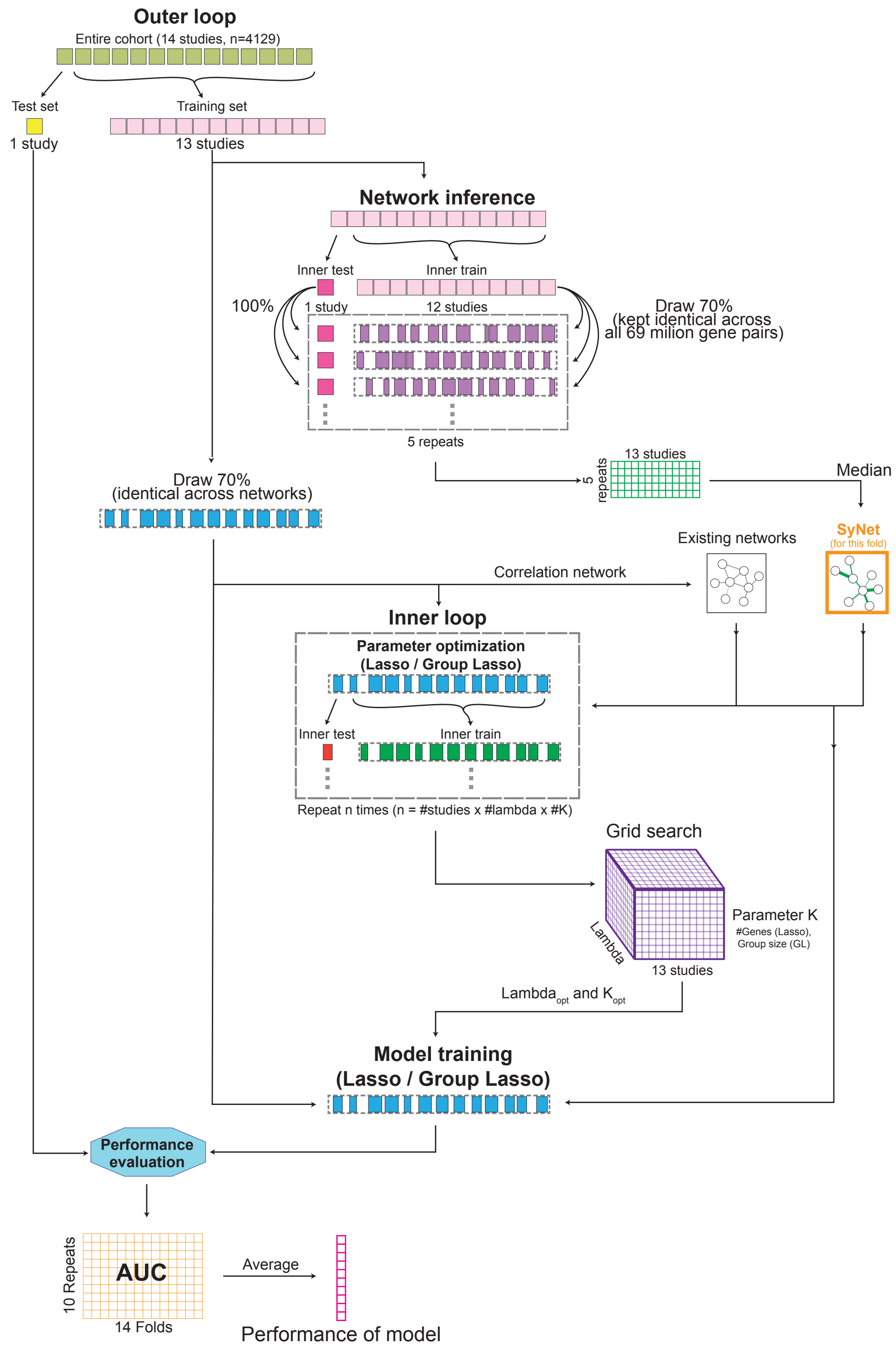


Figure 5.5: The utilized cross-validation scheme to investigate performance of NOPs.

5.1.6. Performance variation in leave one study out cross validation

Even after rigorous batch effect removal (see 5.1.4), we identified a substantial bias for performance of samples collected from individual studies (see Figure 5.6). This indicates that batch effects are still present in our collected dataset. Another explanation could be study specific quality of samples where some studies provided expression profiles with better (more homogeneous expression) compared to others.

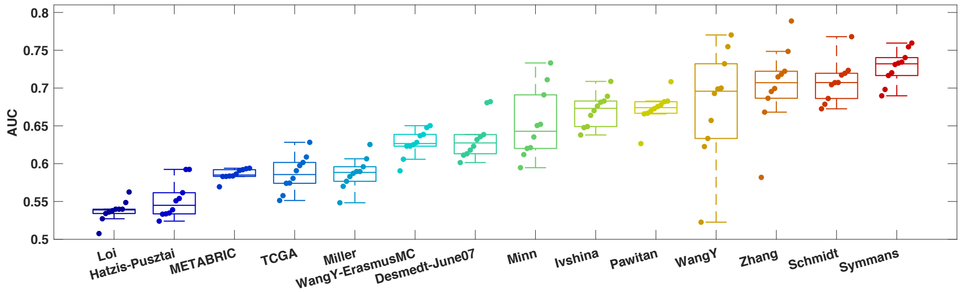


Figure 5.6: Performance of Lasso for predicting survival risk of patients is highly variable across studies. To calculate Lasso performance, this classifier is trained using 70% of samples from 13 studies and tested over the entire samples in the left out study. This procedure is repeated 10 times for each study.

5.1.7. Comparison of performance for networks and groups of identical size

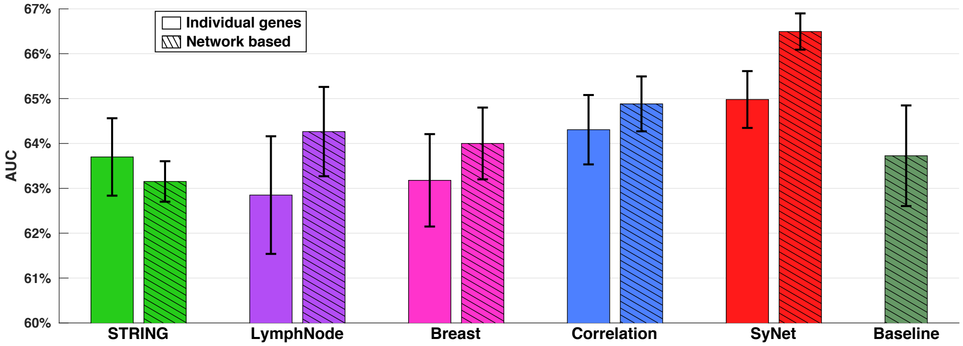


Figure 5.7: Similar trend of performance is observed when number of links in networks are kept identical ($n = 10000$). No grid search for group size is performed in this analysis and group size is kept constant (K=5).

5.1.8. Sparse group lasso performance compared to group Lasso

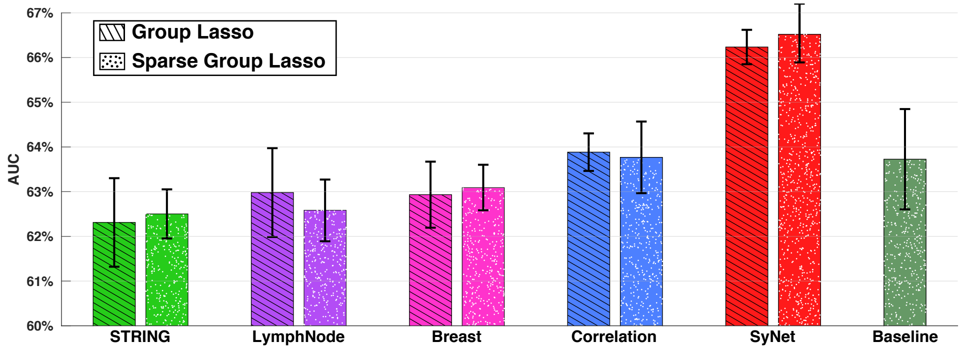


Figure 5.8: While Sparse Group Lasso is computationally more expensive than group Lasso, it does not outperform group Lasso in terms of performance. Identical set of Lambda for both feature level and group level regularization of sparse group lasso is considered. These two parameters are optimized in an inner loop cross-validation fashion as explain in the paper.

5.1.9. PERFORMANCE OF SYNET DOES NOT CHANGE IF #GENES AND GROUP SIZE ARE OPTIMIZED SIMULTANEOUSLY

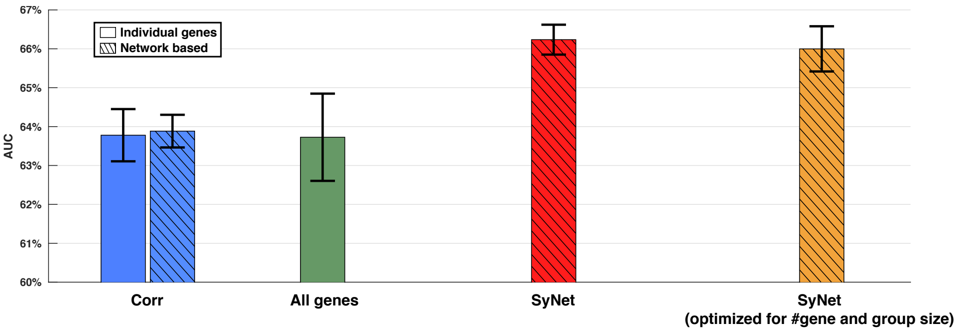


Figure 5.9: Performance of SyNet does not change substantially if number of genes and group size are optimized simultaneously. Instead of optimizing group size and number of genes separately (which is done in the paper, red bar), one can optimize these parameters simultaneously (at the cost of computation time). To this end, a grid search is employed to search across set of group sizes (2, 3, 5, 7 and 10) and number of genes (100, 300, 700, 1000, 1500 and 3000) to compare performance of GL in these two settings (i.e. separate vs. simultaneous optimization). The results indicates that performance of this concurrently optimized model (orange bar) does not change substantially compared to a case when these parameters (group size and number of genes) are optimized separately.

5.1.10. Similarity between biological networks and SyNet without the correlation criterion

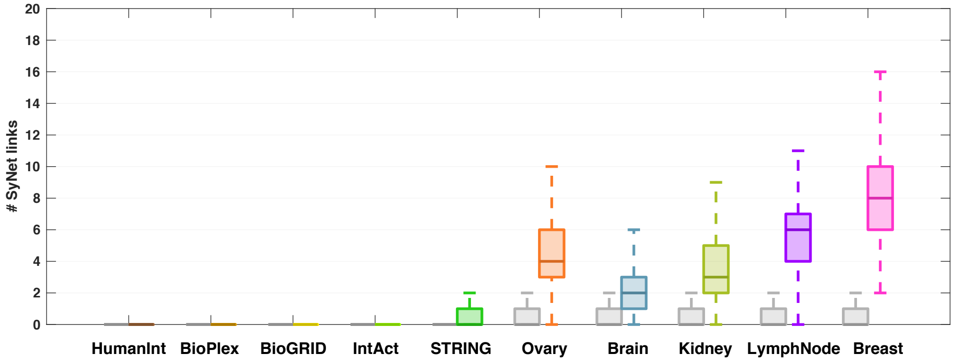


Figure 5.10: Overlap of SyNet links with existing biological networks is not purely driven by correlation criterion. To investigate this, Fitness (F_{ij}) of all pairs are calculated only using synergy (S_{ij}) and average AUC (M_{ij}) while ignoring correlation component (i.e. $F_{ij} = -\sqrt[2]{(1-\overline{S_{ij}})^2 + (1-\overline{M_{ij}})^2}$). Similar to the analysis presented in the main paper, the existence of top SyNet pairs ($n = 3544$, according to the new Fitness) in existing biological networks is assessed by randomly sampling equal number of links ($n = 3544$) in the biological network. The frequency of observing overlapping links are depicted as boxplots. Gray box plots indicate the same analysis performed on the shuffled version of the biological networks.

5.1.11. Precision recall curves for overlap between SyNet and existing networks

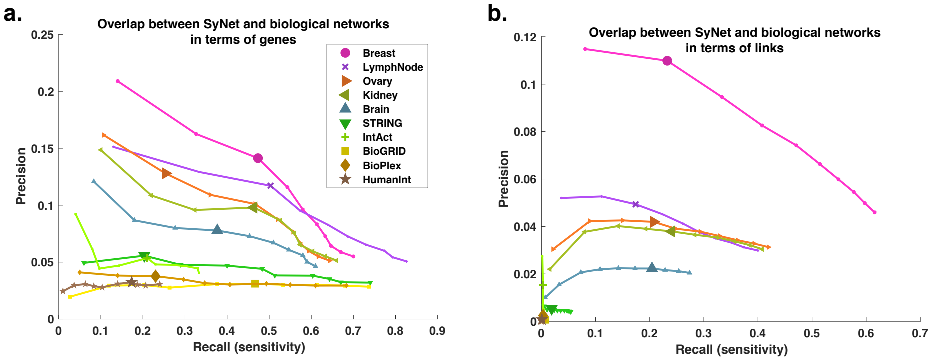


Figure 5.11: Precision recall curves for overlap between SyNet and existing networks. Existing biological networks miss many **a.** genes and **b.** links necessary for outcome prediction. The curves demonstrate the degree of similarity between SyNet genes and links that are also present in biological networks across set of thresholds (5%-100% of total genes/links with steps of 5%). Large markers indicate the threshold with maximum F1-score (computed from precision and recall at each threshold) across considered thresholds.

5.1.12. Performance of Corr network compared to shuffled version

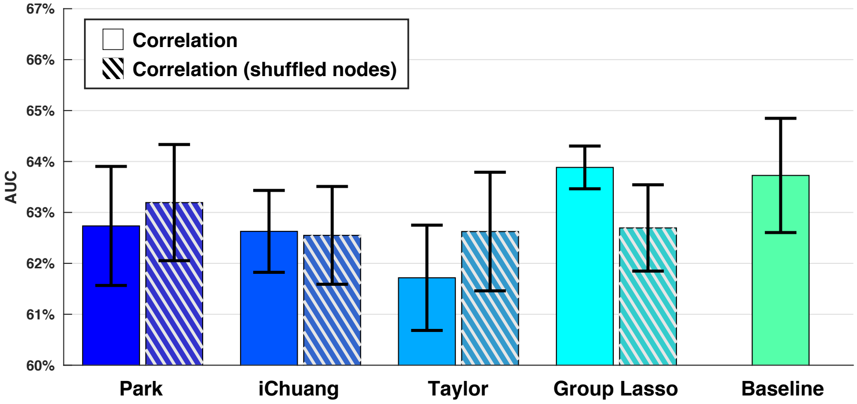


Figure 5.12: Performance of correlation network do not show deterioration in existing NOPs but it does when group lasso used. This indicates that existing NOPs do not effectively incorporate interactions in the given network.

5.1.13. PERFORMANCE OF TOP OUTCOME PREDICTORS WITH LIMITED SAMPLES (SUB-SAMPLING ANALYSIS)

We focused only on the top three predictive networks (i.e. STRING, Correlation and SyNet) and trained a Group Lasso model using an identical set of (training and test) samples as used in the main manuscript for each network. Additionally, we evaluated performance of the baseline model (i.e. Lasso using all genes available in our collected dataset, n=11748). Results of this experiment are represented in Figure 5.13. As expected, our assessment shows reduced classification performance for all models. This reduction is most severe for models that use the training data to infer corresponding network (i.e. Correlation and SyNet). Interestingly, we found that even using 25% of training data, a Group Lasso model guided by SyNet performs better than a model that is guided by Correlation network (the second-best performing network). This shows that, even with a limited number of samples, data-driven gene networks can guide training of outcome predictors.

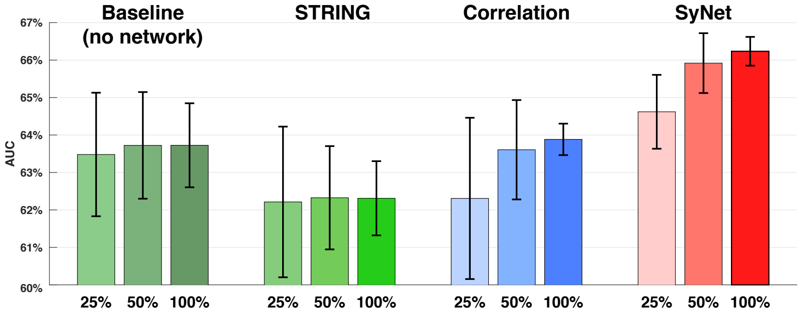


Figure 5.13: Performance of top three predictive networks (as well as the baseline model) when trained using 25% or 50% of samples available in our dataset. SyNet outperforms other models even though it uses only 25% of samples to infer its synergistic network.

5.1.14. Performance of hub genes in SyNet

We selected genes in SyNet with at least 5 neighbors (i.e. degree >=5) and trained Lasso as well as Group Lasso across 14 folds and 10 repeats using identical samples as the analyses in the main manuscript. The median number of genes used in each fold was 175, meaning that nearly half of genes in the original SyNet were given to these "hub-based models". Results of this experiment are visualized in Figure 5.14. Our results show that both Lasso and Group Lasso provide similar performance if they are limited specifically to hub (i.e. degree >=5) genes. However, these models exhibit a slightly larger standard deviation of the performance indicative of a reduced stability of the performance. Therefore, we argue that although "core" genes are important in performance of outcome predictors, proper expression integration of core genes and their (synergistic) neighbors can plays a crucial role in performance and stability of the outcome predictors.

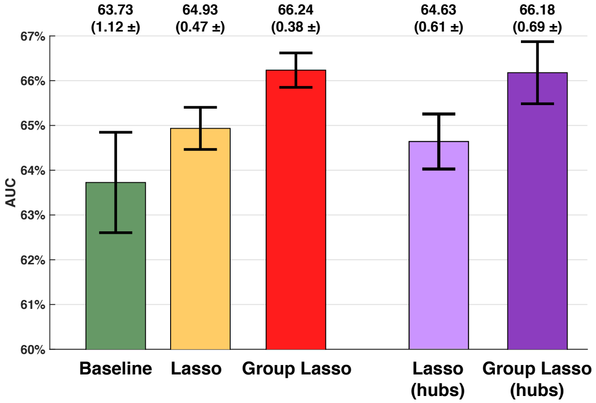


Figure 5.14: Performance of Lasso and Group lasso when only hub (degree >= 5) genes are used. The average and standard deviation of the performance for each model across 14 studies and 10 cross-validation repeats are represented by bars and error bars respectively. These values (i.e. mean and standard deviation) are also represented by numbers above each bar respectively.

5.1.15. PERFORMANCE OF MERGED NETWORKS

To investigate whether a combination of networks would provide a better performance compared to individual networks, we merged 100,000 top links from the top performing networks including STRING, Breast and LymphNode (in total of 202237 links after removal of duplicates) and trained Lasso and Group Lasso to predict survival of unseen patients across independent studies. For this assessment, we utilized an identical set of training and test samples that are also used in the main manuscript. Figure 5.15 represents the result of this experiment. According to these results, combining links from multiple network does not improve performance of classical or network-based outcome predictors. This could be result of excessive number of links that are used from these networks (i.e. curse of dimensionality), or lack of confidence for the included links (as threshold is reduced to include 100k links) which further masks the predictive information in the utilized links. To investigate whether using links with high confidence helps to boost performance of these models, we formed a new network using top one million links in STRING, LymphNode and Breast (in total 3 million links) and selected pairs that are present in all three networks (n=77961 links connecting n=5986 genes). Next, we trained Lasso and Group Lasso using the identical settings as the main manuscript (14 folds, 10 repeats). Corresponding model performances are depicted in Figure 5.15 (dark blue bars). Our analysis shows that using links that are shared by multiple networks have a modest positive impact on Group Lasso's performance. However, reduction of Lasso performance (which only uses genes in this network) hints to lower performance of (individual) genes included in the newly formed network. This could be explained by the fact that intersecting genes from diverse networks would result in selection of broadly active genes that may have a lower specificity to the tissue or even more importantly the disease of interest. In agreement with our conclusion in this paper, we argue that a disease specific measure of selection should be implemented in NOPs to ensure that genes and more crucially their corresponding links contain predictive information that could be extracted by the final classifier.

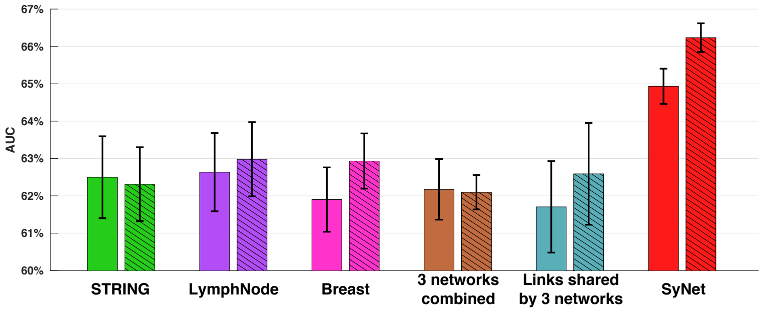


Figure 5.15: Performance of top three networks compared to a case when top 100k links from each network are combined into a single aggregated network (#link= 202k) as well as when top shared links in these networks are used in a cross-study validation procedure. SyNet substantially outperforms these networks although it contains only 300 genes and 3000 links.

5.1.16. Performance of models under study do not change with more populated networks

In the analyses throughout our manuscript, we limited networks under study to have maximum of 50k links to maintain a reasonable network size (in terms of number of links) which results in reducing computational burden. As gene groups in our analyses are formed according to the top weighted neighbors for each gene, reducing confidence threshold (which results in more links in each network) may have little to no influence in the final set of gene groups formed. To demonstrate this effect, we increased number of links for the top three networks (i.e. STRING, LymphNode and Breast) from 50k to 100k, 250k and 500k and trained Lasso and Group Lasso using identical settings as the analyses in the original manuscript (i.e. 14 folds and 10 repeats). Result of this experiment is represented in Figure 5.16. This result demonstrates the minor influence of this threshold on the performance of models under study.

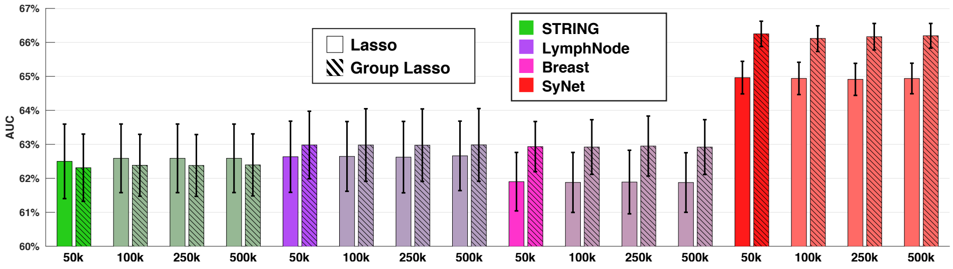


Figure 5.16: Performance of the models under study show minor changes when confidence threshold is reduced to include more links in the network. Mean and standard deviation of performances across 10 repeats are denoted by bars and error bars respectively. Numbers below each bar represent number of links used.

5.1.17. OVERLAP BETWEEN EXISTING NETWORKS AND SYNET ACROSS VARIOUS LINK WEIGHT THRESHOLDS

To investigate whether network size threshold has an impact in our conclusions, we performed an overlap analysis for all 10 networks under study (i.e. BioGRID, Breast, Lymph node, etc.) using four smaller and larger thresholds (namely: 10k, 100k, 250k and 1000k links). To form a binary network for each threshold X, we utilized top X number of links in each network. Next, we randomly selected 3544 links (equal to number of links in SyNet) and asked how many of these selected links are also present in SyNet. This procedure is repeated 1000 times to produce an “observed” number of SyNet links in the networks under study. To estimate an expected distribution of SyNet links, we repeated the same experiment while nodes in each thresholded network were randomly swapped with other nodes in the network in each selected round. A summary result of this analysis is represented in Figure 5.17. As expected, the distribution of observed vs. expected number of SyNet links become more similar when network size increases (due to presence of more irrelevant and low confidence links). According to this result, changing network size threshold has a minor effect in the observed trend in the main paper. Therefore, in corroboration with our argument in the main paper, we conclude that more related tissue-specific networks (i.e. Breast and Lymph node) show larger overlap with SyNet compared to other networks (i.e. other tissue specific or generic networks).

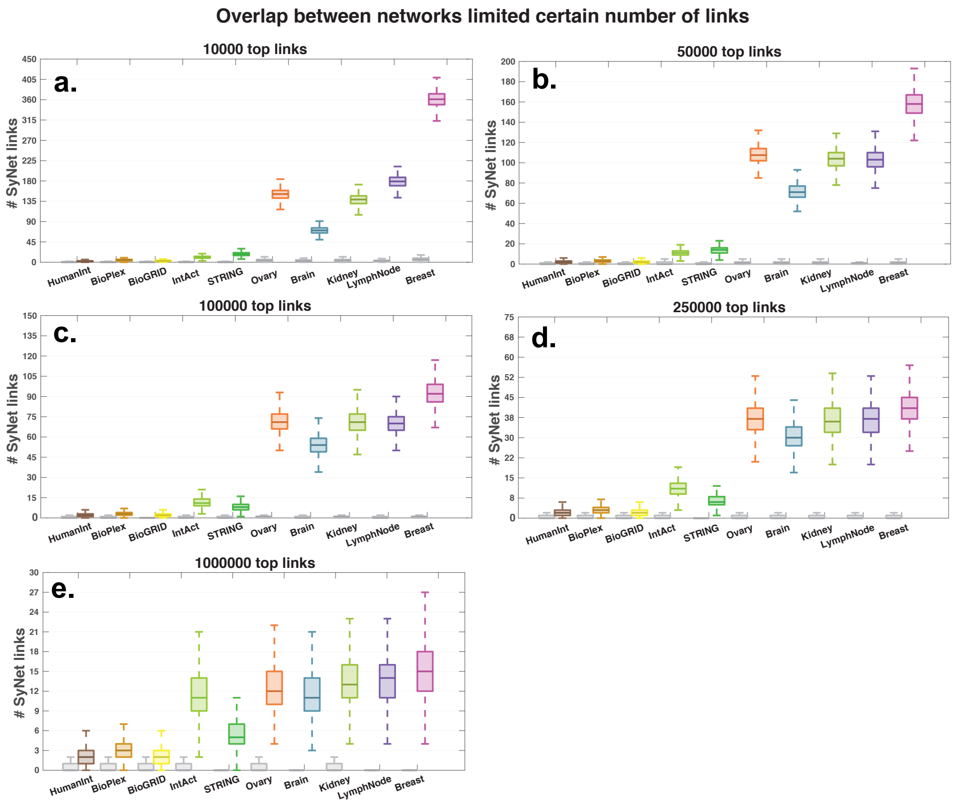


Figure 5.17: Overlaps between phenotype-naive networks and SyNet show a similar trend across smaller or larger network sizes. Overlap analysis between links for network under study when limited to **a.** 10k links, **b.** 50k links, **c.** 100k links, **d.** 250k links and **e.** 1000k links. As expected, observed vs. expected number of SyNet links are more similar for larger network sizes.

5.1.18. Performance of subtype-specific network in outcome prediction

We collected three subtype specific networks (i.e. Basal-A, Basal-B and Luminal) inferred by Zaman *et al.* [24]. These networks are collected from http://www.bri.nrc.ca/wang/ and combined to form a Subtype Specific Breast Cancer network (SSBC). We utilized all genes and links in SSBC network to train and test both Lasso and Group Lasso classifiers across 14 folds and 10 repeats using identical (train/test) samples as the main manuscript. The result of this experiment is represented in Figure 5.18.

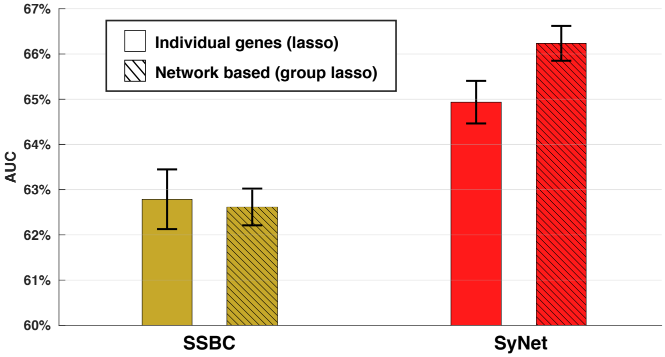


Figure 5.18: Performance of Subtype Specific Breast Cancer (SSBC) network compared to SyNet.

REFERENCES

[1] M. Goldman, B. Craft, J. Zhu, T. Swatloski, M. Cline, and D. Haussler, *Abstract 5270: The UCSC xena system for integrating and visualizing functional genomics,* Cancer Res. **76**, 5270 (2016).

[2] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal, *A proteome-scale map of the human interactome network,* Cell **159**, 1212 (2014).

[3] E. L. Huttlin, R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, J. Szpyt, S. Tam, G. Zarraga, L. Pontano-Vaites, S. Swarup, A. E. White, D. K. Schweppe, R. Rad, B. K. Erickson, R. A. Obar, K. G. Guruharsha, K. Li, S. Artavanis-Tsakonas, S. P. Gygi, and J. W. Harper, *Architecture of the human interactome defines protein communities and disease networks,* Nature **545**, 505 (2017).

[4] A. Chatr-Aryamontri, R. Oughtred, L. Boucher, J. Rust, C. Chang, N. K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld, A. Sellam, C. Stark, B.-J. Breitkreutz, K. Dolinski, and M. Tyers, *The BioGRID interaction database: 2017 update,* Nucleic Acids Res. **45**, D369 (2017).

[5] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. von Mering, *The STRING*

database in 2017: quality-controlled protein–protein association networks, made broadly accessible, Nucleic Acids Res. **45**, D362 (2016).

[6] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, D. I. Chasman, G. A. FitzGerald, K. Dolinski, T. Grosser, and O. G. Troyanskaya, *Understanding multicellular function and disease with human tissue-specific networks,* Nat. Genet. **47**, 569 (2015).

[7] R. J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, P. Kersey, and P. Flicek, *Ensembl BioMarts: a hub for data retrieval across taxonomic space,* Database **2011**, bar030 (2011).

[8] A. Allahyar, J. de Ridder, and J. Ubels, *A data-driven interactome of synergistic genes improves network based cancer outcome prediction,* bioRxiv (2018), 10.1101/349688.

[9] J. L. Gross, J. Yellen, and P. Zhang, *Handbook of Graph Theory, Second Edition* (CRC Press, 2013).

[10] N. Perra and S. Fortunato, *Spectral centrality measures in complex networks,* Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **78**, 036107 (2008).

[11] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,* Proc. Natl. Acad. Sci. U. S. A. **102**, 15545 (2005).

[12] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi, *Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis,* JNCI: Journal of the National Cancer Institute **98**, 262 (2006).

[13] S. Kobayashi, T. Shimamura, S. Monti, U. Steidl, C. J. Hetherington, A. M. Lowell, T. Golub, M. Meyerson, D. G. Tenen, G. I. Shapiro, and B. Halmos, *Transcriptional profiling identifies cyclin d1 as a critical downstream effector of mutant epidermal growth factor receptor signaling,* Cancer Research **66**, 11389 (2006).

[14] K. Shedden, J. M. Taylor, S. A. Enkemann, M.-S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, *et al.*, *Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study,* Nature medicine **14**, 822 (2008).

[15] M. Fischer, P. Grossmann, M. Padi, and J. A. DeCaprio, *Integration of tp53, dream, mmb-foxm1 and rb-e2f target gene analyses identifies cell cycle gene regulatory networks,* Nucleic Acids Research **44**, 6070 (2016).

[16] R. Pescini Gobert, L. Joubert, M.-L. Curchod, C. Salvat, I. Foucault, C. Jorand-Lebrun, M. Lamarine, H. Peixoto, C. Vignaud, C. Frémaux, T. Jomotte, B. Françon, C. Alliod, L. Bernasconi, H. Abderrahim, D. Perrin, A. Bombrun, F. Zanoguera, C. Rommel, and R. H. van Huijsduijnen, *Convergent functional genomics of oligodendrocyte differentiation identifies multiple autoinhibitory signaling circuits,* Molecular and Cellular Biology **29**, 1538 (2009).

[17] C. Rosty, M. Sheffer, D. Tsafrir, N. Stransky, I. Tsafrir, M. Peter, P. de Cremoux, A. de La Rochefordiere, R. Salmon, T. Dorval, *et al.*, *Identification of a proliferation gene cluster associated with hpv e6/e7 expression level and viral dna load in invasive cervical carcinoma,* Oncogene **24**, 7094 (2005).

[18] M. Dutertre, L. Gratadou, E. Dardenne, S. Germann, S. Samaan, R. Lidereau, K. Driouch, P. de la Grange, and D. Auboeuf, *Estrogen regulation and physiopathologic significance of alternative promoters in breast cancer,* Cancer Research **70**, 3760 (2010).

[19] L. E. Dodd, S. Sengupta, I.-H. Chen, J. A. den Boon, Y.-J. Cheng, W. Westra, M. A. Newton, B. F. Mittl, L. McShane, C.-J. Chen, P. Ahlquist, and A. Hildesheim, *Genes involved in dna repair and nitrosamine metabolism and those located on chromosome 14q32 are dysregulated in nasopharyngeal carcinoma,* Cancer Epidemiology and Prevention Biomarkers **15**, 2216 (2006).

[20] M. Kinsey, R. Smith, and S. L. Lessnick, *Nr0b1 is required for the oncogenic phenotype mediated by ews/fli in ewing's sarcoma,* Molecular Cancer Research **4**, 851 (2006).

[21] S. M. Graham, J. K. Vass, T. L. Holyoake, and G. J. Graham, *Transcriptional analysis of quiescent and proliferating cd34+ human hemopoietic cells from normal and chronic myeloid leukemia sources,* STEM CELLS **25**, 3111 (2007).

[22] A. Allahyar and J. de Ridder, *Feral: network-based classifier with application to breast cancer outcome prediction,* Bioinformatics **31**, i311 (2015).

[23] L. v. d. Maaten and G. Hinton, *Visualizing data using t-SNE,* J. Mach. Learn. Res. **9**, 2579 (2008).

[24] N. Zaman, L. Li, M. L. Jaramillo, Z. Sun, C. Tibiche, M. Banville, C. Collins, M. Trifiro, M. Paliouras, A. Nantel, *et al.*, *Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets,* Cell reports **5**, 216 (2013).

6

MC-4C: Multi-Contact Circular Chromosome Conformation Capture

Amin Allahyar
Carlo Vermeulen
Britta A.M. Bouwman
Peter H.L. Krijger
Marjon J.A.M. Verstegen
Geert Geeven
Melissa van Kranenburg
Mark Pieterse
Roy Straver
Judith H.I. Haarhuis
K. Jalink
Hans Teunissen
Ivo J. Renkens
Wigard P. Kloosterman
Benjamin D. Rowland
Elzo de Wit
Jeroen de Ridder
Wouter de Laat

This Chapter is published in Nature Genetics (2018); https://doi.org/10.1038/s41588-018-0161-5 [1].

Locus-Specific Enhancer Hubs And Architectural Loop Collisions Uncovered From Single Allele DNA Topologies

Amin Allahyar, Carlo Vermeulen, Britta A.M. Bouwman, Peter H.L. Krijger, Marjon J.A.M. Verstegen, Geert Geeven, Melissa van Kranenburg, Mark Pieterse, Roy Straver, Judith H.I. Haarhuis, K. Jalink, Hans Teunissen, Ivo J. Renkens, Wigard P. Kloosterman, Benjamin D. Rowland, Elzo de Wit, Jeroen de Ridder, Wouter de Laat

6.1. Abstract

Chromatin folding contributes to the regulation of genomic processes such as gene activity. Existing conformation capture methods characterize genome topology through analysis of pairwise chromatin contacts in populations of cells but cannot discern whether individual interactions occur simultaneously or competitively. Here we present multi-contact 4C (MC-4C), which applies Nanopore sequencing to study multi-way DNA conformations of individual alleles. MC-4C distinguishes cooperative from random and competing interactions and identifies previously missed structures in subpopulations of cells. We show that individual elements of the β-globin superenhancer can aggregate into an enhancer hub that can simultaneously accommodate two genes. Neighboring chromatin domain loops can form rosette-like structures through collision of their CTCF-bound anchors, as seen most prominently in cells lacking the cohesin-unloading factor WAPL. Here, massive collision of CTCF-anchored chromatin loops is believed to reflect 'cohesin traffic jams'. Single-allele topology studies thus help us understand the mechanisms underlying genome folding and functioning.

6.2. Introduction

The invention of chromatin conformation capture (3C) technology [2] and derived methods [3] has greatly advanced our knowledge of the principles and regulatory potential of 3D genome folding in vivo. Insights obtained from genome-wide contact maps derived from Hi-C data include the discovery of topologically associated domains (TADs), structurally insulated units of chromosomes of on average a megabase in size [4–6], and of compartments, nuclear environments in which TADs with similar epigenetic signatures spatially cluster [7]. TADs and nuclear compartments are believed to contribute to genome functioning, whereas chromatin loops are thought to influence genome functioning in a more deterministic, direct fashion. Such loops can only be detected when zooming to a much finer scale than whole chromosomes and TADs, either by ultra-deep Hi-C sequencing or by the application of targeted high-resolution approaches such 4C, 5C or capture-C technologies. Chromatin loops include architectural loops, often anchored by bound CTCF proteins, that form structural chromosomal domains [8, 9] and regulatory chromatin loops that bring distal enhancers in close physical proximity to target gene promoters to control their transcriptional output. Detailed topological studies and genetic evidence have further indicated that individual enhancers can contact and control the expression of multiple genes. Conversely, single genes are often influenced by multiple enhancers [10, 11]. Similarly, in population-based assays, individual CTCF sites can be seen contacting multiple other CTCF sites. Based on such observations it has been hypothesized that DNA may fold into spatial chromatin hubs [12, 13]. However, current population-based pair-wise contact matrices cannot distinguish clustered interactions from mutually exclusive interactions that independently occur in different cells. To investigate the existence and nature of specific hubs formed between regulatory sequences, CTCF-binding sites and/or genes, targeted high-resolution and high-throughput strategies are needed for detection, analysis and interpretation of multi-way DNA contacts.

Recently, several 3C procedures have been modified for the study of multi-way contacts between selected genes and regulatory sequences, but so far these approaches have been inherently limited in contact complexity, complicating the interpretation of their data [14–17]. At the genome-wide level, recent breakthroughs in the analysis of multi-way contacts have been made. These technologies give insight into the types of genomic sequences that tend to co-occupy nuclear compartments. For example, a new genome-wide approach for multi-contact analysis, called C-Walks (chromosomal walks) [17], gave a glimpse of the nuclear aggregation of genomic loci, indicating that, at the compartment level, cooperative aggregation between dispersed intra- and inter-chromosomal sequences may be rare but may occur, for example, at Polycomb bodies. C-walks, three-way Hi-C contact analysis [15] and genome architecture mapping [18] are all genome-wide methods that do not offer the local coverage necessary to study the functionally most relevant fine-scale topologies formed at individual genes, individual regulatory sequences and individual domain anchors. To enable this analysis and to dissect the spatial interplay between multiple individual regulatory DNA elements and genes, we developed multi-contact 4C sequencing (MC-4C).

6.3. Results

6.3.1. MC-4C enables investigation of multi-way DNA conformations.

MC-4C is premised on the fact that 3C-based protocols generate aggregates of DNA segments that reside in each other's 3D proximity in the nucleus. These 'DNA hairballs' are created via in situ formaldehyde cross-linking of chromatin, followed by restriction enzyme-mediated DNA fragmentation and proximity-based re-ligation of cross-linked DNA fragments. The resultant DNA concatemers are characteristically sized >10 kb [19]. Conventional 3C protocols trim these products further to enable efficient analysis of singular ligation junctions only. The MC-4C protocol is designed to keep these concatemers large, enabling the analysis of multi-way contacts for selected genomic sites of interest through third-generation sequencing, such as the Oxford Nanopore Technologies (ONT) MinION. In brief, MC-4C entails the following steps. Like 4C-seq [20] and targeted locus amplification technology [21], MC-4C selectively PCR-amplifies concatemers with primers specific to a fragment of interest (the 'viewpoint'). For this PCR to be sufficiently effective, 3C PCR template in the range of 2-5kb is made by digesting the large concatemers with a six-cutter restriction enzyme and re-ligation under conditions supporting self-circularization. To reduce prevalent rolling circle amplification and eliminate abundant uninformative undigested products, Cas9-mediated in vitro digestion of the viewpoint fragment (between the inverse PCR primers) and its two neighbor fragments is performed before PCR. After PCR, the product is size-selected (>1.5kb) and sequenced on the MinION sequencing platform (Figure 6.1.a).

An integral component of MC-4C is its elaborate computational analysis strategy (explained in detail in the Methods), which provides the necessary pre-processing of the ONT data and downstream analysis to enable meaningful interpretation of allelic co-occurrence frequencies. To appreciate local multi-way contacts at the level of individual alleles, it is key to filter and select for the informative reads that have two or more contacts within a pre-defined chromosomal region of interest. Such analysis requires substantial coverage, as reads having less than two local contacts are not informative for our multi-way analysis. To compute reliable statistics, it is also essential to efficiently remove all reads originating from PCR duplicates. For this, we designed a PCR duplicate removal strategy that is guided by co-captured fragments far outside the region of interest (Supplementary Fig. 1): the chance of independently capturing a given such fragment more than once is extremely small, implying that these sequences can serve as genomically contributed unique molecular identifiers in MC-4C. After this ultra-conservative but very reliable PCR filtering strategy, every remaining read represents a unique micro-topology derived from an individual allele. MC-4C contact profiles are thus a direct reflection of single allele measurements, which in principle makes them quantitative, albeit limited still by technically inherent variation that may arise from differences in cross-linking, digestion, ligation and mapping ability between fragments.

To explore new biology that may be identified by MC-4C we applied the technique to three different genetic systems. We chose the mouse β-globin and Pcdhα loci, both constituting multiple gene promoters and enhancer and superenhancer (SE) elements that act in concert to control defined developmental and cellular expression patterns. We also selected cohesin-looped topological domain boundaries that, upon cohesin stabilization, show extended loops with much more distal anchor sites in population-based Hi-C [22]. We performed a total of 20 MC-4C experiments (27 MinION sequencing runs) to obtain an average of 13,000 individual allelic micro-topologies, spanning an average total of 80,000 spatial contacts, per viewpoint (Supplementary Table 1).

Figure 6.1 summarizes results from a typical MC-4C experiment. Because of PCR, which has a strong bias for small amplicons, and size selection, which we perform to remove the small amplicons before sequencing, the average raw read size is approximately 2 kb (Figure 6.1.b and Supplementary Fig. 2). Most span three or four spatial contacts, some up to ten (Figure 6.1.c and Supplementary Fig. 3), with spatial contacts being scored based on ligation events between restriction fragments that are not immediately juxtaposed in the reference genome. To further reduce the effect of PCR-related over- or under-representation of fragments, we divided the region of interest into 200 bins and quantified the relative interaction frequencies per bin. As in all other 3C methods, the great majority of captured sequences (from raw reads) localize to the immediate chromosomal vicinity of the viewpoint (Figure 6.1.d and Supplementary Fig. 4). The contact profiles derived from sequences directly ligated to the viewpoint (i.e., those that one would analyze in conventional 4C-sequencing) are almost indistinguishable from those created from the indirectly ligated partners (Supplementary Fig. 5). Collectively this indicates that the additional fragments that we capture and analyze by MC-4C are the result of 3D proximity-based ligation events and represent topologically meaningful genomic multi-way contacts made with the viewpoint fragment.

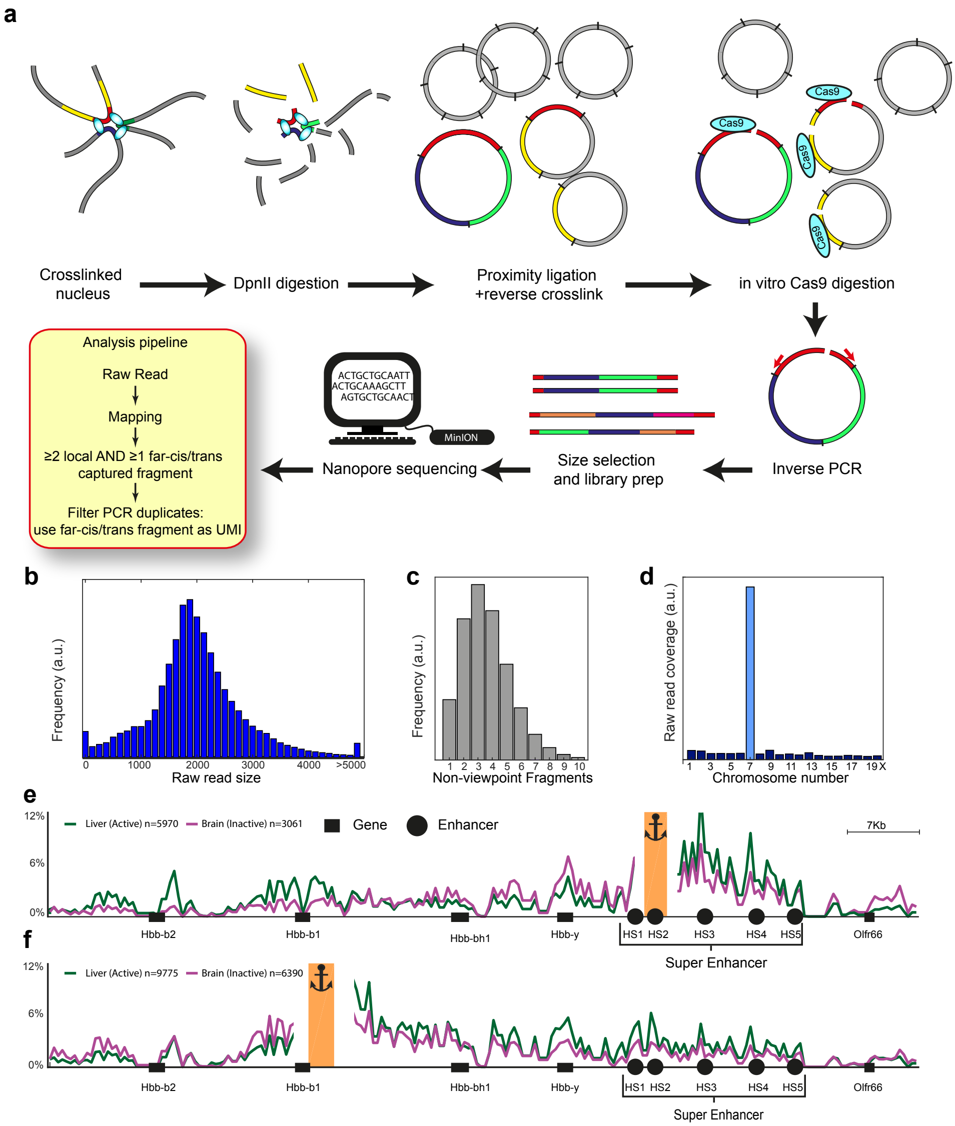


Figure 6.1: Multi-contact 4C technology. **a.** The MC-4C strategy. **b-d.** Statistics of the Hbb-b1 viewpoint in fetal liver cells. UMI, unique molecular identifier. **b.** MC-4C raw read size distribution after ONT MinION sequencing; a.u., arbitrary units. **c.** The number of MC-4C captured fragments per mapped read, excluding the viewpoint fragment. **d.** Chromosomal distribution of captured and mapped fragments. **e, f.** Overall (panallelic) MC-4C contact profile of β-globin HS2 **e.** and Hbb-b1 **f.** in E14.5 fetal liver (green) and brain (purple).

6.3.2. Evidence for an enhancer hub at the β-globin locus.

We first studied higher order conformations of the genetically well-characterized mouse β-globin locus. It carries two embryonic globin genes (Hbby and Hbb-bh1) that compete with two downstream adult globin genes (Hbb-b1 and Hbb-b2) for activation by the upstream β-globin SE [23–25] during development. This SE, also known as the locus control region, is composed of five regulatory elements (hypersensitivity sites (HS) 1-5), of which HS1-HS4 show enhancer activity [26]. Genetic studies in mice further demonstrate that the two developmentally distinct sets of genes compete for activation between sets, but not among members of each set, and that the four enhancer elements of the SE can compensate to a high degree for each other's activity [27, 28]. We performed MC-4C experiments in mouse fetal liver, where the adult genes are highly expressed, and in mouse fetal brain, where the β-globin locus is transcriptionally silent. As viewpoints, we included Hbb-b1, HS2 and HS5, as well as HS3 exclusively in liver. When all fragments captured by the HS2 experiment are aggregated across all individually analyzed alleles in a so-called overall MC-4C contact profile, we find pronounced and precise interactions with the other SE constituents, as well as with the active gene promoters, specifically in expressing (fetal liver) but not in nonexpressing (fetal brain) primary mouse cells (Figure 6.1.e). A similarly detailed and tissue-specific topology is appreciable from the overall MC-4C contact profiles that we obtained when using HS5, Hbb-b1 or HS3 as viewpoints (Figure 6.1.f and Supplementary Fig. 6). MC-4C therefore accurately recapitulates in a qualitative manner the previously observed conformational features of the β-globin locus [13, 29, 30] and additionally specifies contacts within the SE with high precision (see also Supplementary Fig. 6). Results were reproducible between biological replicates, even those sequenced on another third-generation sequencing technology (the Pacific Biosciences sequencing platform) (Supplementary Fig. 7a–d). Nevertheless, in our hands the latter platform provided insufficient reads for the generation of robust contact profiles (Supplementary Fig. 7e), which led us to focus on Nanopore sequencing.

To analyze specific multi-way chromatin conformations adopted by the mouse β-globin locus, we selected from each MC-4C dataset the allelic conformations that contain its viewpoint in contact with a second site of interest (SOI). We then quantified and visualized the contact frequencies with the remaining co-occurring sequences. Figure 6.2.a, b shows two examples of such viewpoint–SOI plots (see also Supplementary Fig. 8). The highly localized peaks exactly at the individual enhancer elements of the SE suggest that alleles that fold to have Hbb-b1 (Figure 6.2.a) or HS5 (Figure 6.2.b) in contact with HS2 are likely to also interact with other SE elements. This would be indicative of enhancer hub formation. We tested this with a statistical method that distinguishes favored from random or disfavored (competitive) multi-way interactions. This method compares through a z-score calculation for each sequence its observed three-way co-occurrence frequency with a given viewpoint–SOI combination to its co-occurrence frequency in conformations where the viewpoint is not in contact with the SOI (6.2a,b and Supplementary Fig. 9). By doing so, we analyze whether the chance of being in contact with any third sequence across the region of interest is increased (favored) or decreased (disfavored) when the viewpoint is interacting with a given SOI. Sequences immediately flanking such SOIs are always found to be enriched in this analysis. This is expected as they cannot be spatially separated from the SOI, but we ignore such immediate neighboring sequences here as their favored detection is not reflective of spatial genome organization. Based on our comparative analysis, we find that contacts with the individual elements of the β-globin SE are significantly favored in conformations that already involve one of them. This preferred co-occurrence is appreciable in allelic conformations involving the distal downstream Hbb-b1 gene, as well as in those involving the upstream HS5 (6.2.c). Particularly for the non-neighboring enhancer elements this seems not the result of mere linear proximity, but a consequence of spatial proximity (Fig. 2c and Supplementary Fig. 8). To further rule out the possibility that preferred co-occurrence is a reflection of linear proximity, we repeated the MC-4C experiments on the same locus in nonexpressing tissue (fetal brain). Here no preferred multi-way interactions were observed beyond the directly neighboring constituents (6.2.d and Supplementary Fig. 8). This shows that the preferred aggregation of β-globin SE constituents seen in expressing cells is not just the consequence of linear proximity. Preferred clustering of active enhancer elements is found even though these sequences are less cross-linkable when active (formaldehyde-assisted isolation of regulatory elements (FAIRE) identifies enhancers through this principle [31]). We thus conclude that the individual elements of the active β-globin SE can form a higher order enhancer hub.

This SE hub will be visited by the globin genes for their activation. To investigate the number of genes the hub can simultaneously accommodate, we analyzed the likelihood of Hbb-b2 and the two embryonic globin genes being in contact with the SE when it is interacting with the adult Hbb-b1 gene (Figure 6.2.f and g). Despite their linear position between the SE and Hbb-b1, the embryonic genes are clearly hindered in contacts with the SE when it is engaged with Hbb-b1, particularly in an active tissue (Figure 6.2.f and g). This suggests that they physically compete with Hbb-b1 for interactions with the active enhancer hub. For Hbb-b2, the other adult globin gene, which is more distal from the SE, we find no indication of physical competition with Hbb-b1 (Figure 6.2.e). Its presence is either normally tolerated or even slightly stimulated in topologies having both SE elements and Hbb-b1 (Figure 6.2.f). MC-4C therefore provides evidence for two higher order topological phenomena. The first is that the individual elements of a single SE, the active β-globin locus control region, can cooperatively interact (i.e., show statistically increased co-occurrence frequencies) to form a spatial enhancer hub. The second is that this single enhancer hub can physically accommodate two genes simultaneously (Figure 6.2.h). We find that, in concordance with detailed gene competition studies at this locus [26–28], partnering at the enhancer hub is allowed between developmentally synchronized genes, but not between genes active at different stages of development. These higher order conformational features therefore provide a topological framework that helps to interpret genetic observations.

Evidence for an enhancer hub at the Pcdhα locus. Higher order topologies may also help control allelic expression patterns in the mouse protocadherin-α (Pcdhα) gene cluster. Per allele, 1 of 12 alternative promoters (those for Pcdha1–Pcdha12) is selected for expression. This ensures that individual neurons express a unique repertoire of membrane-exposed protocadherin molecules, which is essential for axon avoidance [32, 33]. Aside from the variable promoters, two constant promoters are active in every neuron (those for PcdhαC1 and PcdhαC2). The activity of nearly all promoters is regulated by two downstream enhancers, HS7 and HS5-1 (only αC2 seems not to be influenced by HS5-1)

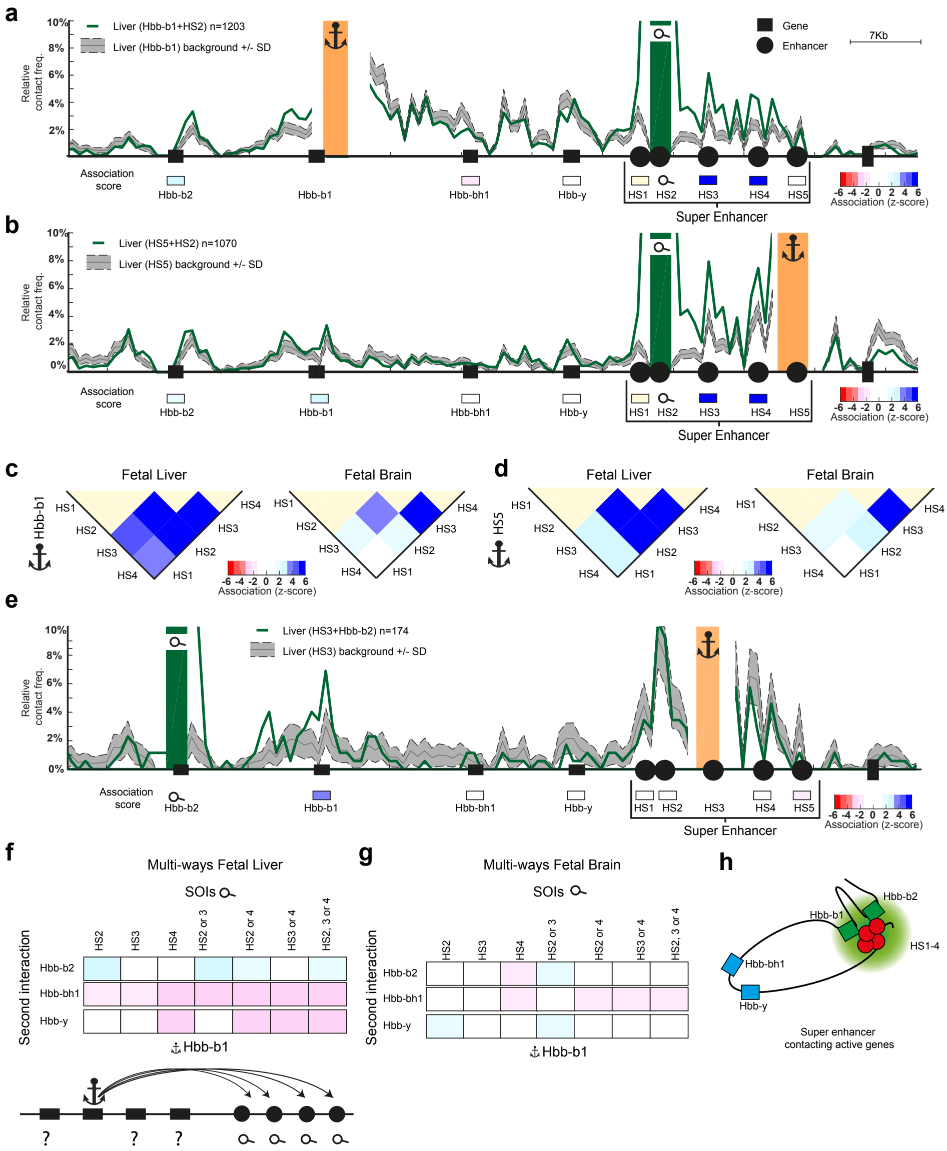


Figure 6.2: A β-globin superenhancer hub that can simultaneously accommodate two genes. **a-b.** Selected microtopologies from fetal liver cells having Hbb-b1 **a.** or HS5 **b.** in contact with HS2 (the number of identified microtopologies is indicated at top left) are specifically enriched for the remainder constituents (HS3 and HS4) of the β-globin superenhancer. Green line shows the observed and gray line the expected (mean ± s.d.) co-occurrence frequency of sites across the locus. z-scores (dark blue indicating significant enrichment, dark red indicating significant depletion of a given site at the interrogated microtopology) are shown for sites of interest in rectangles below each graph. **c-d.**. Summary of all z-scores for all possible pairs of SE elements (HS1 and HS2 are too close to analyze interaction between), when one of them is in contact with the Hbb-b1 gene **c.** or HS5 **d.**, in fetal liver (left; β-globin locus active) and fetal brain (right; β-globin locus inactive). Note the preference for co-occurrence between non-neighboring SE elements specifically in fetal liver cells, revealing an active SE hub. **e.** Selected microtopologies from fetal liver cells having HS3 in contact with Hbb-b2 (the number of microtopologies is indicated in the key). Note that the other active gene, Hbb-b1, is preferentially found at these conformations. **f-g.**. Summary of the clustering behavior (z-scores) of the three remainder β-globin genes when Hbb-b1 is in contact with each of the individual, or combinations of, SE constituents. **h.** Graphic showing the active β-globin superenhancer hub simultaneously contacting the two adult β-globin genes and excluding the two embryonic β-globin genes.

[34, 35]. Forward-oriented CTCF binding to all promoters and reverse-oriented CTCF binding to HS5-1 positively contribute to gene expression [36]. Alternative promoter DNA methylation, which prevents CTCF binding, has been proposed to influence allelic promoter choice [37]. We designed viewpoint primers in both enhancers HS5-1 and HS7 and on the promoters of Pcdhα4 and Pcdhα11 and performed MC-4C analysis in mouse E14.5 fetal brain neurons, which express both Pcdhα variants (Supplementary Fig. 10), and in E14.5 fetal liver cells and NIH-3T3 cells, which do not express from any of the Pcdhα promoters. Data from Pcdhα4 and Pcdhα11 and from HS5-1 and HS-7 were pooled owing to the high similarity between overall profiles. All overall contact profiles showed that contacts between the enhancer and each of the promoter regions were perhaps slightly elevated in brain cells, but overall without dramatic differences in locus topology between fetal brain and inactive cells. This suggests that there is no dominant tissue-specific structure conserved in either fetal brain or inactive cells (Figure 6.3.a and b).

By selectively analyzing the allelic topologies having any of the enhancers in contact with a given alternative promoter in brain cells, we reasoned we could get insight into the specific folding of alleles expressing this particular alternative promoter. As an example, Figure 6.3.c shows how the other sequences of the locus participate in the microtopologies centered around contacts between the Pcdhα4 or Pcdhα11 promoter, when these are contacting HS7. In neurons, these configurations were specifically enriched for the other enhancer, HS5-1 (39kb downstream of HS7), as well as for the constitutively active Pcdhαc2 promoter (34 kb upstream of HS7). In liver cells, the corresponding microtopologies did not specifically engage the HS5-1 enhancer, nor any of the genes, as expected if assuming that here these contacts are a reflection of nonfunctional, random collisions. The brain-specific enhancer hub involving cooperative interactions between HS7 and HS5-1 is similarly appreciable when studying other relevant subsets of allelic conformations (Figure 6.3.d). Additionally, Pcdhαc2 is preferentially found at microtopologies involving interactions between the enhancers and an alternatively transcribed Pcdhα promoter, while Pcdhαc1 is not necessarily evicted from them. The Pcdhα active chromatin hub therefore appears capable of physically accommodating two or more genes at a time. We would have liked to test whether physical competition for enhancer contacts between the Pcdhα1–Pcdhα12 promoters may underlie their mutually exclusive allelic expression in neuronal cells. However, the Pcdha1–Pcdha12 promoters are too close together on the linear chromosome template to observe such mutually exclusive contacts, at least at the current resolution of MC-4C (Supplementary Fig. 11). In summary, as seen for the β-globin SE, the active linearly dispersed individual enhancers HS7 and HS5-1 and the Pcdhαc2 promoter of the Pcdhα locus cooperatively interact to form a tissue-specific active chromatin hub that can simultaneously be contacted by at least one additional gene promoter (Pcdhαc1 or Pcdhα1–Pcdhα12). Notably, our studies on the Pcdhα locus further show that MC-4C can be used to characterize the interaction profiles of rare subpopulations of alleles, identifying topological features that are missed by population-based pairwise contact analysis methods.

WAPL depletion leads to collision of CTCF-anchored loops and to cohesin clustering. As a third model system to study multi-way chromatin interactions, we focused on CTCF and cohesin-anchored chromatin loops. Cohesin is a ring-shaped protein com-

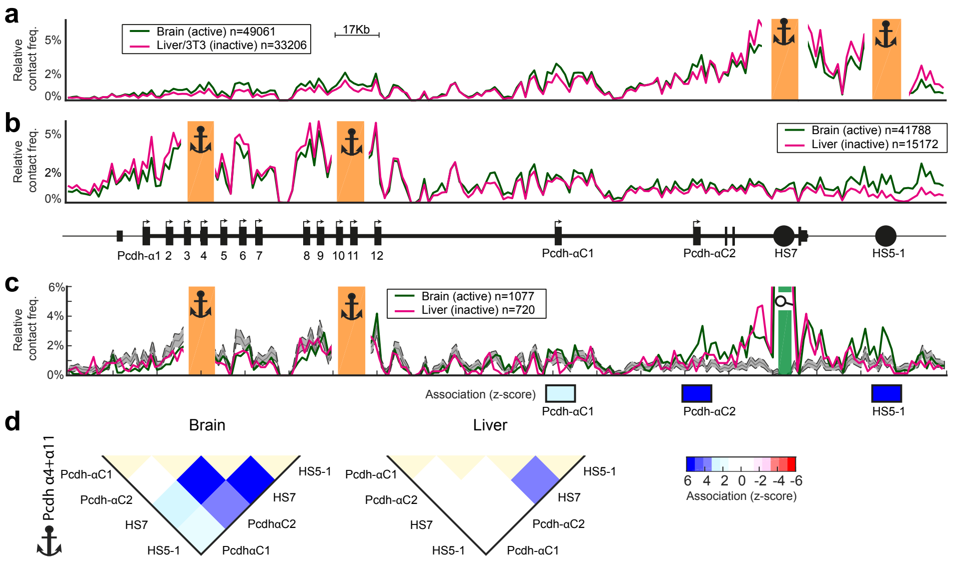


Figure 6.3: MC-4C uncovers Pcdhα hub conformations in tissue-specific subsets of cells. **a**-**b.** Overall (panallelic) MC-4C contact profiles of the combined HS7 and HS5-1 viewpoints **a.** and the combined Pcdhα4 and Pcdhα11 viewpoints **b.** in fetal brain (green; active) and fetal liver or fetal liver and 3T3 cells (purple; inactive). The number of unique reads is indicated in the keys. **c.**, Selected microtopologies from fetal brain (green; active) and fetal liver (purple; inactive) having Pcdhα4 or Pcdhα11 (the viewpoints; see anchors) in contact with HS7 (the SOI; see magnifying glass). The gray line and area show the expected (mean ± s.d.) distribution in fetal brain. In fetal brain these presumably are the rare allelic conformations that transcribe either Pcdhα4 or Pcdhα11. HS5-1 and the pancellularly active Pcdhαc2 gene promoter preferentially cluster at these conformations. The number of identified microtopologies is indicated in the key. **d.**, Summary of the clustering behavior (z-scores) of the neuron-specifically active Pcdhα elements (Pcdhαc1, Pcdhαc2, HS5-1 and HS7) in fetal brain (left, locus active) and fetal liver (right, locus inactive) when any one of these elements is in contact with either Pcdhα4 or Pcdhα11. Notice the preference for co-occurrence between Pcdhαc2, HS5-1 and HS7 specifically in the active Pcdhα locus (brain).

plex that is necessary to form loops between CTCF-bound domain boundaries [38, 39]. The 'loop extrusion' model [40, 41] predicts that cohesin forms loops by a process in which the chromatin fiber is pulled through its lumen. The loop is then progressively enlarged until two compatible roadblocks (convergently oriented CTCF-bound sites) are reached, where the loop is stably anchored. Without WAPL, cohesin remains bound to chromatin for longer periods of time, which enables CTCF sites to engage with new CTCF partners over much larger distances, as measured by Hi-C across the population of WAPL-deficient (ΔWAPL) HAP1 (human chronic myeloid leukemia) cells cells[22]. One possibility is that these additional ultra-long-range interactions are the result of cohesin progressing beyond original CTCF roadblocks to mediate direct pairing between more distal CTCF sites. An alternative explanation would be that distant sites are reeled in through the aggregation of CTCF loop anchors (loop 'collision'), which ultimately brings together distal CTCF sites. Population-based pairwise contact studies cannot distinguish between these two scenarios.

MC-4C, which allows quantification of allelic co-occurrence frequencies, does enable disentanglement of these two scenarios. We selected a region that clearly showed new long-range contacts in ΔWAPL cells based on Hi-C data (Figure 6.4.a) and applied MC-4C to two CTCF sites that anchor these loops. A comparison between their panallelic contact profiles in wild type (WT) and ΔWAPL cells shows that MC-4C recapitulates the published Hi-C results; it also identifies these long-range contacts specifically in the ΔWAPL cell population (Figure 6.4.b). If they occur as a result of the skipping of CTCF roadblocks, we would expect a severe depletion of intervening CTCF sites from the allelic microtopologies having these distal CTCF sites together. We find the opposite: intervening CTCF sites show a strong preference to aggregate with these structures, something we observe irrespective of the combination of new long-range contacts we interrogate at this locus (Figure 6.4.c-d and Supplementary Fig. 12). To exclude the possibility that the effects are locus-specific, we applied MC-4C to another locus showing profound new contacts between distal CTCF sites in ΔWAPL cells. Here as well we find no evidence for mutual exclusivity between CTCF sites that at the cell-population level all seem to interact with each other. Instead, they are again preferentially found clustered at single alleles (Supplementary Figs. 12 and 13). Therefore, rather than—or at least in addition to—the skipping of CTCF roadblocks, our data strongly suggest that WAPL depletion results in loop collision, with distal CTCF sites coming into contact because of progressive aggregation of loop domain anchors. With Hi-C it was also noted that, in the absence of WAPL, contacts between 'illegally' (non-convergently) oriented CTCF sites are more frequently observed [22]. This now seems partially explained as an inevitable result of cluster formation: when three or more CTCF sites form topological aggregates, at least one is in the 'wrong' orientation.

WAPL serves to destabilize, but not to prevent, loop formation, and therefore loop anchor clusters may also exist, albeit less frequently, in WT cells. To investigate this, we selected alleles from WT cells that had the same long-range CTCF contacts interrogated earlier in ΔWAPL cells. Notably, these interactions were too rare in WT cells to stand out in population-based Hi-C and panallelic MC-4C contact profiles (Figure 6.4.a and b). Strikingly, however, in WT cells these rare allelic conformations also showed a strong enrichment of intervening CTCF-based loop anchors. Quantification of alleles showing simultaneous clustering of three or more distinct CTCF anchors showed an increase from 5.6% to 8.6% (for the downstream viewpoint) and from 6.8% to 10.9% (for the upstream viewpoint) in ΔWAPL as compared to WT cells. We therefore conclude that loop collision and anchor aggregation also occur in WT cells, but less frequently, as a result of the counteracting effect of WAPL (Figure 6.4.e,f and Supplementary Fig. 13).

We then searched for an orthogonal methodology that could provide independent evidence for global domain boundary aggregation upon WAPL depletion. For this, we studied the distribution of cohesin in both WT and ΔWAPL cells by means of super-resolution immunofluorescence microscopy. Visual inspection of nuclear images shows a striking reduction of the distance between cohesin molecules in ΔWAPL cells (Figure 6.5.a). A systematic analysis of their distance distribution patterns confirmed the increased proximity between individual cohesin complexes in these cells (Figure 6.5.b). Collectively our data strongly suggest that in the absence of WAPL, cohesin-associated domain boundaries massively collide to form rosette-like chromatin structures in inter-

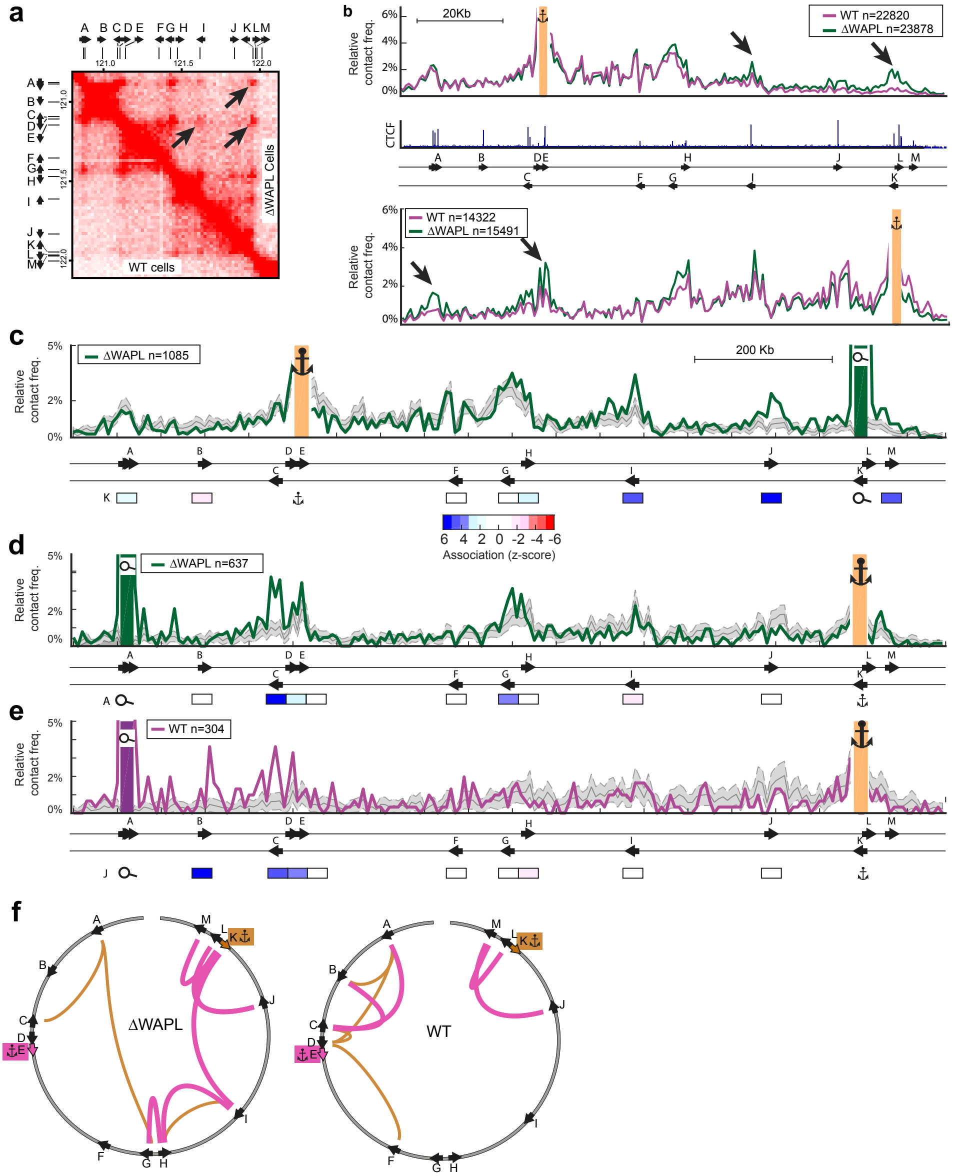


Figure 6.4: Depletion of WAPL stimulates collision of CTCF-anchored domain loops. **a.** Hi-C contact matrix of a genomic region in wild-type (upper right) and ΔWAPL (lower left) HAP1 cells. Position and orientation of CTCF-binding sites are indicated. Arrows point at new long-range contacts that appear upon WAPL knockout. **b.** Overall MC-4C contact profiles of forward-oriented CTCF site E (top) and reverse-oriented CTCF site K (bottom). The number of unique reads for each experiment is indicated in each plot. ΔWAPL cell CTCF chromatin immunoprecipitation (ChIP)-sequencing profile (from Haarhuis et al.20) and CTCF site orientation are shown below. **c.** Microtopologies from ΔWAPL cells having CTCF site E (forward) in contact with CTCF site K (reverse). Gray line and zone indicate negative distribution (mean ± s.d.). z-scores are plotted below, showing preferred clustering of CTCF sites I and J. **d.** Selected microtopologies from ΔWAPL cells having CTCF site K (reverse) in contact with CTCF site A (forward). z-scores are plotted below, showing preferred clustering of CTCF sites C and G. **e.** Selected microtopologies in WT HAP1 cells having CTCF site K (reverse) in contact with CTCF site A (forward). z-scores are plotted below, indicating that the rare allelic conformation wherein K interacts with A co-occurs with interactions with B, C and D, but not with any of the CTCF sites between K and D. **f.** Preferential contacts between CTCF sites. Links are colored with respect to viewpoint and their thickness depicts strength of preferential contacts between CTCF sites.

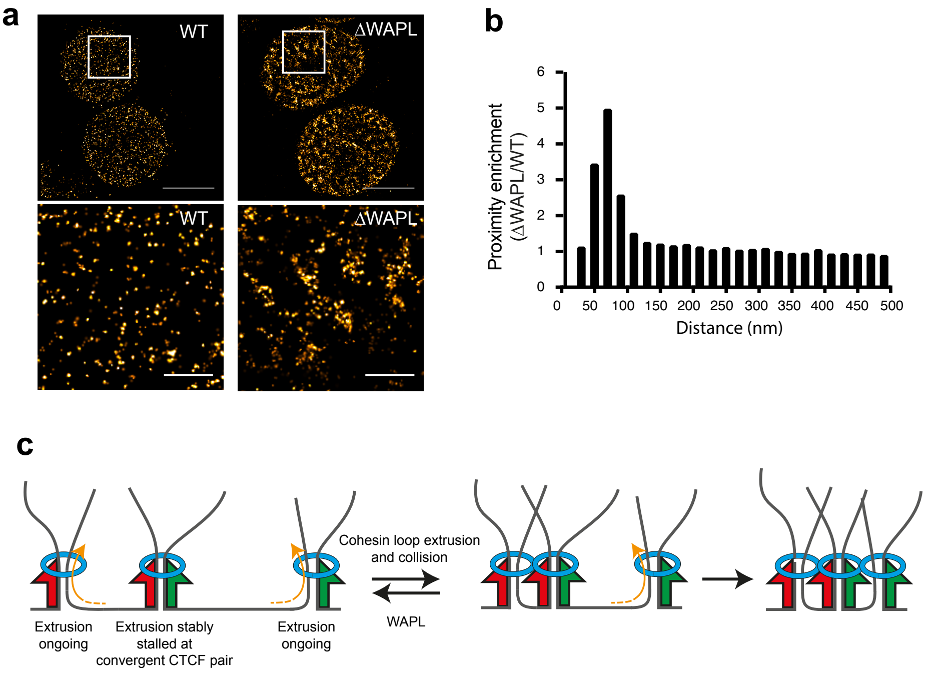


Figure 6.5: Super-resolution microscopy shows cohesin clustering in WAPL-depleted cells. **a.** Representative example super-resolution images of wild-type and ΔWAPL cells. Scale bars 5 μm (top) and 1 μm (bottom, showing magnifications of boxed regions). The experiment was performed twice, with similar results. **b.** Ratio of the proximity between cohesin particles in wild-type and ΔWAPL cells. For each particle, the distance is measured to all other particles per cell. The graph depicts the proximity enrichment up to a distance of 500 nm. The data shown are the distance measured from 5 cells of each genotype. For dot plots of the distances in individual genotypes, see Supplementary Fig. 13. **c.** Schematic of the proposed traffic jam model explaining the increased incidence of CTCF cluster formation in ΔWAPL cells.

phase nuclei. In light of the loop extrusion model, our findings could be explained by assuming a 'cohesin traffic jam'. Any cohesin ring that is extruding a DNA loop (or sliding over the DNA strands) will eventually be released from DNA by WAPL. If not, it will encounter and presumably be stopped by another cohesin ring that was already immobilized at a CTCF roadblock. Subsequent cohesin rings could then start reeling in other CTCF sites from both directions or as nested loops (loops within larger loops), eventually leading to the spatial aggregation of CTCF-bound loop anchors. Collisions from inside and outside an existing loop would then result in a cohesin traffic jam (Figure 6.5.c). Although just a theory, loop collisions resulting in a cohesin traffic jam fit well not only with the high frequency of illegal loops seen by Hi-C in ΔWAPL cells but also with the 'vermicelli' cohesin staining patterns observed in ΔWAPL cells [22, 42].

6.4. Discussion

We present MC-4C, which allows high-resolution analysis of spatial DNA sequence co-occurrence frequencies at the single-allele level. MC-4C contact counts represent relative, not absolute, contact frequencies, as one cannot assume that not being captured (i.e., not being cross-linked, digested, ligated and mapped to the genome) equals not being together. We present a method that, for chosen genomic regions, allows one to statistically distinguish cooperative from random and competitive interactions. In this report we show results directed exclusively toward three-way interactions. Analysis of four-way interactions and beyond poses exponentially increasing demands for the number of analyzed alleles, which is beyond the aims of this study. However, long reads containing more than three fragments are routinely identified, and their content is employed extensively to populate the three-way interaction profiles and to identify PCR duplicates. The data show that, by this method, sequences that directly neighbor each other on the linear chromosome are being scored as obligatorily together in 3D space (cooperative interactions). This is not only as expected (physically connected sequences simply cannot spatially escape each other), but can also be biologically meaningful: it is not without reason that only when transcription factor binding motifs cluster on the linear chromosome can they form functional regulatory motifs. It does emphasize, though, that for correct interpretation of MC-4C results resolution must be high enough to discern spatial clustering as the mere consequence of linear physical proximity from that driven by biological processes. Here we accomplish this by analyzing often more than 10,000 independent allelic conformations per experiment and by comparing allelic co-occurrence frequencies of the same locus in its active versus inactive configuration. The study of higher order chromatin topologies at such high resolution uncovers new biology: individual elements of an SE can aggregate to form an enhancer hub that can accommodate multiple genes simultaneously. Observations such as these highlight the architectural context of SE elements, which combined with their combinatorial deletions will help in understanding their functional hierarchy21–23. Similarly, we also find that cohesin drives aggregation of CTCFbound domain boundaries, which is counteracted by WAPL. Our studies on domain boundary clustering, as well as our work on Pcdhα, further demonstrate that MC-4C can identify and analyze relevant structures missed by population-based contact methods such as Hi-C or 4C because they are present in only a small percentage of cells. High-resolution multi-way contact analysis methods such as MC-4C promise to uncover how the multitude of regulatory sequences and genes truly coordinate their action in the 3D spatial context of the genome. For the visualization of co-occurrence frequencies of any site of interest with a given MC-4C viewpoint and the calculation of the significance of such three-way interactions, we refer the reader to the interactive viewer that we made available, together with the data shown in this manuscript (see URLs).

6.5. URLs

MC-4C processing pipeline, https://github.com/UMCUGenetics/pymc4c/;
MC-4C visualization tool, http://www. multicontactchromatin.nl/;
ImageJ macro and corresponding raw images, https://github.com/aallahyar/MC-4C_SRMl;

Temporal median filter for structured background subtraction,
https://github. com/rharkes/Temporal-Median-Background-Subtraction;
ImageJ, http://imagej.nih.gov/ij/;
Thunderstorm plugin for ImageJ, https://github.com/zitmen/thunderstorm;
raw sequencing MC-4C data, https://www.ebi.ac.uk/ena/data/view/PRJEB23327;
MC-4C processed data, https://doi.org/10.17632/wbk8hk87r2.1.

6.6. Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0161-5.

6.7. Acknowledgements

We thank the Utrecht Sequencing Facility for providing sequencing data and service, E. Schijlen for help with initial Pacific Biosciences sequencing, and D. Leyton-Puig for help with imaging. We thank N. Geijsen and P. Shang (Hubrecht Institute) for providing Cas9 protein. This work was supported by an NWO VIDI grant (639.072.715) to J.d.R. and an NWO/CW TOP grant (714.012.002) and NWO VICI grant (724.012.003) to W.d.L. and by the NIH Common Fund Program, grant U01CA200147, as a Transformative Collaborative Project Award (TCPA, TCPA-2017-DE-LAAT).

6.8. Author contributions

A.A. designed and performed the computational analysis, prepared corresponding plots and wrote the methods and Supplementary Information sections. C.V. and B.A.M.B. designed and performed experiments. C.V. wrote the manuscript and designed figures. P.H.L.K., M.J.A.M.V., M.v.K., M.P. and H.T. performed 'C' methods experiments. R.S. implemented the pipeline in Python. J.H.I.H. generated ΔWAPL cell lines and prepared microscopic slides for super-resolution imaging. K.J. guided acquisition and analyzed super-resolution microscopy data. I.J.R. performed and W.P.K. designed and supervised MinION sequencing experiments. B.D.R. supervised the generation of ΔWAPL cell lines and preparation of microscopic slides for super-resolution imaging. G.G. and E.d.W. helped with computational analysis. E.d.W. performed data analysis on the ΔWAPL Hi-C data. J.d.R. designed and supervised the computational analyses and pipelines and co-wrote the manuscript. W.d.L. conceived and supervised the study and wrote the manuscript.

6.9. Competing Interests

C.V., B.A.M.B., P.H.L.K., M.J.A.M.V. and G.G. are shareholders of Cergentis. E.d.W. is cofounder and shareholder of Cergentis. W.d.L. is founder and shareholder of Cergentis. J.d.R. is cofounder and shareholder of Cyclomics.

6.10. Additional information

Supplementary information for this section can be found in Chapter 5 as well as at: https://doi.org/10.1038/s41588-018-0161-5.

Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.R. or W.L. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

[1] A. Allahyar, C. Vermeulen, B. A. M. Bouwman, P. H. L. Krijger, M. J. A. M. Verstegen, G. Geeven, M. van Kranenburg, M. Pieterse, R. Straver, J. H. I. Haarhuis, K. Jalink, H. Teunissen, I. J. Renkens, W. P. Kloosterman, B. D. Rowland, E. de Wit, J. de Ridder, and W. de Laat, *Enhancer hubs and loop collisions identified from single-allele topologies,* Nature Genetics (2018), 10.1038/s41588-018-0161-5.

[2] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, *Capturing chromosome conformation,* Science **295**, 1306 (2002).

[3] A. Denker and W. de Laat, *The second decade of 3C technologies: detailed insights into nuclear organization,* Genes Dev. **30**, 1357 (2016).

[4] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, *Topological domains in mammalian genomes identified by analysis of chromatin interactions,* Nature **485**, 376 (2012).

[5] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, and E. Heard, *Spatial partitioning of the regulatory landscape of the x-inactivation centre,* Nature **485**, 381 (2012).

[6] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli, *Three-dimensional folding and functional organization principles of the drosophila genome,* Cell **148**, 458 (2012).

[7] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, *Comprehensive mapping of long-range interactions reveals folding principles of the human genome,* Science **326**, 289 (2009).

[8] J. Dekker and L. Mirny, *The 3D genome as moderator of chromosomal communication,* Cell **164**, 1110 (2016).

[9] J. R. Dixon, D. U. Gorkin, and B. Ren, *Chromatin domains: The unit of chromosome organization,* Mol. Cell **62**, 668 (2016).

[10] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C.-L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W.-K. Sung, M. Snyder, and Y. Ruan, *Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation,* Cell **148**, 84 (2012).

[11] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping,* Cell **159**, 1665 (2014).

[12] O. Hanscombe, D. Whyatt, P. Fraser, N. Yannoutsos, D. Greaves, N. Dillon, and F. Grosveld, *Importance of globin gene order for correct developmental expression,* Genes Dev. **5**, 1387 (1991).

[13] B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat, *Looping and interaction between hypersensitive sites in the active beta-globin locus,* Mol. Cell **10**, 1453 (2002).

[14] F. Ay, T. H. Vu, M. J. Zeitz, N. Varoquaux, J. E. Carette, J.-P. Vert, A. R. Hoffman, and W. S. Noble, *Identifying multi-locus chromatin contacts in human cells using tethered multiple 3C,* BMC Genomics **16**, 121 (2015).

[15] E. M. Darrow, M. H. Huntley, O. Dudchenko, E. K. Stamenova, N. C. Durand, Z. Sun, S.-C. Huang, A. L. Sanborn, I. Machol, M. Shamim, A. P. Seberg, E. S. Lander, B. P. Chadwick, and E. L. Aiden, *Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture,* Proc. Natl. Acad. Sci. U. S. A. **113**, E4504 (2016).

[16] A. A. Gavrilov, H. V. Chetverina, E. S. Chermnykh, S. V. Razin, and A. B. Chetverin, *Quantitative analysis of genomic element interactions by molecular colony technique,* Nucleic Acids Res. **42**, e36 (2014).

[17] P. Olivares-Chauvet, Z. Mukamel, A. Lifshitz, O. Schwartzman, N. O. Elkayam, Y. Lubling, G. Deikus, R. P. Sebra, and A. Tanay, *Capturing pairwise and multi-way chromosomal conformations using chromosomal walks,* Nature **540**, 296 (2016).

[18] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. A. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. de Santiago, L.-M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. W. Edwards, M. Nicodemi, and A. Pombo, *Complex multi-enhancer contacts captured by genome architecture mapping,* Nature **543**, 519 (2017).

[19] E. Splinter, E. de Wit, H. J. G. van de Werken, P. Klous, and W. de Laat, *Determining long-range chromatin interactions for selected genomic sites using 4c-seq technology: from fixation to computation,* Methods **58**, 221 (2012).

[20] H. J. G. van de Werken, G. Landan, S. J. B. Holwerda, M. Hoichman, P. Klous, R. Chachik, E. Splinter, C. Valdes-Quezada, Y. Oz, B. A. M. Bouwman, M. J. A. M. Verstegen, E. de Wit, A. Tanay, and W. de Laat, *Robust 4c-seq data analysis to screen for regulatory DNA interactions,* Nat. Methods **9**, 969 (2012).

[21] P. J. P. de Vree, E. de Wit, M. Yilmaz, M. van de Heijning, P. Klous, M. J. A. M. Verstegen, Y. Wan, H. Teunissen, P. H. L. Krijger, G. Geeven, P. P. Eijk, D. Sie, B. Ylstra, L. O. M. Hulsman, M. F. van Dooren, L. J. C. M. van Zutven, A. van den Ouweland, S. Verbeek, K. W. van Dijk, M. Cornelissen, A. T. Das, B. Berkhout, B. Sikkema-Raddatz, E. van den Berg, P. van der Vlies, D. Weening, J. T. den Dunnen, M. Matusiak, M. Lamkanfi, M. J. L. Ligtenberg, P. ter Brugge, J. Jonkers, J. A. Foekens, J. W. Martens, R. van der Luijt, H. K. P. van Amstel, M. van Min, E. Splinter, and W. de Laat, *Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping,* Nature biotechnology **32**, 1019 (2014).

[22] J. H. I. Haarhuis, R. H. van der Weide, V. A. Blomen, J. O. Yáñez-Cuna, M. Amendola, M. S. van Ruiten, P. H. L. Krijger, H. Teunissen, R. H. Medema, B. van Steensel, T. R. Brummelkamp, E. de Wit, and B. D. Rowland, *The cohesin release factor WAPL restricts chromatin loop extension,* Cell **169**, 693 (2017).

[23] S. Pott and J. D. Lieb, *What are super-enhancers?* Nature genetics **47**, 8 (2015), perspective.

[24] N. Dukler, B. Gulko, Y.-F. Huang, and A. Siepel, *Is a super-enhancer greater than the sum of its parts?* Nature genetics **49**, 2 (2017).

[25] J. Huang, K. Li, W. Cai, X. Liu, Y. Zhang, S. H. Orkin, J. Xu, and G.-C. Yuan, *Dissecting super-enhancer hierarchy based on chromatin interactions,* Nature communications **9**, 943 (2018).

[26] M. A. Bender, M. Bulger, J. Close, and M. Groudine, *Beta-globin gene switching and DNase I sensitivity of the endogenous beta-globin locus in mice do not require the locus control region,* Mol. Cell **5**, 387 (2000).

[27] B. Cadiz-Rivera, G. Fromm, C. de Vries, J. Fields, K. E. McGrath, S. Fiering, and M. Bulger, *The chromatin "landscape" of a murine adult β-Globin gene is unaffected by deletion of either the gene promoter or a downstream enhancer,* PLoS One **9**, e92947 (2014).

[28] X. Hu, S. Eszterhas, N. Pallazzi, E. E. Bouhassira, J. Fields, O. Tanabe, S. A. Gerber, M. Bulger, J. D. Engel, M. Groudine, and S. Fiering, *Transcriptional interference among the murine beta-like globin genes,* Blood **109**, 2210 (2007).

[29] J. O. J. Davies, J. M. Telenius, S. J. McGowan, N. A. Roberts, S. Taylor, D. R. Higgs, and J. R. Hughes, *Multiplexed analysis of chromosome conformation at vastly improved sensitivity,* Nat. Methods **13**, 74 (2016).

[30] H. J. G. van de Werken, P. J. P. de Vree, E. Splinter, S. J. B. Holwerda, P. Klous, E. de Wit, and W. de Laat, *4C technology: protocols and data analysis,* Methods Enzymol. **513**, 89 (2012).

[31] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb, *FAIRE (Formaldehyde-Assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin,* Genome Res. **17**, 877 (2007).

[32] S. Esumi, N. Kakazu, Y. Taguchi, T. Hirayama, A. Sasaki, T. Hirabayashi, T. Koide, T. Kitsukawa, S. Hamada, and T. Yagi, *Monoallelic yet combinatorial expression of variable exons of the protocadherin-α gene cluster in single neurons,* Nat. Genet. **37**, 171 (2005).

[33] T. Hirayama and T. Yagi, *The role and expression of the protocadherin-alpha clusters in the CNS,* Curr. Opin. Neurobiol. **16**, 336 (2006).

[34] P. Kehayova, K. Monahan, W. Chen, and T. Maniatis, *Regulatory elements required for the activation and repression of the protocadherin-alpha gene cluster,* Proc. Natl. Acad. Sci. U. S. A. **108**, 17195 (2011).

[35] S. Yokota, T. Hirayama, K. Hirano, R. Kaneko, S. Toyoda, Y. Kawamura, M. Hirabayashi, T. Hirabayashi, and T. Yagi, *Identification of the cluster control region for the protocadherin-beta genes located beyond the protocadherin-gamma cluster,* J. Biol. Chem. **286**, 31885 (2011).

[36] Y. Guo, K. Monahan, H. Wu, J. Gertz, K. E. Varley, W. Li, R. M. Myers, T. Maniatis, and Q. Wu, *CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice,* Proc. Natl. Acad. Sci. U. S. A. **109**, 21081 (2012).

[37] S. Toyoda, M. Kawaguchi, T. Kobayashi, E. Tarusawa, T. Toyama, M. Okano, M. Oda, H. Nakauchi, Y. Yoshimura, M. Sanbo, M. Hirabayashi, T. Hirayama, T. Hirabayashi, and T. Yagi, *Developmental epigenetic modification regulates stochastic expression of clustered protocadherin genes, generating single neuron diversity,* Neuron **82**, 94 (2014).

[38] S. S. P. Rao, S.-C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, X. Huang, M. S. Shamim, J. Shin, D. Turner, Z. Ye, A. D. Omer, J. T. Robinson, T. Schlick, B. E. Bernstein, R. Casellas, E. S. Lander, and E. L. Aiden, *Cohesin loss eliminates all loop domains,* Cell **171**, 305 (2017).

[39] S. Sofueva, E. Yaffe, W.-C. Chan, D. Georgopoulou, M. Vietri Rudan, H. Mira-Bontenbal, S. M. Pollard, G. P. Schroth, A. Tanay, and S. Hadjur, *Cohesin-mediated interactions organize chromosomal domain architecture,* EMBO J. **32**, 3119 (2013).

[40] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, *Formation of chromosomal domains by loop extrusion,* Cell Rep. **15**, 2038 (2016).

[41] A. L. Sanborn, S. S. P. Rao, S.-C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, I. D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, K. P. Geeting, A. Gnirke, A. Melnikov, D. McKenna, E. K. Stamenova, E. S. Lander, and E. L. Aiden, *Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes,* Proc. Natl. Acad. Sci. U. S. A. **112**, E6456 (2015).

[42] A. Tedeschi, G. Wutz, S. Huet, M. Jaritz, A. Wuensche, E. Schirghuber, I. F. Davidson, W. Tang, D. A. Cisneros, V. Bhaskara, T. Nishiyama, A. Vaziri, A. Wutz, J. Ellenberg, and J.-M. Peters, *Wapl is an essential regulator of chromatin structure and chromosome segregation,* Nature **501**, 564 (2013).

7

MC-4C: COMPUTATIONAL ASPECT

Amin Allahyar
Roy Straver
Jeroen de Ridder

This chapter is published in Nature Genetics (2018); https://doi.org/10.1038/s41588-018-0161-5 [1].

7.1. Brief overview of MC-4C library preparation and sequencing

For β-globin experiments, mouse embryos were harvested from surplus pregnant animals at 14.5 days post conception. Their livers and brains were manually dissected. For WPL experiments, wild-type Hap1 cells and WAPL knockout Hap1 cells were cultured and harvested as described in Haarhuis *et al.*. MC-4C template was prepared following the regular 4C protocol (described in van de Werken *et al.*, Splinter *et al.*), with several adjustments. Notably, 6-base pair restriction enzymes (e.g. HindIII) were used to shorten the MC-4C template to approximately 2kb and then circularized.

We amplified (i.e. duplicated) concatemers that contain view point fragment by performing inverse PCR on two primers that were designed on each end of view point fragment. One major difficulty in performing the inverse PCR on circularized templates is that short circles (often formed from view point fragment ligating to itself) tend to amplify extensively compared to other concatemers. Additionally, DNA polymerase continues to copy the strand by "rolling" around the circle over and over again (also known as "rolling circle amplification") which generates a long but uninformative read. By exploiting CRISPR-Cas9 technology, we cut the view point fragment to stop the DNA polymerase from rolling circle amplification of small circles. Additionally, the fragments neighboring the view point are cut to reduce the prevalence of circles containing the view point and its neighbor fragments.

To selectively sequence long reads, Pippin HT size selection within a 1.5-8kb range was performed on PCR products. Subsequently, libraries were prepared using the Oxford nanopore sequencing kits and sequenced with appropriate flow cells. After sequencing, "squiggle" signals were base called using latest versions of either Metrichor or Albacore depending on their market availability. Typically, such a procedure produced approximately 1-2 million reads with average read size distribution around 1.5kb (as expected). Figure 7.2 demonstrates such a distribution for two exemplary runs of MC-4C in liver and brain cells using β-major as the view point.

7.2. Overview of MC-4C data processing pipeline

In order reveal the multi-way DNA interactions captured by MC-4C, sequenced reads need to undergo a few pre-processing steps. These steps ensure read integrity and more crucially filters the reads for PCR duplicates, enabling quantitative analysis of the conformation of micro-topologies (see 7.16 for corresponding schematic).

7.3. Read validity check

To validate fidelity of the sequenced reads, we identified primers as well as their orientations in each read. To this end, Bowtie2 v2.2.6 [9] was employed in local alignment mode (settings: -D 20 -R 3 -N 0 -L 15 -i S,1,0.50 –rdg 2,1 –rfg 2,1 –mp 3,2 –ma 2 -a). We allowed 20% mismatches to take into account errors in Nanopore sequencing. To improve efficiency of this step, we grouped reads into batches of 10,000 reads and mapped primer sequences to reads within batches in parallel. This step is likely to take about 30 seconds on average for each batch.

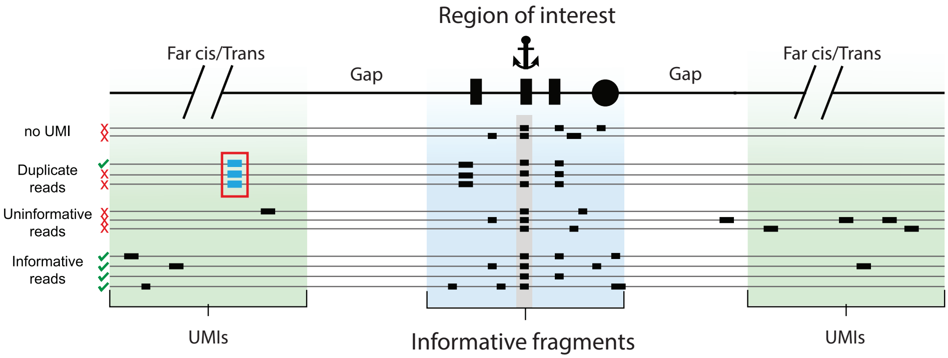


Figure 7.1: Schematic overview of PCR filtering step. Scheme shows the removal of (1) reads with less than two captured fragments in the region of interest (ROI), (2) reads with less than one far-cis/trans fragment and (3) PCR duplicate reads, as guided by the capture of identical far-cis/trans fragments (genomically contributed unique molecular identifiers (UMIs)).

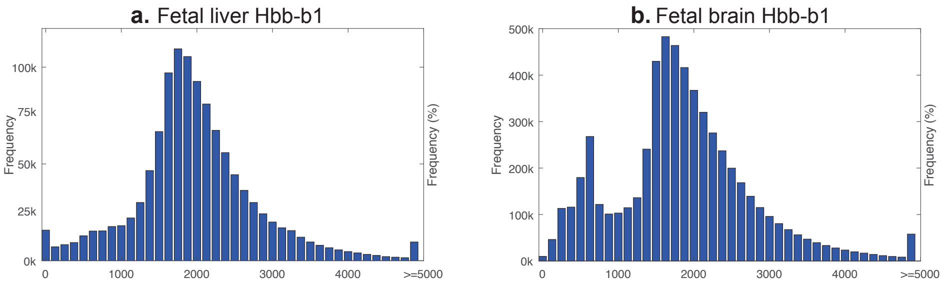


Figure 7.2: MC-4C read size distribution. Read size distribution plots of ten representative MC-4C experiments, using viewpoints inside the β-globin and Pcdhα locus in primary mouse fetal liver and fetal brain cells and viewpoints at CTCF sites in wild type and WAPL knockout human HAP1 cells.

Analyzing primer arrangements in the sequenced reads showed that some reads (1% on average) are formed by ligation of two or more individual molecules. We therefore implemented a correction procedure in which read-ligation events (i.e. two divergent primers within a read) are identified and reads containing such events are cleaved into two sub-reads. The produced sub-reads are treated as independent reads in downstream analysis. We discarded any reads that contained more than four primers or more than one read-ligation event. These requirements ensured that only those configurations that clearly arise as a result of a read ligation event go through the correction procedure. The produced sub-reads were discarded if their primer configuration did not validate (e.g. identification of non-convergent primers on either ends of a read). In this stage, we also discarded any reads (or sub-reads) that were smaller than 500bp as they are unlikely to be of sufficient complexity (i.e. in terms of the number of fragments) to be informative for multi-contact analysis (see table 7.1 corresponding statistics per experiment).

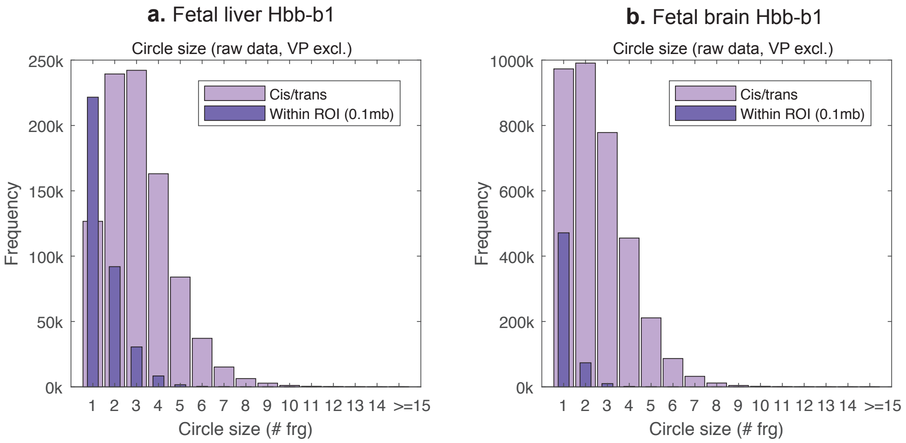


Figure 7.3: Number of captured fragments per read. Plots show number of captured fragments (i.e. number of identified contacts) per read, for representative MC-4C viewpoints inside the β-globin and Pcdhα locus in primary mouse fetal liver and fetal brain cells and viewpoints at CTCF sites in wild type and WAPL knockout human HAP1 cells. Note that restriction fragments which map immediately next to each other on the reference genome are together counted as a single fragment (i.e. single contact).

7.4. Splitting reads into fragments based on restriction enzyme sequences

MC-4C reads are expected to be concatemers of multiple distinct fragments, and should therefore be mapped using an aligner with split-read mapping capabilities (i.e. splitting a single query read and mapping to multiple coordinates). However, as many reads will consist of more than two fragments and splits are expected to occur at known restriction sites in the genome, we pre-split the reads into prospective fragments using restriction enzyme recognition sequence. This procedure showed improved efficacy in mapping fragments compared to relying only on the split-read mapping capabilities of the aligner (see Figure 7.15). Due to sequencing errors, extra restriction sites (i.e. observing GATC while it should have been GAAC) might be erroneously recognized. To consider such cases, the split fragments that map directly adjacent in the reference genome are further fused together in later stages of the pipeline (see section 7.6). For the same reason, restriction sites may be missed. In this case, we relied on the split-read capability of the aligner to correctly identify sub-fragments.

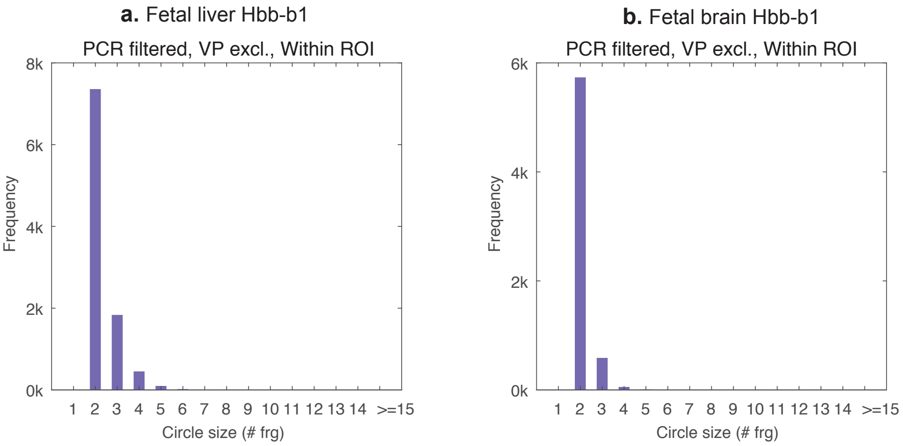


Figure 7.4: Number of captured fragments per informative read. Plots show number of captured non-viewpoint fragments (i.e. number of identified contacts) for all reads that have passed the filtering and selection pipeline.

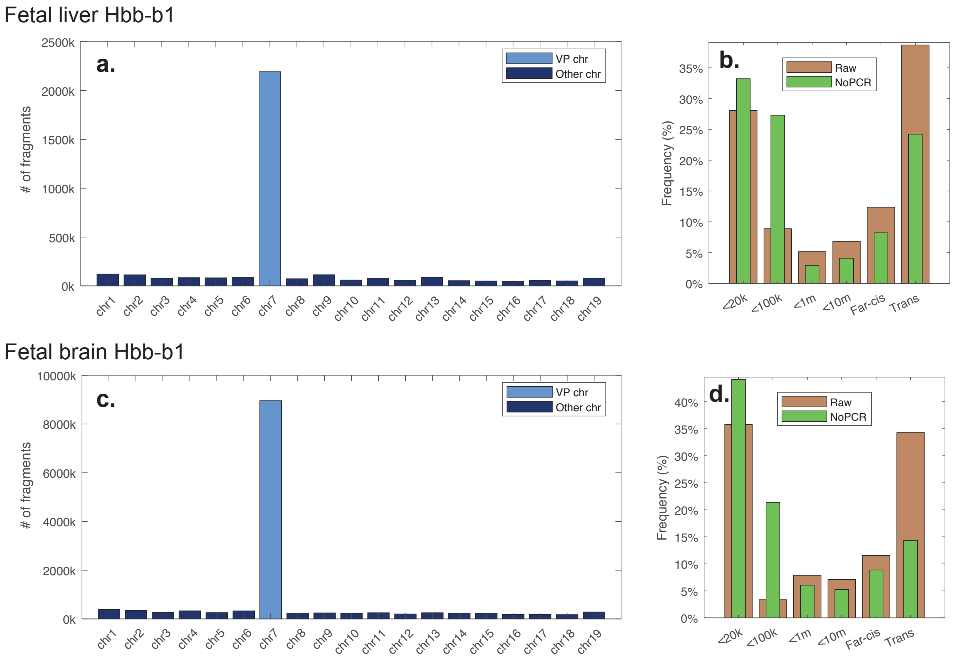


Figure 7.5: Genomic distribution of MC-4C captured fragments. **a.** Chromosomal distribution of captured fragments for ten representative MC-4C experiments. **b.** Distribution of captured fragments across chromosomal intervals at increased distance from the viewpoint, for ten representative MC-4C experiments.

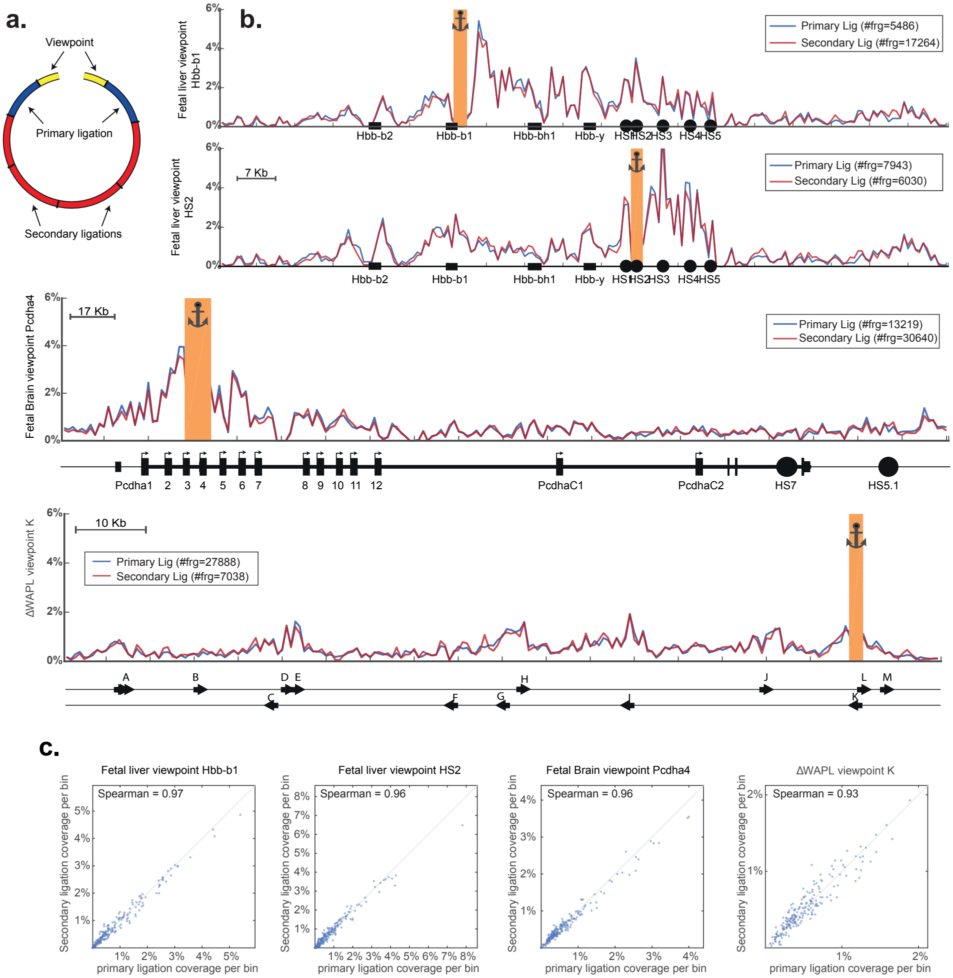
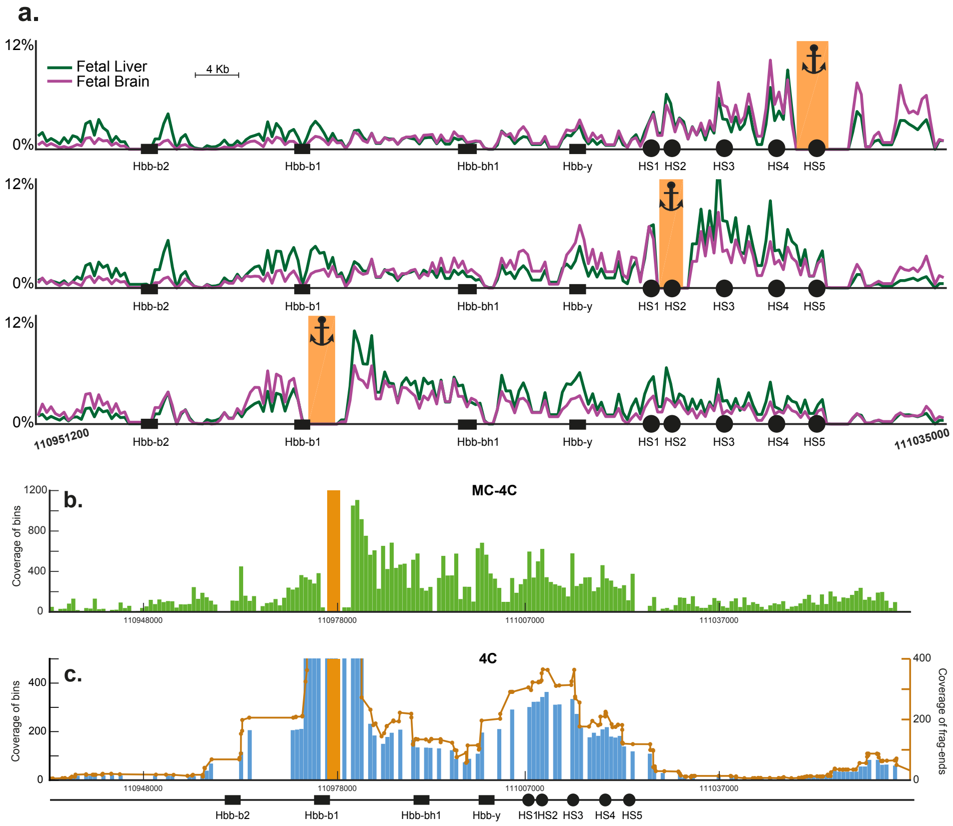


Figure 7.6: Similarity between primary and secondary ligation products. **a.** Cartoon explaining the difference between primary ligation products (as analyzed by Hi-C and 4C-seq) and the secondary ligation products that are additionally captured, sequenced and analyzed in mC-4C. **b.** Overlay of pan-allelic contact profiles of primary and secondary ligation products, for four representative MC-4C experiments. **c.** Comparison of the distribution of primary and secondary captured fragments across chromosomal intervals at increased distance from the viewpoint, for four individual representative, MC-4C experiments.



7

Figure 7.7: Overall (pan-allelic) contact profiles generated by MC-4C and a comparison with 4C-seq. a. In E14.5 fetal liver, the hbb-b1 and hbb-b2 genes are the predominantly active globin genes, while hbb-y expression is being silenced. In fetal brain cells, all globin genes are silenced (but residual expression may come from contaminating circulating blood cells). **b.** MC-4C (top) and 4C-seq (bottom) were applied to the Hbb-b1 gene in E14.5 mouse fetal liver. MC-4C data were collected and plotted (bin sizes 0.7kb) as described in this manuscript. 4C-seq data were collected from GEO (GSE40420) and processed as described in the original paper [3, 5]. 4C-seq PCR amplification biases necessitate data normalization, which is done using a running mean operator with a window size of 21 fragment-ends. As a consequence, the resolution of 4C-seq contact profiles is lower than that of MC-4C. Since 4C-seq involves the mapping and analysis of small fragment ends (instead of complete restriction fragments, as is done in MC-4C), not all sequenced fragment ends can be uniquely mapped to the repetitive sequences of the β-globin locus, hence explaining the gaps in the 4C-seq contact profile.

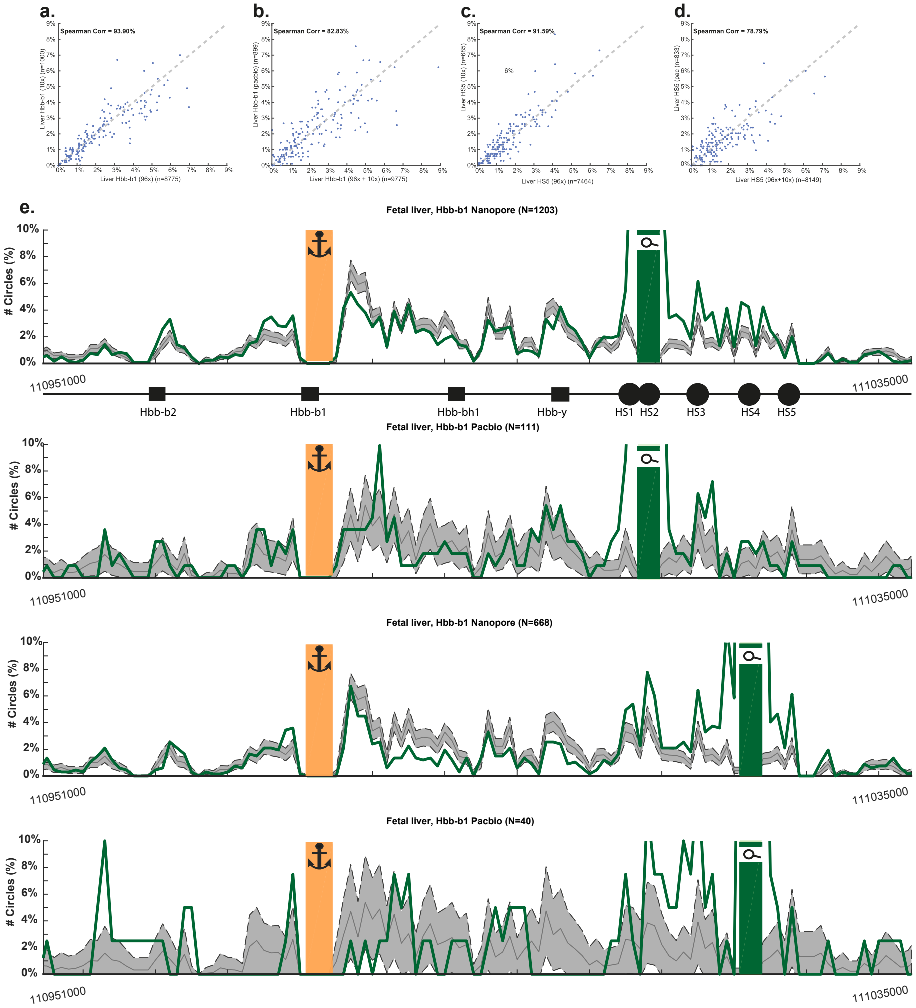
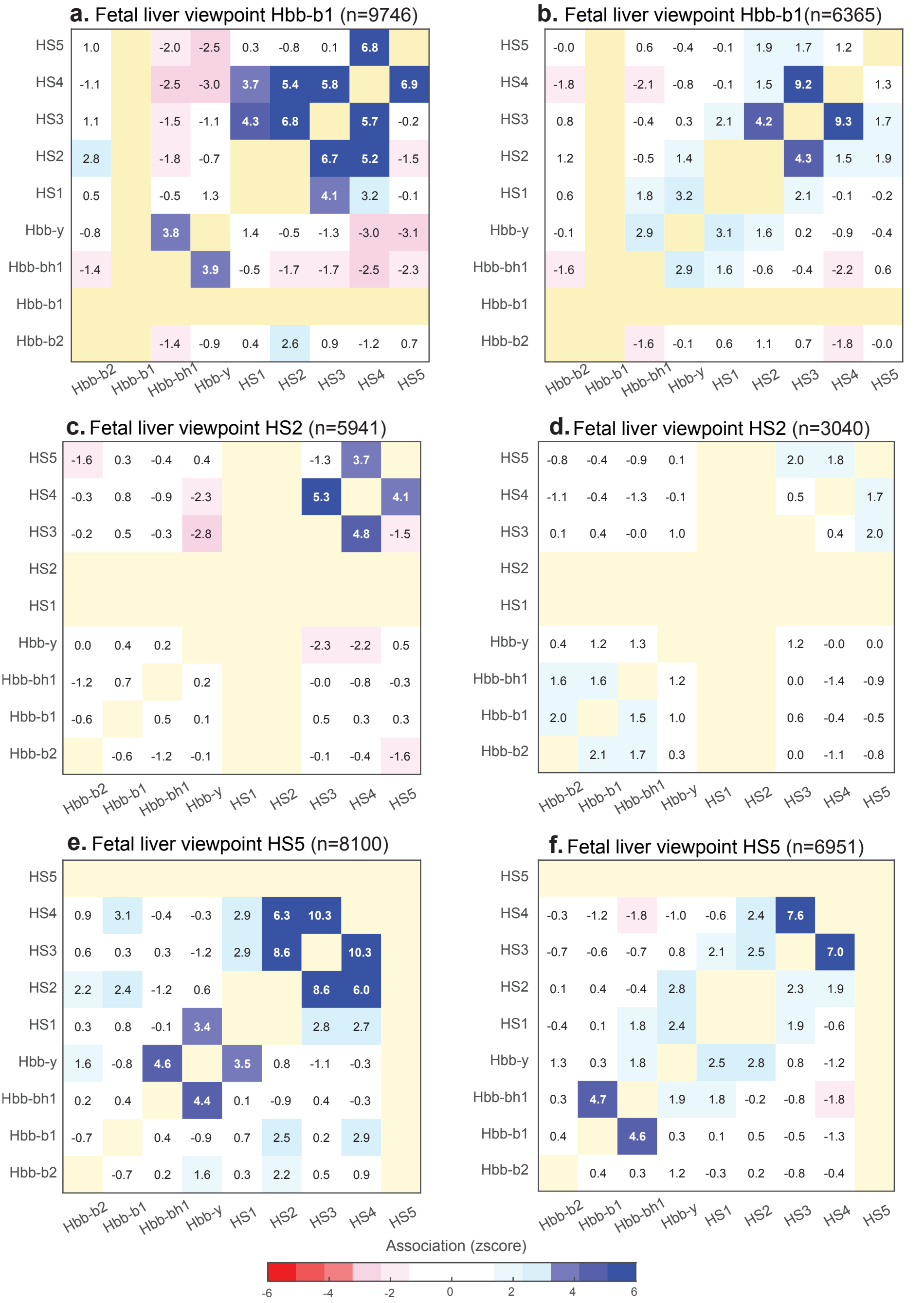


Figure 7.8: Reproducibility between independent experiments and sequencing platforms. **a.** Plots show the correlation between independent replicates using cells from different embryos. **a** and **b** show results obtained in fetal liver cells and the Hbb-b1 viewpoint. **a.** Compares an experiment where 10 PCR reactions were pooled and sequenced, with an experiment where 96 PCR reactions were pooled and sequenced. **b.** Compares an experiment on a Pacific biosciences Single Molecule Real-time sequencer with the pooled 96x and 10x nanopore data. **c** and **d** show the same comparison for the HS5 viewpoint in liver cells. Spearman correlations are shown in the top-left of each plot. **e.** Hbb-b1-HS2 and Hbb-b1-HS4 VP-SOI plots generated by Nanopore (top) and PacBio (bottom) sequencing. Contact profiles look similar overall, but Nanopore sequencing gave us at least 10-fold more sequences, making the profiles more reliable (e.g. represented by smaller standard deviations in the negative set) and the z-score calculations more robust. Grey line and area indicate the negative set (mean ± SD)



a. Fetal liver viewpoint Hbb-b1 (n=9746)

	Hbb-b2	Hbb-b1	Hbb-bh1	Hbb-y	HS1	HS2	HS3	HS4	HS5
HS5	1.0		-2.0	-2.5	0.3	-0.8	0.1	6.8	
HS4	-1.1		-2.5	-3.0	3.7	5.4	5.8		6.9
HS3	1.1		-1.5	-1.1	4.3	6.8		5.7	-0.2
HS2	2.8		-1.8	-0.7			6.7	5.2	-1.5
HS1	0.5		-0.5	1.3			4.1	3.2	-0.1
Hbb-y	-0.8		3.8		1.4	-0.5	-1.3	-3.0	-3.1
Hbb-bh1	-1.4			3.9	-0.5	-1.7	-1.7	-2.5	-2.3
Hbb-b1									
Hbb-b2			-1.4	-0.9	0.4	2.6	0.9	-1.2	0.7

b. Fetal liver viewpoint Hbb-b1(n=6365)

	Hbb-b2	Hbb-b1	Hbb-bh1	Hbb-y	HS1	HS2	HS3	HS4	HS5
HS5	-0.0		0.6	-0.4	-0.1	1.9	1.7	1.2	
HS4	-1.8		-2.1	-0.8	-0.1	1.5	9.2		1.3
HS3	0.8		-0.4	0.3	2.1	4.2		9.3	1.7
HS2	1.2		-0.5	1.4			4.3	1.5	1.9
HS1	0.6		1.8	3.2			2.1	-0.1	-0.2
Hbb-y	-0.1		2.9		3.1	1.6	0.2	-0.9	-0.4
Hbb-bh1	-1.6			2.9	1.6	-0.6	-0.4	-2.2	0.6
Hbb-b1									
Hbb-b2			-1.6	-0.1	0.6	1.1	0.7	-1.8	-0.0

c. Fetal liver viewpoint HS2 (n=5941)

	Hbb-b2	Hbb-b1	Hbb-bh1	Hbb-y	HS1	HS2	HS3	HS4	HS5
HS5	-1.6	0.3	-0.4	0.4			-1.3	3.7	
HS4	-0.3	0.8	-0.9	-2.3			5.3		4.1
HS3	-0.2	0.5	-0.3	-2.8				4.8	-1.5
HS2									
HS1									
Hbb-y	0.0	0.4	0.2				-2.3	-2.2	0.5
Hbb-bh1	-1.2	0.7		0.2			-0.0	-0.8	-0.3
Hbb-b1	-0.6		0.5	0.1			0.5	0.3	0.3
Hbb-b2		-0.6	-1.2	-0.1			-0.1	-0.4	-1.6

d. Fetal liver viewpoint HS2 (n=3040)

	Hbb-b2	Hbb-b1	Hbb-bh1	Hbb-y	HS1	HS2	HS3	HS4	HS5
HS5	-0.8	-0.4	-0.9	0.1			2.0	1.8	
HS4	-1.1	-0.4	-1.3	-0.1			0.5		1.7
HS3	0.1	0.4	-0.0	1.0				0.4	2.0
HS2									
HS1									
Hbb-y	0.4	1.2	1.3				1.2	-0.0	0.0
Hbb-bh1	1.6	1.6		1.2			0.0	-1.4	-0.9
Hbb-b1	2.0		1.5	1.0			0.6	-0.4	-0.5
Hbb-b2		2.1	1.7	0.3			0.0	-1.1	-0.8

e. Fetal liver viewpoint HS5 (n=8100)

	Hbb-b2	Hbb-b1	Hbb-bh1	Hbb-y	HS1	HS2	HS3	HS4	HS5
HS5									
HS4	0.9	3.1	-0.4	-0.3	2.9	6.3	10.3		
HS3	0.6	0.3	0.3	-1.2	2.9	8.6		10.3	
HS2	2.2	2.4	-1.2	0.6			8.6	6.0	
HS1	0.3	0.8	-0.1	3.4			2.8	2.7	
Hbb-y	1.6	-0.8	4.6		3.5	0.8	-1.1	-0.3	
Hbb-bh1	0.2	0.4		4.4	0.1	-0.9	0.4	-0.3	
Hbb-b1	-0.7		0.4	-0.9	0.7	2.5	0.2	2.9	
Hbb-b2		-0.7	0.2	1.6	0.3	2.2	0.5	0.9	

f. Fetal liver viewpoint HS5 (n=6951)

	Hbb-b2	Hbb-b1	Hbb-bh1	Hbb-y	HS1	HS2	HS3	HS4	HS5
HS5									
HS4	-0.3	-1.2	-1.8	-1.0	-0.6	2.4	7.6		
HS3	-0.7	-0.6	-0.7	0.8	2.1	2.5		7.0	
HS2	0.1	0.4	-0.4	2.8			2.3	1.9	
HS1	-0.4	0.1	1.8	2.4			1.9	-0.6	
Hbb-y	1.3	0.3	1.8		2.5	2.8	0.8	-1.2	
Hbb-bh1	0.3	4.7		1.9	1.8	-0.2	-0.8	-1.8	
Hbb-b1	0.4		4.6	0.3	0.1	0.5	-0.5	-1.3	
Hbb-b2		0.4	0.3	1.2	-0.3	0.2	-0.8	-0.4	

Figure 7.9: Allelic co-occur frequencies at the active and inactive β-globin locus. Complete overview of co-occur significance scores found between any pair of genes and/or super enhancer constituents in experiments using Hbb-b1, HS2, HS3 and HS5 as viewpoints, in E14.5 fetal liver and E14.5 fetal brain cells. Numbers in each square represent (in order of appearance): Numbers in each square represent (in order of appearance): z-score of the association test used to assess significance of preferential contacts observed between each genomic pair (in presence of the view point), number of reads containing queried SOI (x-axis) in the positive set (see **Methods**), average number of reads containing queried SOI in 1000 draws of the negative set, percentage of reads containing the queried SOI in the positive set, average percentage of reads containing the queried SOI in the negative set in 1000 draws along with the standard deviation.

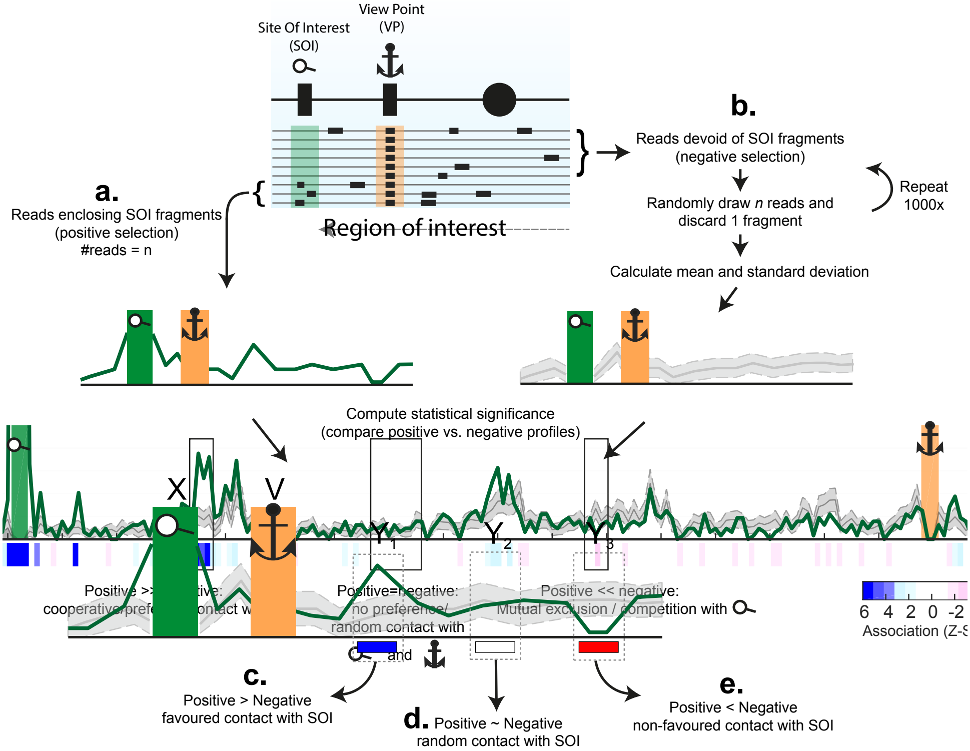
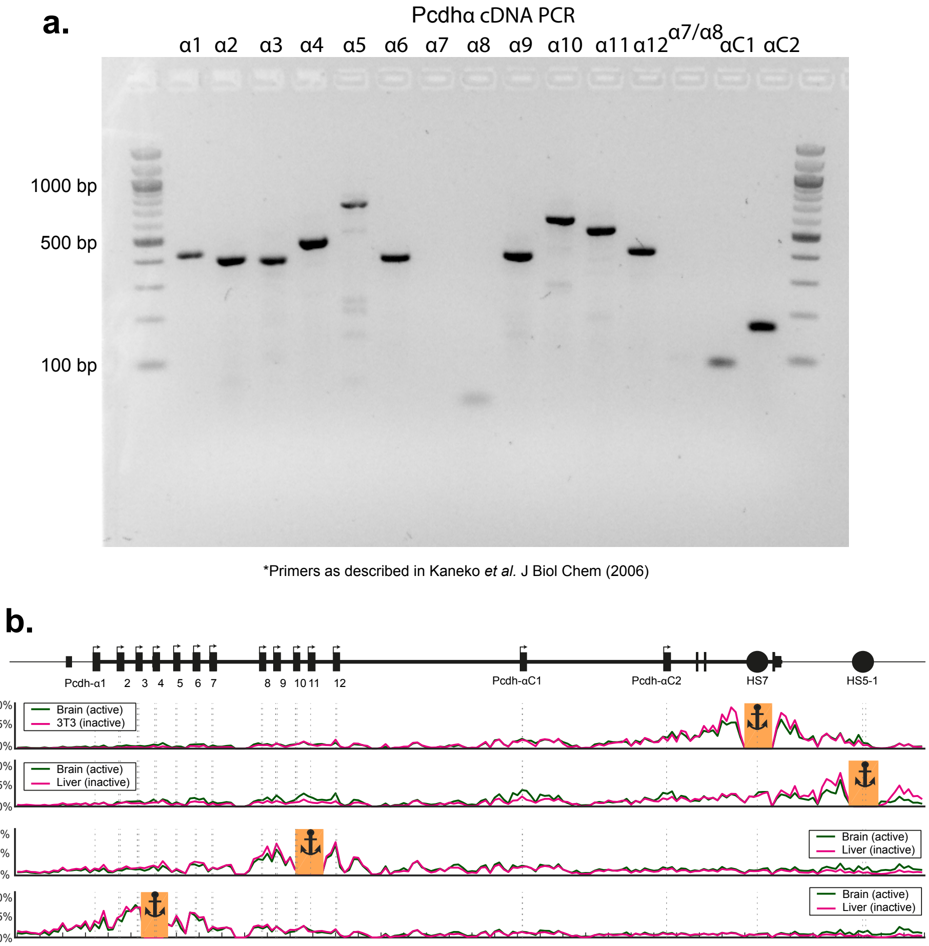


Figure 7.10: Distinguishing cooperative from random and competitive DNA interactions. Schematic overview of association analysis performed for determining preferential contacts formed in the region of interest. a Selection of n reads containing the VP (orange region) and SOI (green region) as positive set. b. Random draw of n reads from the negative set, consisting of reads that contain the SOI, repeated 1000 times. The mean and standard deviation of observing each site in the sub-sampled set are calculated. c. Enrichment of the positive profile (compared to the negative profile) in Y1 which indicate favored contacts between X and Y1 when V is present. d. Random contact frequency between X and Y2 in the case V is present e. Unfavoured contact between X and Y3 when V is present.



*Primers as described in Kaneko *et al.* J Biol Chem (2006)

Figure 7.11: Pcdhα expression in E14.5 fetal brain cells. **a.** Alternative exon-specific primers were used for a PCR (n=1) on cDNA to test which promoters are active in E14.5 fetal brain cells. Primers are listed in Supplementary Table 2. **b.** Overall profiles of Pcdhα4, Pcdhα11, HS5-1 and HS7 viewpoints in liver (inactive) and brain (active) cells.

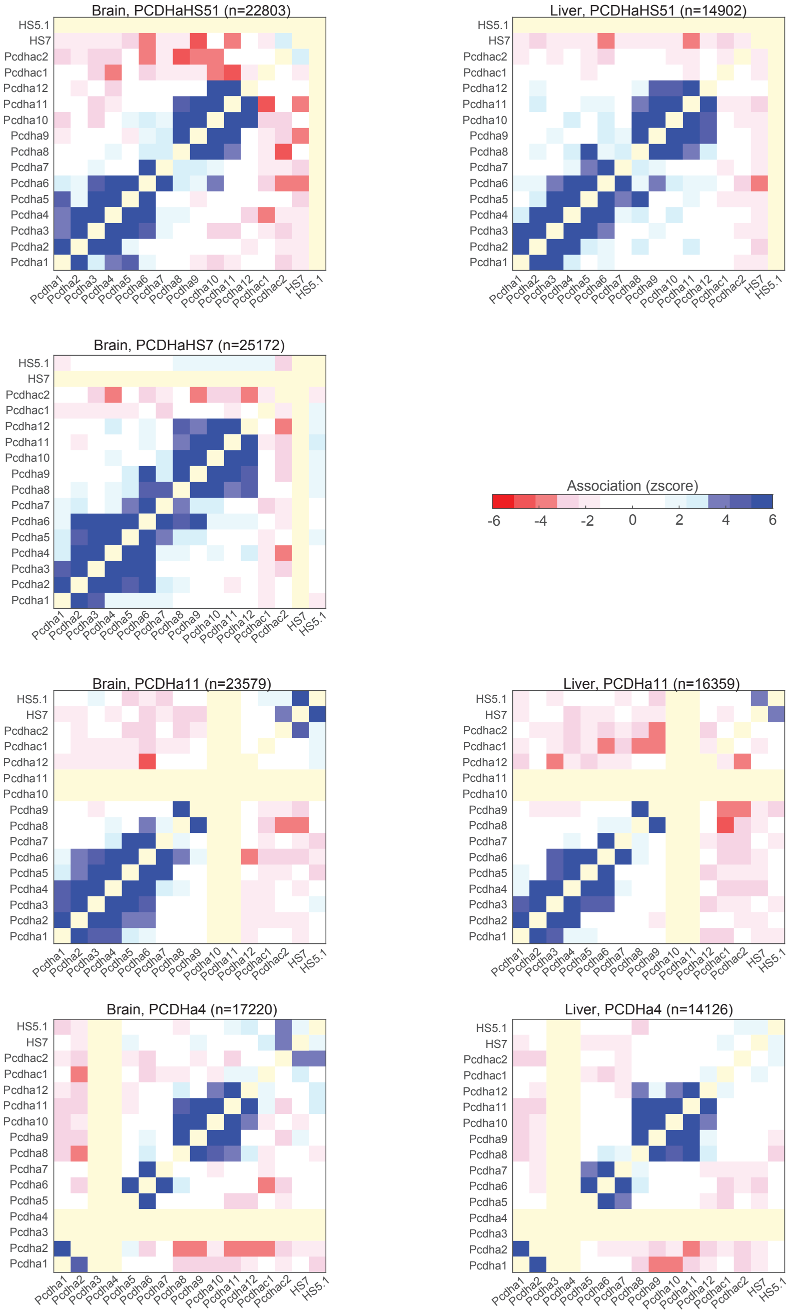


Figure 7.12: Allelic co-occur frequencies at the active and inactive Pcdhα locus. Complete overview of co-occur significance scores found between any pair of genes and/or enhancers constituents in experiments using Pcdhα4, α11, HS7 and HS5-1 as viewpoints, in E14.5 fetal liver (Pcdhα inactive) and E14.5 fetal brain cells (Pcdhα active). Color of each square represents z-score of the association test used to assess significance of preferential contacts observed between corresponding genomic pair (in presence of the view point). Please refer to original publication to see association scores for each pair of sites [1].

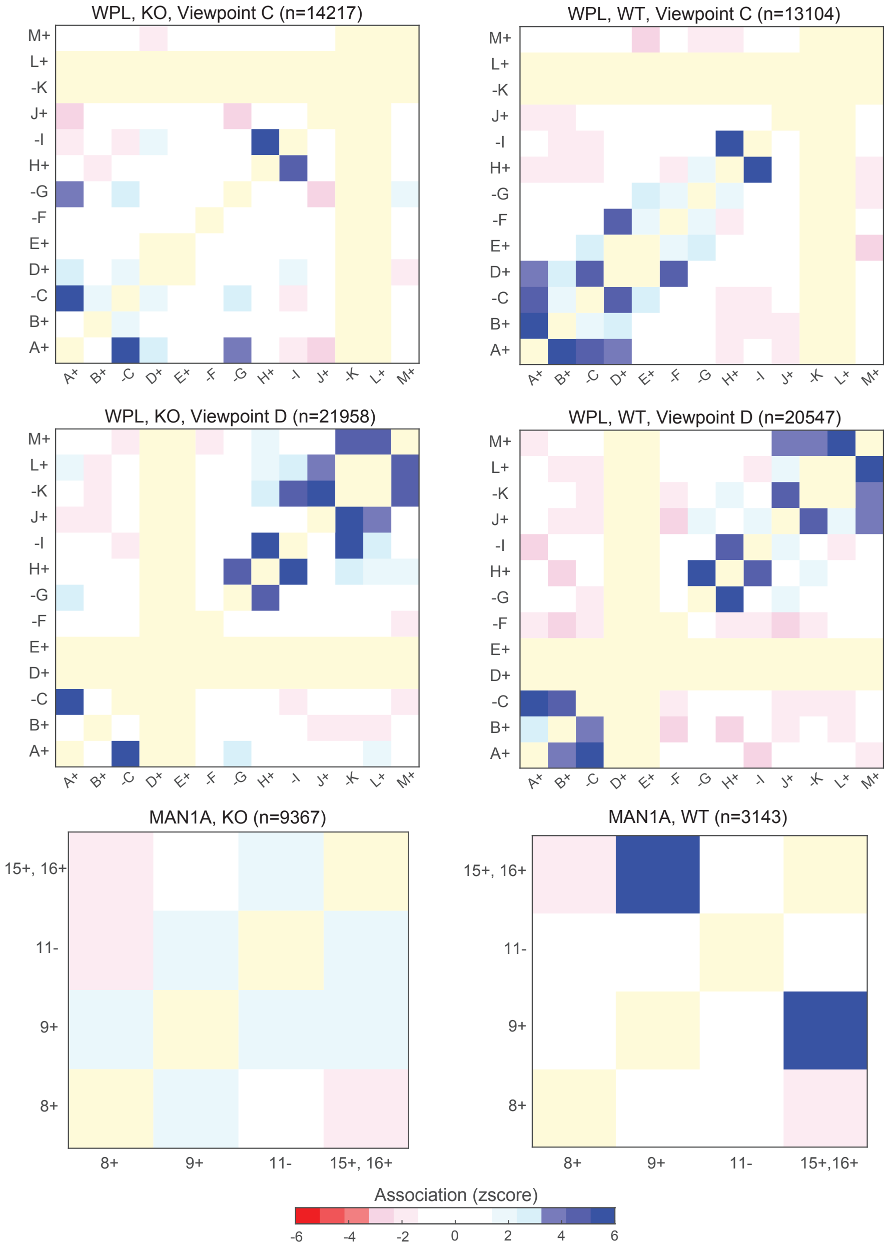


Figure 7.13: Micro-topologies uncovered in the MAN1A locus in ΔWAPL Hap1 cells. **a.** HiC data obtained in WT and ΔWAPL Hap1 cells, in the MAN1A locus, showing multiple novel long-range loops formed exclusively in absence of WAPL. Forward and reverse CTCF sites are indicated, as well as the viewpoint used in MC-4C experiments and the CTCF sites used as SOI **b.** Viewpoint-SOI profiles for the MAN1A viewpoint, using three different CTCF sites as SOI, showing CTCF clustering. z-scores are plotted below the profiles. Color of each square represents z-score of the association test used to assess significance of preferential contacts observed between corresponding genomic pair (in presence of the view point). Please refer to original publication to see association scores for each pair of sites [1].

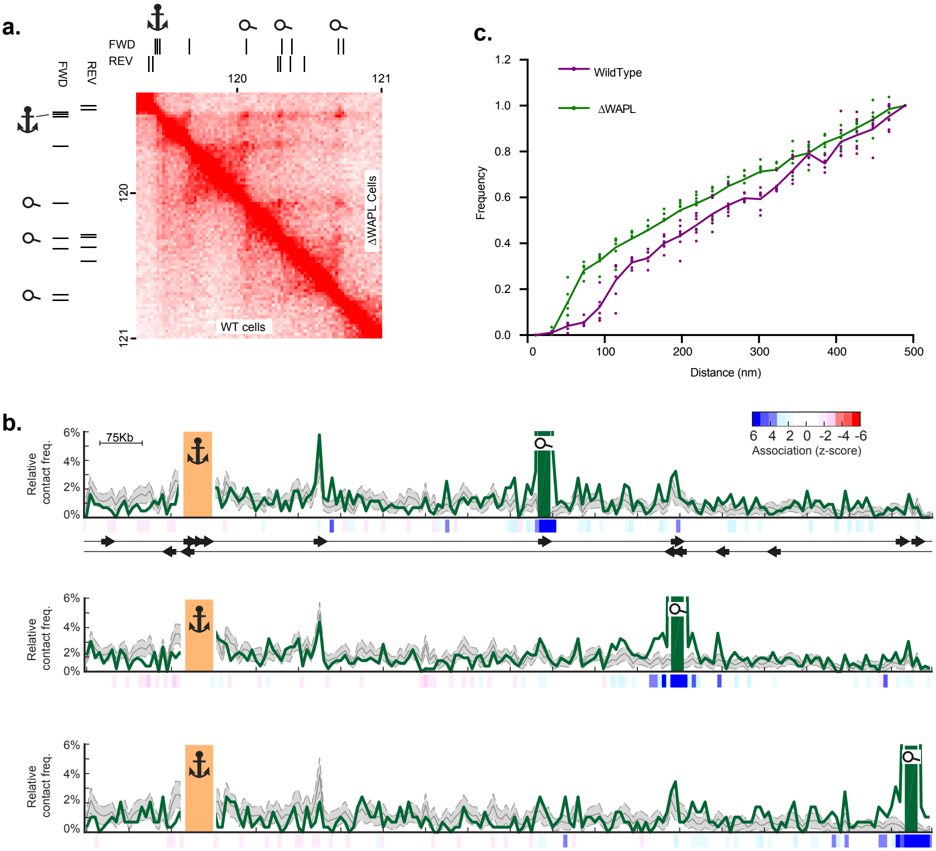


Figure 7.14: Allelic co-occur frequencies of architectural loops in presence or absence of WAPL. **a.** HiC plot of a region of interest on chromosome 6 that shows clear extended loop formation between WT (bottom left) and ΔWAPL Hap1 cells. CTCF sites are separated on orientation and indicated on the top and left axes. An MC-4C viewpoint was chosen (indicated as an anchor) and the three SOIs that are shown in b are indicated with magnifying glasses. **b.** Complete overview of co- occur significance scores found between pairs of CTCF sites in the selected region on chromosome 6, in WT hap1 cells and ΔWAPL Hap1 cells. Green line indicates frequencies observed in ΔWAPL VP-SOI selection, gray areas indicate the background (negative selection) profile (mean ± SD), blue/red bars indicate the z-score for each bin. CTCF sites are indicated below the top-most plot, with arrows indicating their direction. **c.** Dot plot showing the cumulative frequency for the distance between individual Cohesin proteins relative to the total number of distance measurements. Dots indicate frequency for each of the five analyzed cells per genotype, and lines indicate average frequency per genotype.

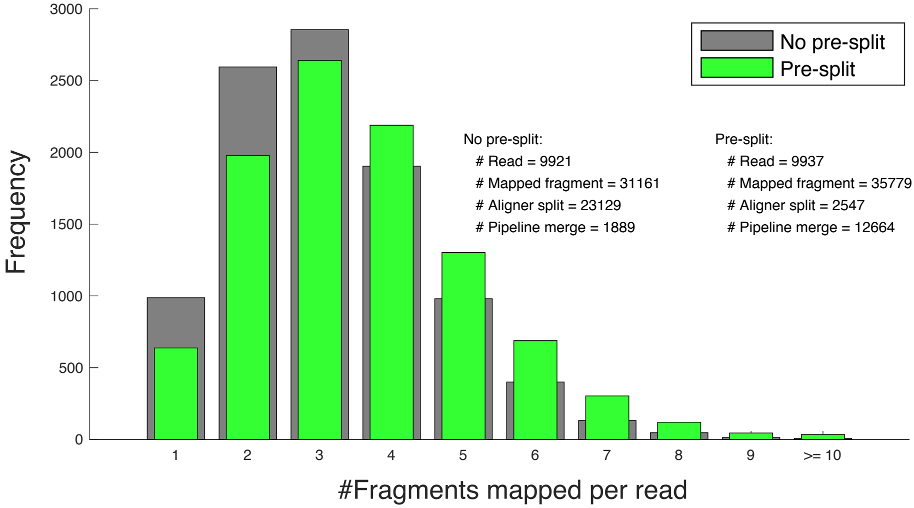


Figure 7.15: Comparison of mapping pre-split reads vs mapping unmodified reads. For this analysis, 10,000 reads were randomly selected from the Hbb-b1/liver experiment and mapped to the reference genome (mm9). For mapping, BWA-SW was used with identical parameters for both methods. Bars represent read size distribution (in terms of the number of mapped fragments). Fragments were considered to be mapped if their Mapping Quality (MQ) is >= 20. The adjacent fragments in reads were merged together if they map closer to 30bp in the reference genome. Based on this result, pre-splitting reads yield 15% more mapped fragments compared to directly mapping reads (31161 vs. 35779). This plot also quantifies the number of times the aligner decided to split a given read or fragment (# aligner split) as well as number of times the MC4C pipeline merged two adjacent fragments (# pipeline merge). These statistics show the decrease in split-read mapping when reads are pre-split (23129+1889 vs. 2547+12664).

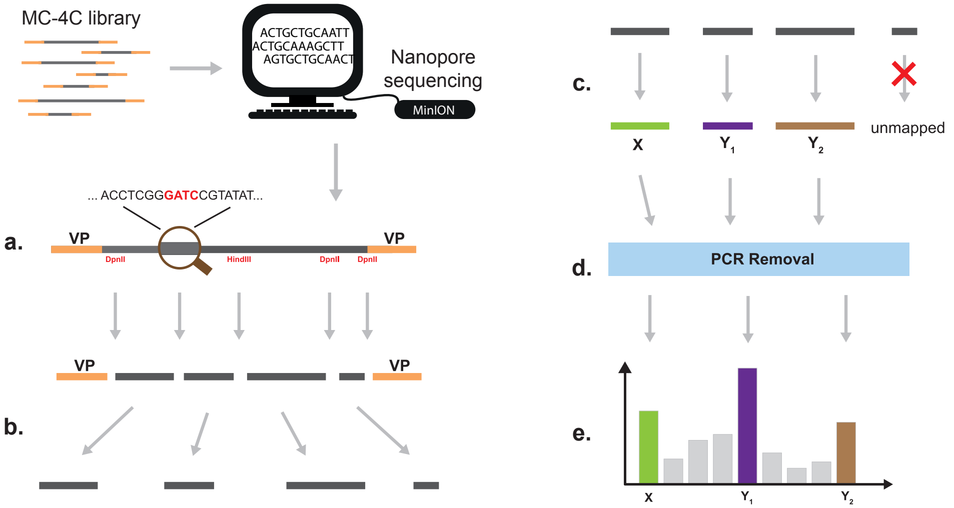


Figure 7.16: Schematic representation of pre-processing steps in MC-4C after sequencing. **a.** Read split: Reads are split into fragments according to the restriction enzyme recognition sequence (only DpnII is depicted in the figure). **b.** Fragment mapping: Fragments are mapped to the reference genome. Due to sequencing errors or short length of fragments, some fragments may not be mapped confidently and are discarded. After mapping, fragments are extended (or shrunk) to the nearest restriction site in the genome. **c.** PCR filter: Reads that have two or more mapped fragments (in addition to the viewpoint) are selected for PCR filtering to ensure that each read represents a single allele (see also Supplementary Figure 1). **d.** Association analysis: PCR filtered reads are employed to assess multi-contact associations between elements in the region of interest (see also Figure 7.10).

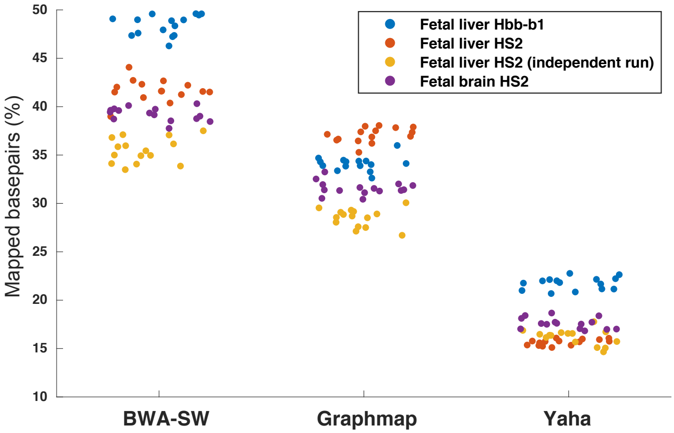


Figure 7.17: Performance assessment of state of the art aligners. 5000 fragments were randomly selected and mapped to a shortened reference genome of ±1MB around the viewpoint using three aligners including BWA-SW [6], Graphmap [7] and Yaha [8] with default settings. Percentage of base pairs mapped (MQ >=20) for each aligner is depicted. This procedure is repeated 15 times to assess sampling bias (represented by dots). To avoid dataset specific performance bias, fragments from different tissue type and viewpoints are used including: Brain-HS2, Liver-HS2, Liver-BMaj. A repeated sequencing of Liver-HS2 is also included to investigate protocol specific variation.

Table 7.1: Statistics of the MC-4C experiments. Shown, per experiment, are the total number of reads sequenced per experiment (Raw read column), number of reads with more than one fragment - excluding VP - in the region of interest (Informative), number of reads with at least one UMI fragment allowing to check for PCR duplicity (Has far cis /trans UMI), number of independent alleles after removing PCR duplicates (PCR filtered unique reads), and the number of MinION sequencing runs that were pooled for each experiment (Sequence runs). The numbers in parentheses are percentage of reads remained after each step of filtering compared to total number of reads sequenced.

Dataset name	Raw reads	Informative	Has far cir/trans UMI	PCR filtered unique reads
Fetal Liver HS5	985391	129239 (13.1%)	74119 (7.5%)	8149 (0.83%)
Fetal Liver HS3	1200081	57462 (4.8%)	26378 (2.2%)	7854 (0.65%)
Fetal Liver HS2	1571364	136600 (8.7%)	105423 (6.7%)	5970 (0.38%)
Fetal Liver Hbb-b1	1154371	128496 (11.1%)	89369 (7.7%)	9775 (0.85%)
Fetal Brain HS5-1	1190411	140087 (11.8%)	89992 (7.6%)	23478 (1.97%)
Feral Brain HS7	1821435	64505 (3.5%)	39454 (2.2%)	25583 (1.40%)
Fetal Brain Pcdh-a11	2031143	312377 (15.4%)	202415 (10.0%)	24186 (1.19%)
Fetal Brain Pcdh-a4	1361631	149871 (11.0%)	95417 (7.0%)	17602 (1.29%)
Fetal brain HS5	2690060	64882 (2.4%)	43679 (1.6%)	7016 (0.26%)
Fetal brain HS2	3329153	90382 (2.7%)	74271 (2.2%)	3061 (0.09%)
Fetal brain Hbb-b1	5777087	81136 (1.4%)	61232 (1.1%)	6390 (0.11%)
Fetal Liver HS5-1	1857223	53952 (2.9%)	38346 (2.1%)	15172 (0.82%)
Fetal Liver Pcdh-a11	2245836	51733 (2.3%)	36817 (1.6%)	16761 (0.75%)
Fetal Liver Pcdh-a4	2888720	41661 (1.4%)	32932 (1.1%)	14341 (0.50%)
WT E	5511511	325057 (5.9%)	108278 (2.0%)	22820 (0.41%)
ΔWAPL E	5239363	588116 (11.2%)	186585 (3.6%)	23878 (0.46%)
WT K	2491602	110958 (4.5%)	43296 (1.7%)	14322 (0.57%)
ΔWAPL K	3339213	199430 (6.0%)	67365 (2.0%)	15491 (0.46%)
Average	2593644	151441 (5.8%)	78632 (3.0%)	14547 (0.56%)

7.5. Mapping reads to reference genome

In order to map the partial reads to the reference genome, we utilized BWA v0.7.16a41 in SW mode (setting: -b 5 -q 2 -r 1 -T 15). Furthermore, the Z-best heuristic of this aligner is set to 10 (i.e. -z 10). This heuristic increases accuracy of the aligner at the cost of speed. On average, mapping one million fragments takes about an hour using a 64 core system running Linux CentOS v7.0. BWA-SW performed best among several tested split-aligners (see Figure 7.17).

7.6. Fragment extension and neighbor fusion

Fragments are extended to nearest restriction site (either the 4-cutter or 6-cutter restriction site) in the reference genome. Extension is continued to next restriction site in the reference genome if a given fragment is mapped more than 10 bases after an identified restriction site. Any fragments that map closer than 30bp in the reference genome are fused together and considered to be a single fragment in the rest of analysis. Finally, any fragment with mapping quality below 20 is considered as unmapped. Figure 7.3 demonstrates read size distribution of two representative experiments in β-globin (i.e. liver and brain cells) in terms of number of contacts after extension and fusion and preservation of confidently mapped fragments.

7.7. Duplicate removal

In order to detect PCR duplicates, we utilized a conservative approach which is based on the premise that in MC-4C, fragments that map far away from the viewpoint are unlikely to be found more than once (see 7.5). Therefore, these far-cis/trans fragments can be directly used as Unique Molecular Identifiers (UMI)s. Therefore, if these UMI fragments are identified in two reads, those reads are far more likely to be the result of a PCR duplication than of two independent ligation events. A schematic representation of this approach is depicted in (see 7.1).

Once a duplicate is found, we removed the read with smaller number of local fragments (i.e. fragments that are mapped within the locus of interest). Locus of interest is defined as a region around the viewpoint that contains expected interacting partners in the locus. Finally, reads that have less than two fragments within the locus of interest are discarded as they are not informative in multi-way contact analysis. Once duplicated reads are filtered, we confirm the validity of MC-4C data by comparing overall profiles with standard 4C (see 7.7)). This is further confirmed by comparing primary vs. secondary ligations (see Figure 7.6). Finally, we compared reproducibility of profiles generated by nanopore sequencing technology to the same profile generated by PacBio sequencing technology. While overall profiles in both platforms show high degree of similarity, nanopore sequencing yielded 10 times more unique reads and was chosen as the primary platform for MC-4C (see 7.8).

7.8. Association analysis

To uncovered contact predisposition between the Viewpoint (V) and two other Sites Of Interest (SOI), say X and Y, we hypothesized that if preferential contact between X and Y (in presence of V) exists, this propensity should be absent when X is not present in the concatemer (V and Y but not X). Accordingly, this preference can be revealed by comparing profiles of reads that contain V and Y and X (positive set, Figure 7.10.a) versus a background profile that is formed from reads that contain V and Y but not X (negative set, Figure 7.10.b). To determine if a read contains a SOI (i.e. V, Y or X), the locus of interest is divided into 200 bins and the SOI perimeter is defined by 3 bins centered around the SOI center. The frequency of observing a SOI is determined by the number of reads that contain a fragment within the SOI or overlapping the SOI boundary.

To account for conformational variation that may occur across the population of cells, we subsampled reads from the negative set to the number of reads in the positive set. This procedure is repeated 1000 times. Moreover, we implemented a correction for the fact that - by definition - reads in the positive set already contributed a fragment to SOI X. Therefore, the positive profile is effectively produced by smaller reads (i.e. #fragments - 1 for each read). Hence, on average, each read in the negative set supplies an extra fragment to the profile compared to reads in the positive set. To compensate for this, and ensure both negative and positive profiles are constructed based on the same distribution in terms of fragments per read, one fragment from each negative read is randomly removed in every random subsampling of the negative set. Finally, the mean and standard deviation of the frequency at which SOI Y is observed in the negative set is calculated. Using these statistics, a z-score can be determined to estimate significance of the (un-)favored contacts formed between V, X and Y. While a modest (close to zero) z-score indicates a random contact frequency between X and Y when V is present (Figure 7.10.d), a positive or negative z-score implies a favored (Figure 7.10.c) or unfavored (Figure 7.10.e) contact between these three elements, respectively.

Table 7.2: Primers used in this study. The primers used for each individual viewpoint, and the coordinates of each region of interest. FW and RV primers are used for MC-4C PCR.

Viewpoint	Region of interest	FWD	REV
Hbb-b1	chr7:110933500-111066500	GCAGTAGTGATTCTATTCAATTTTTGGGATC	CCAGATTTGTGAGCTCAGGGTTTAC
HS2	chr7:110933500-111066500	CAGATGTTTTCAGCTGTGACTGAT	CTTGGACAGTGGTACTGCAATAATT
HS3	chr7:110933500-111066500	CAAAGCAGCCTCTCTCAGTCCC	CTTCTCATTCTCTCAGCTATGTGAAAACAACC
HS5	chr7:110933500-111066500	GGATTTTTCAAAGGCCTGAACTCAAACC	GTCTGTAGGCTCCATAATAATTGTCTTCCC
Pcdhα4	chr18:37060000-37400000	TTCTCACCAGTGACTGTATGTGATC	ATGATGTCGCTCTTTACCGTCAAATA
Pcdhα11	chr18:37060000-37400000	CGCTCTTTACTTGGTGGAAAGA	CCTTAGCTATGTAGGTTTGCATTCT
Pcdhα HS7	chr18:37060000-37400000	TTTGTGGACTGACTGGAGAAGC	AGCCTCTGGATACTCACATGCAA
Pcdhα HS5-1	chr18:37060000-37400000	GGAGGAGGTTAAAGCAAAGACTAAG	ATCTCTGGTATTGTAAAGTGGTCGA
Wapl CTCF E	chr8:120800000-122075000	CAAAGGGAGAGCGCCATCTA	CTCCTGCTCTTCACATCTCAAG
Wapl CTCF K	chr8:120800000-122075000	AGCTGGACATTCTTCAACTGC	GACATGACGTTTGGCTCCATG
Man1A	chr6:119250000-120750000	CACATGTAAAGACTAATTATGAGACGC	GCTCCAGAAATGAAATTTTAGGGAG

REFERENCES

[1] A. Allahyar, C. Vermeulen, B. A. M. Bouwman, P. H. L. Krijger, M. J. A. M. Verstegen, G. Geeven, M. van Kranenburg, M. Pieterse, R. Straver, J. H. I. Haarhuis, K. Jalink, H. Teunissen, I. J. Renkens, W. P. Kloosterman, B. D. Rowland, E. de Wit, J. de Ridder, and W. de Laat, *Enhancer hubs and loop collisions identified from single-allele topologies,* Nature Genetics (2018), 10.1038/s41588-018-0161-5.

[2] J. H. I. Haarhuis, R. H. van der Weide, V. A. Blomen, J. O. Yáñez-Cuna, M. Amendola, M. S. van Ruiten, P. H. L. Krijger, H. Teunissen, R. H. Medema, B. van Steensel, T. R. Brummelkamp, E. de Wit, and B. D. Rowland, *The cohesin release factor WAPL restricts chromatin loop extension,* Cell **169**, 693 (2017).

[3] H. J. G. van de Werken, P. J. P. de Vree, E. Splinter, S. J. B. Holwerda, P. Klous, E. de Wit, and W. de Laat, *4C technology: protocols and data analysis,* Methods Enzymol. **513**, 89 (2012).

[4] E. Splinter, E. de Wit, H. J. G. van de Werken, P. Klous, and W. de Laat, *Determining long-range chromatin interactions for selected genomic sites using 4c-seq technology: from fixation to computation,* Methods **58**, 221 (2012).

[5] H. J. G. van de Werken, G. Landan, S. J. B. Holwerda, M. Hoichman, P. Klous, R. Chachik, E. Splinter, C. Valdes-Quezada, Y. Oz, B. A. M. Bouwman, M. J. A. M. Verstegen, E. de Wit, A. Tanay, and W. de Laat, *Robust 4c-seq data analysis to screen for regulatory DNA interactions,* Nat. Methods **9**, 969 (2012).

[6] H. Li and R. Durbin, *Fast and accurate long-read alignment with burrows–wheeler transform,* Bioinformatics **26**, 589 (2010).

[7] I. Sović, M. Šikić, A. Wilm, S. N. Fenlon, S. Chen, and N. Nagarajan, *Fast and sensitive mapping of nanopore sequencing reads with graphmap,* Nature communications **7**, 11307 (2016).

[8] G. G. Faust and I. M. Hall, *Yaha: fast and flexible long-read alignment with optimal breakpoint detection,* Bioinformatics **28**, 2417 (2012).

[9] B. Langmead and S. L. Salzberg, *Fast gapped-read alignment with bowtie 2,* Nature methods **9**, 357 (2012).

8

Discussion

The more we learn about the universe the simpler it seems, but the cell isn't like that. The more we find out the more complicated things get

Prof. Steve Jones, University College in London

High-throughput expression assays such as microarrays opened a new era in biomedical research. Nowadays, the study of the transcriptional landscape of cancer cells by exploiting massive datasets enclosing several thousand samples is common in genomic research. It is infeasible to manually process, analyze and extract relevant information from this 'Big data'. Computational biology therefore has become an integral part of these investigations. This interdisciplinary collaboration substantially facilitated knowledge discovery and revolutionized modern biology. Nonetheless, combining pieces of information from independent datasets to complete the puzzle and reach the sought knowledge remains challenging. Intending to accelerate knowledge acquisition, a myriad of computational methods were designed in the recent years to integrate data from variety of sources.

In this thesis, we initially focused on integration of gene expression data and biological networks (such as protein-protein interaction networks or gene co-expression networks) and how such a integration could be exploited to estimate survival risk of breast cancer patients. On the second part of this thesis, we took preliminary steps in forming a new biological network that captures multi-way DNA interactions which occur between multiple functional elements of the genome (such as genes or enhancers) in the cell nucleus.

While this thesis explores several aspects of inferring and integrating biological networks for analyzing gene expression data to predict disease outcomes, there are still numerous angles that were not investigated and will be addressed here. On the other hand, whilst proposed methods are developed to tackle specific problems, there application reaches far beyond the initial biological question. Consequently, several other area of research that could benefit from methodologies developed in this thesis will be surveyed. Finally, we will discuss our future perspective on data integration frameworks and network-based models and how we envision advances in such techniques to catalyze our understanding of the underlying working mechanism of cell system.

8.1. NETWORK BASED OUTCOME PREDICTION

To carry out the analysis required for Chapter 2 and 4, diverse types of data and methods were utilized. The corresponding parameters for each method were sought to be based on proper side-analysis to make sure downstream conclusions are as rigorous as possible. At the same time, to make these analyses feasible, choices were needed to be made (or details ignored) without in-depth investigation. In this section, we aim to elaborate on some of these aspects that may have more profound implications on the aforementioned analyses. We approach these details from two different standpoints: data and methods.

8.1.1. CHALLENGES IMPOSED BY DATA CHARACTERISTICS

As described in Chapter 1, large datasets can be formed by pooling expression profiles from independent studies [1]. The expected study-specific expression variations then need to be removed. Due its popularity and ease of use, we utilized COMBAT [2] to achieve this in Chapter 2 and 4. However, many other methods exist that could be used to achieve this (see [3, 4] for survey). Hence it is necessary to investigate whether these models can identify and correct more subtle study-specific variations in the collected cohort.

It is essential to recognize that these "transformative" procedures have detrimental effects as (parts of) the discarded expression variations could have intrinsic (biological) origin. For instance, some studies may have targeted different subtypes of breast cancer which could potentially skew the transformation. Notably, the ACES dataset [5] suffers from this effect since the Desmedt *et al.* cohort investigated outcome markers of node-negative patients [6], while the Loi *et al.* cohort targeted estrogen receptor positive tumors and studied their latent molecular subtypes [7].

Another source of inconsistency can be seen in the "time-to-event" clinical outcome variable. Currently, diverse definitions for this variable are used in clinical trials [8]. Problematically, such variable is often vaguely described in the original article [9]. This diversity and ambiguity in published variable challenges the post interpretation and collection of this variable which in turn induces disagreement in patient outcome [10].

The outcome inconsistency combined with the relatively small publicly available cohorts (fewer patients than variables; genes) made it difficult to form a large and coherent expression dataset which is required for reliable training of the Network-based Outcome Predictors (NOPs). In particular, the collected datasets in this thesis utilized two different “time to event” outcome variables: *Overall* and *Recurrence/Relapse Free* survival. Overall survival indicates the time between diagnosis of cancer and the date in which that patient was last known to be alive (used in METABRIC and TCGA). The recurrence/relapse Free survival denotes the duration between primary treatment and first signs/symptoms for return of that cancer (Used in ACES). Although combining these datasets increased the number of samples by a factor of two, we resorted to ignore the “time to event” differences and regarded this variable as a single entity of *survival time*. This is of course not ideal. Specially because the given labels are considered to be the “truth” in standard classifiers [11].

It should be noted that outcome discrepancy is not the only complicating factor for dataset pooling. Preparation of samples in the lab requires numerous intricate steps which increases the chance of mislabeling samples. Such errors in labeled datasets can be identified by correlation analysis. For example, we utilized the collected cohort in Chapter 4 and probed for highly correlated patients with opposite prognosis. To our surprise, we found several samples satisfying the aforementioned criteria. A notable example was “MB-0228” sample in the METABRIC data with a very poor prognosis (survival time of 325 days). The expression profile of this sample was highly correlated (min(ρ)=0.79) with multiple good prognosis samples (see Figure 8.1.a). Interestingly, the most correlated patient profile in our cohort had ID of “MB-0282” which seems very similar to former patient with ID of “MB-0228” (see Figure 8.1.b), suggestive of incorrect survival time for this patient. One potential solution for this problem could be the incorporation of probabilistic labels representing our confidence in correctness of the sample labels [11, 12]. Such methods explicitly model the uncertainty of labels and by that may improve performance of NOPs. *Transfer learning* is a relatively new concept in machine learning that can rectify such inconsistencies [13]. These models are designed to learn an abstraction from one problem and then “transfer” the extracted information with the aim of solving a different but related problem [14, 15]. Here, we could use this concept to train a predictor using samples with overall survival labels and then transfer the learned model to predict outcome of patients that are represented with recurrence/relapse free survival labels.

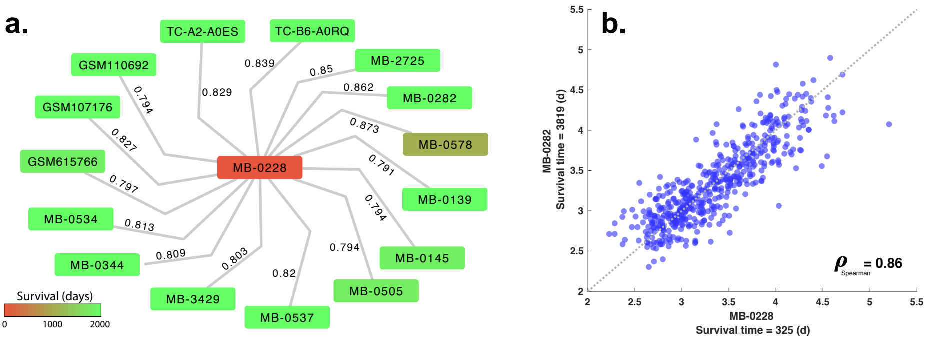


Figure 8.1: Mis-labeled samples can reduce performance of outcome predictors. **a.** Expression profile of a patient with very poor prognosis (red node, survival=325 days) is highly correlated (min()=0.79) with many patients (green nodes) with good prognosis. The top 15 highly correlated patients with this patient in our cohort are visualized. Node colors represent survival of patients and numbers along the edges denote spearman correlation between pair of patients. **b.** Expression profile visualization of top highly correlated patient with "MB-0228" ID.

In this thesis we employed leave-one-study-out cross validation to resemble real-world application of outcome prediction models. However, the entire dataset (training and test set) were analyzed together to identify and remove the batch effects. Yet, in real-world applications, the intrinsic batch effects would be known only in the test phase (i.e. in the clinic). Accounting for this aspect in the assessment of outcome predictors is challenging (if not impossible) as technical variation in the test set can be exerted by innumerable environmental factors with diverse magnitudes. The recent efforts in promoting standardized preparation protocols and processing procedures could have considerable impact in reducing these effects which in turn simplifies assessment of outcome predictors [16].

8.1.2. CHALLENGES RELATED TO MODELS

To analyze the expression data in Chapter 2 and Chapter 4 we focused on the two class (binary) classification problem to discern between patients categorized as poor or good prognosis. The patient prognosis is determined by dichotomizing corresponding survival times according to a (clinically established) five years threshold. This discretization may result in loss of relevant information. Although we utilized Lasso (and its derivatives) as classifiers, they are in fact regression models capable of directly incorporating continuous labels. Therefore, it may seem reasonable to drop the discretization step and directly predict survival time of patients instead. Although, such models may perform poorly due to the unreliability of clinical variables (see section 8.1.1). Another difficulty in utilizing standard classifiers in survival analysis related to the censoring of patients which is very common in clinical trial datasets. Censoring occurs when the time of event (i.e. death) for a patient is not known. For such datasets, Cox proportional hazards models are advised that are capable of incorporating censored survival times as well [17].

To ascertain performance of the proposed models, we utilized the cross-study evaluation to better resemble real world application of these models. We motivated this approach in Chapter 2 by demonstrating that all models under study show inflated performance if standard cross-validation scheme is used. To explain this observation, we argued that standard cross-validation allows the classifier to “see” the intrinsic batch effects in the data which can be further exploited to improve performance. Another likely explanation for this observation can be the utilized “stratified” cross validation scheme which keeps the ratio of subtypes in the training and test set comparable. In contrast, in cross-study validation each study in the collected cohort may contain a different ratio of subtypes and therefore change of subtype composition in the training and test set can be substantial (see Figure 8.2.a). This in turn may deteriorate the performance of the trained models [18]. Similarly, the cost function in classifiers assumes a comparable ratio of classes (good vs. poor outcome) in the training and test set and penalizes miss-classification errors according to this ratio. However, such a property is often violated in pooled cohorts (see Figure 8.2.b). If the test set contains more samples of the “harder-to-classify” class, this will result in the deterioration of performance for that particular test set, even though a good (cross-validation) generalization was achieved during the training phase. One way to mitigate this effect would be to stabilize frequency ratios across training and test set using down sampling of the frequent category. The drawback for this approach is that not all samples are used in the training phase of the classification which could be specially an issue when the number of samples is already small. Another way to circumvent this issue is to explicitly assign weights to samples in the training set according to ratio of samples in the test set. The downside of this approach is that the test set is utilized to assign a parameter in the model which is known to yield over-estimated performance. Another approach would be to have subtype-specific error assignments which balances the occurred error rates according to the frequency of subtypes in the test set.

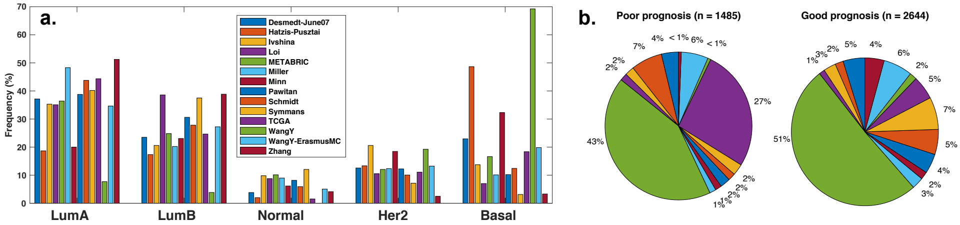


Figure 8.2: Instabilities in the category (either subtype or outcome) frequencies across test and training set. This variation can substantially deteriorate performance of models. **a.** Frequency of subtypes in collected cohort changes across studies. **b.** Number of samples categorized in prognosis vs. poor outcome changes noticeably across datasets.

In Chapter 2 and Chapter 4 we focused our analysis on two different variants of lasso namely group lasso and sparse group lasso to incorporate network information in outcome prediction models. The main restriction in these classifiers is the fixed group size that needs to be defined prior to training. Although in Chapter 4 we addressed this issue by optimizing the group size in the training phase, all gene sets in our model were confined to a constant size (in terms of number of enclosed genes). Yet, number of genes per functional modules or pathways in the cell could vary substantially [19]. One way to circumvent this problem is to infer the number of genes per group in a greedy fashion similar to Chuang *et al.*. In each iteration of greedy expansion, a linear regression model can be trained (using training set) and then its prediction vector could be used as the "meta-gene". In this case, the regularization can be left out as the number of genes in each group are often small compared to number of patients [20]. The drawback of such a model is its tendency to "overfit". This is because sample labels (in training set) are used many times during the greedy procedure and top modules in the training set may not faithfully represent the test set. To circumvent this, the greedy procedure can draw inspiration from boosting approaches utilized in Random Forest [21] where a small (we propose <25% as a role of thumb) number of samples are employed in each iteration. This can be further improved by limiting samples to a single study in each iteration to make sure the observed performance is stable across studies.

Another factor that is not investigated in this thesis is that accuracy of prognosis varies depending on the subtypes of breast cancer [22, 23]. It is known that the Luminal A subtype in breast cancer is associated with good prognosis [24]. This means that there is a higher chance for a patient with Luminal A subtype to be in the good prognosis class. Consequently, a classifier can gain an overall better performance simply by exploiting this property and associating all samples from Luminal A subtype to good outcome [22]. Such a criticism has been made against predictors incorporated in MammaPrint [25] and Oncotype DX [26], two commercially available prognosis predictors in the market [27, 28]. A similar issue can arise when subtypes are not represented with comparable frequencies in the dataset (also known as "class priors" imbalance). A classifier can exploit this property of the dataset and fine tune its parameters to deliver accurate prognosis for frequent subtypes and thereby enhancing its overall performance. Our further investigation into this issue revealed no correlation between the performance of group lasso (guided by SyNet) and different subtypes (see Figure 8.3.a). We speculate that the cross study validation procedure may have prevented this bias as studies in our dataset represented diverse class priors (see Figure 8.2.a) which in turn prevented the classifiers to adapt to a specific subtype.

Another interesting question that was overlooked in Chapter 4 was to investigate the extent to which subtypes of breast cancer benefit from incorporation of network information in the model. To this end, we compared the performance of lasso to its network-based version (GL) that utilizes SyNet to govern its predictions. Figure 8.3.b represents performance gain of (PAM50) subtypes sorted according to their mortality rate [24]. Based on this result, network-based prediction of survival can improve accuracy of Luminal A, B and Normal-like patients much better than Her2 and basal that are often associated with poor survival. This is an exciting finding as many efforts are being done to identify low risk patients with higher accuracy to spare them the toxic effect of chemotherapy [28].

The analyses performed in this thesis were limited to gene expression data. However

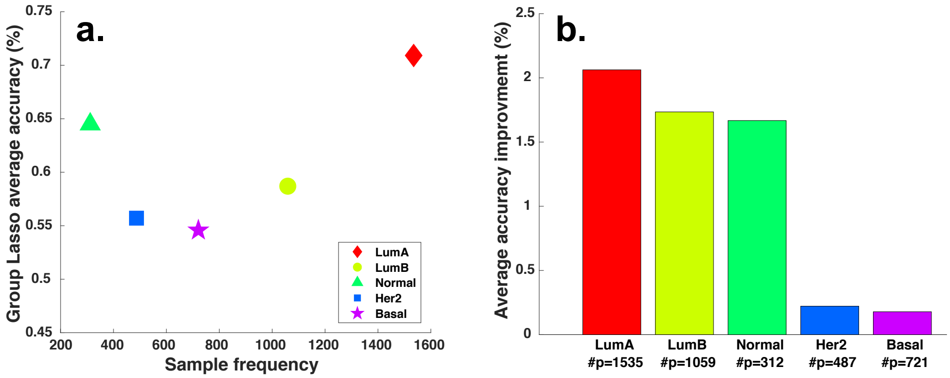


Figure 8.3: Performance of network based outcome predictors to estimate survival risk of patients. **a.** Network based predictions are not biased toward frequency of samples in the cohort. **b.** Network information provides better recognition of markers for patients with low risk of cancer development.

such a single perspective of this complex disease (even after incorporation of network knowledge) may not be enough to untangle the wide range of misregulations that can occur in tumor cells. A notable example is the Basal-like subtype. In the Figure 8.3.b, we showed that the Basal-like subtype has the least benefit from incorporation of network knowledge to the model. Yet, Dai *et al.* reported that combined analysis of gene and *MicroRNA* expression (a non-coding RNA with regulatory activities on target genes) can specifically improve prognosis accuracy of breast cancer patients with basal subtype [29]. Therefore, it can be concluded that for each subtype, we may need to combine different types of data to improve performance of outcome predictors. Recognizing this potential, many aspect of these tumors are currently being measured and the corresponding data are made publicly available. A notable example is the TCGA consortium which apart from gene expression profiles, provides multiple views of the same samples including copy number, SNP profiles and DNA methylation [30]. Utilizing such a comprehensive view of tumors can have a radical impact on the novel insights that can be extracted from computational models.

While the multi-view analysis of tumor profiles is promising, implementing this approach has its own set of challenges. The primary difficulty is normalization of these views into a single coherent dataset [31]. In particular, each view yields features following a particular distributions. For example, while gene expression values are continues and (assumed) normally distributed, SNPs are binary and follow an exponential distribution [32, 33]. One way to circumvent this issue is to design a framework that is aware of these "incompatibilities" across views [34]. To this end and inspired by the Lasso concept, Yang *et al.* proposed a linear regularization model that can perform "multi-view" Lasso [35]. Briefly, this approach learns a low-rank representation of each given view while simultaneously selecting informative features across views. To the best of our knowledge, no group or sparse group regularization is proposed for multi-view analysis.

Another often utilized multi-view approach is Multiple Kernel Learning (MKL) [34]. A kernel is a matrix that represent the pairwise similarity between samples (analogous to a correlation matrix). In MKL, multiple kernels for each view (i.e. mRNA or SNP data) are computed and then optimal weights for a linear combination of these kernels are identified to predict the target label. However, to compute these kernels a suitable similarity function (that preserves discriminative informative) for each particular view needs to be defined which is not a trivial task. Taken together, incorporation of data with diverse nature into a single robust framework is still an open question. It is worth noting that several methods are proposed to perform multi-view outcome prediction by "mapping" gene sets to pathways by averaging expressions [36]. However, as explained in this thesis, this approach may induce substantial information loss and is not recommended.

In this thesis, we essentially regarded wide range of gene relationships as a single entity of "interactions" and did not differentiate between different types of interactions (e.g. physical binding of proteins, protein sequence similarity or their chemical modifications). In addition, each type of interactions may be only meaningful for a particular gene set in our model. This poses a challenging multi-network classification approach where diverse groups are formed according to different network information and then used in outcome prediction.

8.1.3. FUTURE PERSPECTIVES

Owing to the explosion of publicly available data, machine learning is going to be the driving force in personalized treatments of patients and other applications in the clinics [37, 38]. In this section, we will share our future perspective on this topic and speculate on new areas of research that will be possible in the forthcoming years.

Throughout Chapter 2 and Chapter 4, we investigated integration of interaction (such as PPI network) and expression data. To this end, the interaction between two proteins is assumed to be representative of the relationship between their corresponding genes. However, these two types of data are measuring different aspects of intracellular properties. While gene expression represents the relative abundance of mRNA, physical interactions represented in a PPI network describe binding associations between proteins. Within a cell, the abundance of mRNA is related to physical interactions of its protein only after many levels of regulation (e.g. protein translation, folding, transport, stability, etc.) has taken place. To make our analysis feasible in this thesis, such incompatibilities between data types are essentially ignored. How such differences should be incorporated in the model is an open question.

In many (if not all) outcome prediction methodologies, a single profile is considered to represent a complete transcription profile of that patient across all cells. Nonetheless, recent single cell transcriptome profiles of breast cancer tumors revealed a high degree of heterogeneity across cells within a tumor [39]. Similarly, a single time point is utilized to represent the overall cell state of a patient while temporal transcriptome analysis of breast cancer patients has established the dynamic nature of expression profiles through time [40]. This complication is also valid for interaction networks where functional elements in the genome interact dynamically over time. Many methods are proposed that utilize temporal expression data to construct such "time-aware" networks [41]. Considering the heterogeneity across samples, temporal single cell expression data from tumors would be required to infer such a network. Even then, such interactions merely represent correlation associations between genes which is only one aspect of cellular state [42].

Another promising application of machine learning in genomic research is treatment response prediction [43]. Essentially, these models analyze (survival and treatment labeled) expression profiles of patients and aim to determine the best treatment for newly admitted patients to improve their survival [44]. In this context, network based prediction of treatment response is an understudied problem. Our proposed framework in Chapter 4 can be directly incorporated in this problem. To this end, an interaction network can be inferred that connects genes that show synergism with respect to treatment response.

Although, outcome predictor models are well-established within the research domain, they did not find wide application in clinics due to time consuming and labor intensive preparation protocols and computational requirements [45]. With the advent of mobile sequencing technologies like MinION, this desire could be potentially satisfied. Recently, many proof of concept applications for mRNA profiling using these technologies have appeared in the literature [46–48]. We expect that the mobility of these technologies would be a milestone in incorporation of molecular profiling in the clinic in the near future.

8.2. Multi-contact 3D conformation of the genome

In Chapter 6 and 7 we introduced a novel method called Multi-Contact 4C (MC-4C) to investigate the higher order (i.e. more than pairwise) interactions between functional elements in single alleles in a targeted region of interest. Using MC-4C, we demonstrated the existence of an Active Chromatin Hub (ACH) in the β-globin locus and revealed multi-component architectural clusters in Hap1 cells.

Multi-contact analysis of DNA interactions involves complex laboratory routines and computational obstacles that need to be dealt with. With a special attention to the computational aspect, we briefly review these challenges in the following sections and propose solutions to address these hurdles. Additionally, we describe routes that have the potential for future investigations.

8.2.1. Challenges in the data

In our multi-contact analysis, cells were collected from liver or brain tissue of mice, 14.5 days after their conception. However, accurate dissection and collection of these cells are difficult. In fact, harvested cells are often polluted with other cell types (e.g. brain cells are often mixed with many blood cells) which may have entirely different interaction profiles compared to the cells of interest.

In addition, no cycle synchronization for cells has been performed before library preparation. Consequently, fixed cells may be in different cycle states which blurs the measured conformations [49].

Using MC-4C we revealed synergistic hubs that are formed between individual enhancers in the β-globin locus. This is in contrast with recent findings by Olivares-Chauvet *et al.* [50]. One explanation for this discrepancy is the difference between assessed cell types i.e. Olivares-Chauvet *et al.* studied human K562 cell lines, surrogates of erythroid cells expressing the globin genes at considerably lower levels whereas we used primary mouse liver cells that are rich in red blood cells with highly active globin genes.

It is important to note that many (if not all) "C" methodologies (including MC-4C) are unable to measure dynamic interactions that occur through time. Recent preliminary reports supported the existence of these dynamics [51]. There is however, still a great inconsistency between insights generated by proximity ligation methods and microscopy imaging techniques (e.g. fluorescence in situ hybridization (FISH) analyses) and no widely accepted explanation exists in the community [52].

While we showed a clear clustering between β-globin enhancers when contacting β-globin gene, no functional implications could be concluded. In fact, existence of such a promoter-enhancer hub could merely be an alternative pairwise interaction of the β-globin promoter and its enhancers, passively forcing the promoter and enhancer elements to frequently end up in the same nuclear space in the cells with active β-globin gene.

8.2.2. Challenges in the model

Establishing the 3D conformation of the genome, required a precise preparation protocol to form and maintain long concatemers. After nanopore sequencing, a multitude of computational obstacles were needed to be dealt with to reach a robust framework capable of exploring this untouched area of genome organization.

The first challenge rose in converting the raw output of MinION (called *Squiggle*) to corresponding bases through a procedure called *base calling*. Due to rapid developments, new flow cells were quickly depreciating older versions and their corresponding base caller. Additionally, the nanopore community released several new software tools for base calling [53]. It can be expected that each base caller would introduce its own biases [54]. The frequent updates made it however impossible to investigate these subtle biases in MC-4C data or at least identifying the best performing method for base calling. Inevitably, we resorted to use the default software distributed by Oxford Nanopore Technologies, effectively ignoring such technical biases in base calling as is done by many (if not all) other researchers.

A similar challenge was encountered when an appropriate aligner had to be chosen for mapping MC-4C fragments. Through preliminary analysis, we identified BWA-SW to be a suitable candidate for mapping fragments. However, this aligner is designed to map short reads and optimized to deal with biases expected in Illumina sequencing and not for the technical biases that are specifically present in reads sequenced using MinION technology. To address this issue, a variety of aligners have been designed in the recent years to map reads from MinION and thus could potentially perform better than BWA-SW for long reads. However, fragments produced in MC-4C are small which is not often considered in long-read aligners. Consequently, there is a great need for designing aligners that can map small fragments that are affected with high error rate as expected in MinION sequencing.

Sequenced reads in MC-4C contain several fragments that need to be mapped to different locations in the reference genome (known as *split-mapping*). However, standard aligners are not capable of recognizing fragment boundaries within a read and try to map the entire sequence to a single location in the genome. We resolved this by pre-splitting reads into fragments. This strategy is, however, not perfect due to read errors (see section 7.4 and Figure 7.15). It worth noting that the split-mapping is partially supported by aligners such as BWA [55] and Bowtie2 [56]. However, these aligners do not assume multiple fragments to be enclosed in the read and generally split the given read only into two fragments. It is expected that an aligner that supports multi-split-mapping can make a notable contribution to genome conformation research considering the current interest in development of multi-contact methodologies.

Perhaps the greatest challenge on the computational side of MC-4C was to develop a robust statistical model. In many (if not all) 3C based technologies, the chance of capturing a contact between a fragment close to the viewpoint is much higher than a fragment that resides far away from the view point. Unfortunately, the biological elements under investigation (such as enhancers or promoters) are also often close to the view point. Therefore, a robust statistical model is needed to differentiate between a biological interactions and the expected contact of linearly close regions in the genome. A similar (but more complex) bias exists in multi-contact reads. Once a fragment is mapped to a specific region (say A) in the genome, there is a higher chance for other fragments in the read to map in the vicinity of A compared to other locations in the genome. Therefore, in order to segregate biological from random interactions, one needs to compute a null distribution for each fragment representing the expected frequency of observing other nearby fragments. Collectively, these distributions represent the “bendability” of DNA which could be directly used in significance estimation of the multi-contact analysis.

The expected frequency of capturing a neighbor fragment in MC-4C could be exploited to improve mapping efficiency of fragments that show high alignment scores for several locations in the genome. The linear distance of a confidently mapped fragment within a read is a good proxy to identify the correct location for mapping other low quality fragments within that particular read. Caution should be exercised to keep the false discovery of mapped fragments in the vicinity to an acceptable level.

8.2.3. FUTURE OUTLOOK

A genome wide multi-contact view of DNA interactions can boost our understanding of genome architecture and elucidate how individual modules of this organization (e.g. CTCF-CTCF clusters, enhancer-promoter loops, TADs, etc.) work in concert to regulate expression of genes. For example, such an approach is particularly needed to resolve contradicting hypotheses regarding the inter-chromosomal interactions that currently exist in the literature [57]. These debates are specially fueled by FISH experiments that support functional role of interactions between chromosomes [58, 59].

Owing to the similarity of the preparation protocol in many 3C based methodologies, the multi-contact support in MC-4C can be easily extended to its sister approaches. A notable example is Targeted Locus Amplification (TLA) technology which provides read coverage for a targeted region of interest [60]. This method enables robust detection of genetic variations such as single point mutations or structural aberrations [61]. However, as standard TLA uses Illumina sequencing platform which yields short reads, it is still challenging to find allele-specific variations that are required for accurate haplotyping [62]. Nanopore sequencing of TLA products can produce longer reads which in turn increases the chance of identifying variation markers on a single read. Additionally, reads in MC-4C uniquely represent single alleles which allow quantitative assessment of observed haplotype links and their frequencies in the population of cells.

Another interesting application of MC4C is to investigate how the co-localisation of elements changes through time. By performing multiple standard Hi-C through time, Rao *et al.* already showed that interactions are dynamic [51]. However, this approach is again based on measuring pairwise interactions. It is interesting to extend these findings to multi-way interactions by performing MC-4C on the genomic locations that showed high variation through time. Another potential approach to add time dimension to Hi-C data is to progressively cross-link contacting fragments. Inspired by encapsulating chemical drugs in Micelles [63], one can encapsulate and release formaldehyde in the cell nucleus through time [64, 65]. This gradual release immobilizes interactions during a certain period of time which then captures the dynamical interactions within the cell nucleus. Once the cross-linking procedure is finished, the digestion and ligation steps can be followed similar to standard 3C template preparation protocols. The main difficulty in this idea is to identify the time point in which each concatemer is formed after sequencing.

8.3. CONCLUDING REMARKS

Computational methodologies are nowadays a recognized part of genomic research, facilitating knowledge extraction from massive multi-modal datasets [66]. Nonetheless, there is still a vault of treasures in the biological literature that can not be easily exploited by computational methods and has to be handled manually [67]. This is by far the most time-consuming part of computational biology and data science [68]. In Chapter 4 for example, many articles were needed to be read to interpret the top pairs in SyNet making it a tedious and inefficient process. A Persian proverb fits this situation very well: “we are seeking for water in the sea”. We believe that recent initiatives like European Open Science Cloud (EOSC) [69] can fundamentally resolve this issue by promoting the distribution of linked data and semantics where machines can "understand" and query databases, facilitating knowledge discovery in medical and biological research [70, 71].

The pace of development in genomic measurement techniques has never been this high. Within two decades, the genome wide transcriptomic measurements has advanced to the 3rd generation (from microarrays to RNA-seq and now to direct RNA sequencing [72]). While many computational labs (including us) are still investigating the biases and challenges in the first generation techniques, (motivated by competition) sequencing companies are fully focused to the 2nd and 3rd generation sequencing technologies. Furthermore, there is a great motivation for using the latest “hot” technologies in the research community to be able to publish in high impact journals which indirectly devalues older technologies. In this situation, if new technologies keep deprecating older technologies with this momentum, efficiently large expression datasets (in terms of number of samples) for clinical predictions may never be assembled. In parallel, this trend also prevents findings to be reproduced and confirmed by independent groups. Unless a concrete strategy for dealing with this problem is designed, we expect that exploring (technology-specific) biases will get less attention than before in the community. Ignoring the biases in turn reduces robustness of methods that utilize such polluted data. In this situation, the reproducibility in science is going be an even more serious problem soon.

References

[1] M. H. van Vliet, F. Reyal, H. M. Horlings, M. J. van de Vijver, M. J. T. Reinders, and L. F. A. Wessels, *Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability,* BMC Genomics **9**, 375 (2008).

[2] W. E. Johnson, C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical bayes methods,* Biostatistics **8**, 118 (2007).

[3] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solís, R. Duque, H. Bersini, and A. Nowé, *Batch effect removal methods for microarray gene expression data integration: a survey,* Brief. Bioinform. **14**, 469 (2013).

[4] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu, *Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods,* PLoS One **6**, e17238 (2011).

[5] C. Staiger, S. Cadot, B. Györffy, L. F. A. Wessels, and G. W. Klau, *Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis,* Front. Genet. **4**, 289 (2013).

[6] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse, and C. Sotiriou, *Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series,* Clinical Cancer Research **13**, 3207 (2007), http://clincancerres.aacrjournals.org/content/13/11/3207.full.pdf .

[7] S. Loi, B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, A. Harris, J. Bergh, J. A. Foekens, J. G. Klijn, D. Larsimont, M. Buyse, G. Bontempi, M. Delorenzi, M. J. Piccart, and C. Sotiriou, *Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade,* Journal of Clinical Oncology **25**, 1239 (2007), pMID: 17401012, https://doi.org/10.1200/JCO.2006.07.1522 .

[8] S. Gourgou-Bourgade, D. Cameron, P. Poortmans, B. Asselain, D. Azria, F. Cardoso, R. A'Hern, J. Bliss, J. Bogaerts, H. Bonnefoi, E. Brain, M. J. Cardoso, B. Chibaudel, R. Coleman, T. Cufer, L. Dal Lago, F. Dalenc, E. De Azambuja, M. Debled, S. Delaloge, T. Filleron, J. Gligorov, M. Gutowski, W. Jacot, C. Kirkove, G. MacGrogan, S. Michiels, I. Negreiros, B. V. Offersen, F. Penault Llorca, G. Pruneri, H. Roche, N. S. Russell, F. Schmitt, V. Servent, B. Thürlimann, M. Untch, J. A. van der Hage, G. van Tienhoven, H. Wildiers, J. Yarnold, F. Bonnetain, S. Mathoulin-Pélissier, C. Bellera, and T. S. Dabakuyo-Yonli, *Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative (definition for the assessment of time-to-event endpoints in CANcer trials),* Ann. Oncol. **26**, 2505 (2015).

[9] Y. J. Chua, D. Sargent, and D. Cunningham, *Definition of disease-free survival: this is my truth–show me yours,* Annals of Oncology **16**, 1719 (2005).

[10] S. Codish and R. N. Shiffman, *A model of ambiguity and vagueness in clinical practice guideline recommendations,* AMIA Annu Symp Proc **2005**, 146 (2005), amia2005_0146[PII].

[11] P. Peng, R. Wong, and P. Yu, *Learning on probabilistic labels,* in *Proceedings of the 2014 SIAM International Conference on Data Mining*, Proceedings (Society for Industrial and Applied Mathematics, 2014) pp. 307–315.

[12] D. Angluin and P. Laird, *Learning from noisy examples,* Mach. Learn. **2**, 343 (1988).

[13] T. Menzies, E. Kocaguneli, B. Turhan, L. Minku, and F. Peters, *Sharing Data and Models in Software Engineering* (Morgan Kaufmann, 2014).

[14] S. J. Pan and Q. Yang, *A survey on transfer learning,* IEEE Transactions on Knowledge and Data Engineering **22**, 1345 (2010).

[15] S. Thrun and L. Pratt, *Learning to Learn* (Springer Science & Business Media, 2012).

[16] J. Xu, Z. Su, H. Hong, J. Thierry-Mieg, D. Thierry-Mieg, D. P. Kreil, C. E. Mason, W. Tong, and L. Shi, *Cross-platform ultradeep transcriptomic profiling of human reference rna samples by rna-seq,* Scientific data **1**, 140020 (2014).

[17] J. J. Goeman, *L1 penalized estimation in the cox proportional hazards model,* Biom. J. **52**, 70 (2010).

[18] G. Forman and M. Scholz, *Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement,* SIGKDD Explor. Newsl. **12**, 49 (2010).

[19] M. Kanehisa and S. Goto, *Kegg: kyoto encyclopedia of genes and genomes,* Nucleic acids research **28**, 27 (2000).

[20] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, *Network-based classification of breast cancer metastasis,* Molecular systems biology **3**, 140 (2007).

[21] T. K. Ho, *Random decision forests,* in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1 (1995) pp. 278–282 vol.1.

[22] H. M. J. Sontrop, *A Critical Perspective On Microarray Breast Cancer Gene Expression Profiling,* Ph.D. thesis, Delft University of Technology (2015).

[23] A. Taherian-Fard, S. Srihari, and M. A. Ragan, *Breast cancer classification: linking molecular mechanisms to disease prognosis,* Briefings in Bioinformatics **16**, 461 (2015).

[24] *Molecular subtypes of breast cancer,* `http://www.breastcancer.org/symptoms/types/molecular-subtypes` (2018), accessed: 2018-2-9.

[25] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, *A gene-expression signature as a predictor of survival in breast cancer,* New England Journal of Medicine **347**, 1999 (2002).

[26] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark, *A multigene assay to predict recurrence of Tamoxifen-Treated, Node-Negative breast cancer,* N. Engl. J. Med. **351**, 2817 (2004).

[27] C. Fan, D. S. Oh, L. Wessels, B. Weigelt, D. S. A. Nuyten, A. B. Nobel, L. J. van't Veer, and C. M. Perou, *Concordance among gene-expression-based predictors for breast cancer,* N. Engl. J. Med. **355**, 560 (2006).

[28] A. Prat and C. M. Perou, *Deconstructing the molecular portraits of breast cancer,* Mol. Oncol. **5**, 5 (2011).

[29] X. Dai, A. Chen, and Z. Bai, *Integrative investigation on breast cancer in ER, PR and HER2-defined subgroups using mRNA and miRNA expression profiling. sci. rep. 4, 6566,* (2014).

[30] Cancer Genome Atlas Network, *Comprehensive molecular portraits of human breast tumours,* Nature **490**, 61 (2012).

[31] E. Taskesen, S. M. Huisman, A. Mahfouz, J. H. Krijthe, J. De Ridder, A. Van De Stolpe, E. Van Den Akker, W. Verheagh, and M. J. Reinders, *Pan-cancer subtyping in a 2d-map shows substructures that are driven by specific combinations of molecular characteristics,* Scientific reports **6**, 24949 (2016).

[32] Y.-T. Huang, T. J. VanderWeele, and X. Lin, *Joint analysis of snp and gene expression data in genetic association studies of complex diseases,* Ann. Appl. Stat. **8**, 352 (2014).

[33] Q. Xiong, N. Ancona, E. R. Hauser, S. Mukherjee, and T. S. Furey, *Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets,* Genome Research **22**, 386 (2012), http://genome.cshlp.org/content/22/2/386.full.pdf+html .

[34] C. Xu, D. Tao, and C. Xu, *A survey on multi-view learning,* arXiv (2013), arXiv:1304.5634 [cs.LG] .

[35] W. Yang, Y. Gao, Y. Shi, and L. Cao, *Mrm-lasso: A sparse multiview feature selection method via low-rank analysis,* IEEE transactions on neural networks and learning systems **26**, 2801 (2015).

[36] D. Kim, R. Li, A. Lucas, S. S. Verma, S. M. Dudek, and M. D. Ritchie, *Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma,* J. Am. Med. Inform. Assoc. **24**, 577 (2017).

[37] C. X. Ma and M. J. Ellis, *The cancer genome atlas: clinical applications for breast cancer,* Oncology **27**, 1263 (2013).

[38] D. Shyr and Q. Liu, *Next generation sequencing in cancer research and clinical application,* Biological Procedures Online **15**, 4 (2013).

[39] W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, Z. Kan, W. Han, and W.-Y. Park, *Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer,* Nature Communication **8**, 15081 (2017).

[40] M. Mutarelli, L. Cicatiello, L. Ferraro, O. M. Grober, M. Ravo, A. M. Facchiano, C. Angelini, and A. Weisz, *Time-course analysis of genome-wide gene expression data from hormone-responsive human breast cancer cells,* BMC Bioinformatics **9 Suppl 2**, S12 (2008).

[41] L. Ou-Yang, D.-Q. Dai, X.-L. Li, M. Wu, X.-F. Zhang, and P. Yang, *Detecting temporal protein complexes from dynamic protein-protein interaction networks,* BMC Bioinformatics **15**, 335 (2014).

[42] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering, *STRING v10: protein–protein interaction networks, integrated over the tree of life,* Nucleic Acids Res. **43**, D447 (2015).

[43] J. J. Keats, G. Speyer, A. Christofferson, C. Legendre, J. Aldrich, M. Russell, L. Cuyugan, J. Adkins, A. Blanski, M. Hodges, D. Rohrer, S. Jagannath, R. Vij, G. Orloff, T. Zimmerman, R. Niesvizky, D. Liles, J. W. Fay, J. L. Wolf, R. M. Rifkin, N. C. Gutierrez, M. CoMMpass Network, J. Yesil, M. Derome, S. Kim, W. Liang, P. G. Kidd, S. Jewell, J. D. Carpten, D. Auclair, and S. Lonial, *Molecular predictors of outcome and drug response in multiple myeloma: An interim analysis of the mmrf commpass study,* Blood **128**, 194 (2016), http://www.bloodjournal.org/content .

[44] J. Ubels, E. H. van Beers, A. Broijl, P. Sonneveld, M. H. van Vliet, and J. de Ridder, *TOPSPIN: a novel algorithm to predict treatment specific survival in cancer,* in *HAEMATOLOGICA*, Vol. 102 (2017) pp. 524–525.

[45] D. J. Toft and V. L. Cryns, *Minireview: Basal-like breast cancer: from molecular profiles to targeted therapies,* Mol. Endocrinol. **25**, 199 (2011).

[46] A. Byrne, A. E. Beaudin, H. E. Olsen, M. Jain, C. Cole, T. Palmer, R. M. DuBois, E. C. Forsberg, M. Akeson, and C. Vollmers, *Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells,* Nat. Commun. **8**, 16027 (2017).

[47] D. R. Garalde, E. A. Snell, D. Jachimowicz, A. J. Heron, M. Bruce, J. Lloyd, A. Warland, N. Pantic, T. Admassu, J. Ciccone, and Others, *Highly parallel direct RNA sequencing on an array of nanopores. biorxiv 068809,* (2016).

[48] S. Oikonomopoulos, Y. C. Wang, H. Djambazian, D. Badescu, and J. Ragoussis, *Benchmarking of the oxford nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations,* Sci. Rep. **6**, 31602 (2016).

[49] L. Giorgetti, N. Servant, and E. Heard, *Changes in the organization of the genome during the mammalian cell cycle,* Genome Biol. **14**, 142 (2013).

[50] P. Olivares-Chauvet, Z. Mukamel, A. Lifshitz, O. Schwartzman, N. O. Elkayam, Y. Lubling, G. Deikus, R. P. Sebra, and A. Tanay, *Capturing pairwise and multi-way chromosomal conformations using chromosomal walks,* Nature **540**, 296 (2016).

[51] S. S. P. Rao, S.-C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, X. Huang, M. S. Shamim, J. Shin, D. Turner, Z. Ye, A. D. Omer, J. T. Robinson, T. Schlick, B. E. Bernstein, R. Casellas, E. S. Lander, and E. L. Aiden, *Cohesin loss eliminates all loop domains,* Cell **171**, 305 (2017).

[52] S. A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. Lai, A. Shishkin, P. Bhat, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, P. McDonel, M. Garber, and M. Guttman, *Higher-order inter-chromosomal hubs shape 3-dimensional genome organization in the nucleus,* bioRxiv (2017), 10.1101/219683, https://www.biorxiv.org/content/early/2017/11/18/219683.full.pdf .

[53] M. Ratković, *Deep Learning Model for Base Calling of MinION Nanopore Reads,* Ph.D. thesis, Fakultet Elektrotehnike i Računarstva, Sveučilište u Zagrebu (2017).

[54] R. Krishnakumar, A. Sinha, S. W. Bird, H. Jayamohan, H. S. Edwards, J. S. Schoeniger, K. D. Patel, S. S. Branda, and M. S. Bartsch, *Systematic and stochastic influences on the performance of the minion nanopore sequencer across a range of nucleotide bias,* Scientific reports **8**, 3159 (2018).

[55] H. Li and R. Durbin, *Fast and accurate long-read alignment with burrows–wheeler transform,* Bioinformatics **26**, 589 (2010).

[56] B. Langmead and S. L. Salzberg, *Fast gapped-read alignment with bowtie 2,* Nature methods **9**, 357 (2012).

[57] J. Dekker and T. Misteli, *Long-range chromatin interactions,* Cold Spring Harbor Perspectives in Biology **7** (2015), 10.1101/cshperspect.a019356, http://cshperspectives.cshlp.org/content/7/10/a019356.full.pdf+html .

[58] Z. Wei, D. Huang, F. Gao, W.-H. Chang, W. An, G. A. Coetzee, K. Wang, and W. Lu, *Biological implications and regulatory mechanisms of long-range chromosomal interactions,* Journal of Biological Chemistry **288**, 22369 (2013), http://www.jbc.org/content/288/31/22369.full.pdf+html .

[59] D. Noordermeer, E. De Wit, P. Klous, H. Van De Werken, M. Simonis, M. Lopez-Jones, B. Eussen, A. De Klein, R. H. Singer, and W. De Laat, *Variegated gene expression caused by cell-specific long-range dna interactions,* Nature cell biology **13**, 944 (2011).

[60] P. J. P. de Vree, E. de Wit, M. Yilmaz, M. van de Heijning, P. Klous, M. J. A. M. Verstegen, Y. Wan, H. Teunissen, P. H. L. Krijger, G. Geeven, P. P. Eijk, D. Sie, B. Ylstra,

L. O. M. Hulsman, M. F. van Dooren, L. J. C. M. van Zutven, A. van den Ouweland, S. Verbeek, K. W. van Dijk, M. Cornelissen, A. T. Das, B. Berkhout, B. Sikkema-Raddatz, E. van den Berg, P. van der Vlies, D. Weening, J. T. den Dunnen, M. Matusiak, M. Lamkanfi, M. J. L. Ligtenberg, P. ter Brugge, J. Jonkers, J. A. Foekens, J. W. Martens, R. van der Luijt, H. K. P. van Amstel, M. van Min, E. Splinter, and W. de Laat, *Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping,* Nature biotechnology **32**, 1019 (2014).

[61] Q. P. Hottentot, M. van Min, E. Splinter, and S. J. White, *Targeted locus amplification and next-generation sequencing,* in *Genotyping* (Springer, 2017) pp. 185–196.

[62] C. Vermeulen, G. Geeven, E. de Wit, M. J. Verstegen, R. P. Jansen, M. van Kranenburg, E. de Bruijn, S. L. Pulit, E. Kruisselbrink, Z. Shahsavari, *et al.*, *Sensitive monogenic noninvasive prenatal diagnosis by targeted haplotyping,* The American Journal of Human Genetics **101**, 326 (2017).

[63] M. Ugarenko, C.-K. Chan, A. Nudelman, A. Rephaeli, S. M. Cutts, and D. R. Phillips, *Development of pluronic micelle-encapsulated doxorubicin and formaldehyde-releasing prodrugs for localized anticancer chemotherapy,* Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics **17**, 283 (2009).

[64] M.-A. Flyvholm, *Formaldehyde and formaldehyde releasers,* in *Handbook of Occupational Dermatology*, edited by L. Kanerva, J. E. Wahlberg, P. Elsner, and H. I. Maibach (Springer Berlin Heidelberg, Berlin, Heidelberg, 2000) pp. 474–478.

[65] D. G. Anton, W. I. R., F. Mari-Ann, L. Gerda, and C. Pieter-Jan, *Formaldehyde-releasers in cosmetics: relationship to formaldehyde contact allergy,* Contact Dermatitis **62**, 18 (2009).

[66] V. Marx, *Biology: The big challenges of big data,* (2013).

[67] M. Saleem, M. R. Kamdar, A. Iqbal, S. Sampath, H. F. Deus, and A.-C. N. Ngomo, *Big linked cancer data: Integrating linked tcga and pubmed,* Web Semantics: Science, Services and Agents on the World Wide Web **27-28**, 34 (2014), semantic Web Challenge 2013.

[68] Gil Press, *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says,* (2016).

[69] *The european open science cloud for research pilot project,* (EOSC2018).

[70] H. Chen, T. Yu, and J. Y. Chen, *Semantic web meets integrative biology: a survey,* Briefings in Bioinformatics **14**, 109 (2013).

[71] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson, *Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud,* Information Services & Use **37**, 49 (2017).

[72] D. R. Garalde, E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, M. Bruce, N. Pantic, T. Admassu, P. James, A. Warland, *et al.*, *Highly parallel direct rna sequencing on an array of nanopores,* Nature methods (2018).

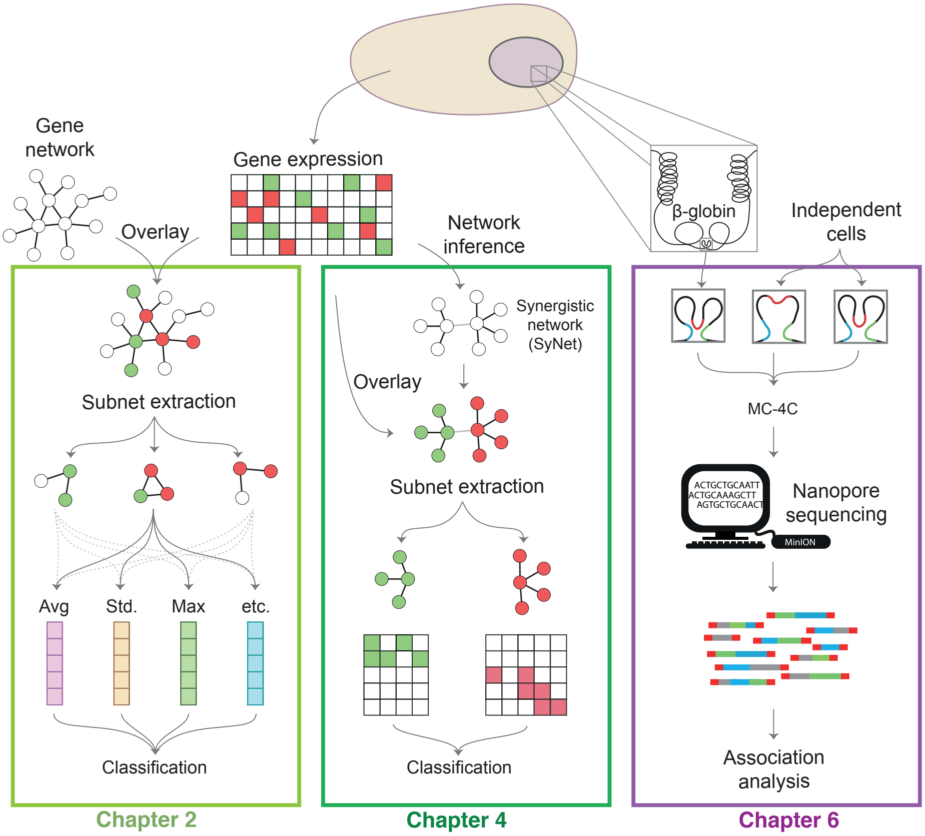
SUMMARY

In the last two decades, our understanding of the molecular mechanisms within the cell has witnessed a great leap forward. For the most part this is due to the fast innovation of the genomic measurements technologies and widespread usage of computational methods which enables knowledge extraction from the massive datasets produced by these measurements. A notable example of a field that has substantially benefitted from this progress is cancer patient outcome prediction, in which the aim is to predict patient prognosis from common clinical variables such as tumor size, age or histological parameters. With the application of machine learning methods to gene expression profiles of the tumor a major improvement of the prediction accuracy could be realized. These models are later succeeded by Network based Outcome Predictors (NOP) that consider the cellular wiring diagram of cell in the model to identify stable and relevant markers that can accurately estimate outcome of patients. Problematically, after a decade of research in this area, NOPs did not find extensive application compared to the classical models due to contradicting reports regarding their performance, stability and relevance of markers in the literature.

In this thesis, we introduce a new NOP - called FERAL - that alleviates several fundamental issues in state-of-the-art NOPs which prevented these models to reach the optimal prediction performance, stability and marker relevance. We furthermore demonstrate that generic biological networks do not contain sufficiently informative interactions to truly aid NOP. We therefore infer a phenotype-specific network called SyNet which connects pairs of genes that together achieve patient outcome prediction performance beyond what is attainable by individually genes. We show that a NOP that use identical gene expression datasets, yields superior performance merely by considering groups of genes suggested by SyNet. We, moreover, show that model performance is severely reduced if nodes in SyNet are shuffled, which confirms that also the links in SyNet are relevant to outcome prediction.

An important limitation of current biological networks is that they are restricted to pairwise interactions. We show that higher order interactions between functional elements in the cell are relevant in outcome prediction. We later introduce a novel genomics method called Multi-Contact 4C (MC-4C) to measure and investigate multi-way interactions between functional elements. In contrast to existing methods, MC-4C exploits long-read 3rd generation sequencing technologies and detects higher order interactions that occur in a region of interest at the level of a single allele. We further devise a well-founded statistical model that is required for significance estimation of observed interactions. Using MC-4C, we experimentally confirm a 26 years old hypothesis regarding the looping and co-localization of enhancers in the β -globin region in the mouse genome. Additionally, we provide the first experimental explanation for the "vermicelli" phenomenon that was observed through microscopic inspection of cells depleted of WAPL (the element responsible for unwinding of loops in mammalian cells). Therefore,

targeted multi-way conformation analysis methods like MC-4C promise to uncover how the multitude of regulatory sequences and genes coordinate their activity in the spatial context of the genome.



Schematic overview of chapters in the thesis

ACKNOWLEDGEMENTS

Jeroen, my first and still closest friend in the Netherlands. This is just a small text to sincerely appreciate your long-standing support during the journey of my PhD. I am grateful for every (nasty or otherwise positive :D) comments/feedbacks/remarks you shared with me. I assure you that every bit is used as a brick to build up a new/better Amin. I am glad that I still have my scientific (as well as non-scientific) connection with you. It would be a shame to let go of such a great source of inspiration which I extensively benefited from during my PhD. Apart from a notable scientific growth, this period substantially changed me personally. I never thought it could. For both aspects, thank you! You were used to mock me for saying "lets do X and Y and then problem solved!". Interestingly, I noticed that you sometimes use this sentence nowadays. Its now my turn to mock you! I am glad that I could have a minor contribution in your already extensive vocabulary. Please stay like this man, you have been "the best supervisor" ever!

Marcel, frankly, I would be lost without your guidance, the most effective supervision I ever got. Your remarks/comments were always short and snappy but right to the point and (embarrassingly enough for me) correct! It was often terrifying to prepare a presentation for you, knowing that you will find the problem in my idea in the second slide. Thank you for your contributions during my PhD, they were a necessary light in my adventure.

Wouter (de Laat), it is often difficult to talk about scientists with great contributions. It is even more difficult when they are also considered as "awesome" in their personalities. Thank you for letting me contribute to MC-4C. It was a challenge to learn biologists language which I really missed in my Bioinformatics toolbox. I don't know what future holds for me, but I am sure I continue to love my contributions to your lab. Great people, great environment, thank you.

Lodewyk, thank you for being an inspiring example of an excellent scientist. Through many interactions with you, I could see how a research work should be presented, how constructive critics should be formulated, and even how a scientific network should be maintained and used to evolve my research career. You were truly a galvanizing figure in the past five years.

Thomas, I was always amazed how you could contribute to every topic that was being discussed (scientific or otherwise life related). Similarly, when talking about my own project, it was always fun/inspiring to hear your comments as they were coming from a new angle that I never thought about.

David and **Marco**, you are among a handful of group leaders who know the secrets of combining exceptional science and endless fun/joy. I can not remember any of our discussions which was not accommodated with frequent jokes and laughs. At the same time, thank you for being the "go-to" dictionary in pattern recognition for me. I never left your office empty handed.

I should thank all of my colleagues in TU Delft including **Marc**, **Joana**, **Erik**, **Sepideh**, **Ahmed**, **Wouter (Kouw)**, **Marcel (van den Broek)**, **Jasper**, **Christian (Groß)**, **Mamun**, **Sjoerd**, **Christine**, **Arlin**, **Stavros**, **Alex**, **Tamim**, **Ekin**, **Laura**, **Taygun** and **Gorkem**. It was shame that I had to leave Delft. I hope I get to work with you soon. **Ahmed**, it was often terrifying for me to think about the challenge of keeping a balance between work and life if I choose a future career in science, specially for an expat. Your astounding success in this regard was on the one hand soothing and on the other hand envying. Additionally, you were able to beautifully cultivate a fruitful scientific network around yourself to prove me that its possible for an expat to reach there. Furthermore, I appreciate every scientific discussions we had during the past five years. Thank you! **Joana**, it is encouraging to see how you passed many obstacles in the (currently imbalanced) academic environment and quickly climbing up the scientific ladder. With such a dedication from you and your peers, I am very positive that these issues will soon disappear. I should also thank you for your caring attitude toward your colleagues. At times, this was missing in our environment. Thank you for adding it back. **Marc**, sometimes during our countless discussions, I felt that my brain is just too slow to process all the information you are sharing with me. At this point, I am quite sure your brain is cheating somehow. Or at least there is a secret auxiliary brain you have somewhere, remotely "ssh"ing to you to acquire/analyze/interpret on your behalf (!). Thank you for opening for me a door to a whole new way of thinking about scientific problems we discussed and also providing an amazing source of insights which I frequently profited from. **Erik**, you were one of the few colleagues who spent a considerable amount of time on the outcome prediction models. Thank you for sharing many experiences and insights regarding these models with me. I hope you keep using your (awesome) green bag. I loved that iconic color! Good luck in Leiden, although you don't need one. **Sjoerd**, at any point when I did not understand some statistical model, you were always there to help me out. Thank you for being an awesome colleague. And thank you for reminding me that there is a life beside PhD. Looking at you keeping yourself busy with music and site-seeing was definitely a poke in the ribs to wake me up and stopping me from becoming a single dimensional graduate. I am grateful for this. **Wouter (Kouw)**, mate I can finally tell you this. Thank you for making my PhD an envious experience. There wasn't even a single occasion when I was brainstorming with you and I could feel that I am contributing something, or think to myself "oh, I am as good as Wouter in this". Zero, zilch, zip, nada, none! In every meeting we had, I had to sit and watch you writing a complex set of formulas and cost functions and later try to make an "I-understand" face when you were calculating the corresponding derivatives. Meanwhile I had no clue what you are doing! I am sure you are going to amaze people in Copenhagen. I hope to have another opportunity to work with you. Undoubtedly, its going to be another fun project. **Marcel (van den Broek)**, I always loved the early morning conversations we had on Tuesdays. During my PhD, you were the only colleague who I could spend several hours seriously discussing a non-scientific and life/society related topic. Thank you for sharing your thoughts with me. Those deep thoughts coming from an individual who was born and raised in the Netherlands culture helped me to open my mind. With that, I could take a new look at many problems in this planet which bugged for me years. I am grateful, I learned a lot from you. I also remember the Tony's Chocolonely story and the chocolate bar you brought for me. These chocolate bars are forever a reminder of you. **Sepideh**, having a colleague from your own land was a blessing. Thank you for sharing everything you knew about the Biology as well as the Netherlands and sorry that I knew nothing! It is a shame that I missed the opportunity to work on a Hi-C project in collaboration with you. I think your idea in Hi-C matrix normalization was a powerful yet simple approach to beautifully take care of the systematic bias in these matrices. Well done! **Jasper**, thank you for sharing your experience and expertise regarding long-read sequencing and its corresponding analysis. Your critical view on this newly born technology and your substantial enthusiasm to resolve the issues around it (even the smallest ones) were always notable to me. Please stay like this. **Christian (Groß)**, I am glad that I got to know you man. The short coffee breaks to share our frustration about a particular topic/problem was a fail safe for me. I hope it was the same for you. I am quite sure you will do well in your PhD. A friendly advice, don't worry too much, all these hassles will be history soon. While you forget most of them, the honorable outcome will stay with you forever! **Alexey**, I appreciate every minute you spent to teach me how to handle highly parallel jobs in a complex grid infrastructure. Additionally, I loved all non-scientific but heated discussions we had during coffee breaks. I learn so much about the world I live in through those arguments. Please enjoy your time at Google. You worked so much to reach there! **Mamun**, you are just too good with toolboxes and data analysis man. I don't get how you were doing it, but somehow you were able to build a complete pipeline in half a day and make me hate myself. Thanks for that! **Saskia**, **Bart**, **Ruud** thank you for being a "true" support staff. A sense of humor was all I needed at moments where something broke down or I frantically needed something. **Robbert**, sorry for every ill-submitted jobs to INSY cluster :). Just to be sure, we are clear that those mistakes were not intentional, right? Maybe those angry emails were necessary to make sure we are taking our utmost caution with what we are doing.

To all my colleagues in Jeroen's lab, **Sara**, **Joep**, **Roy**, **Joske**, **Marleen**, **Alexandra**, **Joanna (von Berg)**, **Joanna (Wolthuis)**, **Buys**, **Tilman**, **Luca**, it was (and still is) an honor to work with you. You are simply the best colleague that someone can have (I am serious about this!). **Sara**, I can not enumerate many bright scientists who are also humble in their interactions with their colleagues. A quick-witted and caring supervisor is the best supervisor, you are (and continue to be) a colleague to be inspired from for the rest of my career in science. Lets be honest here, I am also jealous. I don't know if I should thank you about that :D **Joep**, there are only few people in my network who I am terrified of having a (non-)scientific argument with. As soon as I idiotically start one with you, I know the outcome. So I am not going to appreciate that here! Meanwhile, you are exceptionally good at science and more importantly in the collaborations underneath it. I am still trying to learn how to do that from you. Please stay around. **Joske**, considering the observed capabilities/talents you showed so far, I impatiently look forward to the impressive PhD thesis you are going to have. It was fascinating to see how you started from a purely biological background and quickly became an expert in machine learning. Thank you for being an invaluable colleague during my PhD journey. I learned many basics of biology from you and you were always there to help. Thank you. At the same time, any expat needs a close friend to help him/her survive in desperate moments. Thank you for being open, caring and willing to help at all times. Finally, I loved our collaboration in SyNet which was undoubtedly impossible to finish without your contributions. This is the most precious part of my PhD as every brick of it is laid out by us. **Roy**, I appreciate your help regarding MC-4C. Without you, MC-4C would never find applicability in a large-scale applications.

To all my colleagues at UMC, **Chris**, **Alessio**, **Christina**, **Mircea**, **Jose**, **Glen**, **Ivo**, **Mark (van Roosmalen)**, I can not thank you enough. I said that before but I really felt home there. This level of hospitality and openness is not common and could not be realized without you! **Glen** and **Chris**, thank you for being a caring colleague during my time in UMC. Whenever I encountered a difficulty in the Dutch bureaucracy, I knew I can count on you to help me out. **Alessio**, man your impressive talent in combining wet and dry lab was always fascinating to me. I hope I can find a project with which I get the opportunity to work with you. **Mircea**, you proved me that it is possible to have a life while doing a PhD. You made this challenging process (at least for me) enjoyable in a level that I have never imagined to be feasible. I am 100% sure that I cant do it like you did and for that I envy your life style. I hope at some point I learn to live like you. Well done man. **Ivo**, **Mark (van Roosmalen)**, thank you for helping out with MC-4C and nowadays MC-HC. I knew nothing about 3rd generation sequencing and without you it was impossible to reach here.

To biomedical genomics group (de Laat's lab) in Hubrecht Institute, **Marjon**, **Peter**, **Geert**, **Mark (Pieterse)**, **Valerio**, **Christian (Valdes)**, **Angelica**, **Erica**, **Floor**, **Carien**, **Milan**, **Niels**, the impressive achievements in this group stems from an exceptional talents that are gathered together combined with a friendly (and welcoming) environment that fuels its inter (and intra) group collaborations. Thanks to every one of you for having me and letting me enjoy working with you. In MC-4C work, I am specially grateful to **Britta**, **Carlo** for sharing their biological expertise and knowledge with me which helped me to learn the basics of genome conformation techniques in the beginning of our collaboration. Also I must express my gratitude to **Marjon** and **Mark (Pieterse)** for being an amazing support staff/technician during our many projects including MC-4C and now MC-HiC and 4C translocation detection. Your impressive skills in genomic research and laboratory techniques were key ingredients for the success of these projects. Furthermore, I appreciate every tips and tricks shared by **Geert** and **Valerio** during my involvements with 3D genome interaction projects and the related computational analysis for either 4C, Hi-C or their multi-contact extensions.

Throughout this dissertation, I babbled about science (as I am supposed to!). Yet, I must not (and will not) forget that my research career has became only possible through continuous support of my family and their colossal sacrifices. This has specially reinforced tremendously during my stay in the Netherlands. Therefore, I feel obliged to heartedly express my gratitude toward these folks. First and foremost my beautiful and caring wife, **Maryam**. My darling, it has been and will be impossible to express in words how grateful I am in every day of my life for every bit of love you selflessly shared with me. I am utterly confident that nobody on this planet could have filled up my heart the way you did. Your absolute devotion to our small family is the primary reason for its evident prosperity. This assures me a bright future ahead and I am impatiently looking forward to it. Let me also apologize for every morning that I left early as well as every late evening that I came back home. I am deeply sorry for every one of them! Honestly, it was quite embarrassing to see that in return how easy you could ignore those days and kept fueling me with your passion during my PhD journey. I am proud to be your partner and will do my best to be "the one" for you for the rest of my life. Meanwhile, I am thankful to my larger family. My father **Hossein**, from whom I learned how to be a man. I am always in your debt dad. Thank you for everything you did for me. You are unequivocally the best dad in the world. My mother **Zahra**, who showed me how far should I go, if I truly care about somebody. You are the whitest color in my universe mom, thank you! And last but not least, my sister **Elham**. From whom, I learned countless life-saving lessons, scientific or otherwise. Thank you for taking care of mom and dad while I was away. I know how difficult this responsibility has been and I am sorry that you were single-handed along the way. My dearest family, you all are indisputably the primal root of my success and nowadays the only incentive for my persistence and endeavor.

Taken together, I am so lucky to have you all as my family/friend/colleague and I will do my best to keep myself connected to this amazing network of clever, caring and awesome people. Admittedly, without you comrades, I could not be in this position ...

Amin Allahyar
Utrecht, Oct 2018

CURRICULUM VITÆ

Amin ALLAHYAR

14-02-1987 Born in Shiraz, Iran.

EDUCATION

2004–2008 Bachelor of Science in Software Engineering
Bahonar University of Shiraz
Shiraz, Iran

2010–2012 Master of Science in Artificial Intelligence
Ferdowsi University of Mashhad
Mashhad, Iran
Thesis: Online semi-supervised learning by self-organizing maps
Promotor: Prof. dr. ir. H. Sadoghi

2013 PhD. Bioinformatics
Delft University of Technology
Delft, The Netherlands
Thesis: Molecular interactomes: network-guided cancer prognosis prediction & multi-way chromatin interaction analysis
Promotor: Prof. dr. ir. M.J.T Reinders

AWARDS

2016 Best poster award: BioSB 2016

2012 Cum laude (MSc degree)

2009 Rank 1 in Fars province Informatics Olympiad

2004 Rank 4 in National University Entrance Exam

List of Publications

Thesis research:

- **Allahyar, A.** [1], Vermeulen, C. [1], Bouwman, B., Krijger, P., Verstegen, M., Geeven, G., ... & Teunissen, H. (2017). Locus-Specific Enhancer Hubs And Architectural Loop Collisions Uncovered From Single Allele DNA Topologies. bioRxiv, https://doi.org/10.1101/206094.

- **Allahyar, A.**, Ubels, J. , de Ridder, J. (2018). A data-driven interactome of synergistic genes improves network based cancer outcome prediction. bioRxiv: https://doi.org/10.1101/349688

- **Allahyar, A.**, & de Ridder, J. (2015). FERAL: network-based classifier with application to breast cancer outcome prediction. Bioinformatics, 31(12), i311-i319.

Auxiliary research:

- Ranzani, M., Alifrangis, C., Thompson, N., Rust, A., **Allahyar, A.**; Lyer, V., Price, S., Ellis, P., Turner, G., De Ridder, J., McDermott, U., Adams, D. (2018). A lentiviral vector-based insertional mutagenesis screen identifies mechanisms of resistance to MAPK inhibitors in melanoma. Accepted in Pigment Cell & Melanoma Research: 10.1111/pcmr.12737.

- Gilroy, K. L., Terry, A., Naseer, A., de Ridder, J., **Allahyar, A.**, Wang, W., Carpenter, E., Mason, A, Wong, G., Cameron E., Kilbey, A., Neil, J. (2016). Gamma-retrovirus integration marks cell type-specific cancer genes: a novel profiling tool in cancer genomics. PloS one, 11(4), e0154070.

- **Allahyar, A.**, Sadoghi Yazdi, H., & Harati, A. (2015). Constrained Semi-supervised Growing Self-organizing Map. Neurocomputing, 147, 456-471.

- **Allahyar, A.**, & Sadoghi Yazdi, H. (2014). Online Discriminative Component Analysis Feature Extraction From Stream Data With Domain Knowledge. Intelligent Data Analysis, 18(5), 927-951.

- **Allahyar, A.**, Sadoghi Yazdi, H., Mansori, F. (2013). Semi-Supervised Content Based Music Recommendation. Proceedings of the 11th Iranian Conference on Intelligent Systems (ICIS2013), February 2013.

- **Allahyar, A.**, Sadoghi Yazdi, H., Toosi, A., Kahani, M.. (2012). English-Persian Sentence Aligning, A Semi-Supervised Growing Self Organizing Approach. Proceedings of the International IEEE Conference on Persian Language Processing, July 2012.

- **Allahyar, A.**, Sadoghi Yazdi, H. (2012). Eigenvector Selection in Spectral Clustering Using Data Boundary Separation. Proceedings of the 4th Conference on Information and Knowledge Technology (IKT2012), Babol Noshirvani University of Technology, May 2012. (In Persian).

[1] Contributed equally

Propositions

accompanying the dissertation

MOLECULAR INTERACTOMES

NETWORK-GUIDED CANCER PROGNOSIS PREDICTION & MULTI-WAY CHROMATIN INTERACTION ANALYSIS

by

Amin ALLAHYAR

1. **Expression averaging to represent the expression of a set of genes in outcome prediction removes most (if not all) predictive power (chapter 2).**
2. **Network-based outcome predictors guided by a generic interaction network can never outperform standard classifiers (chapter 4).**
3. **Unless effective batch effect removal methods are developed, clinical application of outcome predictors is meaningless (chapter 2 & 4).**
4. **Significant progress in chromatin conformation research is heavily hampered by excessive competition (chapter 6).**
5. **The possibility of "fooling" deep learning models necessitates the introduction of an "unknown" class with several order of magnitude more samples.**
6. **Having a supplementary section called "attempts that did not work" does not impose redundancy to a scientific paper.**
7. **Scientific funds should be allocated to ideas chosen through a "Reddit/Wikipedia" like server where scientists can comment, contribute or vote to submitted proposals.**
8. **Considering the continuous increase in treatment cost, soon it will be cheaper to clone yourself and start over after being diagnosed with a serious disease.**
9. **Surprisingly, political practices do not require an ethics committee.**
10. **A hierarchical voting system can solve the problem of uninformed people electing unfit candidates.**

These propositions are regarded as opposable and defendable, and have been approved as such by the promoter prof. dr. ir. M.J.T Reinders.