Characterization of complex genetic interactions in cancer cells

Minor Research Project Report

Jannes Kromhout

Daily supervisor: Aram Nikolai Andersen

Enserink lab, Department of Molecular Cell Biology, Oslo University Hospital

Supervisor host institute: Prof. Jorrit Enserink

Examiner UMC Utrecht: Dr. Jeroen de Ridder

Table of contents

Abstract

Cancer dependencies are genes required for proliferation and survival in cancer cells, making them potential therapeutic targets. Cancer dependencies are often selective for subtypes of cancers, and are measured as conditional changes in fitness caused by genomic and molecular aberrations and are known as genetic interactions (GIs). However, cancer dependencies can be influenced by several factors, resulting in higher-order GIs, which are difficult to predict for any given cancer type. This has resulted in a varied success rate for the development and application of new targeted therapies. In order to systematically identify cancer dependencies and GIs, genome-wide CRISPR/Cas9 knock-out (KO) screens across pan-cancer libraries have been performed. Due to the limitations of large-scale screenings with multiple KOs, advanced computational strategies have to be used for inferring higher-order interactions and predicting cancer dependencies based on the molecular characteristics of the cancer cells.

In this study, we show a robust method to infer GIs from pan-cancer CRISPR screens based on the genetic and transcriptional background of the cancer cells. Pairwise GIs were inferred by predicting fitness change from CRISPR/Cas9-mediated gene deletion using multivariate penalised linear regression, combined with null-hypothesis testing. Furthermore, we developed an XGBoost approach, where regression tree structures were mined for variable interactions, to discover potential complex higher-order interactions from transcriptional changes between cancer cells. Novel GIs were subsequently mapped and analysed in a genome-wide GI network.

In conclusion, our study shows a robust framework for predicting complex GIs involved in the regulation of cancer fitness.

Layman's summary

Cancer is a complex genetic disease and extensive research has been done to find novel treatments. Despite the amount of work and money put into this research, a lot of promising potential treatments fail to become an approved drug. Due to high genomic and molecular variability between cancers, research is shifting towards personalised medicine. One approach for a more personalised treatment is to identify the so-called cancer dependencies. Cancer dependencies are genes that are required for proliferation and survival in cancer cells. These dependencies are often selective across subtypes of cancers.

To identify cancer dependencies, a so-called CRISPR/Cas9 screen can be used. A CRISPR screen investigates the effect of the loss of a single gene on the fitness (survival) of a specific cancer. To systematically study this, a CRISPR screen can be used that will consider the whole genome, by subsequentially knocking out every single gene in a specific cancer. This has been performed over a broad range of cancer cell types. Although these screens are powerful tools to find cancer dependencies and genetic interactions, it won't show complex interactions between the genes. Therefore, computational approaches have to be used to identify complex genetic interactions involved in cancer fitness.

In this study, we used statistical and computational methods to infer genetic interactions across a broad range of cancer cell types. We obtained data from a public database that contained results of a whole-genome CRISPR screen of many cancer cell lines. In addition, the molecular characteristics of these cancer cell lines, including gene expression, mutations and gene copy numbers were available. Using all data we inferred genetic interactions with a technique called multivariate penalised linear regression. The identified genetic interactions appeared to be highly enriched for functional relations, indicating that our approach is useful for predicting genetic interactions from molecular and genomic data of cancer cells.

The second part of this study was to predict second-order genetic interactions. For this approach we used a machine learning technique which uses decisions trees for its prediction (XGBoost). The XGBoost algorithm assumes a hierarchical structure between the variables (genes), making it possible to extract variables as an pairwise interaction from the trees. The interactions obtained from the XGBoost trees are second-order interactions. The identified interactions appeared to be also highly enriched for functional relations. We selected the top hits of the second-order interactions for characterization and evaluation.

In conclusion, our study shows a robust framework for predicting complex GIs involved in the regulation of cancer fitness.

Acknowledgements

First of all, I want to thank Aram for the guidance and supervision throughout the internship. It has been such a pleasure working with you and you truly are a great supervisor. Not only did I learn a lot from you academically, but you were also a personal mentor for me. I really appreciated that. Secondly, I want to thank Jorrit for giving me the opportunity to do this internship in his lab and give me the experience to work in such a pleasant working environment. I would like to thank Jeroen for being my examiner from Utrecht. Next, Aram, Nathalia, Andrea, Eirik and Alberto I enjoyed the interesting discussions we had during our bioinformatics journal clubs, I learned a lot from you. Eirik, Alberto and Arno, thank you guys for the wonderful time in and outside the lab. In addition, I want to thank the rest of Jorrit's group and the 6th floor for the pleasant time in the lab. Feyza it was nice to have you as an office buddy. Finally, I want to thank my Norwegian 'family'. I am really grateful for how you incorporated me into your family. My time in Oslo would not have been the same without you. So thank you, Kristian, Ingrid, Heidi, Hanne, Solveig, Sigur, Nicolai, Peter, Sverre and a very special thank you for Hilde.

Introduction

Cancer is a complex disease caused by genomic instability[1]. Despite the extensive research put into the development of cancer drugs, the success rate of cancer drugs being approved for therapy remains low (~3%) [2]. Due to high genetic variability within and between patient groups harbouring different types of cancer, treatment is shifting towards personalized medicine. In the field of personalized cancer medicine, the (molecular) characteristics of the tumour are being identified and exploited for treatment. The profiling of the tumour focuses on alterations that drive tumour progression, which covers a broad spectrum of aberrations. Among these characteristics are mutations, epigenetic and proteomic markers, transcriptional profiles, and drug sensitivity. Treatment is based on these tumour characteristics, which results in a targeted approach with less side effects. In recent years, the use of genomics in cancer treatment has taken a more significant role[3,4]. The research for novel treatments is heading towards learning about the complexity of tumours and the predicted response to a given treatment option[3,5,6]. Genes that are selectively required for cancer cell proliferation and survival, are so called cancer dependencies, which makes them potential therapeutic targets[7]. Cancer dependencies are often selective for subtypes of cancers, and are measured as conditional changes in fitness caused by genomic and molecular aberrations and are known as genetic interactions (GIs).

A GI between two genes occurs when the fitness consequence of a double loss-of-function (LOF) mutant is different from the expected fitness based on the single LOF mutants alone. If the fitness of the double mutant is lower than expected, it is called a negative GI. An extreme case of a negative GI is called synthetic lethality, in which a double mutant is not viable, although the single mutants are. In case two mutations lead to a higher fitness than what is expected, it is considered a positive GI[8,9]. The concept of synthetic lethality has been around for a long time, and was described already one century ago[10,11]. However, it was not till 1997 when Hartwell and his colleagues proposed to use synthetic lethality for designing cancer drug treatments[12]. This research led to novel genetic-drug interactions, which are currently used in cancer treatment. The use of synthetic lethality in cancer treatment has been extensively and excellently reviewed[13,14]. The most well-studied case of this concept which has successfully been implemented in the clinic is for breast cancer patients with a LOF BRCA1 or BRCA2 mutation. Breast cancer tumours with this mutation are susceptible to Poly (ADP-ribose) polymerase (PARP) inhibitors. BRCA1 and BRCA2 are both tumour suppressor genes involved in the repair of double stranded breaks (DSBs) of the DNA[15]. PARP1 and PARP2 are enzymes that activate the DNA-damage response after sensing DNA damage[16]. The loss of BRCA1/2 results in a dependency of PARP1/2 for the DNA damage response. Healthy cells which do not harbour a mutation in their BRCA1/2 genes remain healthy after PARP-inhibition[17], resulting in a targeted cancer therapy. The BRCA and PARP genes are involved in the same functional process, which is often the case for negative GIs, making them susceptible for targeted therapy.

Although the concept of synthetic lethality is promising for personal treatment, its use has been limited in the clinic[13]. In recent years, extensive research has been conducted and many novel synthetic lethal interactions have been found in cancer cells[18–21]. Recent advances in high-throughput assays have enabled researchers to conduct large genome-wide screens with many cancer cell lines to identify selective dependencies in cancer cells. Genome-wide CRISPR/Cas9 knock-out (KO) and drug screens are powerful tools to identify novel synthetic lethal pairs. In addition to the CRISPR and drug screens, cancer cell lines can be characterised for their mutations, transcriptional and epigenetic profile. Finding novel synthetic lethal interactions can thus be inferred from genetic and chemogenetic screens, together with the characterization of the cancer cell line[22].

As it comes to genome-wide KO screens in model organisms, a large amount of work has been performed over the last years in the budding yeast S*accharomyces cerevisiae.* Systematically high-throughput screens revealed a global GI network[9]. Further attempts to expand the GI network in yeast were initiated and more recently, trigenic and digenetic-environmental interactions have been revealed[23,24]. Whereas systematic screening in yeast for di- and trigenic interactions is feasible and has been carried out, in human cell lines this is more challenging. The number of human genes is one of the limiting factors for performing systematically screens for double (or higher up) CRISPR KOs. A more cancer specific burden is the high variety of different cancer cell lines, including differences in the tissue of origin, but also in mutations. Nevertheless, a recent study conducted a CRISPR interference screening over more than 200.000 gene-pairs in two human cancer cell lines and established a template for the genetic landscape of human cells[25]. Furthermore, a CRIPSR interference screening was performed for identifying synergistic drug targets and pairwise genetic interactions in a leukaemia cell line[26].

As cancer is a complex genetic disease there is a huge genetic variation between the different cancer subtypes. The variation occurs in many molecular and genomic factors and may influence the GIs. Understanding these variations could improve the development of drugs and their success. To address this problem, molecular and genomic characterization have been performed with a broad range of cancer cell lines to create a pan-cancer genome-wide analysis[7,27,28]. This research is focused on finding cancer dependencies across cancer cell lines. In 2017, Tsherniak and colleagues published an initial framework for a human cancer dependency map with a loss of function pan-cancer screen. The cancer dependency map can be used to identify genes essential for cell survival across cancer cell lines and these dependencies can be exploited for targeted treatments[7]. The research to establish a cancer dependency map has been continued by the Broad Institute in collaboration with the Wellcome Sanger Institute, and every quarter of a year they release an updated pan-cancer screening which is made public through the DepMap portal. The DepMap portal also provides genetic background of the cancer cell lines, provided by the Cancer Cell Line Encyclopedia (CCLE). The genetic background includes gene copy numbers, gene expression and mutations. Whereas the initial Cancer Dependency Map was created with RNAi screens, the current LOF screens are nowadays performed with CRISPR/Cas9. Although both institutes perform their own screens, they made it possible to combine their screens to generate a larger pan-cancer dataset[29]. The use of these large pan-cancer datasets has been valuable in novel targets for in the clinic and has been recently reviewed[30].

Due to its limitation of performing systematic double (or higher up) KO screens in human cell lines, advanced computational strategies can be exploited to find potential GIs, such as novel synthetic lethal interactions and higher-order interactions. The genetic variability and the complexity of cancer cells makes it important to look further than pairwise interactions. Higher-order interactions are important for cancer survival due to the multivariate nature of cancer cells. These higher-order interactions can also play a role in drug resistance. The initial treatment can influence the gene expression which could lead to drug resistance cancer cells. Therefore, the transcriptional modulation of higher-order interactions could be important for drug sensitivity. Whereas finding synthetic lethal interactions in cancer has been subject of interest over the last years, less research has been conducted in finding higher-order interactions in genome-wide cancer screens[31,32].

In this study we investigated potential novel synthetic lethal interactions from a pan-cancer CRISPR screen and a genetic LOF dataset. Another aim of this study was to find higher-order GIs. We focused on higher-order GIs in gene expression, to study transcriptional modulation of the fitness outcome and its potential regulatory influence on GIs. We created a robust machine learning algorithm to extract second-order interactions from the structure of boosted tree models. In addition, we mapped a global GI network across pan-cancer cell lines with second-order GIs and characterised the functional processes of these genes.

Methods

Data availability

All data was obtained from the DepMap portal (www.depmap.org). For initial model optimizations and testing we used the data from the 21Q1 release which included datasets of pan-cancer CRISPR screens and molecular characteristics of cell lines[33]. We obtained the following files: 'Achilles_gene_effect.csv', 'CCLE_expression.csv', 'CCLE_gene_cn.csv' and 'CCLE_mutations.csv'. The Achilles gene effect dataset contains the results of a pan-cancer CRISPR KO screen of 808 cell lines. For the CRISPR screen the genome-wide Avana sgRNA library[34] was used to target 18,119 genes. The effect of the gene copy numbers on the gene effect score (fitness perturbation) was corrected by the CERES algorithm[35] and corrected for the batch effect. The gene effect score was normalized for a set of essential and nonessential genes[36] resulting in a score of 0 for a nonessential gene KO and -1 as the median for an essential gene KO[33]. The whole pipeline of the Achilles CRISPR screen has previously been described[37].

The three CCLE datasets contain molecular characteristics of the cancer cell lines with respectively, gene expression, gene copy numbers and mutations. The gene expressions were measured of 19,177 genes in 1,376 pan-cancer cell lines with RNA sequencing. The data has been processed with a pseudo-count of 1 and a log2 transformation[28,33]. The gene copy numbers were measured of 27,563 genes in 1,470 pan-cancer cell lines and obtained from whole genome sequencing (WGS), whole exon sequencing (WES) or single nucleotide polymorphism (SNP) arrays. The gene copy number data has also been processed with a pseudo-count of 1 and a log2 transformation[28,33]. The mutations data contained information from 1,747 cancer cell lines and 18,788 genes and was obtained from WGS, WES or RNA-sequencing[33]. The pipelines for the CCLE data can be obtained from the GitHub of the Broad Institute www.github.com/broadinstitute/depmap_omics.

For the final analysis of the second-order interactions, we used the data from the 21Q4 DepMap release and obtained the following files: 'CRISPR_gene_effect.csv' and 'CCLE_expression.csv'[38]. The CRISPR gene effect dataset also contains the fitness perturbation of a CRISPR screen, the same as the Achilles gene effect, but the CRISPR gene effect data is a combined dataset of the CRISPR screens of both the Wellcome Sanger and the Broad Institute[29]. Another difference is that the CRISPR gene effect has been processed with the Chronos algorithm instead of the CERES algorithm[39].

All data was processed and analysed with statistical programme R (version 4.1.1) and RStudio version (2021.09.0)[40,41].

Data preparation

The CRISPR dataset was first standardized over each cell line to align the fitness distributions of the individual CRISPR screens. In order to perform genome-wide regression over each CRISPR gene target using a common set of hyperparameters, we also standardized the fitness distribution of each gene. Subsequently for the linear regression, the mean and standard deviation of each CRISPR gene target were kept to re-scale the estimated coefficients. The gene expression dataset was also standardized over each gene in order to reduce uneven penalization and variable selection based on variations in gene expression magnitude.

A binary LOF dataset was created from the mutation and the gene copy number datasets. Genes with deleterious mutations or copy-number aberrations resulting in homozygous gene deletions were counted as a LOF event for a given cell line. The final LOF dataset was filtered to only include genes with a coverage of LOF events in at least 1% of all the cancer cell lines.

Multivariate penalised linear regression

A multivariate penalised linear regression was performed to predict the fitness perturbation of the CRISPR screen with two different sets of independent variables, the gene expression and the LOF data. The outcome of the linear regression showed GIs between the CRISPR genes and the gene expression or the LOF genes as the magnitude of the fitted coefficients. Before performing the penalised linear regression, we removed cell lines if they were not present in both datasets. The R package 'glmnet' was used for carrying out the regressions[42].

To optimise regularisation parameters for the penalised linear regressions, we first randomly selected 200 genes of the CRISPR dataset and performed a five-fold cross-validation (CV) pilot run. The CV was used for selecting the optimal penalty (λ) for three regularisation approaches. The L1-penalisation of Lasso regression, and the L2-penalisation of Ridge regression, and Elastic Net regression which combines both L1- and L2-penalisation. We selected the penalty with the best average CV-error over the 200 genes. Subsequently, Ridge, Elastic Net and Lasso regressions were performed over the whole genome using their complementary pre-selected penalties.

To obtain a null distribution for the estimated interactions we subsequently performed additional linear regressions by scrambling either the independent variables (H0 X) or the predicted outcome (H0 y). These null distributions were estimated the same way as the regression as described above.

Synthetic lethal interactions

To find potential synthetic lethal interaction we used the LOF dataset to predict the fitness perturbation of the CRISPR screens with the Lasso and Elastic Net regressions. To validate our findings, we computed a precision-recall and ROC curves with validated human synthetic lethal interactions from the SynLethDB database[43]. These were evaluated against a potential negative reference set of interactions with non-essential genes described by Hart and colleagues[36]. We computed a gene-wise p-value and a q-value (false discovery rate (FDR)-adjusted p-values) for the interactions inferred from the regressions. The p-value and q-value were computed against their null-distribution generated with the scrambling for each covariate in the model and were used for adjusting the rank order of the interaction coefficients.

Gene set enrichment analysis

A gene set enrichment analysis (GSEA) of gene ontology (GO)-terms was performed over the number of GIs obtained from the Elastic Net per CRISPR gene. We selected the class of 'Biological Processes' of the GO-terms. The GSEA was accomplished with the R package 'clusterProfiler'[44]. GO-terms with a p-value < 0.05 were selected. To reduce the redundancy of the obtained GO-terms, we performed semantic similarity using the Wang method [45] with a similarity cut-off < 0.5.

XGBoost hyperparameter searches

To find non-linear and higher-order interactions between the genes of the gene expression, we used extreme gradient boosting (XGBoost) with the R packages 'xgboost' and 'EIX'[46,47]. Variables with non-zero coefficients from the Elastic Net regression of the gene expression dataset served as a soft variable pre-selection for the XGBoost in order to reduce the dimension of evaluated variables.

For optimizing the test error of predicting the fitness perturbation with the preselected variables with the XGBoost algorithm, we ran a hyperparameter over a broad range of values of the different parameters to establish an approximate range for the best solution. These parameters consisted of the learning rate (eta), minimal child weight and the maximum depth of the trees. This hyperparameter search was performed with a five-fold CV and the top 10 ranked CRISPR genes with the highest absolute coefficient sum of the Elastic Net regression. Per CRISPR gene we selected the number of iterations based on the lowest CV-error and performed a final XGBoost model with all the samples of the preselected variables and fitness perturbations.

Another hyperparameter search was conducted with the top 200 ranked CRISPR genes with the highest absolute coefficient sum of the Elastic Net. This hyperparameter search focussed on the subsampling and the column sample per tree and used a fixed learning rate, minimal child weight and maximum depth of the trees. This hyperparameter search was performed the same way as previous described above.

Second-order interactions

To extract interactions between the covariates of the gene expression genes, we mined the structure of the boosted trees with the EIX package. An interaction pair is considered between parent and child nodes in a tree. The XGBoost algorithm uses a Gain score to evaluate the split of a growing tree. The Gain is used by the EIX package to calculate a sumGain which sums the Gain of the interaction pair present in all the trees of the model. If the Gain of a child node is higher than its parent node, this is considered to be a strong interaction. Another evaluation measurement of a single features is the Frequency. The Frequency score measures the frequency of a single feature in all the trees of the model. In addition to the sumGain, the EIX package also provides the frequency of a given pair in all the trees of the model.

For further optimization of finding interactions within the structure of the boosted trees, we ran 40 random XGBoost models for the same fitness perturbation prediction and created ensembles of the extracted interactions, to test the pooling of models trained under random subsampling of the training data for the gene expression in the XGBoost model. The data was divided into a test and training dataset, to evaluate model performance in relation to model structure. The test dataset was established by selecting 10 cell lines from different clusters based on their gene expression after dimension reduction with the Uniform Manifold Approximation and Projection (UMAP) algorithm. The 10 different clusters were created with k-means clustering[48].

The second-order interactions selected for preliminary evaluation in a genome-wide GI network with hypernodes were obtained after a five-fold CV for the whole genome for preselecting CRISPR targets based on a cut-off of 0.9 of the CV round mean squared error (RMSE). After preselecting the CRISPR targets, we performed the same procedure with 40 XGBoost models as described above. The obtained interactions were ranked by their sumGain. We performed a Fisher exact test for cumulative PPI enrichment and set a cut-off at a p-value of 0.001. The selected second-order interactions were used to expand the gene expression dataset and were modelled as an interaction term of the two single genes of the second-order interaction. This extended gene expression dataset was used for the multivariate penalised linear regression and the same procedure was performed as previously described.

Protein-protein interactions enrichment

The interactions obtained from the linear regressions and the XGBoost trees were scored for their enrichment of experimentally validated protein-protein interactions (PPIs) from the STRING database[49]. First, all the unique genes which formed an interaction were extracted and the chance of a randomly validated PPI interaction was established. Secondly, the interactions found were ranked by the sum of the value of interest (coefficient, sumGain or frequency). The percentage of validated PPIs of the ranked interaction were calculated for multiple cut-offs, based on a quantile or rank. The enrichment score was then obtained by calculating the fold-change of the percentage of found interactions versus the random chance.

Genome-wide genetic interaction map

The UMAP algorithm was used to create genome-wide GI maps. The UMAP algorithm is a technique for dimension reduction [48], and we used the Pearson squared correlation as a distance metric. We used the interactions and the coefficients obtained from the Ridge regression of the gene expression data for the initial GI map and the GI map with the second-order interactions. In the projected genome-wide GI map, clusters were created by k-means clustering with 100 clusters and annotated for their enriched ‘Biological process’ from the Gene Ontology (GO) Term Enrichment database[44,50].

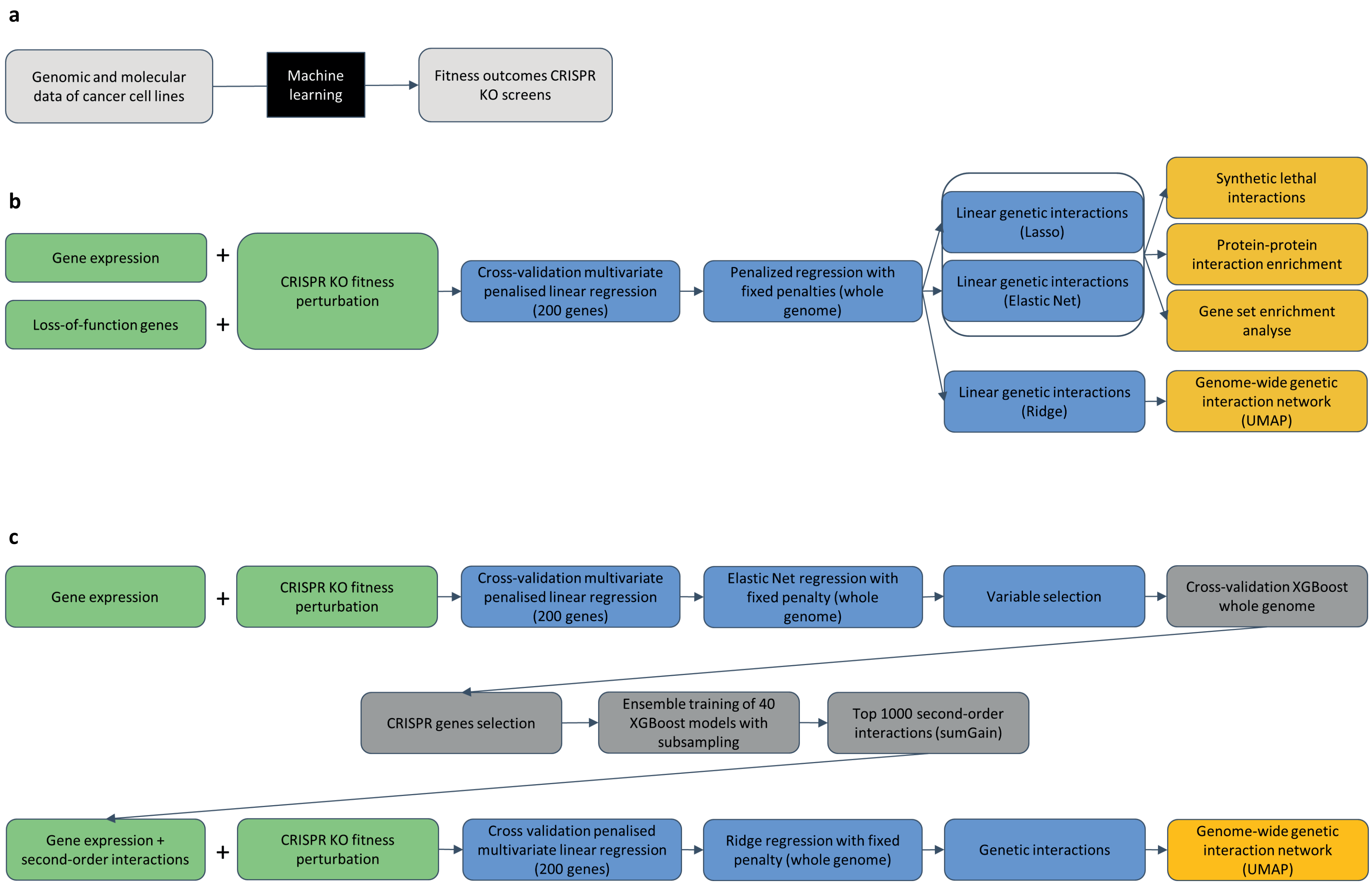
Results

Workflow

In this study, we used the L1-regularised linear regression with Lasso and Elastic Net which combines both L1- and L2-penalisation, to infer GIs from pan-cancer CRISPR screens based on the genetic and transcriptional background of the cancer cells (figure 1a). This technique was used for both the gene expression and LOF datasets and predicted the fitness perturbation of the CRISPR screen. L1-regularisation using Lasso is a reliable method for finding causal inference by strongly penalising multicollinear structures that exist in biological omics data[51–53]. Additionally, the Elastic Net regression was used to preselect candidates for inferring second-order GIs, being a softer variable selection method compared to Lasso. These methods were benchmarked for enrichment of PPIs and detection of synthetic lethal interactions between the selected covariates and their respective CRISPR targets (figure 1b).

On the other hand, we used the L2-regularisation Ridge regression to map a genome-wide GI network based on the correlative fitness structure of gene expression. Functional similar genes cluster together within the network with a correlated fitness interaction profile[9,54]. We used the Ridge regression in order to preserve a continuous parametric fitness interaction spectrum for the whole genome. The UMAP algorithm was used for dimension reduction to create a two-dimensional network (figure 1b).

The XGBoost algorithm was used for predicting higher-order non-linear fitness dependencies. We mined the structures of the XGBoost trees to establish second-order interactions. To improve the inference of true interactions, we benchmarked the methodology for enrichment of PPIs under various hyperparameter conditions. Second-order interactions discovered by the XGBoost algorithm were subsequently modelled as polynomials within the linear regression scheme in order to infer the directionality of the predicted fitness perturbations. In addition, we used the second-order interaction for hypergraph mapping and functional annotation within a global GI network using Ridge regression and UMAP (figure 1c).


a
Genomic and molecular data of cancer cell lines
Machine learning
Fitness outcomes CRISPR KO screens
b
Gene expression
+
Loss-of-function genes
+
CRISPR KO fitness perturbation
Cross-validation multivariate penalised linear regression (200 genes)
Penalized regression with fixed penalties (whole genome)
Linear genetic interactions (Lasso)
Linear genetic interactions (Elastic Net)
Linear genetic interactions (Ridge)
Synthetic lethal interactions
Protein-protein interaction enrichment
Gene set enrichment analyse
Genome-wide genetic interaction network (UMAP)
c
Gene expression
+
CRISPR KO fitness perturbation
Cross-validation multivariate penalised linear regression (200 genes)
Elastic Net regression with fixed penalty (whole genome)
Variable selection
Cross-validation XGBoost whole genome
CRISPR genes selection
Ensemble training of 40 XGBoost models with subsampling
Top 1000 second-order interactions (sumGain)
Gene expression + second-order interactions
+
CRISPR KO fitness perturbation
Cross validation penalised multivariate linear regression (200 genes)
Ridge regression with fixed penalty (whole genome)
Genetic interactions
Genome-wide genetic interaction network (UMAP)

Multivariate penalised linear regression

In order to infer GIs from the CRISPR screen and the independent variables, the gene expression or the LOF data, we performed a multivariate penalised linear regression with different penalties. For each of the linear regression, we performed a five-fold cross validation over 200 randomly selected CRISPR genes to establish a general penalty that was subsequently used for the regressions over the whole genome. In order to create a null-hypothesis reference point for the inferred interactions, the same optimization schemes were performed by either randomising the sample order of either the CRISPR dataset, the dependent variables, (H0 y) or the independent variable dataset (H0 X). The results of the Elastic Net regression with the gene expression as the independent variables showed that the sum of the absolute coefficients per CRISPR KO gene is significantly higher than both scrambled null hypotheses (figure 2b and supplement 1). This was also applicable for the absolute sum of the coefficients per gene of the gene expression. The scrambling of H0 X destroyed the correlative pattern of the gene expression, which resulted in a more conservative null distribution for hypothesis testing. In contrast to the H0 X scrambling, the randomization of H0 y maintained the correlative structure of the independent variables and was therefore subjected to a stronger penalization for highly correlated variables, that could arise due correlated gene deletion patterns for genes on the same chromosomes. The clearance between the true data and the null-distributions was stronger for the gene expression data than the LOF data after Elastic Net regression.

The null-distributions for the LOF dataset also showed that randomization results in a significant increase of the absolute sum of coefficients per CRISPR gene or a LOF gene (figure 2b and supplement 1). In contrast to the gene expression data, the H0 y of the LOF data yielded a more conservative hypothesis testing, due to the preserved correlative structures of the mutations and gene copy number losses. As the correlations in the mutation patterns are maintained and variables with better coverage in the dataset tend to be biasedly selected (in the absence of any fitness association). The random associations are potentially more evenly distributed when randomizing the X data, thus resulting in more numerous small coefficients.

To show the distribution at the level of the individual genes (gene expression or LOF genes) we picked the 12 highest ranked genes based on their absolute sum of coefficients per gene of the Elastic Net regression. The distributions showed the clear distinction between the true linear regression compared to the null distributions (figure 1c and 1d).

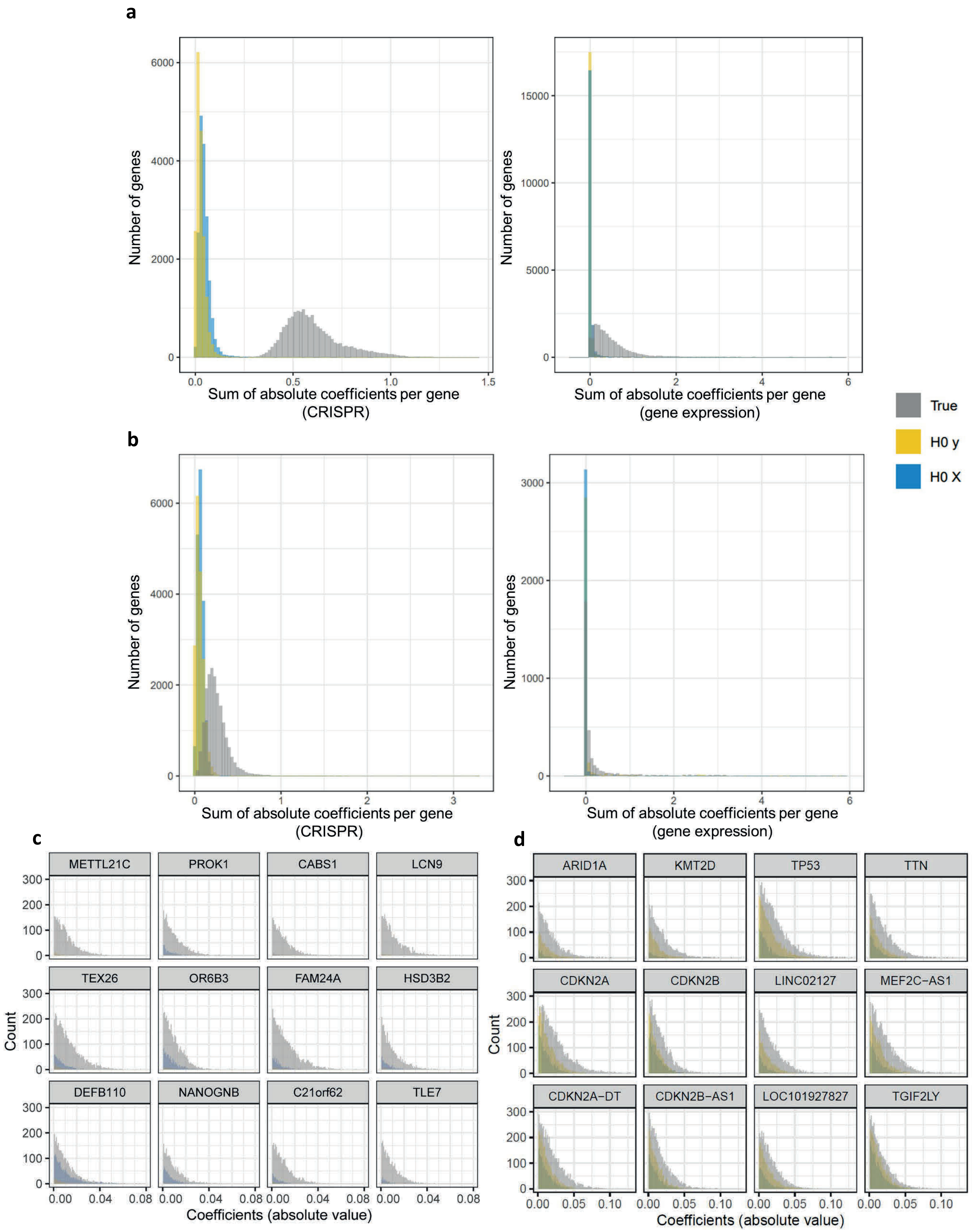


Figure 2. Distribution of the Elastic Net linear regression coefficients. **(a)** Distribution of the absolute sum of the coefficients of the Elastic Net regression with the gene expression data. **(b)** Distribution of the absolute sum of the coefficients of the Elastic Net regression with the LOF data. **(c)**. The top 12 genes with the highest absolute coefficient sum for the gene expression data. **(c)**. The top 12 genes with the highest absolute coefficient sum for the LOF data.

Gene level enrichment

To validate the GIs found by the Lasso and Elastic Net regressions we looked at the enrichment of experimentally validated PPIs. We found that in the linear regressions with the gene expression and the CRISPR data, the GIs are highly enriched at the top hits in both the Elastic Net and Lasso regressions. The top 100 hits are ~13-fold enriched (Elastic Net) and ~17.5-fold enriched (Lasso) (figure 3a). Interestingly, the positive GIs were higher enriched compared to the negative GIs. The top 100 positive GIs for the Elastic net regression were ~19.5-fold enriched versus ~2.5-fold enrichment of the negative GIs. For the Lasso regression, we found a similar trend. A ~22.5-fold enrichment of the top 100 positive GIs compared to ~5-fold enrichment of the negative GIs. These trends could suggest that gene expression causes an increase in gene dependencies which are not commonly manifested through physical interactions. On the other hand, if we look at the PPI enrichment of the predicted GIs between LOF mutations and the CRISPR KOs, there is a smaller difference. The top 100 ranked hits of the Elastic Net showed an enrichment of ~20-fold for the positive GIs and ~21-fold for the negative GIs (figure 3b). For the Lasso regression, the enrichment was ~17.5-fold for the top 100 ranked positive GIs versus ~14-fold enrichment of the negative GIs.

An interesting observation was that the enrichment of the positive GIs from the gene expression after performing the Elastic Net and Lasso were higher enriched in PPIs than negative GIs (figure 2a). A positive GI between a gene expression and a CRISPR gene indicate that a lower gene expression is associated with a lower fitness. For a negative GI it is the other way around, a lower gene expression is associated with a higher fitness than expected. This is contrary to the meaning of a GI between the LOF and CRISPR genes, where a negative GI results in lower fitness than expected. An enrichment of PPIs in positive GIs corresponds with previous findings in yeast[55].

The negative GIs inferred by the Elastic Net and Lasso regressions of the LOF dataset were used for the prediction of synthetic lethal interactions between the LOF and the CRISPR genes. To validate these synthetic lethal interactions, we computed a receiver operating characteristic curve (ROC) with respectively 235 true positives and 11,339 potential false positives with non-essential genes[36] (Elastic Net), and 106 true positive and 1,704 potential false positive interactions with non-essential genes[36] (Lasso). Because we found a substantial bias in inferring an interaction of genes with strong coverage in the LOF data, we also tested the adjustment of the rank order based on gene-wise hypothesis tests by computing p-values and q-values (FDR-adjusted p-values) for interactions of genes in the LOF data against their respective null-distributions (figure 3c and 3d).

The sensitivity of detecting true positive synthetic lethal interactions increased after gene-wise p-value adjustments with both null distributions for both the Elastic Net and the Lasso (figure 3c and 3d). The sensitivity of detecting true synthetic lethal interactions with the Elastic Net was 0.98 with an FDR of 0.05 for both H0's (figure 3c). The sensitivity of detecting true synthetic lethal interactions with the Lasso was 0.99 with an FDR of 0.05 when using H0 y (figure 3d). On the other hand, the sensitivity dropped to 0.95 with an FDR of 0.05 when using H0 X (figure 3d). These findings suggest that the Elastic Net and Lasso regression combined with gene-wise null hypothesis testing results in a high sensitivity for detecting true GIs.

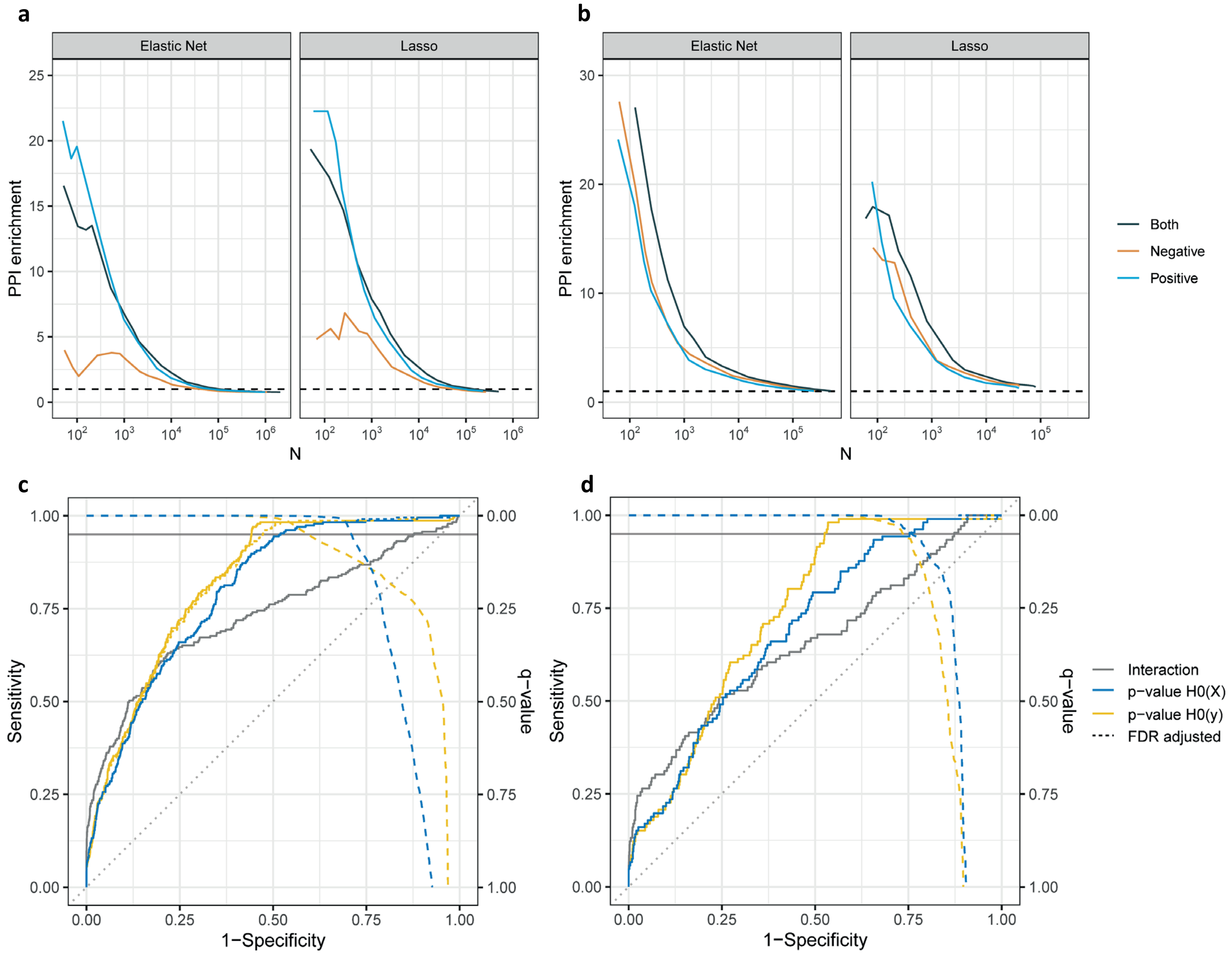


Figure 3. Functional enrichment of the linear genetic interactions. (a) PPI enrichment of GIs found with the Lasso and Elastic Net regressions. The linear regression was performed with the independent variable set of gene expression genes. The positive GIs are coloured in blue, the negative GIs in orange and combined in black. The GIs are ranked based on the linear regression coefficient. The black dotted line indicates an enrichment of 1-fold. **(b)** PPI enrichment of GIs found with the Lasso and Elastic Net regressions. The linear regression was performed with the independent variable set of LOF. **(c)** Receiver operating characteristic (ROC) curve of synthetic lethal interactions of the Elastic Net regression with the LOF dataset. The continuous lines indicate the ROC curves based on the order of the regression coefficients (grey), and the coefficients' order re-adjusted with the gene-wise p-values computed for H0 y (yellow) and H0 X (blue). The dotted lines are ROC curves based on the coefficients' order re-adjusted with the FDR-adjusted p-values (q-values). The dashed lines indicate the q-values for the potentially false positives for both null-hypotheses, with H0 y in yellow and the H0 X in blue. **(d)** ROC curve of synthetic lethal interactions of the Lasso regression with the LOF dataset.

Process level enrichment

In order to investigate which biological processes were most strongly associated with a high degree of GIs, we performed a GSEA with the GO-terms of 'Biological processes' on a list of the number of GIs identified per CRISPR gene with Elastic Net regression. For GIs inferred from the gene expression data, these associations represent cellular dependencies that are most highly influenced by changes in gene expression. Here we found mainly enrichment of GO-terms associated with cell core-specific processes, such as *protein-DNA complex subunit, establishment of RNA localization* and *regulation of G2/M transition of mitotic cell cycle* (figure 4a). Although the genes with the highest number of GIs were mainly involved in cell core specific processes, the biological process with the highest GIs was *viral gene expression*. This could be a side effect of the CRISPR screen, which was performed with viral transfection of the sgRNA to the cancer cells**.** Compared to GO-terms associated with a high number of GIs inferred from the LOF data, we identified mainly processes of the mitochondrion, including the following processes, *mitochondrial respiratory chain complex assembly*, *respiratory electron transport chain* and *mitochondrial translation* (figure 4b).

The enriched processes with the least GIs from the gene expression and the LOF data were mainly involved in tissue-specific processes. The enriched processes of the LOF data were involved in regulation of tissue- or sense-specific processes, such as *negative regulation of response to food*, *negative regulation of appetite* and *negative regulation of response to nutrient levels* (figure 4b). The gene expression data was enriched for other tissue-specific processes such as *regulation of macrophage activation, regulation of coagulation,* and *sensory perception of pain* (figure 4a). The genes with the lowest number of GIs were less dependent on gene expression or a LOF for the fitness perturbation.

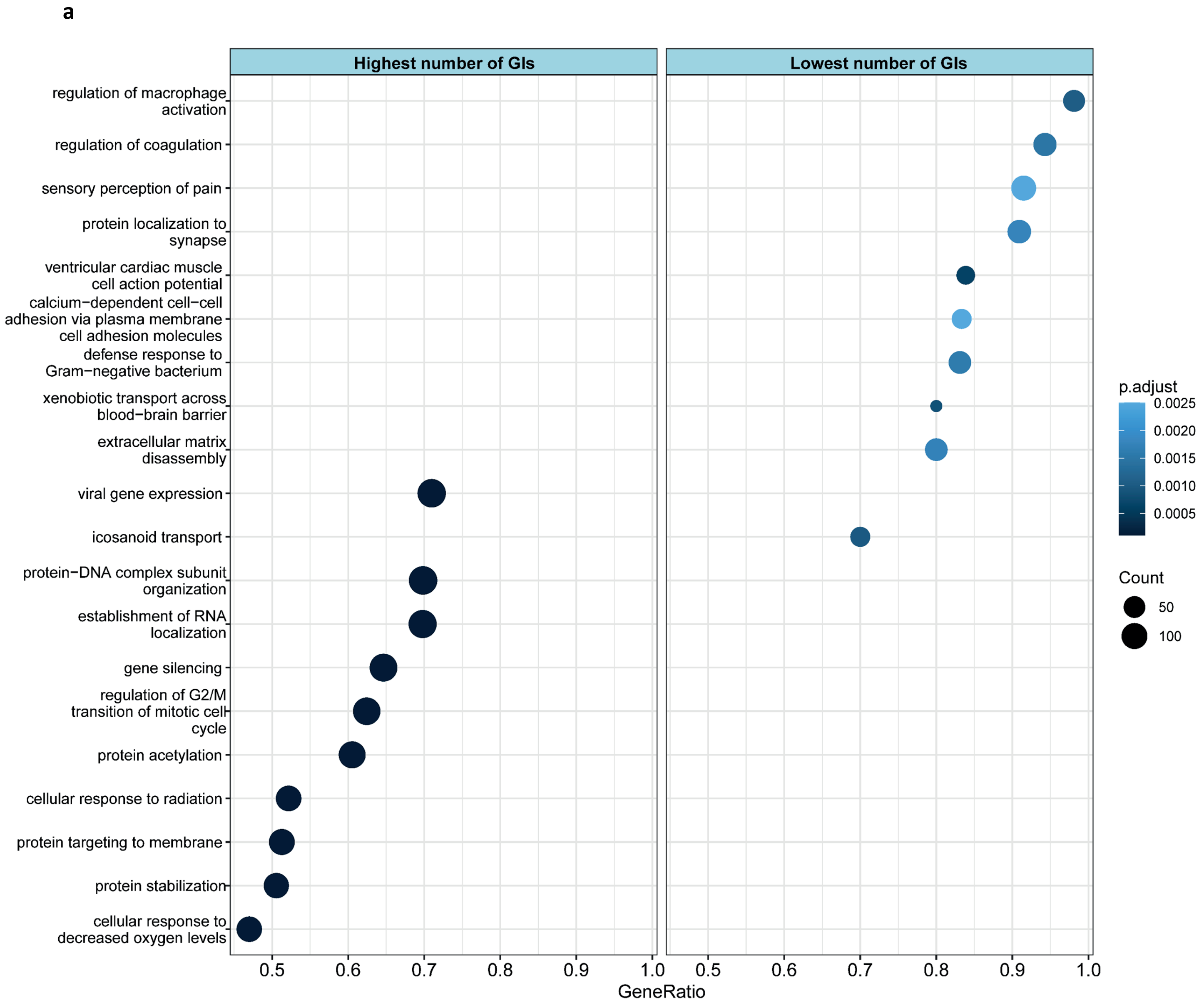


Figure 4. Gene set enrichment analysis of the linear genetic interactions. (a) Gene set enrichment plots for the number of GIs per CRISPR gene after Elastic Net regression with the gene expression data. The GeneRatio indicates the ratio of the genes present in the enriched process.

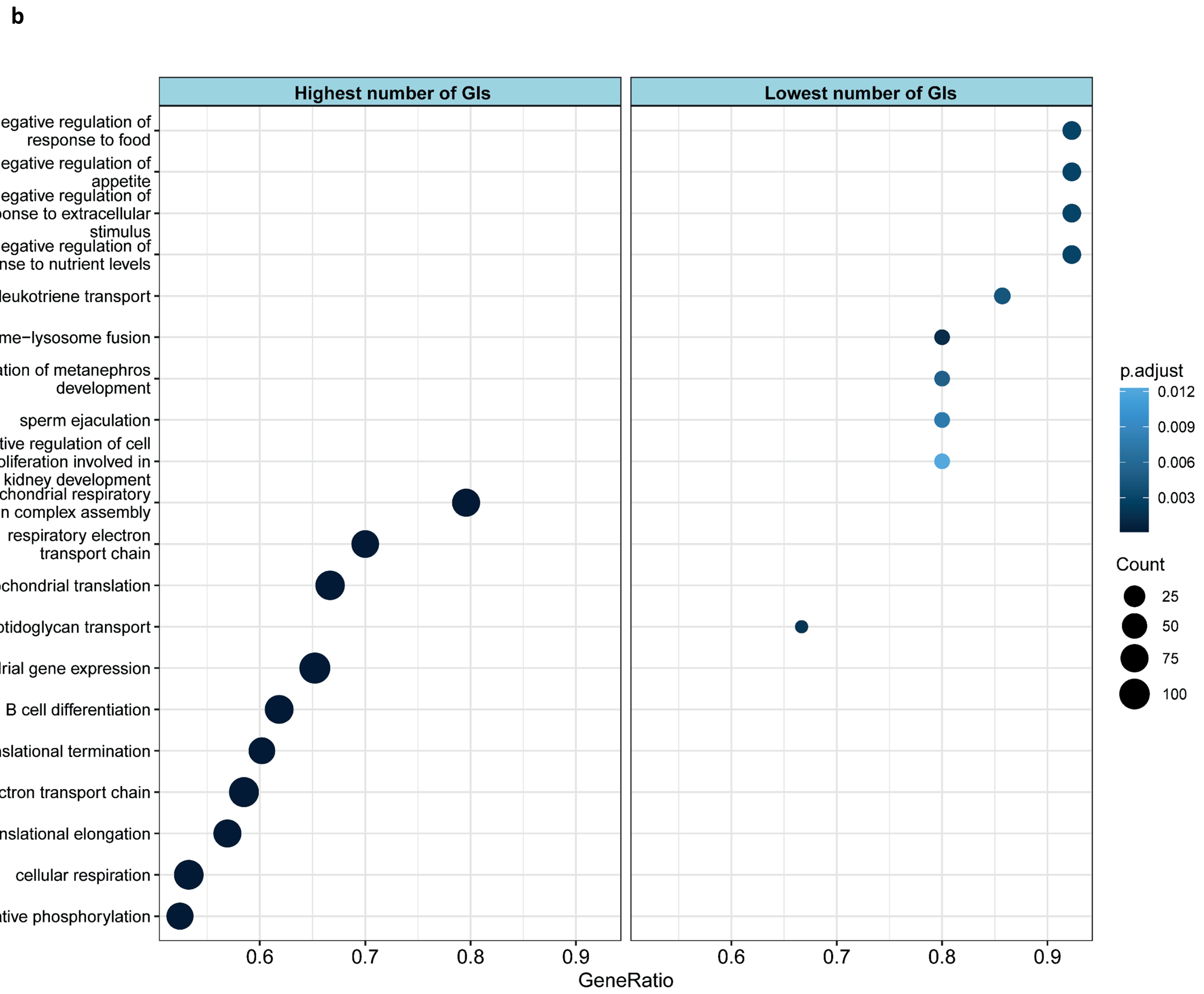


Figure 4. Gene set enrichment analysis of the linear genetic interactions. (b) Gene set enrichment plots for the number of GIs per CRISPR gene after Elastic Net regression with the LOF data. The GeneRatio indicates the ratio of the genes present in the enriched process.

Genome-wide genetic interaction network

Correlations in GIs have been shown to cluster genes together with similar functionality in a genome-wide GI network[9,54]. To establish a genome-wide network, we used the L2-regularisation based Ridge regression. The L2-regularisation forces coefficients for non-predictive association towards zero but maintains a non-zero coefficient for all independent variables, thus maintaining a genome-wide fitness interaction profiles. Therefore, it can establish a framework for computing a correlative global GI network. For the mapping of the GIs correlations in a two-dimensional space, we used the UMAP algorithm to reduce the dimensions from the whole-genomic interactions. We established two global networks, either based on the genes from the gene expression or the CRISPR screen. The network of the gene expression genes is based on the correlative structure of genetic fitness interactions of the CRISPR screen. The other network, based on the genetic fitness interactions of the CRISPR genes is based on the correlative structure of the regulatory genes.

In order to identify cluster-wise enrichment of functional processes we performed k-means clustering with a total of 100 clusters, followed by GO-term enrichment of the genes in the clusters (figure 5a and 5b). The results of the genome-wide GI map showed distinct functional clusters of different enriched biological processes. First of all, the network of the gene expression genes showed that cell core specific processes and tissue specific biological processes were distinct from each other (figure 5a). The cell core processes were mainly at the edges and the centre of the upper part of the network. Among these cell core processes we found: *RNA splicing, protein targeting, ribonucleoprotein complex biogenesis,* and *regulation of GTPase activity.* The tissue-specific processes are forming a bigger cluster at the centre of the network, including the following processes: *skin development, detection of chemical stimulus involved in sensory perception, lymphocyte mediated immunity,* and *blood coagulation* (figure 5a).

For the global GI network for the genes of the CRISPR targets we observed the opposite compared to the genes of the gene expression. At the edges of the network, we found that it is mainly enriched with tissue-specific processes (figure 5b). For instance, *epidermis development, sensory perception of smell, humoral immune response,* and *defence response to bacterium.* The cell core specific processes were mainly found in the centre of the network, forming a bigger cluster. Among these processes we found: *Golgi vesicle transport, mRNA catabolic response*, *tRNA modification,* and *actin filament organisation* (figure 5b).

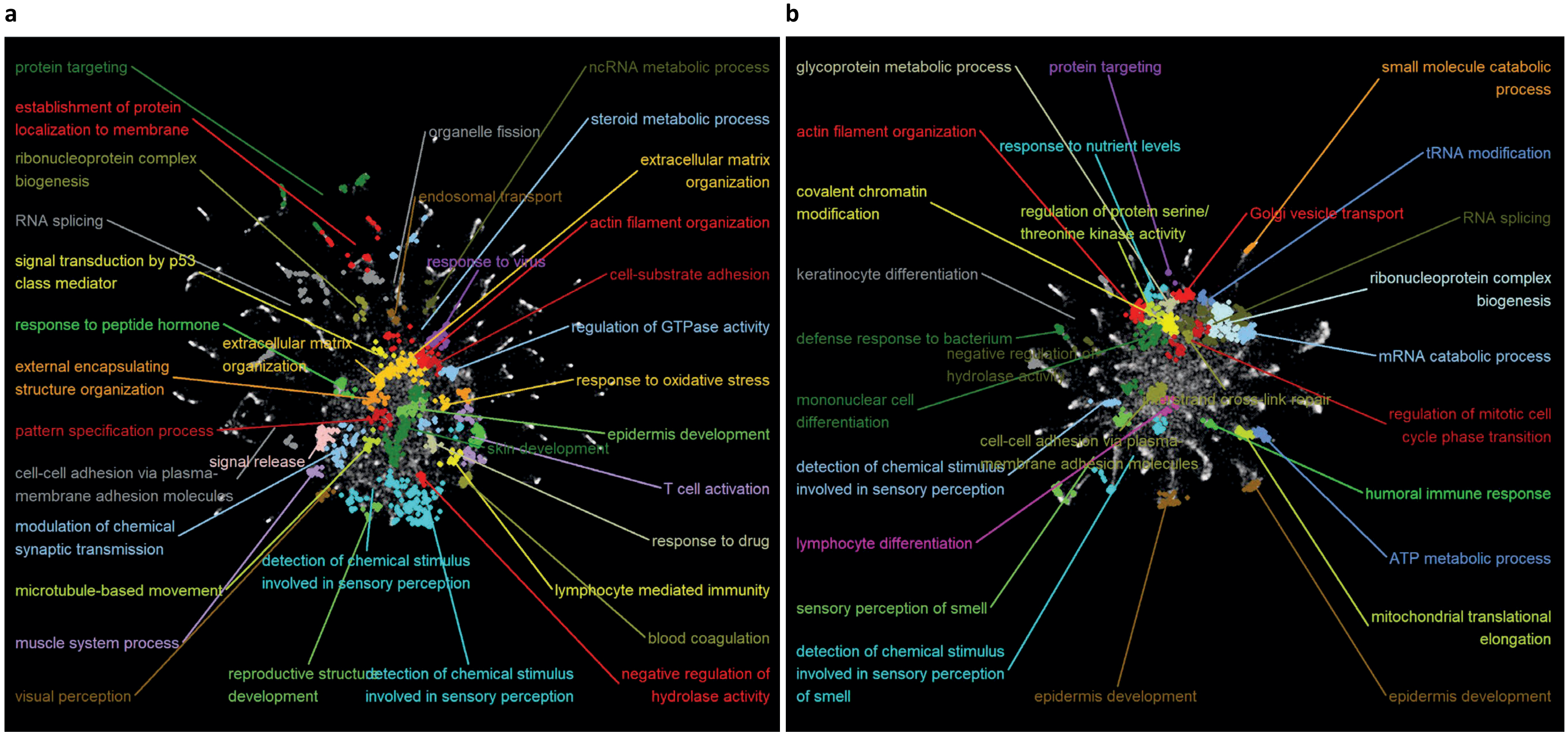


Figure 5. Genome-wide genetic interaction networks. **(a)** The white dots represent the genes from the gene expression. Genes were clustered by k-means clustering with 100 clusters. Enriched clusters are indicated with different colours and annotated for the enriched biological process GO-term. **(b)** The white dots represent the genes from the CRISPR screen. **(b)** Genes were clustered by k-means clustering with 100 clusters. Enriched clusters are indicated with different colours and annotated for the enriched biological process GO-term.

Second-order interactions

While the mapping and identification of pairwise interactions is a powerful tool to explore fitness interactions across a panel of cancer cells, it limits the analysis to an average value with varying validity in different cancer subtypes. GIs may vary depending on other factors and understanding such variations may be crucial for drug development and response. One step to address this complexity is to map potential second or higher variables that change the fitness interaction of an initial pair. To find such higher-order dependencies we focused on the regulatory role of gene expression.

In order to find the second-order interactions we used the machine learning technique XGBoost. The XGBoost algorithm uses boosted gradient trees for its prediction, making it suitable to mine non-linear interactions between the covariates within the trees of the model. The variables for predicting the fitness perturbations with the XGBoost algorithm were preselected with the Elastic Net regression. We used the Elastic Net for preselecting the variables with the highest predictive linear value on the fitness perturbation. The regularisation of the Elastic Net is weaker compared to the Lasso, and therefore more variables are accounted for the prediction. The soft variable selection of the Elastic Net still holds enough variables to find nonlinear interactions.

In order to optimize the XGBoost algorithm for predicting fitness perturbations, we ran a hyperparameter search to identify the important parameters of the algorithm to minimize the CV-error. The first hyperparameter runs were to find the approximate regions of the minimal child weight, maximum depth, and the learning rate (eta) (supplement 2). After establishing these parameters for minimizing the CV-error, we conducted another hyperparameter search for two other parameters, subsampling of the training data (subsample), and the column sample per tree (figure 6a). The subsample indicates the fraction of training data available for every iteration of a growing tree in the XGBoost algorithm. The column sample per tree is another subsampling method, but this regulates the fraction of available variables evaluated per growing tree. A subsample of 0.5 prevented overfitting of the models compared to no subsampling at all (figure 6a).

We mined the interactions between the covariates from the XGBoost trees. The interactions were classified as a strong interaction pair when the child node had a higher Gain than the parent node. To investigate the functional enrichment of the obtained interactions we performed a PPI enrichment test (figure 6b). The interactions were ranked by the sum of their sumGain or frequency. The PPI enrichment was improving for all the interactions ranked on either their sumGain or frequency when the column sample per tree was increasing from 0.05 to 0.4 for both the subsampling fractions of 0.5 and 1 (figure 6b). Increasing the column sample from 0.4 upward, the interactions did not show an ascending PPI enrichment anymore for both subsample values. For the strong interactions ranked on their sumGain, there was an improvement of PPI enrichment under subsampling of 0.5. These results indicate that subsampling prevents model overfitting, and this is associated with a high enrichment of functional GIs.

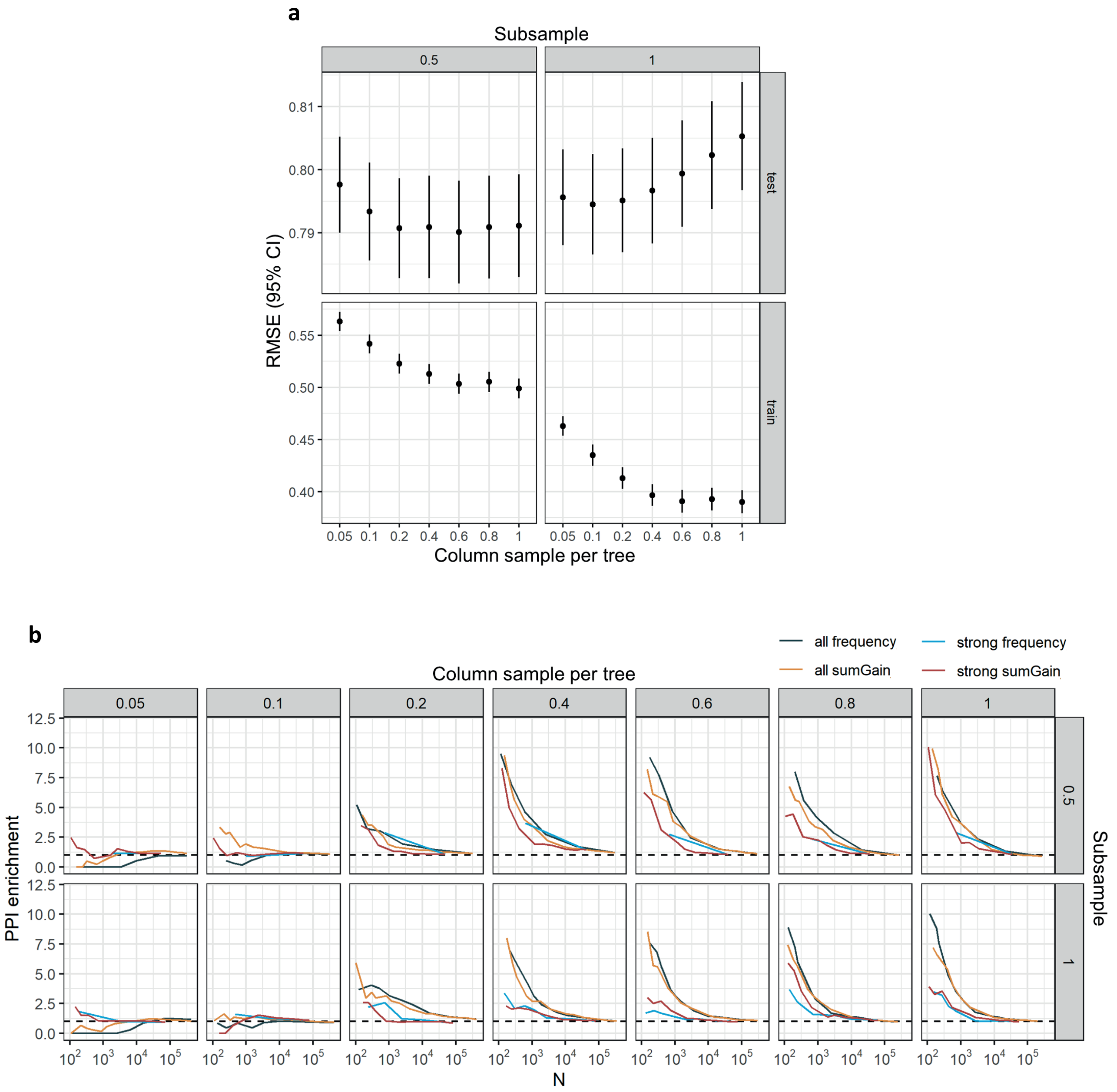


Figure 6. XGBoost hyperparameter search. Hyperparameter search of the subsample and column sample per tree. **(a)** The round mean squared error (RMSE) and the 95% confidence interval (black lines) are shown for the test and training data of the XGBoost models. **(b)** The PPI enrichment for different subsamples and column samples per tree values, based on the ranking of the interactions. The coloured lines indicate the type of interactions and by which metric they are ranked. The black line is based on the ranking of the sum of frequencies, in orange the sum of the sumGain, in blue the sum of the strong frequencies and in red the sum of strong sumGain. The black dotted line represents an enrichment of 1-fold.

XGBoost models binning

The results from the hyperparameter search with the subsampling showed an improvement with respect to overfitting and finding GIs enriched for PPIs, that was dependent on the coverage of evaluating covariate sampling. Therefore, we wanted to investigate the effect of pooling several models with subsampling trained under random sampling. In order to investigate the effect of binning models, we ran 40 models with random subsampling. We found that random subsampling is responsible for variability between the enrichment of PPIs. The 40 models we ran were performed with 50% of subsampling of the training data for predicting the fitness perturbation. The findings showed that the enrichment of PPIs was improving, indicating that ensembles of multiple models trained under random subsampling is improving the findings of PPIs (figure 7a and 7b).

For further exploration, we investigated the possibility to extract the XGBoost models which showed the highest enrichment of PPIs. Therefore, we looked at the relation between the enrichment of PPIs from an ensemble of models and their average test RMSE. To enhance the enrichment of the interactions, we binned the different XGBoost models based on their ranking of the test RMSE value. We binned the outcomes of the interactions on two different methods. The first binning method was to bin the interactions found per XGBoost model by their ranking per CRISPR gene. We used 200 CRISPR genes, so 200 models were binned, one for every CRISPR gene. Therefore, the results show 40 different enrichments with a different test RMSE score. The enrichment of all the 40 models was higher than the reference model of no subsampling at all (figure 7a). In addition, we binned the models with sets of 5 models per CRISPR gene, so 1000 models were binned together. This showed an improved enrichment for all the binned models compared to the reference model (figure 7a). However, the binning strategy based on the RMSE score per CRISPR gene did not show a significant correlation between the test RMSE and the PPI enrichment for the frequency (p-value = 0.53 and p-value = 0.83, respectively 200 and 1000 models) and for the sumGain (p-value = 0.43 and p = 0.62, 200 and 1000 models) (figure 7c).

The other binning strategy was based on ranking the different XGBoost models by overall ranking and not per CRISPR gene. This was performed with the binning of 200 and 1000 models (figure 7b). This binning strategy showed a negative correlation between the test RMSE and the PPI enrichment (figure 7d). The negative correlation was higher for the 1000 models (R = -0.96 and p-value < 0.001, and R = -0.94 and p-value = 0.019) respectively for the frequency and sumGain, compared to the 200 models (R = -0.5 and p-value= 0.0011, and R = -0.52 and p-value < 0.001) respectively for the frequency and the sumGain. Therefore, the binning of models trained under random subsampling improved the detection of PPIs and selective binning of the models with high prediction accuracy improved the detection of PPIs even further.

After these findings, we proceeded with the latest data from the DepMap consortium. We used these datasets from the 21Q4 for the final global GI map, because more cancer cell lines were included and were not published yet when we started the study**.** We performed the same pipeline with the linear regression as previously described. For the XGBoost we performed a whole genome CV-run to preselect CRISPR targets based on their CV-error. The CV RMSE cut-off was set to <

0.9, which resulted in 2,870 CRISPR genes (supplement 3 figure 1). The selected CRISPR genes and its gene expression variables preselected by the Elastic Net were used and 40 models with 50% of random subsampling were performed for each CRISPR target. The top 1000 GIs ranked on the sumGain were used for further analysis.

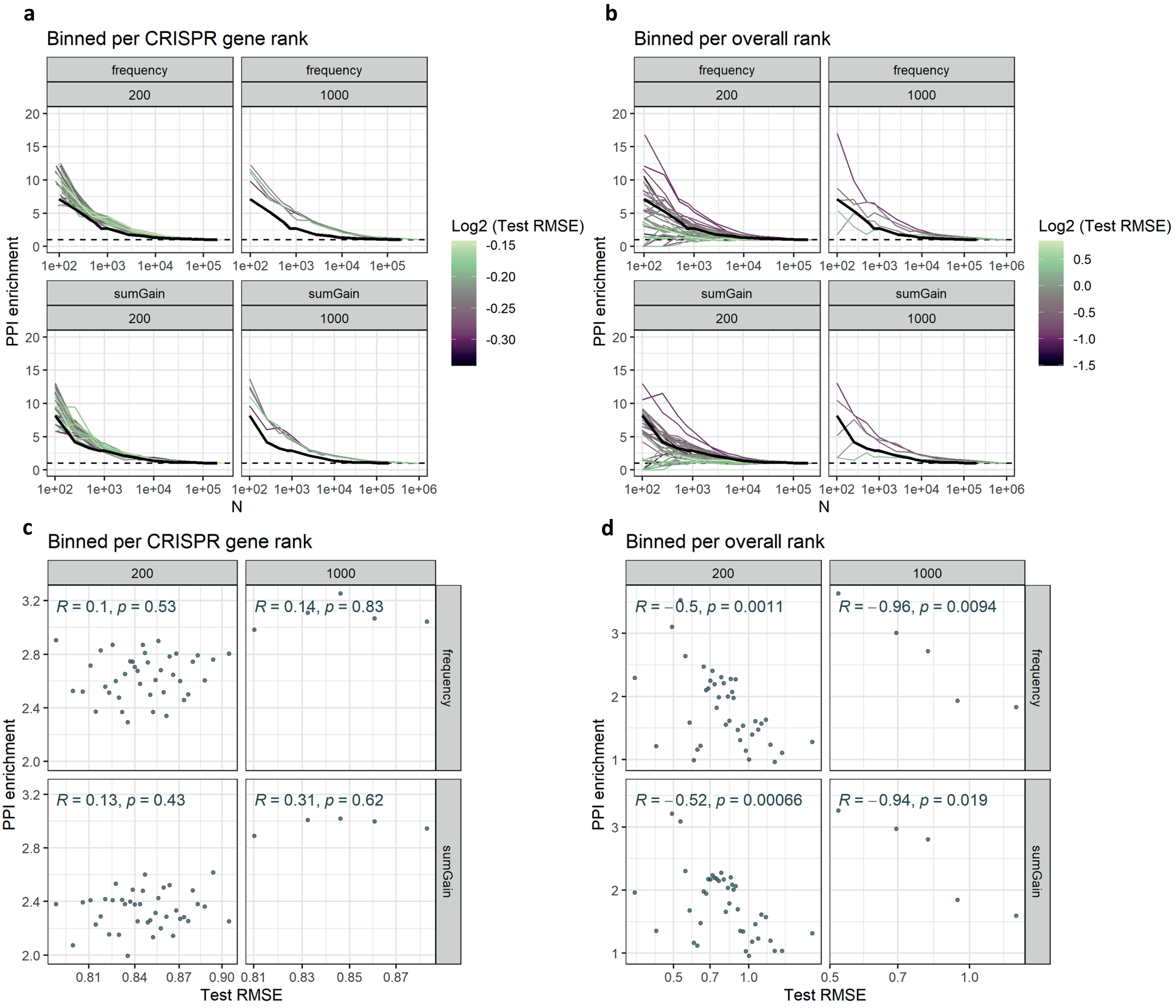


Figure 7. XGBoost model binning strategies. (a) PPI enrichment of interactions obtained from an ensemble of XGBoost models (200 or 1000) binned per CRISPR gene ranking. The test RMSE is indicated by a colour gradient. The black line indicates the XGBoost model without subsampling. **(b)** PPI enrichment of interactions obtained from an ensemble of XGBoost models (200 or 1000) binned by overall ranking. The test RMSE is indicated by a colour gradient. The black line indicates the XGBoost model without subsampling. **(c)** Correlation between the PPI enrichment and the test RMSE for different number of models per CRISPR target ranking. **(d)** Correlation between the PPI enrichment and the test RMSE for different number of models binned by overall ranking.

Genome-wide genetic interaction network with second-order interactions

In order to investigate the complex regulatory fitness interactions of gene expression, we generated a global GI network that included the second-order interactions. The location of the second-order interactions within the global GI network could provide information about the functional process of the interaction. The second-order interactions were extracted from the XGBoost models and modelled as an interaction term of the gene expression dataset. The same workflow was performed with the Ridge regression, but with the second-order interactions included. The UMAP algorithm was used to generate a GI network for the fitness correlations along the transcriptome dimension.

In the GI network with the functional enrichment, we observed a distinction between the cell core specific processes, such as *RNA splicing, actin filament organization, peptidyl-lysine modification,* and *proteosomal protein catabolic process*, which were enriched in the right lower corner of the network (figure 8a). On the other side of the network, we found biological processes involved in tissue-specific processes, including *T cell activation, epidermis development, visual perception* and *regulation of membrane potential.*

The distribution of the second-order interactions and the two single paired genes were mapped on the GI network (figure 8b). We found that the second-order interactions were mainly occurring in the centre and left corner of the network, whereas the single genes of the interactions were mainly distributed on the other side of the network (figure 8b). The distribution of the single genes was therefore mainly located in the more cell core processes region, whereas the second-order interactions were more located at the tissue-specific part of the network (figure 8a and 8b).

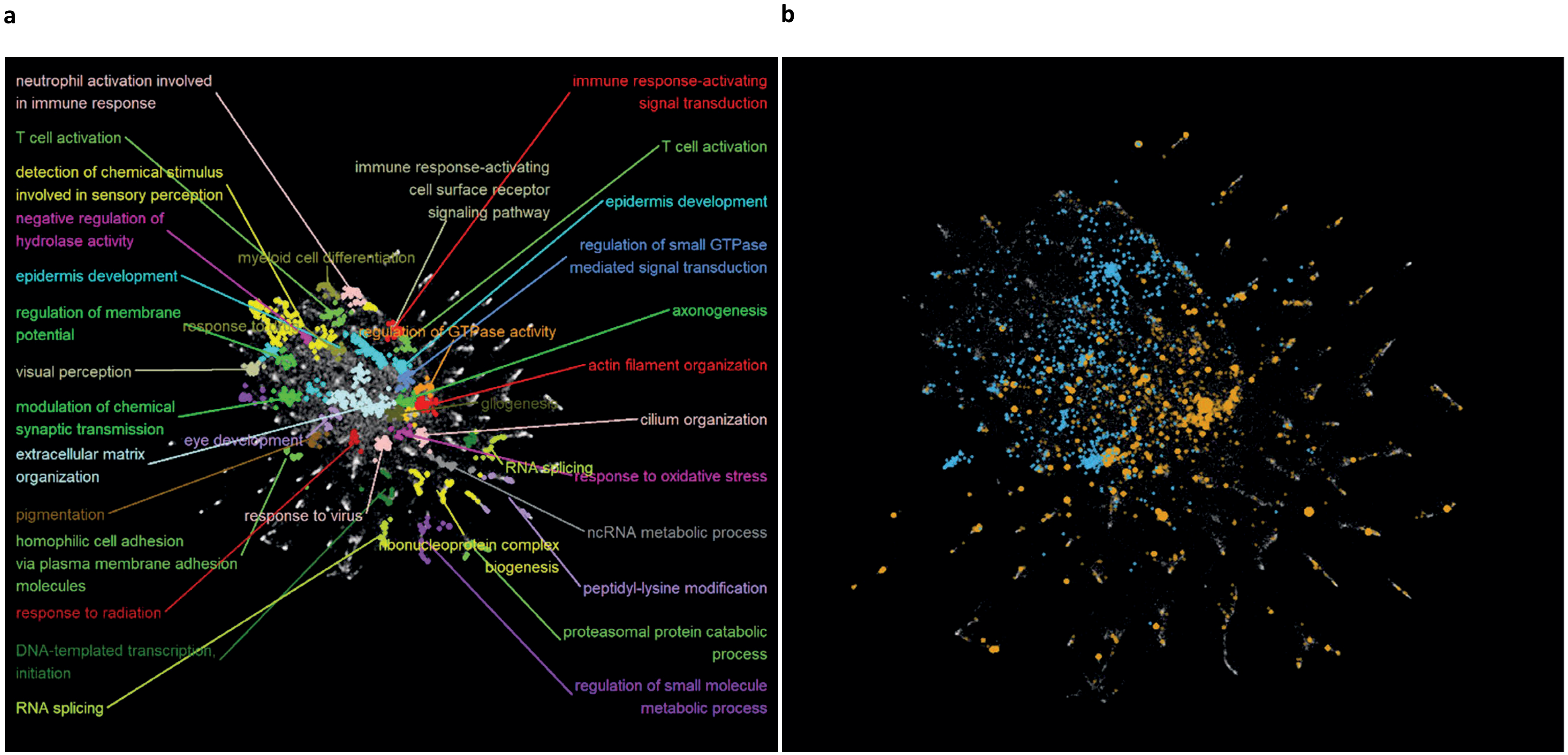


Figure 8. Genome-wide genetic interaction network with second-order interactions. (a) The white dots represent the genes from the gene expression. The genes were clustered by k-means clustering with 100 clusters. Enriched clusters are indicated with different colours and annotated for the enriched biological process GO-term. **(b)** Distribution of the second-order interactions (blue) and the single genes of the second-order interactions (orange) in the genome-wide GI network.

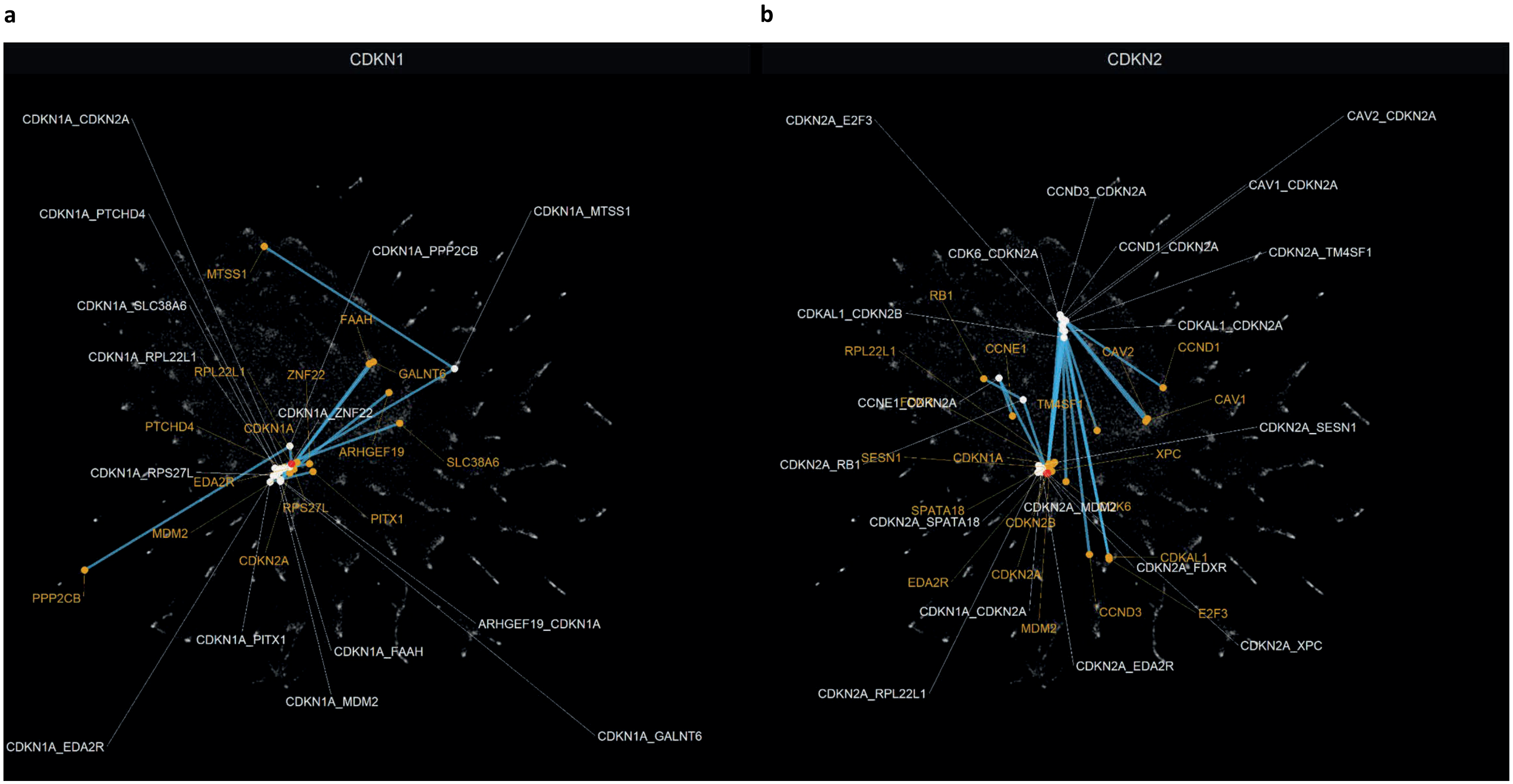


Figure 9. Second-order interactions in the genome-wide interaction network. (a) Mapping of the second-order interactions (white) of CDKN1A in the genome-wide interaction network of figure 8a. The single genes of the second-order interactions are given in orange. The blue lines are in between the second-order interaction and the single genes it consists of. The red dot represents the location of CDKN1A. **(b)** Mapping of the second-order interactions (white) of CDKN2A and CDKN2B in the genome-wide interaction network of figure 8a. The single genes of the second-order interactions are given in orange. The blue lines are in between the second-order interaction and the single genes it consists of. The red dots represent the location of CDKN2A and CDKN2B.

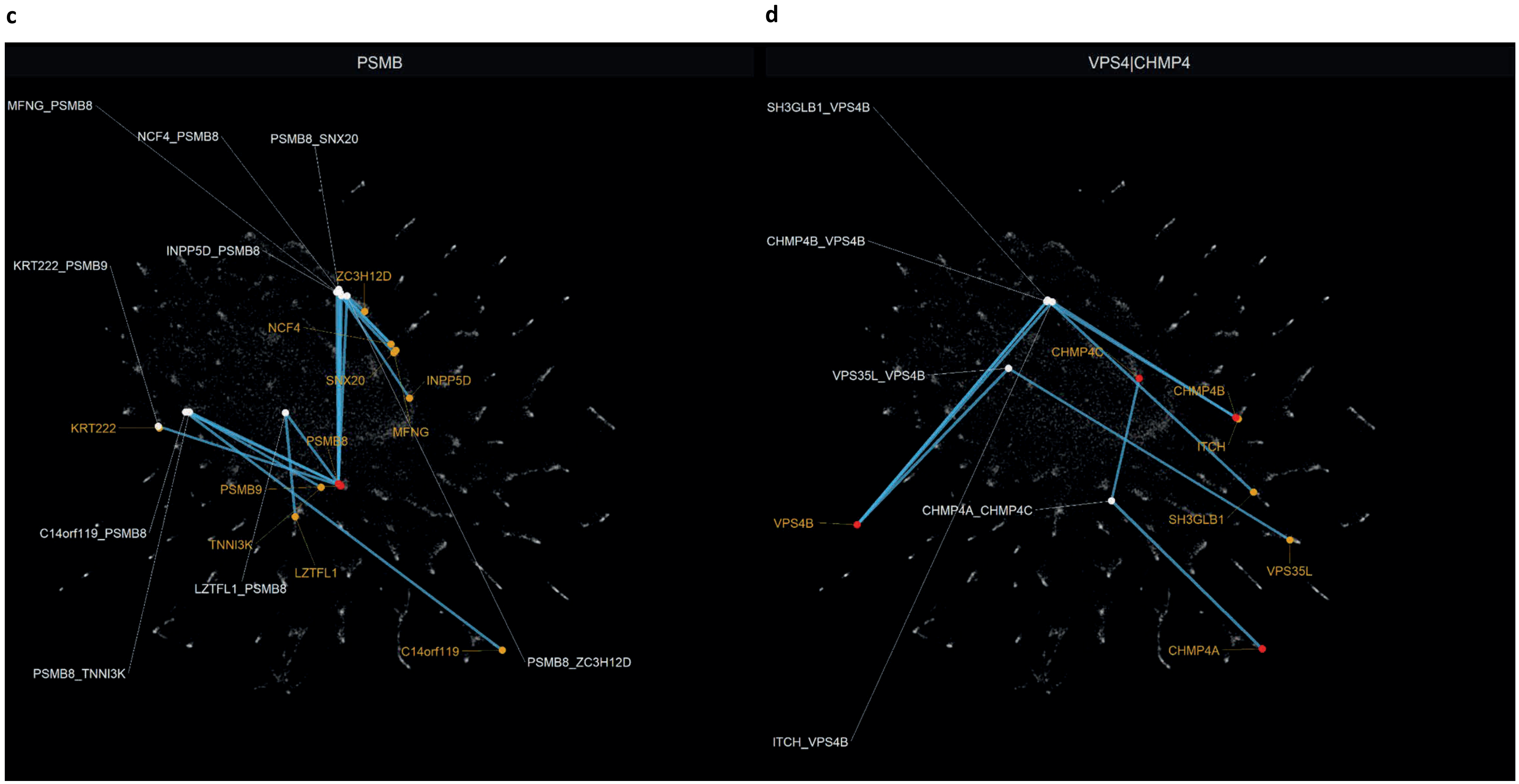


Figure 9. Second-order interactions in the genome-wide interaction network. (c) Mapping of the second-order interactions (white) of PSMB8 and PSMB9 in the genome-wide interaction network of figure 8a. The single genes of the second-order interactions are given in orange. The blue lines are in between the second-order interaction and the single genes it consists of. The red dots represent the location of PSMB8 and PSMB9. **(d)** Mapping of the second-order interactions (white) of VPS4, CHMP4A, CHMP4B, and CHMP4C in the genome-wide interaction network of figure 8a. The single genes of the second-order interactions are given in orange. The blue lines are in between the second-order interaction and the single genes it consists of. The red dots represent the location of VPS4, CHMP4A, CHMP4B, and CHMP4C.

For demonstration, we selected some genes and its interaction partners which were an established second-order GI and located their position in the global GI network. The first gene we selected was CDKN1A, which encodes for the protein Cyclin Dependent Kinase Inhibitor 1A. CDKN1A is a kinase inhibitor involved in the cell cycle progression. In the global GI network, CDKN1A was located at the enriched biological process of *response to radiation* (figure 8a and 9a). Interestingly, the GIs of CDKN1A were clustering in proximity to CDKN1A itself in the network. In this cluster we see the GI between CDKN1A and CDKN2A. The proteins of these two genes have similar functions in the cell cycle regulation and play a role in numerous types of cancer [56]. The two genes encode for the proteins p21 (CDKN1A) and p16 (CDKN2A). Both genes also had overlapping GIs, namely with MDM2, EDA2R and RPL22L1. MDM2 is a negative regulator of the tumour suppressor gene TP53[57] and has been shown that CDKN1A and CDKN2A both interact with MDM2. The p16 protein interacts and downregulates MDM2[58], whereas MDM2 is a negative regulator of p21[59]. However, the other two genes with a second-order interaction with both CDKNs, have not been previously shown to interact with them. RPL22L1 encodes for the Ribosomal Protein L22 Like 1 protein and is a paralog of the RPL22, which is a 60S ribosomal subunit. A recent study linked both paralogs and its pathways to patients with colorectal cancer[60]. The last gene is EDA2R, and this gene encodes for the transmembrane Ectodysplasin A2 Receptor that is involved in NF-kappa-beta and JNK pathways[61].Other genes with an established GI with the CDKN1A gene within the cluster were genes mainly involved in general core cell processes. This includes the amino acid transporter SLC38A6, the polypeptide transferase GALNT6, the fatty acid hydrolase FAAH, and the guanine nucleotide exchange factor ARHGEF19 (figure 9a).

The second example is from CDKN2A, the gene of the p16 protein**,** and CDKN2B, the gene of p15 protein. These two Cyclin Dependent Kinase Inhibitors are both tumour suppressor genes, which regulate the cell cycle. Most of the GIs found with CDKN2A, clustered at two sites of the network (figure 9b). One of the clusters was close to the genes of CDKN2A and CDKN2B, which was also located at the enriched site of *response to radiation* (figure 8a and 9b). In this cluster we found second-order interactions of CDKN2A with multiple genes involved in regulation of the tumour suppressor gene TP53. The MDM2 proto-oncogene encodes for a protein which is a negative regulator of tumour suppressor p53[57]. Another GI partner of CDKN2A with the second-order interaction in the cluster was the Xeroderma pigmentosum, complementation group C (XPC) gene. The XPC protein has been shown that it can degrade p53 via the MDM2 pathway[62]. The gene SESN1 is on the other side regulated by p53[63]. Two other genes with a GI in the same cluster were Ferredoxin Reductase (FDXR) and Spermatogenesis Associated 18 (SPATA18). These two genes are both involved in cellular processes within the mitochondria[64,65].

The other cluster of the second-order interactions was located near the biological process of *response to drugs* and the *epidermis development* (figure 8a and 9b). In this cluster of GIs, the second-order interaction of CDKN2A showed interesting interactions around the retinoblastoma (RB1) gene, although the interaction between CDKN2A and RB1 was located outside this cluster (figure 9b). It has been shown that CDKN2A negatively regulates the RB1 pathway by suppressing CDK4/6[66]. In this cluster four other second-order interactions with CDKN2A were present with more genes that regulate RB1. These four were the E2F3 transcription factor, Cyclin D1/D3 functional subunits (CCND1 and CCND3) and Cyclin Dependent Kinase 6 (CDK6)**.** CCND1 and CCND3 are functional subunits that form a complex together with CDK6 and this complex regulates RB1[67]. E2F3 encodes for a transcription factor by the same name and this protein is directly suppressed by RB1, therefore it is important for cell cycle regulation[68]. Two other genes in the cluster of second-order GIs with CDKN2A were Caveolin 1 and Caveolin 2 (CAV1 and CAV2). These two genes encode for two proteins which form a hetero-oligomeric scaffolding complex. This complex is functional in caveolar membranes across different cell types. It has been suggested that this complex acts as a tumour suppressor, due to its role in the Ras-ERK pathway[69,70]. Another interesting finding was the interaction of CDKN1A and CDKN2A with CDK5 Regulatory Subunit Associated Protein 1 Like 1 (CDKAL1). It is noteworthy that it had a interaction with both CDKN1A and CDKN2A, and that the function of CDKAL1 is not completely understood. It has been shown that it plays a role in type 2 diabetes [71], but it has not been associated with cancer so far.

The third example consists of two genes, the Proteasome 20S Subunit Beta 8 (PSMB8) and PSMB9. Both genes were located in the enriched cluster of *response to virus* (figure 8a and 9c). Both genes encode for subunits of the 20S proteasome complex. Interestingly, a part of the second-order interactions formed a small cluster at the enriched *immune response-activating signal transduction* cluster (figure 8a and 9c), thus a biological process involved within the immune system. The second-order GIs in this cluster consisted of PSMB8 and other genes. Among these other genes is Neutrophil Cytosolic Factor 4 (NCF4). NCF4 encodes for a protein that is part of an enzyme complex involved in the host defence of neutrophils[72]. Another gene is Sorting nexin-20 (SNX20), which is involved in regulation of the innate immune system. A recent study showed that SNX20 is a potential therapeutic target and biomarker in lung adenocarcinomas[73]. The last gene which also formed a GI in the cluster is MFNG, which encodes for the enzyme Beta-1,3-N-acetylglucosaminyltransferase manic fringe, unlike the other partner genes, this enzyme is not involved in the immune system, but in the Notch signalling pathway[74].

As final example, we showed the second-order interactions of VPS4B, CHMP4A, CHMP4B and CHMP4C (figure 9d). The CHMP genes encode for proteins which are part of the ESCRT-III complex. ESCRT-III is part of a bigger complex, namely the endosomal sorting complex required for transport (ESCRT) complex and the ESCRT complex is involved in membrane fission[75,76]. A recent study showed a synthetic lethal pair interaction between VPS4A and VPS4B in absence of SMAD4 and CDH1[77]. In addition, the same study showed that VPS4A and VPS4B are co-essential with CHMP4A. The authors suggested that the VPS4B expression and dependency could be due to a paralog interaction with CHMP1A and CHMP4B[77]. In the GI network map, the VPS4B and CHMP4 genes were located in a different cluster. However, their second-order interactions were located in a small cluster close to the enriched cluster of *response to drug* (figure 8a and 9d). A recent study suggested that ESCRT-III could be exploited as a potential therapeutic target in drug resistant cancer therapy[78]. The second-order interaction between VPS4B and ITCH was also located in this cluster. The ITCH gene encodes for the Itchy E3 ubiquitin ligase which is an important player for the innate immune system. This gene has been of interest as a therapeutic target in different types of cancer[79]. Interestingly, in a VPS4A suppressed cancer cell line, ITCH KO increased the cell survival[77]. The other gene with a second-order interaction with VPS4B is the autophagy related gene SH3GLB1[80].

Discussion

In this study, we showed a robust framework to establish a genome-wide GI network from a pan-cancer CRISPR screen. The network was characterised by biological functional clusters and was based on the regulatory role of gene expression on fitness perturbation in cancer. Furthermore, we found second-order interactions by mining the structures of XGBoost trees and mapping them in the genome-wide network.

Our genome-wide network was focused on interactions between gene expression and the fitness perturbations of a CRISPR screen. The second-order interactions derived from our technique were established between gene expression covariates. We focused on gene expression, due to the regulatory role of gene expression on the fitness of cancer cells and the influence on GIs. Due to the complexity of cancers, gene expression variability occurs between different cancer and normal cells. Overexpression of many genes has been associated with different types of cancers and may act as an oncogene. Regulation of the gene expression takes place through different types of mechanisms, including mutations of oncogenes and tumour suppressor genes. Gene expression variability in cancer cells can also be caused by epigenetic regulation, including DNA methylation and chromatin remodelling[81,82]. Profiling the gene expression yields a high predictiveness for drug-responses and can be used for drug discoveries[83,84]. Transcriptional modulation is thereby important for cancers to become drug-resistance, due to the changes in expression after initial treatment. Therefore, the variability and the transcriptional modulation of GIs and fitness changes in response to gene deletions are of great interest for personalised medicine[85,86].

For finding second-order interactions, we used an Elastic Net regression for preselecting variables before the XGBoost algorithm. Nowadays, high-throughput screens have become the standard in genetic screens. Due to the huge number of genes and relatively low observations, predicting fitness dependencies in large CRISPR screens is a statistical burden. Not only is the set of large variables a statistical concern, but it also comes with a computational challenge. Therefore, we reduced the number of variables and selected the variables with the highest predictive linear value. The preselected variables can still be used for finding non-linear relationships within the XGBoost trees and in addition it improves the computational speed.

We showed that the second-order interactions from mining the XGBoost trees were enriched for PPIs. The enrichment of PPIs was also improved by running multiple rounds of random XGBoost models under random subsampling and binning these models. In this study we ran 40 random models with 50% subsampling, but it could be interesting to see if including more XGBoost models would improve the PPI enrichment. The interactions established from the XGBoost trees showed a negative correlation between PPI enrichment and the test RMSE. However, there was only a correlation between them if the models were ranked on their overall score (sumGain and frequency) and not ranked per CRISPR gene target. This implies that not all CRISPR targets have many predictive regulatory genes for their fitness after preselecting variables with Elastic Net. Alternatively, the fitness spectrum of these genes may be difficult to predict and require more data due to low signal-to-noise ratio. Genes with a lower number of GIs were associated with more tissue-specific processes, and thus may require more data or more experimental variation in the culturing conditions for the CRISPR screens (figure 4b). Indeed, a recent study showed that novel fitness dependencies tend to arise in multicellular culture environments[87].

In continuation of this study, we selected the CRISPR targets with the most predictive fitness spectrums. Therefore, doing our final run**,** we set a cut-off based on the test RMSE. This allowed us to look at the CRISPR targets with the most reliable interaction structures between regulatory genes and reduce the computational burden. Furthermore, the hyperparameter optimization indicated that optimal detection of GIs (assessed by PPI enrichment) was associated with a minimization of model overfitting. This could suggest that the most profound higher-order interactions in fitness perturbation are predicted from models that can generalise well across the pan-cancer cell library. A recent study by Costanzo and his colleagues, showed that in yeast most GIs in a global GI network remains the same in different environmental conditions[23].

With our technique we found potential second-order GIs and showed a few examples of the interactions within the genome-wide network. The second-order interactions of the CDKN2A, CDKN2B and CDKN1B formed different distinct clusters. These clusters were in proximity of other functionally enriched clusters, which could indicate a similar biological function. Another possibility could be that these higher-order interactions form their own functionally distinct clusters. Characterizing these clusters could show how genes from different processes of the network could contribute to emergent functions through the means of higher-order interactions. Notably, the distribution of higher-order interactions in regions predominantly associated with tissue-specific functions could suggest a path for mapping how complex interactions between genes yield the vast amount functional variability seen in different human cells.

A utility of our genome-wide GI network is to characterise genes for their function. The function of many genes is still poorly understood. Previous work in genome-wide GI networks showed that genes cluster together with similar functionality and could be used to assign functions to uncharacterized genes[9,25,54]. Therefore, it would be interesting to use our method and look more into depth of a certain biological process within the network to characterise genes which are still poorly understood.

It would be also interesting to validate potential findings of higher-order interactions with our technique. Validation of higher-order interactions involved in cancer fitness would be of great interest for personalised treatments. Treatment could focus on multiple drugs targeting different genes involved in the same biological processes, also known as combination drug therapy. Combination drug therapy is already important in the clinic for many types of cancer[88] and could be exploited even further. Besides the potential benefits for improving cancer treatments. Our research could contribute to the understanding of the complexity of genetics, and how it contributes to variations in drug sensitivity in different cancer subtypes. One of the key challenges in the field of genetics remains the ability of genes to create a wide variety of differential tissues and cells[54,89].

For further development of this technique, more genetic characteristics of the cell lines should be included to find additional higher-order fitness dependencies. These should be gain-of-function mutations, gene fusions, loss-of-functions, and gene expression. Due to the limitations of this study, we mainly focused on the gene expression as a proof of concept of our method.

Many recent studies focused on synthetic lethal interactions from large CRISPR screens and novel interactions are investigated as potential drug targets[13]. In this study, we also looked at synthetic lethal interactions after establishing a LOF dataset from mutations and gene copy numbers data. We selected only deleterious mutations which were present in more than 1% of the cancer cell lines to obtain sufficient coverage for the regression. The results from the ROC curve of the potential synthetic lethal interactions showed that when adjusting for FDR coming from the coverage and distribution of mutations in the cancer library, L1-penalised linear regression serves as a robust method for detecting true positive interactions. Therefore, it would be interesting to investigate the top hits from the Lasso as potential synthetic lethal interactions. For further development of this technique, we could include more mutations and explore the higher-order dependencies of synthetic lethality on gene expression variation.

In conclusion, our work showed a framework for establishing a genome-wide GI network with functional clusters. We also showed a method for predicting higher-order GIs and this framework can be further developed and be used to gain a better understanding of complex GIs and its regulation of fitness in cancer cells.

Literature

1. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

2. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).

3. Francies, H. E., McDermott, U. & Garnett, M. J. Genomics-guided pre-clinical development of cancer therapies. *Nature Cancer* **1**, 482–492 (2020).

4. Chin, L., Andersen, J. N. & Futreal, P. A. Cancer genomics: from discovery science to personalized medicine. *Nature Medicine* **17**, 297–303 (2011).

5. Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer medicine. *Nature Reviews Clinical Oncology* **15**, 353–365 (2018).

6. Du, W. & Elemento, O. Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. *Oncogene* **34**, 3215–3225 (2015).

7. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).

8. Boucher, B. & Jenna, S. Genetic interaction networks: better understand to better predict. *Frontiers in Genetics* **4**, (2013).

9. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).

10. Bridges, C. B. The Origin of Variations in Sexual and Sex-Limited Characters. *The American Naturalist* **56**, 51–63 (1922).

11. Dobzhansky, T. Genetics of natural populations; recombination and variability in populations of Drosophila pseudoobscura. *Genetics* **31**, 269–290 (1946).

12. Hartwell, L. H., Szankasi, P., Roberts, C. J., Murray, A. W. & Friend, S. H. Integrating genetic approaches into the discovery of anticancer drugs. *Science* **278**, 1064–1068 (1997).

13. Huang, A., Garraway, L. A., Ashworth, A. & Weber, B. Synthetic lethality as an engine for cancer drug target discovery. *Nature Reviews Drug Discovery* **19**, 23–38 (2020).

14. O'Neil, N. J., Bailey, M. L. & Hieter, P. Synthetic lethality and cancer. *Nature Reviews Genetics* **18**, 613–623 (2017).

15. Chen, C.-C., Feng, W., Lim, P. X., Kass, E. M. & Jasin, M. Homology-Directed Repair and the Role of BRCA1, BRCA2, and Related Proteins in Genome Integrity and Cancer. *Annual Review of Cancer Biology* **2**, 313–336 (2018).

16. Ashworth, A. & Lord, C. J. Synthetic lethal therapies for cancer: what's next after PARP inhibitors? *Nature Reviews Clinical Oncology* **15**, 564–576 (2018).

17. Fong, P. C. *et al.* Inhibition of Poly(ADP-Ribose) Polymerase in Tumors from BRCA Mutation Carriers. *New England Journal of Medicine* **361**, 123–134 (2009).

18. Mengwasser, K. E. *et al.* Genetic Screens Reveal FEN1 and APEX2 as BRCA2 Synthetic Lethal Targets. *Molecular Cell* **73**, 885-899.e6 (2019).

19. Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* **568**, 511–516 (2019).

20. Chan, E. M. *et al.* WRN helicase is a synthetic lethal target in microsatellite unstable cancers. *Nature* **568**, 551–556 (2019).

21. Nie, M. *et al.* Genome-wide CRISPR screens reveal synthetic lethal interaction between CREBBP and EP300 in diffuse large B-cell lymphoma. *Cell Death & Disease* **12**, 419 (2021).

22. Setton, J. *et al.* Synthetic Lethality in Cancer Therapeutics: The Next Generation. *Cancer Discovery* **11**, 1626–1635 (2021).

23. Costanzo, M. *et al.* Environmental robustness of the global yeast genetic interaction network. *Science* **372**, eabf8424 (2021).

24. Kuzmin, E. *et al.* Systematic analysis of complex genetic interactions. *Science* **360**, eaao1729 (2018).

25. Horlbeck, M. A. *et al.* Mapping the Genetic Landscape of Human Cells. *Cell* **174**, 953-967.e22 (2018).

26. Han, K. *et al.* Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nature biotechnology* **35**, 463–474 (2017).

27. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

28. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

29. Dempster, J. M. *et al.* Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. *Nature Communications* **10**, 5817 (2019).

30. Chen, F. *et al.* Moving pan-cancer studies from basic research toward the clinic. *Nature Cancer* **2**, 879–890 (2021).

31. Basu, S., Kumbier, K., Brown, J. B. & Yu, B. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences* **115**, 1943 (2018).

32. Park, S., Supek, F. & Lehner, B. Higher order genetic interactions switch cancer genes from two-hit to one-hit drivers. *Nature Communications* **12**, 7051 (2021).

33. DepMap, B. DepMap 21Q1 Public. (2021) doi:10.6084/m9.figshare.13681534.v1.

34. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34**, 184–191 (2016).

35. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics* **49**, 1779–1784 (2017).

36. Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular Systems Biology* **10**, 733 (2014).

37. Dempster, J. M. *et al.* Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. *bioRxiv* 720243 (2019) doi:10.1101/720243.

38. DepMap, B. DepMap 21Q4 Public. (2021) doi:10.6084/m9.figshare.16924132.v1.

39. Dempster, J. M. *et al.* Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biology* **22**, 343 (2021).

40. RStudio Team. RStudio: Integrated Development Environment for R. (2021).

41. R Core Team. R: A language and environment for statistical computing. Vienna (2021).

42. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1–22 (2010).

43. Wang, J. *et al.* SynLethDB 2.0: A web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. *bioRxiv* 2021.12.28.474346 (2021) doi:10.1101/2021.12.28.474346.

44. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, (2021).

45. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281 (2007).

46. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 785–794 (2016).

47. Maksymiuk, S., Karbowiak, E. & Biecek, P. EIX: Explain Interactions in "XGBoost." (2021).

48. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

49. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607–D613 (2019).

50. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

51. Lu, J. *et al.* Causal network inference from gene transcriptional time-series response to glucocorticoids. *PLoS computational biology* **17**, e1008223–e1008223 (2021).

52. Shojaie, A. & Michailidis, G. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics* **26**, i517–i523 (2010).

53. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances* **5**, eaau4996–eaau4996 (2019).

54. Wainberg, M. *et al.* A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. *Nature Genetics* **53**, 638–649 (2021).

55. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).

56. Zhao, R., Choi, B. Y., Lee, M.-H., Bode, A. M. & Dong, Z. Implications of Genetic and Epigenetic Alterations of CDKN2A(p16INK4a) in Cancer. *eBioMedicine* **8**, 30–39 (2016).

57. Michael, D. & Oren, M. The p53–Mdm2 module and the ubiquitin system. *Seminars in Cancer Biology* **13**, 49–58 (2003).

58. Pomerantz, J. *et al.* The Ink4a Tumor Suppressor Gene Product, p19Arf, Interacts with MDM2 and Neutralizes MDM2's Inhibition of p53. *Cell* **92**, 713–723 (1998).

59. Zhang, Z. *et al.* MDM2 Is a Negative Regulator of p21WAF1/CIP1, Independent of p53 *. *Journal of Biological Chemistry* **279**, 16000–16006 (2004).

60. Rao, S. *et al.* RPL22L1 induction in colorectal cancer is associated with poor prognosis and 5-FU resistance. *PLOS ONE* **14**, e0222392- (2019).

61. Sinha, S. K., Zachariah, S., Quiñones, H. I., Shindo, M. & Chaudhary, P. M. Role of TRAF3 and -6 in the Activation of the NF-3B and JNK Pathways by X-linked Ectodermal Dysplasia Receptor *. *Journal of Biological Chemistry* **277**, 44953–44961 (2002).

62. Krzeszinski, J. Y. *et al.* XPC promotes MDM2-mediated degradation of the p53 tumor suppressor. *Molecular Biology of the Cell* **25**, 213–221 (2013).

63. Budanov, A. v & Karin, M. p53 Target Genes Sestrin1 and Sestrin2 Connect Genotoxic Stress and mTOR Signaling. *Cell* **134**, 451–460 (2008).

64. Shi, Y., Ghosh, M., Kovtunovych, G., Crooks, D. R. & Rouault, T. A. Both human ferredoxins 1 and 2 and ferredoxin reductase are important for iron-sulfur cluster biogenesis. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1823**, 484–492 (2012).

65. Dan, X. *et al.* DNA damage invokes mitophagy through a pathway involving Spata18. *Nucleic Acids Research* **48**, 6611–6623 (2020).

66. Witkiewicz, A. K., Knudsen, K. E., Dicker, A. P. & Knudsen, E. S. The meaning of p16ink4a expression in tumors. *Cell Cycle* **10**, 2497–2503 (2011).

67. Sherr, C. J., Beach, D. & Shapiro, G. I. Targeting CDK4 and CDK6: From Discovery to Therapy. *Cancer Discovery* **6**, 353–367 (2016).

68. Chen, H.-Z., Tsai, S.-Y. & Leone, G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nature Reviews Cancer* **9**, 785–797 (2009).

69. Volonte, D. *et al.* Caveolin-1 promotes the tumor suppressor properties of oncogene-induced cellular senescence. *Journal of Biological Chemistry* **293**, 1794–1809 (2018).

70. Kortum, R. L. *et al.* Caveolin-1 is required for kinase suppressor of Ras 1 (KSR1)-mediated extracellular signal-regulated kinase 1/2 activation, H-RasV12-induced senescence, and transformation. *Molecular and cellular biology* **34**, 3461–3472 (2014).

71. Palmer, C. J. *et al.* Cdkal1, a type 2 diabetes susceptibility gene, regulates mitochondrial function in adipose tissue. *Molecular metabolism* **6**, 1212–1225 (2017).

72. Lehman, H. K. & Segal, B. H. The role of neutrophils in host defense and disease. *Journal of Allergy and Clinical Immunology* **145**, 1535–1544 (2020).

73. Wu, G. J. *et al.* SNX20 Expression Correlates with Immune Cell Infiltration and Can Predict Prognosis in Lung Adenocarcinoma. *International journal of general medicine* **14**, 7599–7611 (2021).

74. Castro, R. C., Gonçales, R. A., Zambuzi, F. A. & Frantz, F. G. Notch signaling pathway in infectious diseases: role in the regulation of immune response. *Inflammation Research* **70**, 261–274 (2021).

75. Christ, L., Raiborg, C., Wenzel, E. M., Campsteijn, C. & Stenmark, H. Cellular Functions and Molecular Mechanisms of the ESCRT Membrane-Scission Machinery. *Trends in Biochemical Sciences* **42**, 42–56 (2017).

76. McCullough, J., Frost, A. & Sundquist, W. I. Structures, Functions, and Dynamics of ESCRT-III/Vps4 Membrane Remodeling and Fission Complexes. *Annual Review of Cell and Developmental Biology* **34**, 85–109 (2018).

77. Neggers, J. E. *et al.* Synthetic Lethal Interaction between the ESCRT Paralog Enzymes VPS4A and VPS4B in Cancers Harboring Loss of Chromosome 18q or 16q. *Cell Reports* **33**, (2020).

78. Liu, J., Kang, R. & Tang, D. ESCRT-III-mediated membrane repair in cell death and tumor resistance. *Cancer Gene Therapy* **28**, 1–4 (2021).

79. Yin, Q., Wyatt, C. J., Han, T., Smalley, K. S. M. & Wan, L. ITCH as a potential therapeutic target in human cancers. *Seminars in Cancer Biology* **67**, 117–130 (2020).

80. Takahashi, Y., Meyerkord, C. L. & Wang, H.-G. Bif-1/Endophilin B1: a candidate for crescent driving force in autophagy. *Cell Death & Differentiation* **16**, 947–955 (2009).

81. Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics* **3**, 415–428 (2002).

82. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics* **43**, 768–775 (2011).

83. Iorio, F. *et al.* Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences* (2010).

84. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology* **32**, 1202–1212 (2014).

85. Chen, B. *et al.* Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nature Communications* **8**, 16022 (2017).

86. van 't Veer, L. J. & Bernards, R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* **452**, 564–570 (2008).

87. Han, K. *et al.* CRISPR screens in cancer spheroids identify 3D growth-specific vulnerabilities. *Nature* **580**, 136–141 (2020).

88. Bayat Mokhtari, R. *et al.* Combination therapy in combating cancer. *Oncotarget* **8**, 38022–38043 (2017).

89. Chuang, H.-Y., Hofree, M. & Ideker, T. A Decade of Systems Biology. *Annual Review of Cell and Developmental Biology* **26**, 721–744 (2010).

Supplement 1

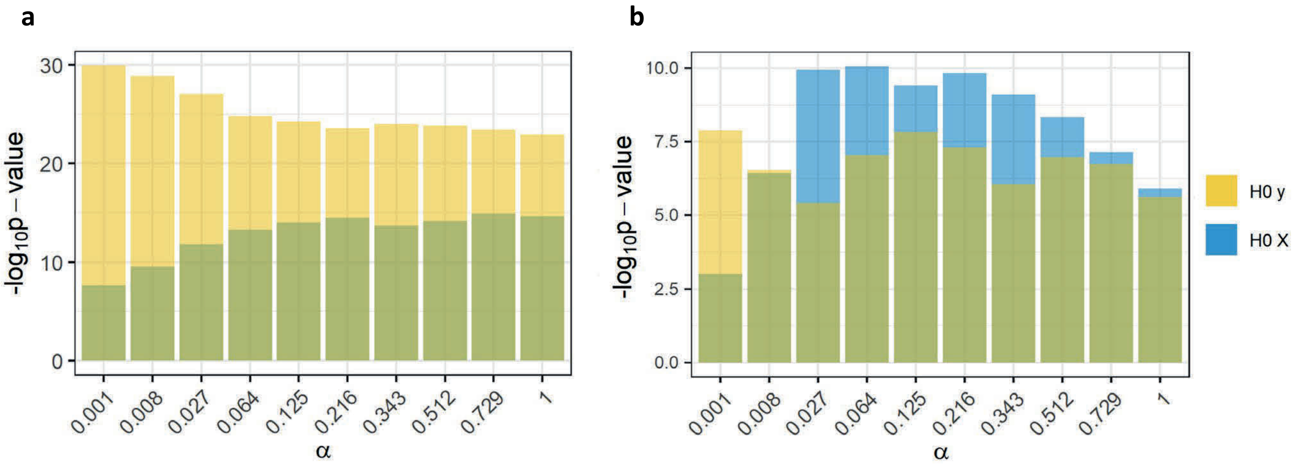


Figure 1. Statistics of the multivariate penalised linear regression cross-validation. (a) Paired Wilcoxon-signed rank test for the coefficients obtained from the different Elastic Net-regularisation (0 < α < 1) and the Lasso-regularisation (α = 1) CV regressions. The coefficients obtained from the true regressions with the gene expression as independent variables were compared to the coefficients of their estimated null distributions of H0 y (yellow) and H0 X (blue). **(b)** Paired Wilcoxon-signed rank test for the coefficients obtained from the different Elastic Net-regularisation and the Lasso-regularisation CV regressions. The coefficients obtained from the true regressions with the LOF as independent variables were compared to the coefficients of their estimated null distributions of H0 y (yellow) and H0 X (blue).

Supplement 2

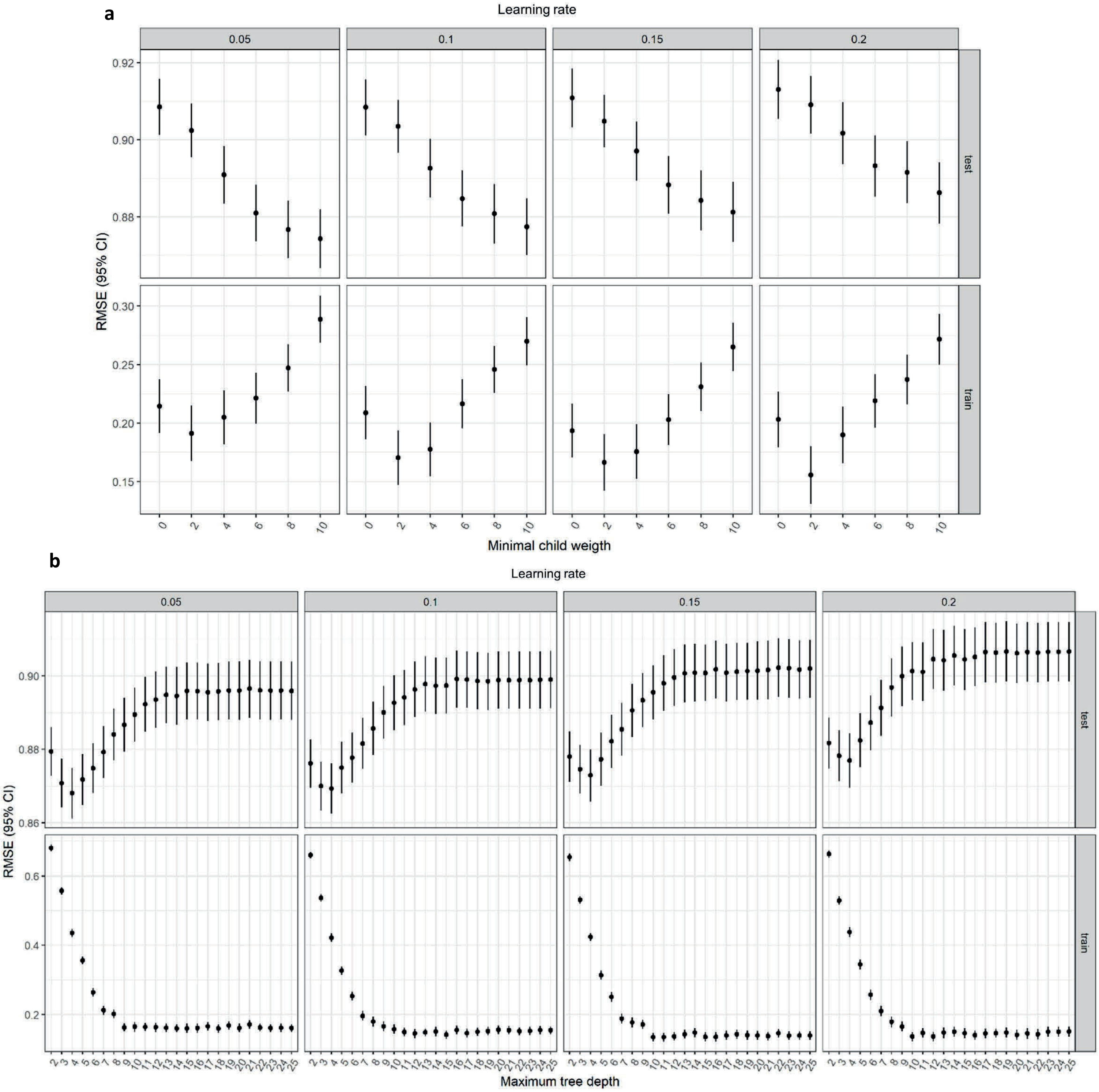


Figure 1. XGBoost hyperparameter search. (a) Hyperparameter search for the learning rate and minimal child weight. **(b)** Hyperparameter search for the learning rate and the maximum tree depth. The round mean squared error (RMSE) and the 95% confidence interval (black lines) are shown for the test and training data of the XGBoost models.

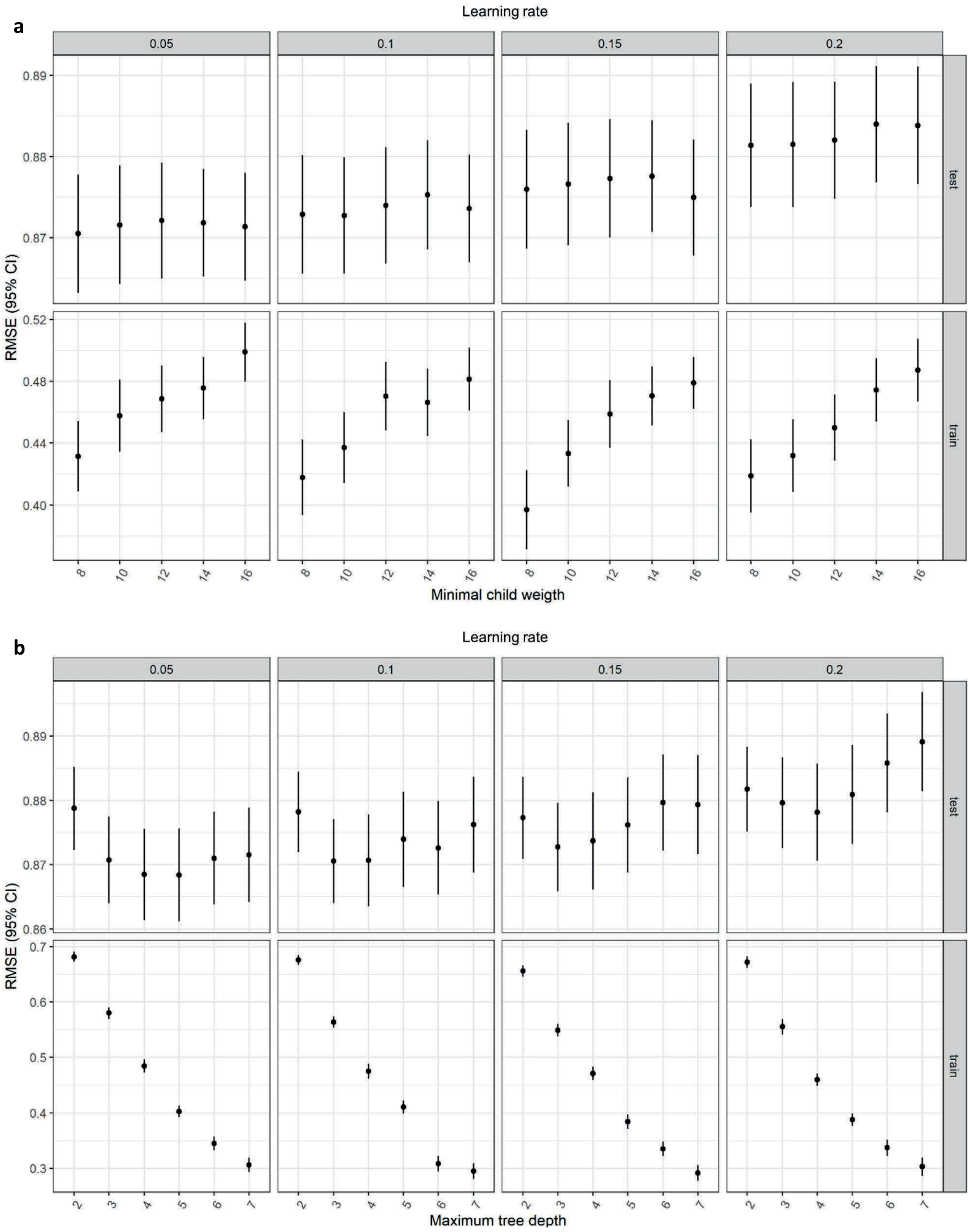


Figure 2. XGBoost hyperparameter search. (a) Hyperparameter search for the learning rate and minimal child weight. **(b)** Hyperparameter search for the learning rate and the maximum tree depth. The round mean squared error (RMSE) and the 95% confidence interval (black lines) are shown for the test and training data of the XGBoost models.

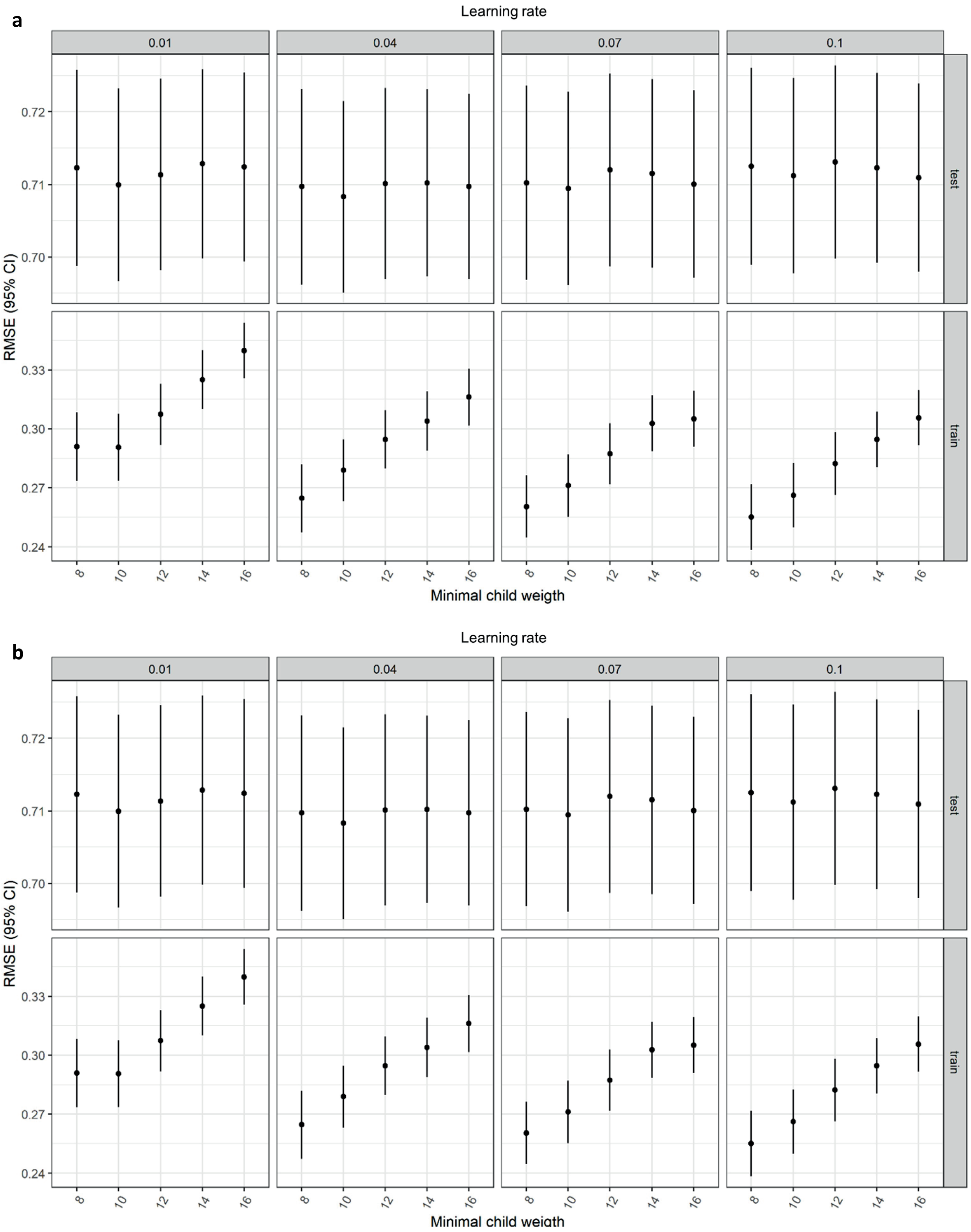


Figure 3. XGBoost hyperparameter search. (a) Hyperparameter search for the learning rate and minimal child weight. **(b)** Hyperparameter search for the learning rate and the maximum tree depth. The round mean squared error (RMSE) and the 95% confidence interval (black lines) are shown for the test and training data of the XGBoost models.

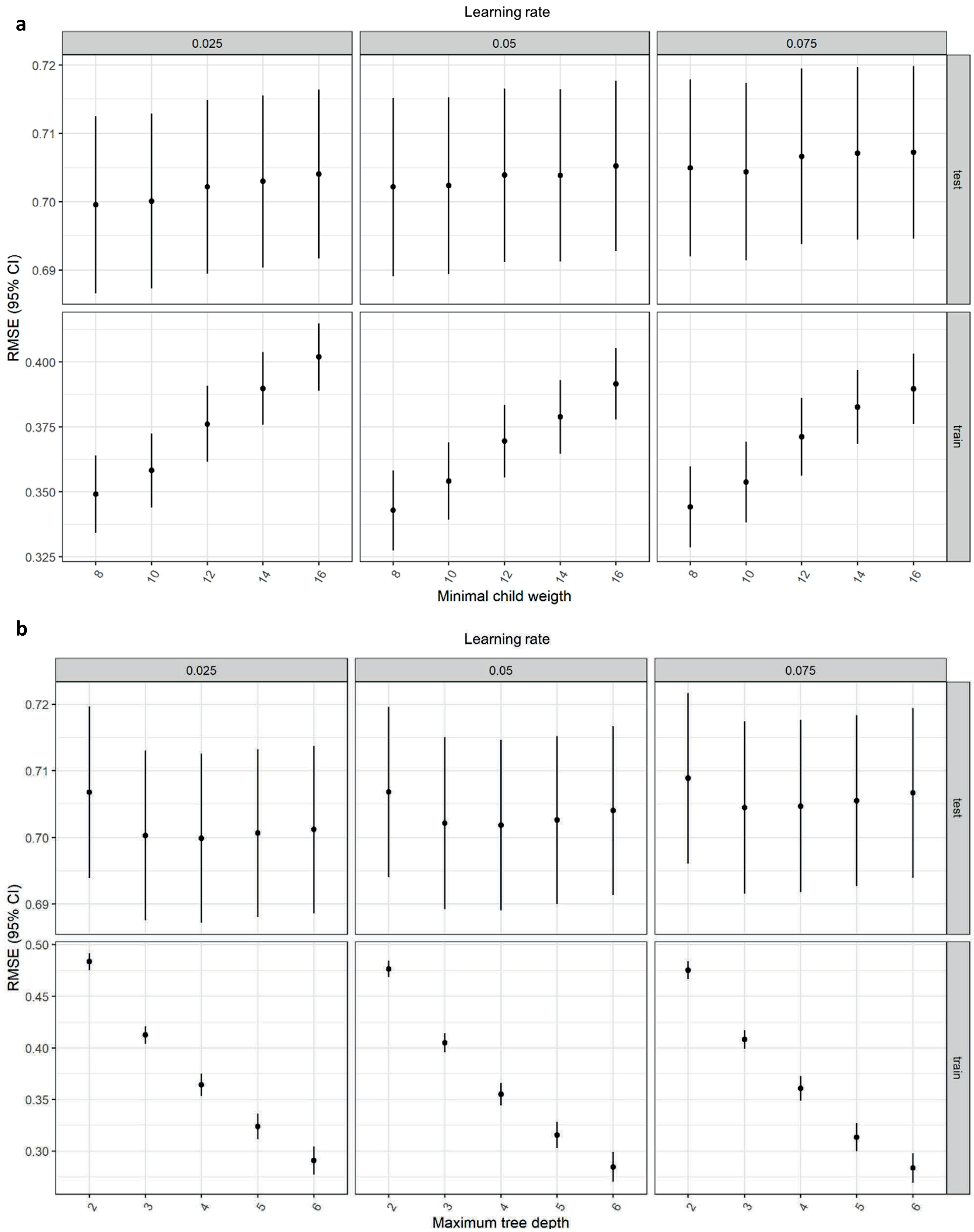


Figure 4. XGBoost hyperparameter search. (a) Hyperparameter search for the learning rate and minimal child weight. **(b)** Hyperparameter search for the learning rate and the maximum tree depth. The round mean squared error (RMSE) and the 95% confidence interval (black lines) are shown for the test and training data of the XGBoost models.

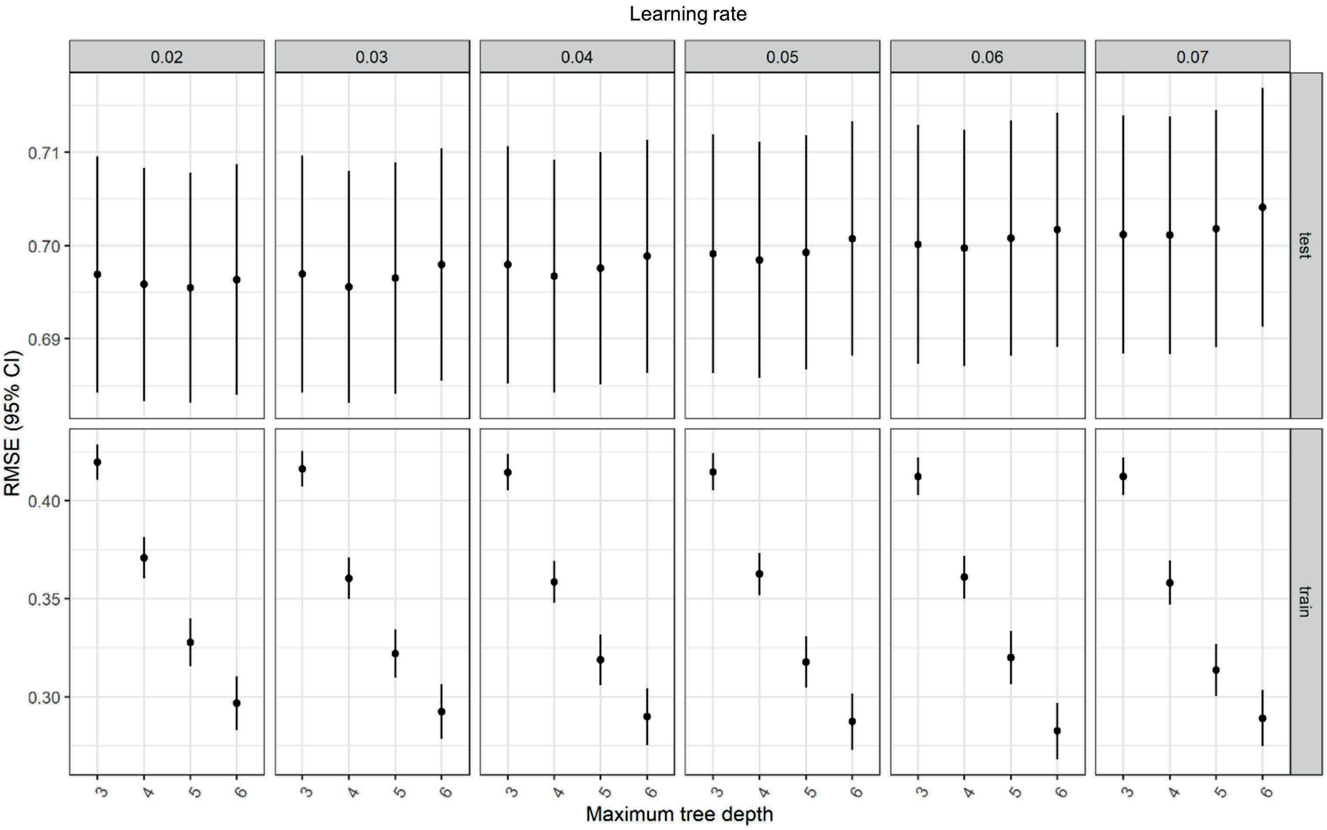


Figure 5. XGBoost hyperparameter search. Hyperparameter search for the learning rate and the maximum tree depth. The round mean squared error (RMSE) and the 95% confidence interval (black lines) are shown for the test and training data of the XGBoost models.

Supplement 3

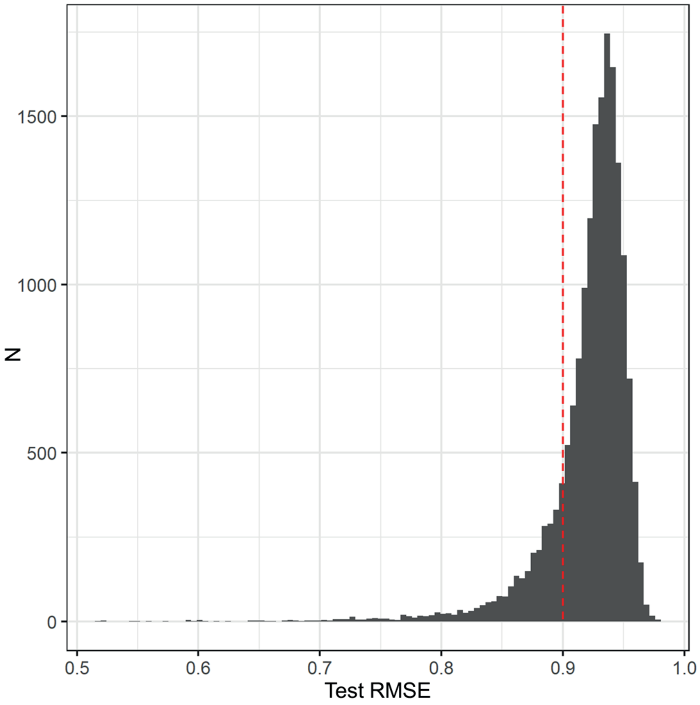


Figure 1. Histogram of the whole-genome XGBoost cross-validation run. The red dotted line indicates the cut-off for the test round mean squared error (RMSE).