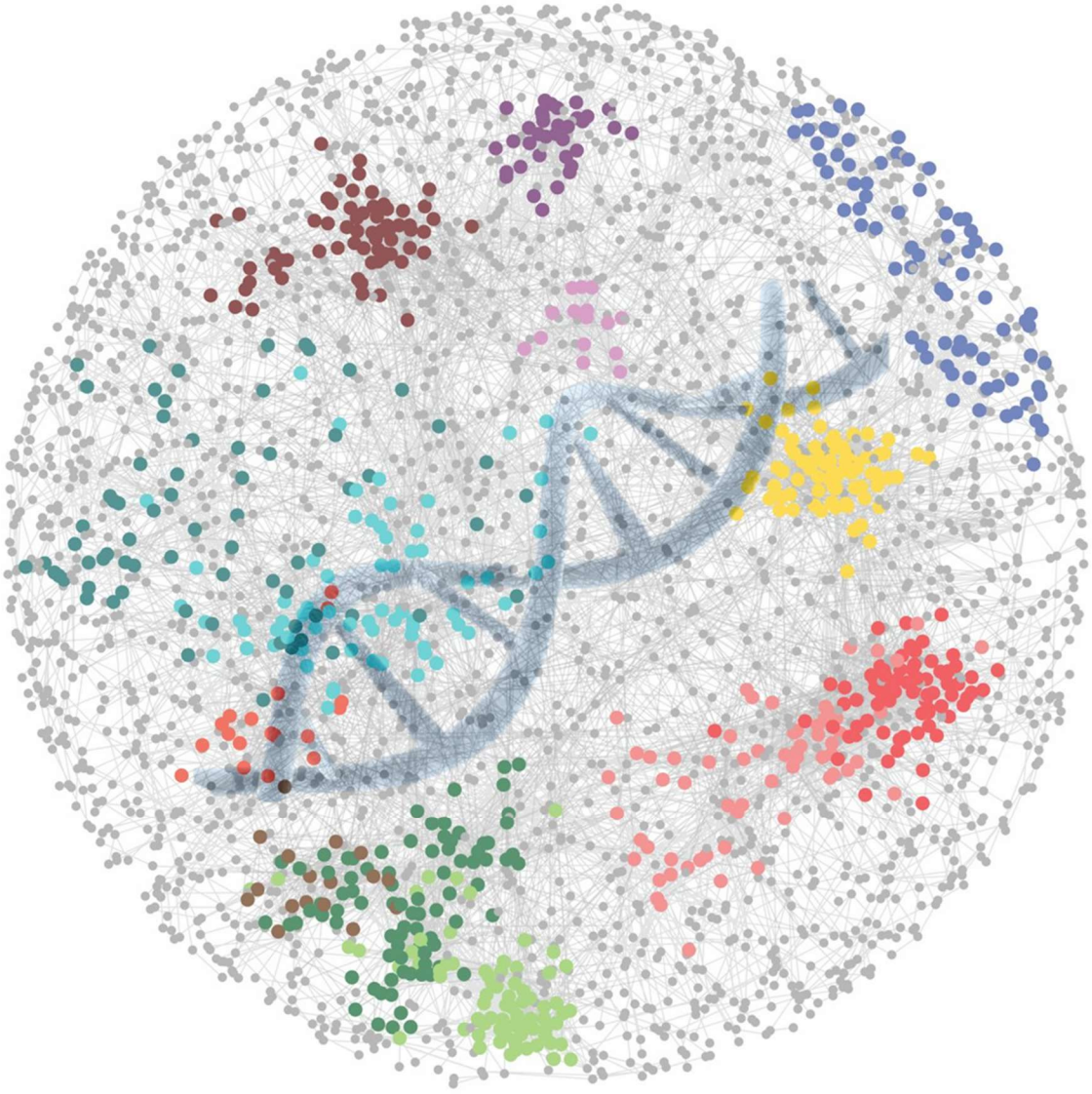
Algorithms for identifying eQTL modulated interactions between genes

January – July 2021 *Benjan Karnebeek (s3157318)*



Prof. dr. O.P. Kuipers (examiner)
P. Deelen (daily supervisor)
Research Group: Department of Genetics UMCG

Picture composed using pictures from https://www.parool.nl/nederland/dna-van-duizenden-strafzaken-opnieuw-onder-de-loep~b383b6ce/?referrer=https%3A%2F%2Fwww.google.com%2F and https://www.eurekalert.org/pub_releases/2019-03/uot-ugi032219.php.

Abstract:

A Gene Regulatory Network (GRN) is a network of genes regulating each other, mostly through transcription factors. Expression Quantitative Trait Loci (eQTLs) also affect expression of genes. Recently, a type of interaction was identified where the Cis-eQTLs of a gene affected the binding affinity of a transcription factor. The aim of this project was to design algorithms which would identify more of those interactions using genotype data and gene expression data, so that a GRN may be reconstructed from those interactions. The results of the algorithms were not replicated in experimental data, indicating the need for further development.

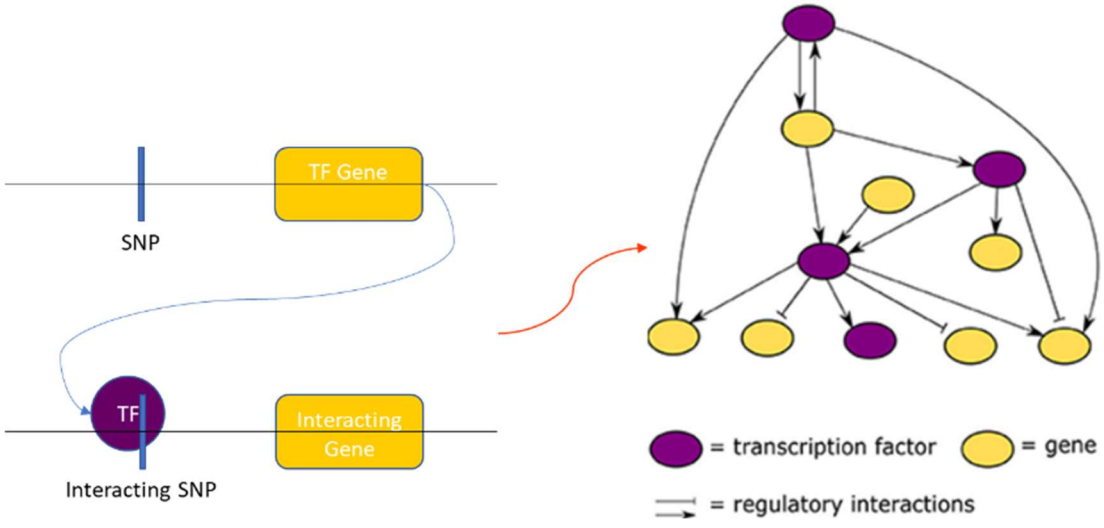
Inhoud

Introduction

Since the discovery of gene inheritance by Mendel[1,2], there has been research into how exactly alleles affect various traits[3]. After 2006, Genome Wide Association Studies (GWAS) became prevalent, entire genomes were sequenced revealing countless alleles associated with diseases[4]. This also revealed that most inherited diseases are complex diseases and do not follow mendelian patterns of inheritance. The disease associated alleles only slightly increase or decrease the risk for one or more complex diseases[3].

Disease associated alleles are most often Single-Nucleotide Polymorphisms (SNPs) found in noncoding regions of the genome, and therefore most likely affect the level of gene expression instead of gene function[5]. Most of these SNPs are expression Quantitative Trait Loci (eQTLs) which are loci that explain a fraction of the genetic variance of a gene expression phenotype.
In literature, eQTLs are typically divided in Cis-eQTLs, which are in close proximity to the gene they regulate, and Trans eQTLs which are further away or even on a different chromosome. Trans-eQTLs regulate genes indirectly i.e., through a transcription factor, this means they are often also Cis-eQTLs for another gene. Standard eQTL analysis involves a direct association test between markers of genetic variation and gene expression levels and is typically measured in a large group of individuals. Most notably, no knowledge about regulating regions is required for the GWAS based identification of new eQTLs[5].

Interestingly, most traits and the expression of most genes are influenced by many causal SNPs with small effect sizes[3,6]. Most of these causal SNPs are spread across the genome and are not located near genes with disease related functions. These observations lend to the hypothesis of the "omnigenic" model[3], which states that all Gene Regulatory Networks (GRNs)[7] are highly interconnected with each other. The Cis-eQTL of one Transcription Factor (TF) gene can also be the Trans-eQTL of another gene interacting with the TF. In the "omnigenic" model, these interacting genes can be located anywhere in the genome and can have functions seemingly unrelated to that of the TF gene. According to model GRNs should contain "core genes", nodes in the GRNs which are central parts in the regulation process according to this model. These node would either regulate a lot of other genes, be regulated by a lot of other genes, or both[3,8]. Nodes completely upstream in the gene regulatory network, and thus not subject to external regulation, should also exist within GRNs. These upstream nodes would be good places to start untangling the gene regulatory network, and eQTLs could be used to determine interactions between genes[8].



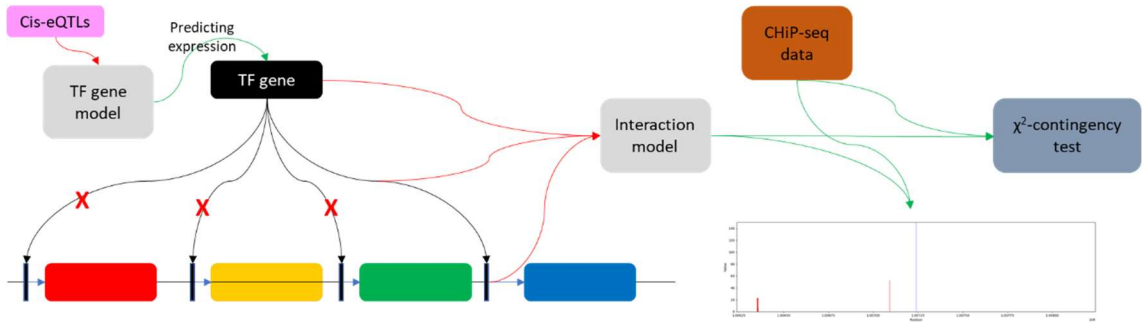


Figure 1: Overview of concepts and research project. **Top,** The main concept behind the research project; using interactions where the effect of the transcription factor is modulated by a Cis-eQTL SNP[9], to reconstruct Gene Regulatory Networks (GRNs). Right hand side of the picture is adapted from *Vandereyken et al., 2018*[10]. **Bottom,** Approach of the research project; generating interaction models for the interactions a TF gene has with the Cis-eQTL SNP of a genes, and selecting the best model. For that model use CHiP-seq data to generate peak plots and perform a chi test. Expression of TF gene is predicted using only Cis-eQTL SNPs.

Zhernakova et al. 2017[9] identified Cis-eQTL SNPs using a hypothesis free strategy. Some of these Cis-eQTL SNPs modulate the binding of a TF is to the interacting gene[11,12], thus there the interaction between the SNP and the TF is what controls the expression of the interacting gene. If all these interactions between the genes/nodes in the GRN are identified and quantified, then a GRN could be reconstructed as series of equations. This network of equations could be used to make accurate predictions; e.g., how the GRN would respond to medication, or how disruption in the network can lead to pathology.

The aim of this project was to develop new algorithm to identify Cis-eQTL modulated interactions between two genes through predicting the expression of those genes. Interactions between two genes are modeled using genotype data of Cis-eQTL SNPs and gene expression data. The expression of the TF gene being predicted based on the Cis-eQTL SNPs alone, the expression of the interacting gene is predicted based on the predicted expression of the TF gene. These models are then evaluated to determine if such an interaction is likely to occur in vivo. As far as I know currently no algorithms exist which predict interactions between genes using only Cis-eQTLs.
Algorithms for identifying Cis-eQTLs were developed to aid in achieving the aim of the project. These algorithms use the residuals of each previously identified Cis-eQTL SNP for the identification of the next Cis-eQTL SNP. The idea was to correct this way for Linkage Disequilibrium (LD)[13] between multiple Cis-eQTLs. A pipeline from the eQTLgen consortium[14,15] was used for validation of the results of the Cis-eQTL identification algorithms.

None of the top potential interactions were replicated in CHiP-seq data[16,17] and any predictions using identified Cis-eQTLs showed a high probability that any association was due to chance. This indicates that the algorithms need to be developed further or that the used data needs further preprocessing to correct for nongenetic differences between the samples.

Materials and methods

See supplementary files for scripts themselves. Scripts are also available on GitHub:
https://github.com/molgenis/GeneticsPrivateScripts/tree/master/umcg-bkarnebeek

Cohort descriptions

In this project a total of six cohorts were used, these are collectively called the BIOS cohorts because all of the six cohorts are part of the BIOS cosortium[14]. The cohorts are described below.

CODAM

As described by *Zhernakova et al. 2017*[9], the Cohort on Diabetes and Atherosclerosis Maastricht (CODAM)[9,14,18] consists of: *"[..] a selection of 547 subjects from a larger population-based cohort.*

Inclusion of subjects into CODAM was based on a moderately increased risk of developing cardiometabolic diseases such as type 2 diabetes and/or cardiovascular disease. Subjects were included if they were of European descent, over 40 years of age and additionally met at least one of the following criteria: increased body mass index (BMI; >25), a positive family history of type 2 diabetes, a history of gestational diabetes and/or glycosuria, or use of antihypertensive medication."

LLD

The LifeLines-DEEP (LLD)[9,14,19] cohort is a subcohort of the LifeLines[20] cohort with 1,500 participants for which additional molecular data is available. As explained by *Zhernakova et al. 2017*[9]: *"LifeLines is a multidisciplinary prospective population-based cohort study examining the health and health-related behaviors of 167,729 individuals living in the northern parts of the Netherlands using a unique three-generation design. It employs a broad range of investigative procedures assessing the biomedical, sociodemographic, behavioral, physical and psychological factors contributing to health and disease in the general population, with a special focus on multi-morbidity and complex genetics."*

LLS

As stated by *Zhernakova et al. 2017*[9], the Leiden Longevity Study (LLS)[9,14,21] aims to: *"[..]identify genetic factors influencing longevity and examine their interaction with the environment to develop interventions by which to increase health at older ages. To this end, long-lived siblings of European descent were recruited together with their offspring and their offspring's partners, on the condition that at least two long-lived siblings were alive at the time of ascertainment. For men, the age criterion was 89 years or older; for women, the age criterion was 91 years or older. These criteria led to the ascertainment of 944 long-lived siblings from 421 families, together with 1,671 of their offspring and 744 partners."*

RS

The Rotterdam Study (RS)[9,14,22] is a single-center, prospective population-based cohort study conducted in Rotterdam, the Netherlands. As explained by *Zhernakova et al. 2017*[9]: *"The cohort has data on 14,926 subjects (both male and female) aged 45 years or older. The main objective is the investigation of the prevalence and incidence of chronic diseases as well as the prevalence and incidence of risk factors for those diseases. This is to improve prevention and treatment of these diseases in the elderly."*

NTR

The Netherlands Twin Register (NTR)[14,23] is a national register in which twins, multiples and their parents, siblings, spouses and other family members participate. A stated in the 2019 report on the register: *"Since the early 1980s, the NTR has enrolled around 120,000 twins and a roughly equal number of their relatives. The majority of twin families have participated in survey studies, and subsamples took part in biomaterial collection (e.g. DNA) and dedicated projects, for example, for neuropsychological, biomarker and behavioral traits. The recruitment into the NTR is all inclusive without any restrictions on enrollment. These resources — the longitudinal phenotyping, the extended pedigree structures and the multigeneration genotyping — allow for future twin-family research that will contribute to gene discovery, causality modeling, and studies of genetic and cultural inheritance."*. By using twin pairs as the probands recruitment initially included parents, but later on also siblings, spouses and children of the twins. According to the report this has resulted in: *"[..] a database with roughly equal proportions of participants who are and who are not twins. Over the years, a total of 280,569 participants were registered at the NTR, 231,088 of whom are still contactable. The total group of participants includes 255,785 members of twin families and 24,784 participants contributing as a teacher of a child registered at the NTR."*

PAN

The Prospective ALS Study Netherlands (PAN)[14,24] is a prospective study for patients suffering from amyotrophic lateral sclerosis (ALS). As stated in Võsa et al. 2018[14]: *"Since 2006, PAN aims to include all Dutch patients with ALS and similar phenotypes to correlate potential lifestyle, genetic and environmental risk factors with the onset and prognosis of ALS. To date, 3,400 patients have been included, and genotypes and expression data have been generated for a subset of these patients."*

The number of samples per cohort used in this study can be found in *Figure S1*. Gene expression and SNP genotype data from the blood samples of the subjects in the cohorts were used. The information of the cohorts was combined into one *.vcf*[25] file per chromosome.

Handling of VCF files

The genotype data from different cohorts were extracted as VCF files. All VCF[25] files were read using the *cyvcf2* package[26] of the Python programming language. The genotype data of the SNPs is loaded as following: the data of samples for that SNP is selected and recoded to single number; 0 if the sample is homozygote for the major allele of that SNP, 1 if the samples is heterozygote for that SNP and 2 if the sample is homozygote for the minor allele of that SNP. This creates a Pandas *DataFrame*[27] with the samples as the index and the SNPs as the columns.

Preparation of general files

The original gene expression matrix had the expression code of the samples as the columns and the ENSG/Ensembl[28] code of the genes as the rows. This matrix was loaded as a Pandas *DataFrame*[27] and a genotype to expression coupling file was used to alter the names of the columns to the genotype code of the samples. The matrix was then transposed to put the samples as the index and the genes as the columns. This matrix was stored as both a *.txt* text file and a pickle file[29], called *gene_expression.txt/.pkl*. Prior to usage in calculations the samples present in the gene expression matrix were compared to those in the VCF file; any samples which are not present in both files were discarded before calculations were initiated.

The annotation file called annotation_meta_all_2018-01-31.txt was read and converted to a pickle file called *gene_mappings.pkl*. The start and stop positions in this annotation file were the same as in *ProbeAnnotation_STARv2.3.0e_Ensembl71.txt*. Both annotation files originated from the BIOS consortium[14].

toRegressOut.log.gz contains data of an earlier experiment of the BIOS consortium[9,14], it also contains information about all Cis-eQTL SNPs which were conditionally determined to be independent. The Cis-eQTL SNPs determined to be independent were extracted and stored per gene in a dictionary which was stored in a file called *ind_snps.pkl*

Graphs generation

Different types of plots were generated to enable evaluation of results. The different types of plots each enable the evaluation of a different type of results, or a different aspect of the results.

The interaction plots are graphs visualizing the interactions between the protein of a gene and the SNP of the interacting gene. Interaction plots are plots of the same type as figures 2B and 4B of *Zhernakova et al. 2017*[9]. The samples were split along their genotype for the SNP modulating the interaction between TF gene and interacting gene. The linear regression models were generated per set of samples in the following form:

$$interacting\ gene\ =\ b0\ +\ b1 * TF\ gene$$

The plots are named as following {code of SNP}_{ENSG of interacting gene}_{ENSG of TF gene}*.png*.

The predicted vs measured plots are graphs plotting the measured expression of a gene against the expression of that same predicted by a model. The predicted vs measured plots were generated to test the accuracy of the model. The predicted vs measured plots are named as *{ENSG of gene}.png*.

The comparison plots are graphs plotting the predict and measured values of a TF gene and a gene it interacts with. Comparison plots are similar to the interaction plots, and are produced to check if the predictions have pattern similar to that of the measurements. A linear regression model was generated for both measured and predicted sets in the following form:

$$interacting\ gene\ =\ b0\ +\ b1 * TF\ gene$$

The plots are named as following *{code of SNP}_{ENSG of TF gene}_{ENSG of interacting gene}_comparison.png*.

The graphs plotting peaks of DNA binding by TFs detected in vivo (peak graphs in short), were generated to enable checking if identified potential interactions actually exist in vivo. These plots are of the same type as figure 4C of Zhernakova et al. 2017[9]. The plots were stored as *{experiment code}_{replicate code}_chr{chromosome of SNP}_{start position of SNP}_{stop position of SNP}_B/BB.png* (*B* is when a .Bed file is used for the generation of the plot, *BB* when a .bigBed file is used).

Identification of significant Cis-eQTL SNPs

One method for identifying significant Cis-eQTL SNPs(*analysis.py*) utilizes series of single linear regressions. This method works as following: the *.vcf*[25] file for the chromosome on which the gene of interest is located is opened using the *VCF_reader* function of *cyvcf2*[26] and a section of the chromosome encompassing from 250 kb before the start of the gene until 250 kb after the end of the gene is selected. All SNPs with a Minor Allele Frequency (MAF) lower than 0.05 are dropped; any SNPs which have a missing value for any of the samples are also dropped. Both the range of the selection and the MAF cutoff can be adjusted if required. For each of these SNPs a linear regression model is generated (using *scipy.stats.linregress*) in the form of:

$$gene\ expression\ =\ b0\ +\ b1 *\ genotype\ of\ SNP$$

The models are sorted in descending order by their p-value. The model with the lowest p-value is selected and used to predict the values of the gene of interest, following that the residuals between the prediction and the actual expression are calculated. Using these residuals new linear regression models are generated for each of the selected SNPs, and these will again be used to predict the expression and to calculate the residuals of the residuals. This process will continue until no more models can be generated with a p-value below the set threshold. All the SNPs selected up until that point are considered significant.

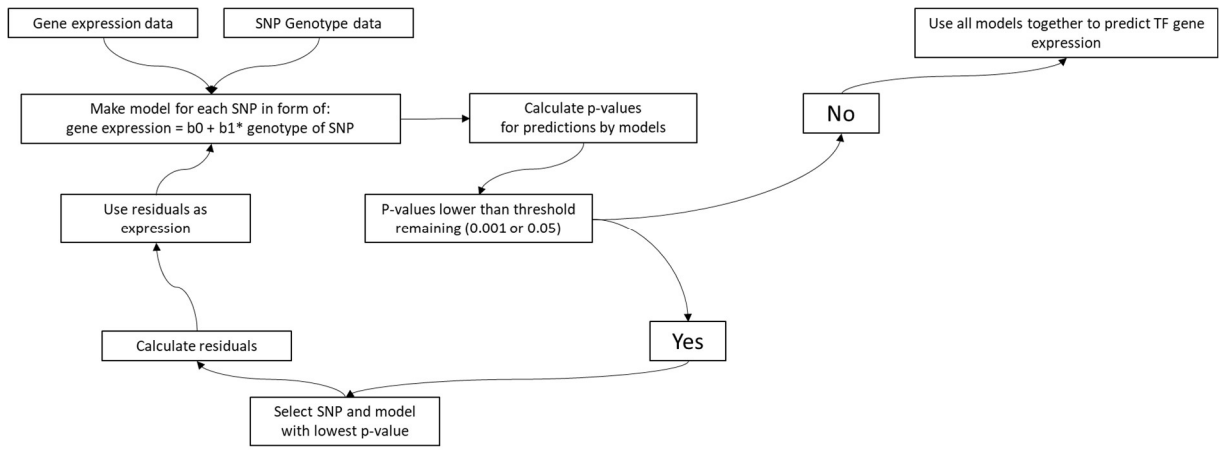


Figure 2: overview of the workflow of identifying Cis-eQTL SNP utilizing series of single linear regression models.

Another method was attempted which generated a single multivariate linear regression model instead. The SNPs were selected in the same manner as with the series of single linear regression models. For each of the SNPs for each of these SNPs a linear regression model is generated (using *scipy.stats.linregress*) in the form of:

$$gene\ expression\ =\ b0\ +\ b1 * \ genotype\ of\ SNP$$

and the models are sorted in descending order by their p-value. The SNP of the model with the lowest p-value is selected and added to a list of components SNPs. The list is then used to generate a multivariate linear regression model using *sklearn.linear_model.LinearRegression* by using the genotypes of the SNPs in the components list as the x-values and the expression of the gene of interest as the y-values. This model is then used to predict the expression of the gene of interest and the residuals between this prediction and the actual expression is calculated. Using these residuals new linear regression models are generated for each of the selected SNPs, of these the SNP with the lowest p-value will be selected again and will be added to the list of component SNPs. The updated list is used to generate a new model and the residuals between the prediction of the new model and the measured expression and the cycle is repeated. This cycle will continue until no more SNPs with a p-value below the set threshold. All the SNPs that are part of the components list are considered significant.

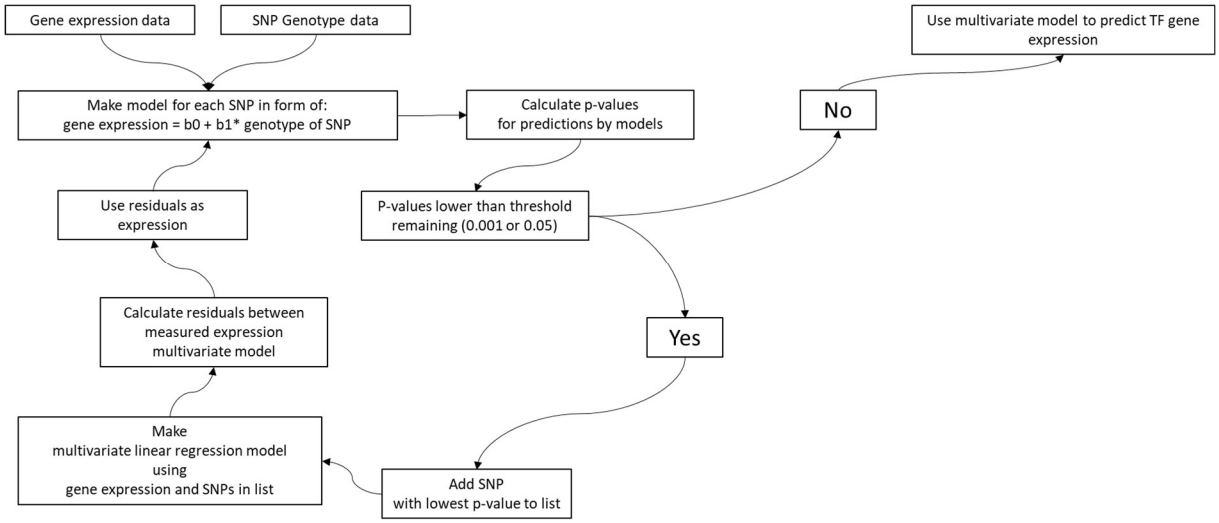


Figure 3: overview of the workflow of identifying Cis-eQTL SNP using a multivariate linear regression model.

For both methods predicted vs measured plots were generated. In case of the first method the predicted expression of all the different models was summed to get the total prediction, in case of the second the single multivariate linear regression model was used for the prediction. Interaction plots were also generated in case of both methods using the predictions being used for the TF gene and the measured values for the interacting gene. The model or list of models both methods generated were saved as well as both a *.txt* text file and a *.pkl*/pickle file.

Run of the eQTLgen pipeline

The eQTLgen pipeline[14,15] was used to have fully developed algorithm to compare the Cis-eQTL identification algorithm with. The pipeline ran a conditional Cis-eQTL analysis, both MetaQTL and Iterative modes were attempted. Exact settings are described in figure S2.

Generation of models from SNPs determined by external methods

SNPs identified by external methods were used to generate multivariate linear regression model. This refers to either the Cis-eQTL SNPs identified using the eQTLgen pipeline or the Cis-eQTL SNPs determined to be independent in a previous study[9,14]. The SNPs from the file were extracted and the genotype information was retrieved from the VCF[25] file of the chromosome on which the SNPs were located and were used as the x-values to generate a multivariate linear regression model using

sklearn.linear_model.LinearRegression with the expression of the corresponding gene as the y-values. In most cases this method was used the results were saved as both a .pkl and *.txt* file

Two-step predictions

A two-step prediction means that the predicted values of the TF gene are used to predict the values of the interacting gene. The predictions for the interacting gene are made by interaction models which have the following formula:

$$Interacting\ gene\ expression\ = \\ Intercept/b0\ +\ genotype\ interacting\ SNP * b1\ +\ TF\ gene\ expression * b2 \\ +\ genotype\ interacting\ SNP * TF\ gene\ expression * b3$$

Both measured and predicted values of the TF gene can be used for the initialization of the multivariate linear regression model. The model is created using *sklearn.linear_model.LinearRegression* with the expression of the interacting gene as the y-values and genotype interacting SNP, expression of TF gene and genotype interacting SNP*expression of TF gene as the x-values. The p-values of this model are named as following: the p-value of b0 is called the intercept p-value, that of b1 is called the SNP p-value, that of b2 the gene p-value and that of b3 the interaction p-value. In most cases this method was used the model was saved as both a .pkl and *.txt* file.

Identification of best predicted genes

In order to determine which genes were best capable of being predicted by their Cis-eQTLs alone, a multivariate linear regression model was generated for all of the genes for which expression has been measured in the BIOS data. The models were generated using the genotypes of Cis-eQTL SNPs of a certain gene determined to be independent in earlier experiment[9,14] (extracted from *ind_snps.pkl* and VCF[25] files) and the expression of that gene. The Pearson Rank correlation/r-value between the predicted and the measured values of the genes were calculated and the genes were sorted in ascending order by their r-value. The models of the top 20 best ranked genes were saved and had predicted vs measured plots generated for them.

Identification of interactions

All potential Gene Cis-eQTL SNP pairs that would be tested for, were contained in *2019-12-11-Cis-eQTLsFDR0.05-ProbeLevel-CohortInfoRemoved-BonferroniAdded.txt.gz* from the eQTLgen consortium[14], file was downloaded from *https://eqtlgen.org/Cis-eQTLs.html* on Tue 23-03-2021. For each gene, the SNP with the highest p-value was selected and the rest was discarded. The gene Cis-eQTL SNP combination were split per chromosome on which they were located, and the information was stored as both *Cis-eQTLs*.txt* and *Cis-eQTLs*.pkl* where * is the number of the chromosome.

Two methods were used to search for gene Cis-eQTL SNP pairs the TF gene can interact with. Both methods would discard any SNPs which did not have all three genotypes (major-major, major-minor and minor-minor)

Method 1: First the samples were split along their genotype for the Cis-eQTL SNP of the gene SNP pair and an independent t-test[30] is performed between each of the three groups to test if a different genotype for a certain SNP gave differences in expression for the corresponding gene of the gene Cis-eQTL SNP pair. This generated three t-values t-01(between homozygote major and heterozygote), t-02(between homozygote major and homozygote minor) and t-12(between heterozygote and homozygote minor). The sum of these three t-values is called the combined t-value. Gene Cis-eQTL SNP pairs were sorted in descending order by their combined t-values and the top 50 were selected. For these 50 pairs the p-value was calculated for the linear regression model:

$$interacting\ gene\ expression\ =\ b0\ +\ b1 * expression\ of\ TF\ gene$$

The 50 pairs were sorted in ascending order by that p-value and the top 10 of those were selected. For these top 10, the two step prediction models were generated using predicted values of the TF gene,

these models were saved. In addition, interaction, predicted vs measured and comparison plots were generated.

Method 2: For all gene Cis-eQTL SNP pairs, two step prediction models were generated and the pairs were sorted by the interaction p-values of that model in ascending order and the top 10 models were selected. For these top 10, the two step prediction models were generated using predicted values of the TF gene, these models were saved. In addition; interaction, predicted vs measured and comparison plots were generated.

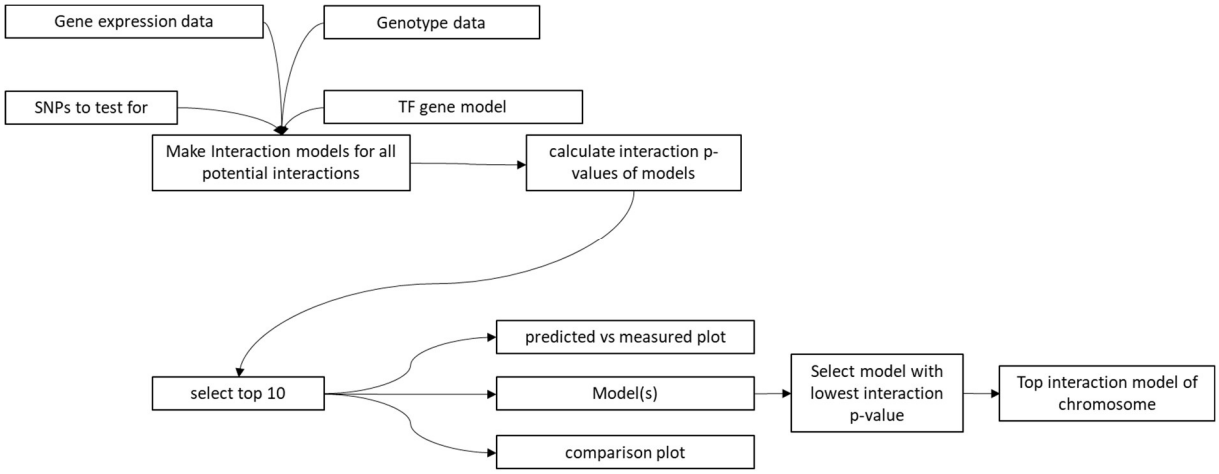


Figure 4: overview of the workflow of method 2 for identification of potential interactions between a TF gene and an interacting gene. Workflow is shown for the run process for a single chromosome, The algorithm performs runs for chromosome 1 to 22.

The workflow of the process as a whole works as following (script used: *system_analysisA.py* for method 1 or *system_anaslysisB.py* for method 2):

1. The file *settings.txt* is read, this file contains the settings used to run the script. The file *settings.txt* contains the following parameters (full path required):
   expression_file (the file containing the gene_expression, generated earlier under name gene_expression.pkl)
   mappings_file (the pickle file of the annotation file, generated earlier under name gene_mappings.pkl)
   independent_snps_file (the file containing the Cis-eQTL SNPs determined to be independent, generated earlier under name ind_snp.pkl)
   vcf_directory (the directory with subdirectories containing the *.vcf* files with the genotype information about the SNPs on that chromosome)
   SNP_selection_directory (the directory containing Cis-eQTLs*.pkl which contains all potential Gene Cis-eQTL SNP pairs on chromosome * to be tested for)
   gene_to_test (the ENSG code of the gene of interest)
   gene_overlap_range (the number of bp beyond the start and stop positions of the gene of interest which is the range wherein SNPs are considered to overlap with the gene of interest)
   display_counting (if set to true the iterations of loop will be printed to the screen, which enables it to be check if the run has frozen)
   check_data_dir (name of directory to check for files with CHiP-seq data, example folder is included)
   check_plim (maximum interaction p-values allowed before an interaction model is no longer considered to have predicted an interaction)
   check_xrange (The range surrounding a SNP in bp which needs to be checked if CHiP-seq peaks are present in that area)

chi2_bar (the alpha value for the chi^2 contingency test).
The expression file and gene mappings file are immediately read as is the file called *ENSG_dict.txt* which contains the name of a gene per ENSG code.

2. The chromosome and start and stop positions of the gene of interest as well are determined using the annotation file, and are used to determine which SNPs overlap with the gene of interest and these SNPs are stored within a list. With the default settings, all SNPs within 1 Mb of the start and stop positions of the gene of interest are considered to be overlapping.
3. It is then checked if a file containing the genotype information of the SNPs determined to be independent Cis-eQTLs for the gene of interest exists. If this file does not exist it will use the *ind_snps.pkl* file and the .vcf file of the chromosome on which the gene of interest is located to generate this file.
4. The genotype data of the independent SNPs of the gene of interest is used to generate a multivariate linear regression model with the genotypes of the SNPs as the x-values and the expression of the gene of interest as the y-values. The model is saved as both *model*.txt* and *model*.pkl* where * is the ENSG code of the gene. In addition, a predicted vs measured plot is generated for the model.
5. A loop is performed for chromosomes 1 to 22.
    a. A directory is made for that chromosome if it does not already exist.
    b. From the SNP selection directory, the file *Cis-eQTLs*.pkl* is loaded where * is the number of the chromosome.
    c. Then it is checked if the files *snp_cat_chr*.pkl* (containing genotype information for SNPs on chromosome *) and *alleles_chr*.pkl* (containing the names of the alleles for the SNPs on chromosome *) exist. If they do not exist then the .vcf file of chromosome * will be used to generate the files for the SNPs stored in *Cis-eQTLs*.pkl*. if the files do exist make sure they have information for all the SNPs mentioned in *Cis-eQTLs*.pkl* or the program will crash further down the line.
    d. If the chromosome is the same as the one containing the gene of interest any SNPs which are considered to overlap with the gene of interest are dropped.
    e. For all remaining Gene Cis-eQTL SNP pairs either method 1 or 2 is performed to see if the gene of interest interacts with them. The model with the smallest interaction p-value is copied to the main directory of the chromosome as *chr*_itop_model*{ENSG code of interacting gene}*.txt* where * is the number of the chromosome.
6. The top models for all chromosome are evaluated and the one with the smallest interaction p-value is copied to the main directory of the gene of interest.
7. A summary file and an extended summary file are generated containing information about all the models that were saved.
    a. The summary files contain in the columns from left to right: rank/nr, name of gene of interest, chromosome of Gene Cis-eQTL SNP pair, ENSG of gene of Gene Cis-eQTL SNP pair, name of gene of Gene Cis-eQTL SNP pair, SNP of Gene Cis-eQTL SNP pair, interaction p-value.
    b. The extended summary files contain in the columns from left to right: rank/nr, ENSG of gene of interest, name of gene of interest, chromosome of gene of interest, chromosome of Gene Cis-eQTL SNP pair, ENSG of gene of Gene Cis-eQTL SNP pair, name of gene of Gene Cis-eQTL SNP pair, SNP of Gene Cis-eQTL SNP pair, intercept p-values, SNP p-values , gene p-values, interaction p-values.

This can be done for multiple genes simultaneously by using the several scripts:

1. *prep_runs.py* which reads a file containing information of several genes, the columns in this file from left to right should be: rank/nr of how well they are predicted by Cis-eQTLs alone, ENSG code of gene, name of gene. This information is used to make a copy of the template folder for each of the genes. The template folder should contain the script used for the analysis (either system_analysisA.py or B.py), systems_analysisB.sh which is used to load the run as a SLURM job to the queue, *ENSG_dict.txt* and *settings.txt*. *settings.txt* has gene_to_test altered to the ENSG of the gene of interest of that folder. It is advisory but not required to add one folder for each chromosome to the template folder which contains the *snp_cat_chr*.pkl* (containing genotype information for SNPs on chromosome *) and *alleles_chr*.pkl* (containing the names of the alleles for the SNPs on chromosome *) files, by doing this it improves the running time by at least 2 hours because step 5C is skipped. In addition, a file called *run_all.sh* is created in the directory that contains the *prep_runs* script, this script will load the runs of all the genes to the SLURM queue.
2. *make_overview.py* which make a directory called overview and generates several files in that directory: overview and extended overview, summary and extended summary, overlap information and the top model of each gene saved as {gene name}*_chr*_itop_model*{ENSG code of interacting gene}*.txt* where * is the number of the chromosome of the interacting gene. The overview and overview extended files only contain information about the best interaction for each gene of interest, while the summary and summary extended files contain information about the top 10 interactions of each gene of interest. The overlap information file contains information about whether or not there is overlap between the SNP and the gene of interest.
    a. The overview file contains in the columns from left to right: name of gene of interest, name of gene of Gene Cis-eQTL SNP pair, SNP of Gene Cis-eQTL SNP pair, interaction p-value.
    b. The summary file contains in the columns from left to right: rank/nr, name of gene of interest, chromosome of Gene Cis-eQTL SNP pair, ENSG of gene of Gene Cis-eQTL SNP pair, name of gene of Gene Cis-eQTL SNP pair, SNP of Gene Cis-eQTL SNP pair, interaction p-value.
    c. Both the extended overview and the extended summary file contain in the columns from left to right: rank/nr, ENSG of gene of interest, name of gene of interest, chromosome of gene of interest, chromosome of Gene Cis-eQTL SNP pair, ENSG of gene of Gene Cis-eQTL SNP pair, name of gene of Gene Cis-eQTL SNP pair, SNP of Gene Cis-eQTL SNP pair, intercept p-values, SNP p-values , gene p-values, interaction p-values
    d. The overlap information file contains in the columns from left to right: rank/nr, ENSG of gene of interest, name of gene of interest, chromosome of gene of interest, start and stop position of gene of interest, chromosome of Gene Cis-eQTL SNP pair, ENSG of gene of Gene Cis-eQTL SNP pair, name of gene of Gene Cis-eQTL SNP pair, SNP of Gene Cis-eQTL SNP pair, start and stop position of SNP, does overlap exist under run settings criteria.
3. *find_snps_locs.py* which determines the coordinates of the SNPs of the top interactions of each gene and stores it as both *snp_locs.txt* and *.bed*. If the overview directory exists, the files will be stored there, if not they will be stored in same directory as the script itself. *snp_locs.txt* contains in the columns from left to right: name of gene of interest, SNP of Gene Cis-eQTL SNP pair, location of SNP on chromosomes, coordinates of SNP in *chr*:start_pos-end_pos* format.

<u>Check of identified interactions</u>

To check whether or not the identified potential interactions exist in vivo, the IDR thresholded peaks and conservative IDR thresholded peaks data from as many as possible CHiP-seq experiments of the gene of interest were collected from the ENCODE database[17]. This data was stored *.bed*[31] or *.bigbed*[31]

files, the files were stored in a folder called ENCODE_IDR. The structure of the ENCODE_IDR folder was as following: (gene of interest)<(experiment)<(replicate of experiment) (Also see *figure S3*). The folder can be changed from ENCODE_IDR to any other in the settings file, as long as folder structure and file types are ok.

The script *ENCODE_check.py* reads all the *.bed* and *.bigbed* files and collect the data about the peaks within. The *.bigbed* files were read using a module called *pyBigWig*[32]. All peaks within 100 kb (range can be adjusted) of the SNP of the identified potential interactions were selected and put into a Pandas *DataFrame*[27]. This *DataFrame* was used to generate a peaks graph. In case of the .bed files the files were load directly into a *DataFrame* and afterwards selected all the peaks within 100 kb of the of the SNP of the identified potential interactions, which were then used to generate the peaks graphs. If there were no peaks within 100 kb of the SNP then no graph would be generated. The graphs were saved in the main folder of the corresponding gene of interest within the ENCODE_IDR folder. *snp_locs.txt* and *snp_locs.bed* were used to determine the coordinates of a SNP, the .bed file has columns for chromosome, start position and stop position. It is important that these coordinates are of the same build as the experiments on the ENCODE database[17]. If the builds are different, the data was converted with the online Assembly converter (https://www.ensembl.org/Homo_sapiens/Tools/AssemblyConverter). The annotation files in this experiment made use of build GRCh37, while the experiments on the ENCODE database were of build GRCh38. Conversions were performed on Friday 23-03-2021.

To perform a more statistical check of the results produced by *system_anaslysisB.py* find_all_locs.py and ENCODE_chi2_test.py were used in sequence.
find_all_locs.py first checks if the directory *CHiP-seq_comparison* exist and if does not this directory will be created. Following that for all genes the *all_models.pkl* file is copied to the CHiP-seq_comparison directory as *{TF gene}_all_models.pkl*. Finally, if *all_snp_locs.pkl* does not exists it will be generated by collecting the locations info about all SNPs in the SNP_selection_directory in a similar manner as *find_snp_locs.py* and stores it in *all_snp_locs.txt* and *.pkl*. in addition, a .bed file for the same SNPs will be generated called *all_snp_locs.bed*.
ENCODE_chi2_test.py performs a χ^2 contingency test for the all the genes for which a *{TF gene}_all_models.pkl* exists in the CHiP-seq_comparison directory. First it is checked if *all_snp_locs.pkl* or *all_snp_locs.txt* exist in the CHiP-seq_comparison directory, it is also checked if *all_snp_locs_new.bed* or all_snp_locs.bed exist in the CHiP-seq_comparison directory. If none of these files exist, the calculations cannot occur. A file called defective_incomplete_missing.txt is created and opened, any errors concerning files, SNPs etc. will be written in this file. Afterwards the *{TF gene}_all_models.pkl* files are opened in sequence, and all models with an interaction p-value smaller than 10**-6 will be considered as having a predicted interaction. For all files in *ENCODE_IDR* the TF gene will be checked for all SNPs to see if there are any peaks within 10 kb of the SNP. A contingency matrix *(Figure 2)* is made and a chi2 contingency test is performed on that matrix. The information of each file/replicate is collected and gathered and put into a table in a file called *{TF gene}_chi2-test* for each gene. The threshold p-value for the test, the threshold for the interaction p-values and the range to check for a binding peak can be adjusted in the settings file.

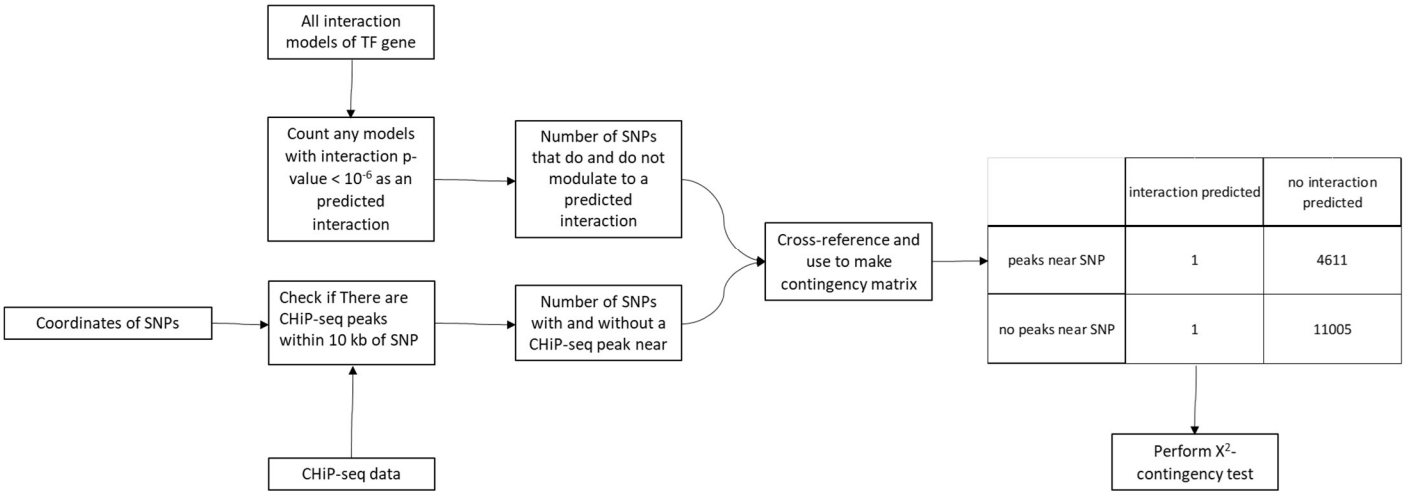


Figure 5: Workflow of the χ2 contingency test. On the right-hand end of the figure is an example of the contingency matrix used for the χ2 contingency test; numbers may vary from replicate to replicate. The index of the matrix indicates whether or not the SNP has a CHiP-seq binding peak with 10 kb range, the columns indicated whether or not the SNP was part of an interaction models which was considered to be part of a predicted interaction.

Results

All algorithms developed in this project were fully based on Python programming language[29].
Initially algorithms for identifying Cis-eQTLs were developed in order to gain a better understanding of the concepts. The identified Cis-eQTL SNPs would be used to predict the expression of the TF genes to validate the results. Cis-eQTL SNP identification would also be performed using the eQTLgen pipeline. Following that the interaction identification algorithm was developed, which predicts the expression of TF genes based on their Cis-eQTL SNPs. The algorithm then models the modulation of the interaction between the TF and a Cis-eQTL SNP of an interacting gene. The models are evaluated and used to determine if the interaction is likely to occur in vivo. CHiP-seq data is used to confirm whether or not these interactions actually do occur in vivo.

Predictions of gene expression using identified Cis-eQTL SNPs have low accuracy

The methods for identifying Cis-eQTLs are based on using the residuals of the previously identified Cis-eQTLs to identify new Cis-eQTLs. An effect on gene expression can be falsely attributed to a SNP as a result of a strong LD between that SNP and a Cis-eQTL SNP with an actual effect. By identifying the eQTL with which the false effect SNPs are in LD, the false effects can be removed by calculating the residuals between the identified Cis-eQTL and the expression. This is because the SNPs in LD with the Cis-eQTL would have virtually identical effects on expression and the residuals miss the effect of the identified Cis-eQTLs. Therefore the effects of SNPs in LD with the Cis-eQTL would be absent in the residuals as well. Algorithms based on this approach are covered here.
The identified Cis-eQTL SNPs would be used to predict the expression of the TF genes to validate the results. The concept being; the more accurate the prediction the higher the probability the right Cis-eQTL SNPs were identified.

First series of single linear regressions were performed *(figure 6B, figure S4)* using for each of the BIOS samples the gene expression of either STX3 or SREBF2 and the genotype data. The Cis-eQTL SNP and model with the lowest p-value is saved and added to a list. Once no more SNPs have p-values lower than the threshold, all saved models are used together to calculate the expression of the TF gene. Runs were performed with the p-value set at 0.001 and 0.05. The interaction plots show that with the p-value threshold set to 0.05 STX3 has resemblance to the reproduction figure counterpart *(figure 6A)*, but this is not the case for SREBF2. With the p-value threshold set to 0.001, the resemblance with the reproduction figures decreases for STX3 but this remains mostly the same for SREBF2. In addition, the

number of Cis-eQTL SNPs identified decreases from 9 to 4 for STX3; and from 4 to 3 for SREBF2. For both the genes, the p-values were rather high (0.1-0.8) and at both threshold values. This indicates that any correlation between the TF gene and the interacting gene has a high probability to be the result of random noise. In addition, the predictions for both STX3 and SREBF2 covered a much smaller range of values than the measurements for both p-value thresholds.

The predicted vs measured plots *(figure S5)* generally show poor correlation between the prediction and the measurements at both p-value thresholds, with the prediction covering a much smaller range of values than the measurements. The Pearson r-values of the correlations are typically ~0.1, which is not considered a strong correlation[33,34]. This further confirms the poor quality of the predictions.

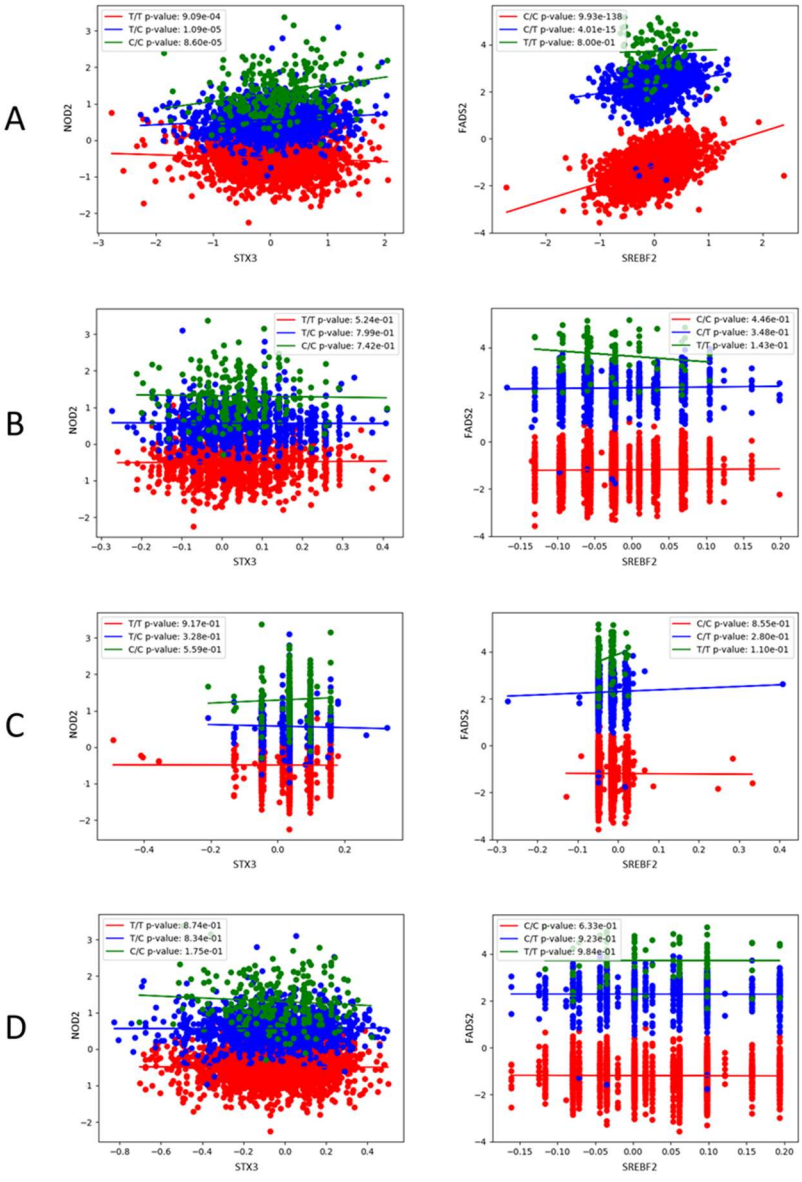


Figure 6: Interaction plots of the results of the Cis-eQTL identification algorithms. The figures on the left show the modulation of the interaction of STX3 with NOD2 by rs1981760 (same as figure 2B of *Zhernakova et al.2017*[9]), the one on the right shows the modulation of the interaction of NOD2 with FADS2 by rs968567 (same as figure 4B of *Zhernakova et al.2017*[9]). The expression of the TF gene (either STX3 or SREBF2) is on the x-axis while that of the interacting gene (either NOD2 or FADS2) is on the y-axis, red dot indicates the sample is homozygous for the major allele, blue indicates heterozygosity, and green indicates the sample is homozygous for the minor allele. In the legend of each plot are the names of the genotypes along with the p-values[35] (the probability that the association between x and y is due to chance) of the regression line for the genotype set of samples. **A,** interaction plots functioning as reproductions of figures 2B and 4B of *Zhernakova et al.2017*[9] using data from all six BIOS cohorts. **B,** interaction plots of the predictions of STX3 and SREBF2 made using the cis-eQTLs identified using the series of single linear regressions. **C,** interaction plots of the predictions of STX3 and SREBF2 made by the multivariate linear regression models which use the SNPs

identified by the pipeline of the eQTLgen consortium. **D,** interaction plots of the predictions of STX3 and SREBF2 made by the multivariate linear regression models made by the multivariate Cis-eQTL identification algorithm.

To get Cis-eQTL SNPs which give a more accurate prediction, the pipeline of the eQTLgen consortium was used. The pipeline was used for a conditional Cis-eQTL analysis in the iterative form to identify SNPs. The genotype data of the BIOS samples for the identified SNPs was used in combination with gene expression data to initialize one multivariate linear regression model for each TF gene. The interaction plots *(figure 6C, figure S6)* of these models showed poorer resemblance with the reproduction figures than either of the figures for the series of linear regressions. The p-values are in the same order of magnitude as the series of linear regressions while the correlations between measured and predicted become lower. Thus, the performance of the eQTLgen pipeline was not an improvement over the series of linear regressions algorithm.

Inspired by the eQTLgen pipeline a Multivariate linear regression method was created which does not generate a set of single linear regression models, instead using all identified SNPs to generate a single linear regression model. This method was performed for both STX3 and SREBF2 using two different thresholds for p-values (0.05 and 0.001). With the p-value threshold set 0.001 the interaction plot of STX3 closely resembles the one generated by the series of single linear regressions, only the range covered by the predictions is about twice as broad. The same holds true for SREBF2; the major difference being that the range of the predictions is narrower instead of broader. The multivariate linear regression model of STX3 uses the same SNPs which were identified by the series of linear regressions. The model SREBF2 uses 3 SNPs in total, these three are also part of the 4 SNPs identified by the series of linear regressions *(Supplementary data 1)*.
With the p-value threshold set 0.05 the interaction plot *(figure 6D, figure S4)* of STX3 resembles the pattern of the reproduction figures rather closely, but the predictions cover a much narrower range, although this range is still the broadest range seen for predictions thus far. In case of SREBF2 the pattern does not change much in interaction to the series of linear regressions the range of the predictions also remains the same. The multivariate linear regression model of STX3 uses the same SNPs which were identified by the series of linear regressions and three SNPs identified by the new algorithm. The model of SREBF2 uses two of the SNPs identified the series of linear regressions and one SNP which is new *(Supplementary data 1)*. For both p-value thresholds the p-values remain in the same range and slightly increase for the SREBF2 gene.
The predicted vs measured plots *(figure S5)* mostly do not show improvement over those from the series of single linear regressions. The Pearson r-values all decrease by an order of 10-100, the only exception being SREFB2 at 0.001 threshold value which remains around ~0.1.

Interactions between genes identified using Cis-eQTL SNPs have a low probability of occurring in vivo
Algorithms for the identification of interactions of the same type as described *Zhernakova et al. 2017*[9] were developed. Since the identification process of Cis-eQTLs did not work properly, Cis-eQTL SNPs determined to be independent in a previous study[9,14], were used as the Cis-eQTLs for the models of TF genes. The SNPs to be tested as potential modulators of interactions were retrieved from *https://eqtlgen.org/Cis-eQTLs.html*. NF-κB was chosen as the initial TF gene as it is a Transcription Factor with over a 100 known targets[36]. The initial run with NF-κB were performed using method 1 for interaction identification. The lowest interaction p-value this yielded was 0.00069 *(figure 7A, figure S7)* which is higher than the p-value threshold of $5*10^{-8}$ for GWAS studies[37]. Thus, with no significant interaction p-values found using this method, method 2 was developed which only selects for interaction p-values. Interaction analysis for NF-κB was run using method 2 for interaction identification. All identified potential interactions were different from the run with the method 1. The lowest interaction p-value was $6.23 * 10^{-7}$ *(figure 7B, figure S8)* which comes closer to the p-value threshold of $5*10^{-8}$ but is still higher than the threshold.

A

Rank	interacting chrom	interacting ENSG	interacting gene	interacting snp	interaction p-value	Rank	interacting chrom	interacting ENSG	interacting gene	interacting snp	interaction p-value
1	17	ENSG00000141295	SCRN2	rs62076106	0.000694171	12	19	ENSG00000104936	DMPK	rs1799894	0.072024793
2	10	ENSG00000107771	CCSER2	rs1343111	0.008714595	13	2	ENSG00000144580	CNOT9	rs2303566	0.086933116
3	21	ENSG00000159256	MORC3	rs11909469	0.015490838	14	1	ENSG00000116213	WRAP73	rs9662052	0.090663654
4	12	ENSG00000079337	RAPGEF3	rs145192203	0.01557548	15	22	ENSG00000183473	N.A.	rs13053175	0.09228508
5	6	ENSG00000135587	SMPD2	rs7754650	0.019084121	16	20	ENSG00000130590	SAMD10	rs2427581	0.095651988
6	15	ENSG00000128944	KNSTRN	rs34034104	0.021872941	17	8	ENSG00000147439	BIN3	rs7005025	0.10430304
7	11	ENSG00000236304	AP001189.1	rs11236839	0.035729649	18	5	ENSG00000152684	PELO	rs139420269	0.107721895
8	13	ENSG00000132932	ATP8A2	rs9578952	0.050837379	19	14	ENSG00000198208	RPS6KL1	rs11159109	0.168156993
9	3	ENSG00000170248	PDCD6IP	rs9311031	0.066689171	20	18	ENSG00000166347	CYB5A	rs12458414	0.213250437
10	7	ENSG00000188707	ZBED6CL	rs10952249	0.069064166	21	16	ENSG00000159648	TEPP	rs2241771	0.224272242
11	9	ENSG00000050555	LAMC3	rs10901333	0.069204098	22	4	ENSG00000151725	CENPU	rs6552800	0.236061223

B

Rank	interacting chrom	interacting ENSG	interacting gene	interacting snp	interaction p-value	Rank	interacting chrom	interacting ENSG	interacting gene	interacting snp	interaction p-value
1	1	ENSG00000233355	CHRM3-AS2	rs12021900	6.23E-07	12	16	ENSG00000007520	TSR3	rs11865640	0.001072636
2	9	ENSG00000147883	CDKN2B	rs2069426	1.39425E-05	13	3	ENSG00000163376	KBTBD8	rs2364281	0.001242143
3	15	ENSG00000259703	N.A.	rs144197278	6.75427E-05	14	2	ENSG00000242766	IGKV1D-17	rs2162488	0.001617933
4	12	ENSG00000258546	CENPUP2	rs10842464	7.75708E-05	15	8	ENSG00000104361	NIPAL2	rs10103296	0.001899512
5	17	ENSG00000108309	RUNDC3A	rs2011895	0.000123872	16	20	ENSG00000101138	CSTF1	rs6024857	0.001969193
6	6	ENSG00000146285	SCML4	rs6568505	0.000135871	17	4	ENSG00000179979	N.A.	rs28481697	0.002153895
7	14	ENSG00000203485	INF2	rs4072285	0.000204883	18	11	ENSG00000254750	CASP1P2	rs1792755	0.002824272
8	19	ENSG00000169136	ATF5	rs78331666	0.000297953	19	10	ENSG00000119906	SLF2	rs10883567	0.003711664
9	5	ENSG00000154153	RETREG1	rs35004	0.00067312	20	13	ENSG00000213995	NAXD	rs61969228	0.005775403
10	22	ENSG00000100241	SBF1	rs76275199	0.001003543	21	18	ENSG00000266053	NDUFV2-AS1	rs4797359	0.006867808
11	7	ENSG00000002726	AOC1	rs28891172	0.001046796	22	21	ENSG00000228107	AP000692.1	rs28385572	0.014090076

Figure 7: The top interactions with NF-κB (ENSG00000109320) identified for each chromosome. Columns are as following: *Rank* indicates the how highly the interaction was ranked out of the 22 based on the interaction p-value; *interacting chrom* indicates the chromosome on which the interacting gene is located; *interacting ENSG* gives the ENSG/Ensembl database code of the interacting gene; *interacting gene* gives the name of the interacting gene; *interacting SNP* gives the code of the Cis-eQTL SNP which modulates the interaction; and the *interaction p-value* column contains the interaction p-value. **A,** the top interactions for the chromosomes identified using method 1. **B,** the top interactions for the chromosomes identified using method 2.

While methods 1 and 2 did not generate any significant interaction p-values for NF-κB, this could be the result of a statistical anomaly. In order to check whether this was truly the case, runs were performed for other genes. First this was attempted for STX3 and SREBF2 as these have the type of interaction the algorithm is supposed to identify. Instead of searching for potential interactions the interactions identified by *Zhernakova et al. 2017*[9] were put into interaction models. For the model to predict the expression of STX3 and SREBF2 were initialized twice once using the SNPs identified by the eQTLgen pipeline and once using the SNPs in *toRegressOut.log.gz*. These models were then used to initialize the interaction models which were then used to predict the expression of NOD2 and FADS2. The Comparison plots of both versions *(figure 8)* again show a smaller range of values for the predictions in

comparison to the measurements. The predicted vs measured plots (figure S9, S10) show the same small range for the TF genes (STX2 or SREBF2) but not for the interacting genes (NOD2 or FADS2). The interacting genes show a range of values for the predictions which rather closely resembles that of the measurements. The key difference being that the predicted roughly cluster in three areas while the measured values are more continuously distributed. This further supported by the Pearson r-values between the measurements and predictions, as those of the TF genes are typically in the order of ~0.06-0.1 while those of the interacting genes are in the order of ~0.8-0.9. In general, the models made using the SNPs from toRegressOut.log.gz had higher r-values and are thus more accurate. None of the interaction models generated had interaction p-values below the threshold of $5*10^{-8}$, thus it did not work as intended.

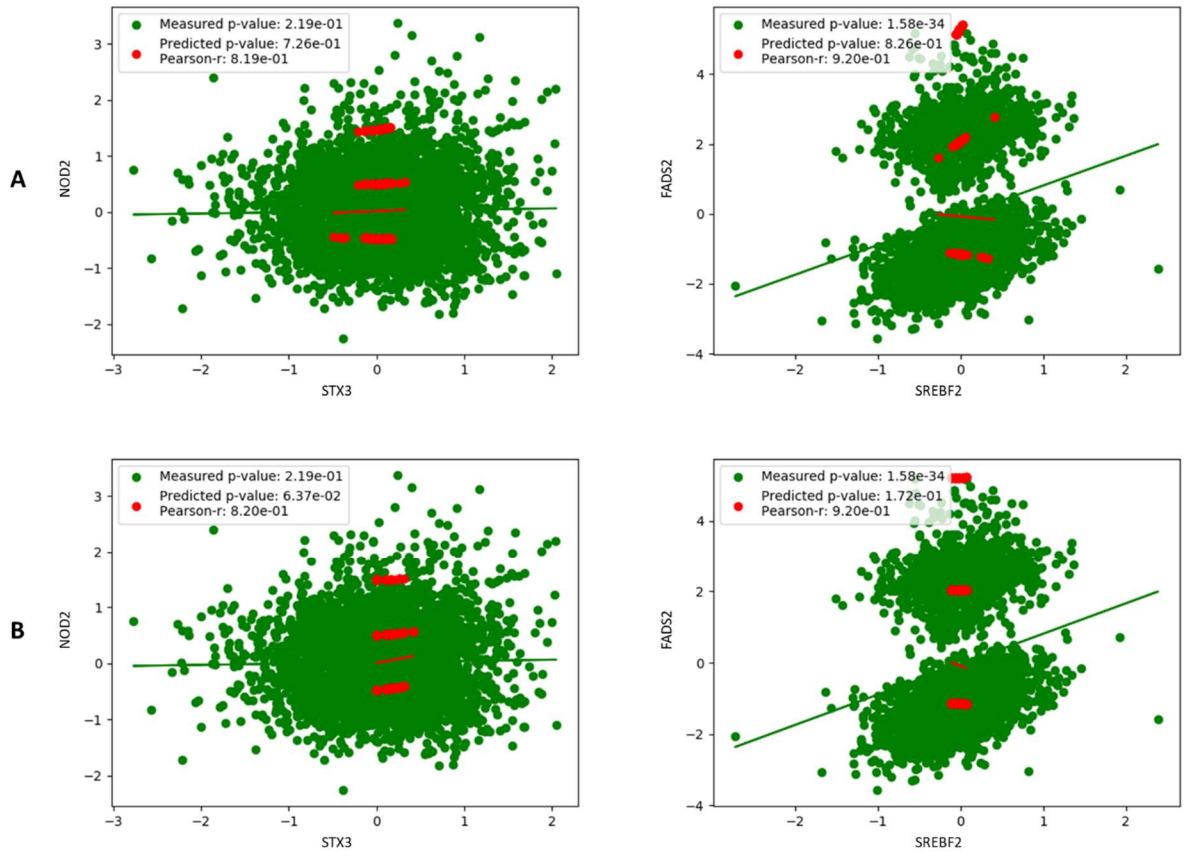


Figure 8: The comparison plots of the predictions of the interaction models for NOD2 and FADS2. On the x-axis is the expression of STX3 or SREBF2, on the y-axis is the expression of NOD2 or FADS2. Green dots indicate measured values of expression while red dots indicated predicted values of expression, the regression lines are drawn using the dots of the same colors and the p-values are of those regression lines. In the legend are p-values of the regression lines for both the measured and predicted values as well the Pearson-r value[33] between the predicted and measured values of the samples[34]. **A,** Results when the models of STX3 and SREBF2 use SNPs identified using the eQTLgen pipeline. **B,** Results when the models of STX3 and SREBF2 use the SNPs in the toRegressOut.log.gz.

	Rank best predicted	1	2	3	4	5	6	7	8	9	10
A	name	FAM118A	CUTALP	ERAP2	PEX6	DDX11	PSPHP1	ACCS	RPS26	CHURC1	CYP26B1
	ENSG	ENSG00000100376	ENSG00000226752	ENSG00000164308	ENSG00000124587	ENSG00000013573	ENSG00000226278	ENSG00000110455	ENSG00000197728	ENSG00000258289	ENSG00000003137
	Rank best predicted	1	2	3	4	5	6	7	8	9	10
B	name	ZNF266	NKX3-1	NFXL1	ATOH8	ZNF589	ZNF83	PAX8	NR2F6	ZNF132	MYBL2
	ENSG	ENSG00000174652	ENSG00000167034	ENSG00000170448	ENSG00000168874	ENSG00000164048	ENSG00000167766	ENSG00000125618	ENSG00000160113	ENSG00000131849	ENSG00000101057

Figure 9: The tables giving the two sets of ten TF genes for which the interaction identification algorithms were run on large scale. Rows are as following: *Rank* best predicted indicates how high the genes were ranked based on the Pearson rank correlation between the measurements and the predictions by a multivariate linear regression model using the SNPs in toRegressOut.log.gz, *name* gives the name of the gene and *ENSG* gives the ENSG/Ensembl database code for that gene. **A,** the first set of 10 genes for which the interaction algorithm was run. **B,** the second set of 10 genes for which the interaction algorithm was run, this set was created by cross-referencing the list of genes considered to be best capable of being predicted using cis-eQTLs alone, with the list of all genes for which CHiP-seq experiment were present in the ENCODE Database.

Following that it was decided to find the genes which are best capable of being predicted using only their Cis-eQTLs. This was done because such genes are less likely to have their expression strongly influenced by Transcription factors thus making enabling better predictions when using these genes as TF genes. multivariate linear regression models were generated for all genes in toRegressOut.log.gz using the SNPs in that file. The genes were ranked by their Pearson r-values of the correlation between measured and predicted values. For the top 10 genes *(figure 9A)* identification of potential interactions using method 2 was performed. The interaction p-values of the top interactions *(figure 10A)* are generally lower than was the case with NF-κB. All of the p-values of the top interaction are below 10^{-5} and four of the top interactions (FAM118A, CUTALP, PEX6, CHURC1) have interaction p-values lower than the threshold of $5*10^{-8}$. These top interactions could not be validated using experimental data as the ENCODE database[17] had no CHiP-seq experiment with any of these genes as the target.

	Rank	ENSG	name	chrom of gene	interacting chrom	interacting ENSG	interacting gene	interacting snp	interaction p-value
	1	ENSG00000100376	FAM118A	22	1	ENSG00000233355	CHRM3-AS2	rs12021900	8.88527E-12
	2	ENSG00000226752	CUTALP	9	1	ENSG00000233355	CHRM3-AS2	rs12021900	2.03121E-08
	3	ENSG00000164308	ERAP2	5	10	ENSG00000188690	UROS	rs10901520	4.44586E-05
	4	ENSG00000124587	PEX6	6	5	ENSG00000248874	C5orf17	rs77824697	5.0313E-10
A	5	ENSG00000013573	DDX11	12	3	ENSG00000144820	ADGRG7	rs140174756	3.75024E-07
	6	ENSG00000226278	PSPHP1	7	19	ENSG00000268433	MTDHP3	rs8110320	1.20396E-06
	7	ENSG00000110455	ACCS	11	6	ENSG00000137312	FLOT1	rs115114376	5.87095E-06
	8	ENSG00000197728	RPS26	12	17	ENSG00000108352	RAPGEFL1	rs117330222	3.4562E-05
	9	ENSG00000258289	CHURC1	14	7	ENSG00000240859	AC093627.4	rs76029673	1.20962E-12
	10	ENSG00000003137	CYP26B1	2	20	ENSG00000125885	MCM8	rs236106	9.9263E-06

B

Rank	ENSG	name	chrom of gene	interacting chrom	interacting ENSG	interacting gene	interacting snp	interaction p-value
1	ENSG00000174652	ZNF266	19	3	ENSG00000144820	ADGRG7	rs140174756	3.76148E-05
2	ENSG00000167034	NKX3-1	8	5	ENSG00000113119	TMCO6	rs3138076	3.24198E-06
3	ENSG00000170448	NFXL1	4	8	ENSG00000104361	NIPAL2	rs10103296	1.76769E-14
4	ENSG00000168874	ATOH8	2	8	ENSG00000104361	NIPAL2	rs10103296	2.23268E-19
5	ENSG00000164048	ZNF589	3	3	ENSG00000144820	ADGRG7	rs140174756	3.09512E-18
6	ENSG00000167766	ZNF83	19	1	ENSG00000171680	PLEKHG5	rs10779790	1.52446E-06
7	ENSG00000125618	PAX8	2	11	ENSG00000255860	AP000812.2	rs7939676	3.81003E-05
8	ENSG00000160113	NR2F6	19	3	ENSG00000173905	GOLIM4	rs4266235	1.42068E-05
9	ENSG00000131849	ZNF132	19	19	ENSG00000204920	ZNF155	rs62116613	1.39749E-07
10	ENSG00000101057	MYBL2	20	3	ENSG00000144820	ADGRG7	rs140174756	1.97817E-08

Figure 10: Two lists containing the top interactions identified for both sets of 10 TF genes, The columns are as following: *Rank* is the rank of the TF gene based on the Pearson rank correlation between the measurements and the predictions by a multivariate linear regression model using the SNPs in toRegressOut.log.gz; *ENSG* is the ENSG/Ensembl database code of the TF gene; *name* is the name of the TF gene; *chrom of gene* is the chromosome on which the TF gene is located; *interacting chrom* is the chromosome on which the interacting gene is located; *interacting ENSG* is the ENSG/Ensembl database code of the interacting gene; *interacting gene* gives the name of the interacting gene; *interacting snp* gives the code of the cis-eQTL SNP modulating the interaction; *interaction p-value* gives the interaction p-value. **A,** the best interactions identified for the first set of 10 TF gene. **B,** the best interactions identified for the first set of 10 TF gene.

In order to find genes which could be validated using experimental data, a list of all the genes for which the ENCODE database[17] did have CHiP-seq experiment data was made. This list was cross-referenced with the list of best predictable genes to create a list of the genes which are best capable of being predicted by their Cis-eQTLs and have CHiP-seq data available in the ENCODE database. For the top 10 genes *(figure 9B)* of this list identification of potential interactions using method 2 was performed. The interaction p-values of the top interactions *(figure 10B)* were in the same order of size as with the first set of 10 genes, and in a several case the p-values were even lower. Four of the top interactions (NFXL1, ATOH8, ZNF589, MYBL2) had interaction p-values lower than the threshold, just as with the first set of genes. To validate the top interactions peak plots were generated *(figure 11, Supplementary data 5)* using the IDR thresholded peaks and conservative IDR thresholded peaks, only the area within 100 kb of the SNP of the potential interaction was checked. The best peak plot generated belonged to NFXL1 which had the peak closest to the location of the SNP of the interaction. In all other cases, the peak(s) were further away from the SNPs. For three of the ten genes (ATOH8, ZNF266, ZNF589) there were no peaks found near the SNP in any file and thus no graphs were generated for those genes.
A test was performed to see if method 2 interaction identification increases the probability of finding interactions occurring in vivo. All interaction models generated for the 10 genes during interaction identification runs were used in combination with the IDR data from the ENCODE database to generate contingency matrices. These matrices were then each subjected to a χ^2-contingency test. For only five of the ten genes (ATOH8, MYBL2, NFXL1, ZNF132, ZNF589) were any interactions considered to have been predicted when setting the interaction p-value threshold to 10^{-6}. The best test outcome was for ATOH8 *(figure 12)*; for two replicates of an experiment was there a significant difference between SNPs for which an interaction was predicted and SNPs for which no interaction was predicted. There were genes *(figure S12-S15)* where a significant difference was found more often or with a greater significance than for ATOH8, but in all those cases the experiment from which the data originated had replicates missing

TF binding information about one or more chromosomes, which makes the outcome less reliable. In all cases where a significant difference was found, all predicted interactions had no peaks near them according to CHiP-seq data. This means that if these χ^2-contingency tests are correct that the interactions identified have a lower probability to have that interaction occur in vivo as well.

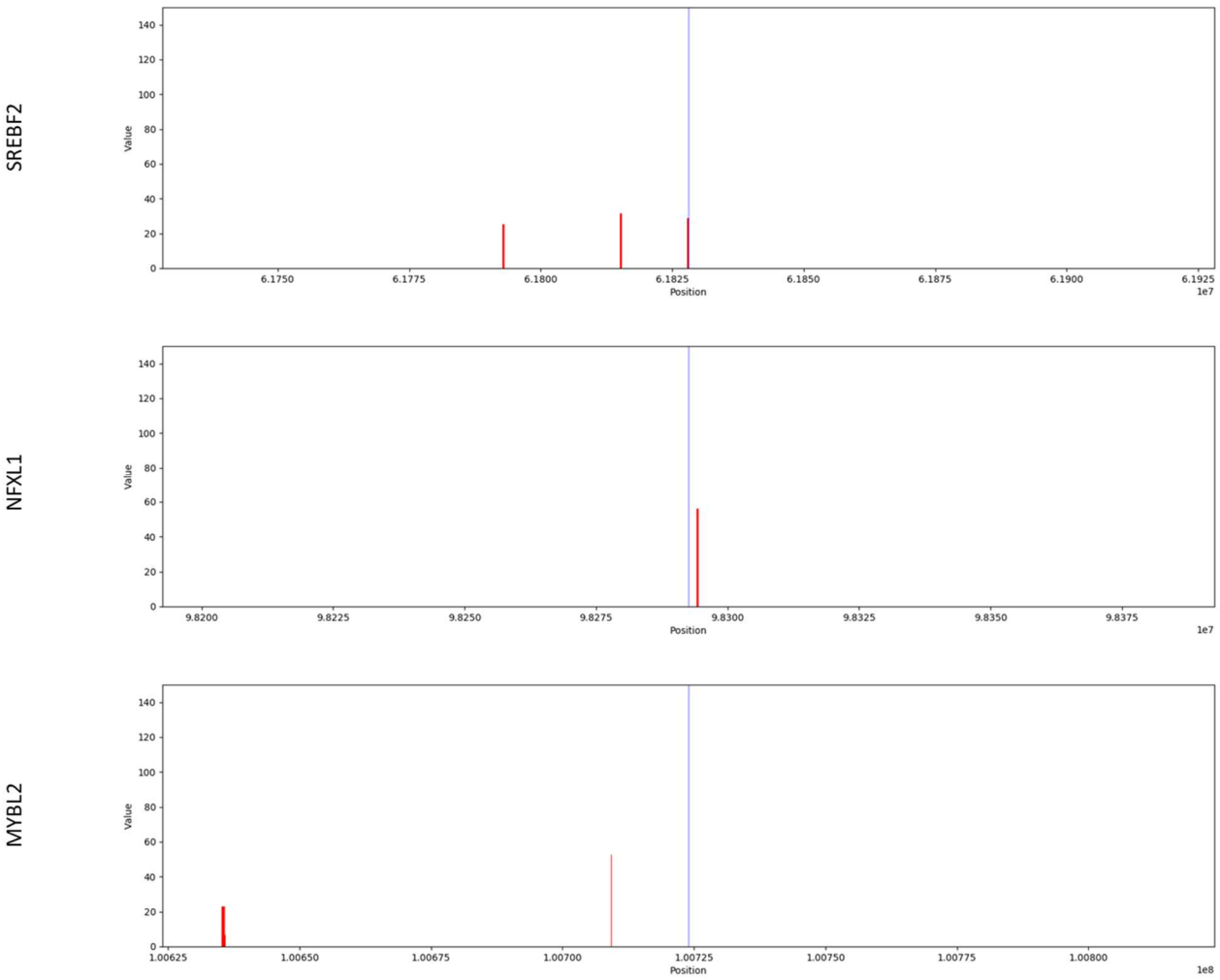


Figure 11: The peak graphs generated using the IDR Thresholded peaks and Conservative IDR Thresholded peaks data of CHiP-seq experiment on the ENCODE database. The peak graphs are bar plots with a blue line in the middle indicating the location of the Cis-eQTL SNP modulating the top potential interaction of a TF gene. The red bars/peaks indicate that the TF binds at that location according to CHiP-seq data. A range of 200 kb is displayed, 100 kb on each end of the SNP. On the y-axis is the value of the TF binding. The SREBF2 graph visualizes the interaction of SREBF2 with FADS2 and it modulation by rs968567, it was included as a positive control as it was already determined that this interaction occurs in vivo[9]. NFXL1 graph visualizes TF binding of NFXL1 very close to a SNP, data is from experiment ENCSR746XEG replicate ENCFF513WDR which is IDR thresholded peak data. MYBL2 graph visualizes the TF binding of MYBL2 which is more distant from the SNP, data is from experiment ENCSR581KCO replicate ENCFF848JYU which is IDR thresholded peak data. More Peak graphs are in Supplementary data 5.

A

	filetype	interaction peak	interaction no peak	no interaction peak	no interaction no peak	chi^2-statistic	p-value	significant with a probability of 95.0%
ENCSR161CZA_ENCFF772HNB	bigBed	0	1	1557	14045	1.783125448	0.181766007	False
ENCSR161CZA_ENCFF650DTU	bigBed	0	1	1067	14535	2.924349097	0.087252406	False
ENCSR161CZA_ENCFF951PHQ	bigBed	0	1	593	15009	5.838244178	0.015681468	True
ENCSR161CZA_ENCFF723JPN	Bed	0	1	1058	14544	2.955283611	0.085597287	False
ENCSR161CZA_ENCFF846XXK	Bed	0	1	584	15018	5.93946812	0.014805486	True
ENCSR161CZA_ENCFF520POQ	Bed	0	1	1536	14066	1.816965162	0.177674947	False

B

	filetype	interaction peak	interaction no peak	no interaction peak	no interaction no peak	chi^2-statistic	p-value	significant with a probability of 95.0%
ENCSR161CZA_ENCFF337ZRI	bigBed	0	1	1419	14183	2.024083217	0.154822333	False
ENCSR161CZA_ENCFF096THK	Bed	0	1	1406	14196	2.049249975	0.152281193	False

Figure 12: The results of the χ2 contingency tests for ATOH8 from the output of interaction identification with initialization using predicted values. The index is composed as {experiment}_{replicate}, the columns are as following: *filetype* indicates in what type of file the data was stored, a .Bed or .bigBed file; *interaction peak* indicates the number of interaction models for which an interaction was predicted and had a CHiP-seq binding peak near them; *interaction no peak* indicates the number of interaction models for which an interaction was predicted but had no CHiP-seq binding peak near them; *no interaction peak* indicates the number interaction model for which no interactions were predicted but had a CHiP-seq binding peak near them; *no interaction no peak* indicates the number interaction model for which no interactions were predicted and had no CHiP-seq binding peak near them; *chi^2-statistic* gives the χ2-statistic of the test; *p-value* gives the p-value of the test; *significant with a probability of 95.0%* indicates whether or not there is a significant difference between the models of with predicted interactions and those without, in regard to how often they have a CHiP-seq binding peak near their modulating SNP. **A,** the results for the data of the IDR thresholded Peaks. **B,** the results for the data of the Conservative IDR thresholded peaks.

These issues may have been caused by the fact that the interaction models were initialized using the predicted expression for the TF genes instead of the measured expression. To see if this is indeed the case the interaction models for STX3-NOD2 and SREBF2-FADS2 and the runs for both sets of 10 genes were performed again, this time initializing the interaction models using the measured values of the TF genes. For the interaction models for STX3-NOD2 and SREBF2-FADS2, the comparison plots (figure 13), predicted vs measured plots *(figure S10)* and the Pearson r-values were not subject to any significant changes. The Interaction models however have some significant change most notably in all cases the interaction p-value is in the order of 10^{-13}-10^{-12} which is below the threshold and thus the predictions are less likely to be accurate because of random noise.

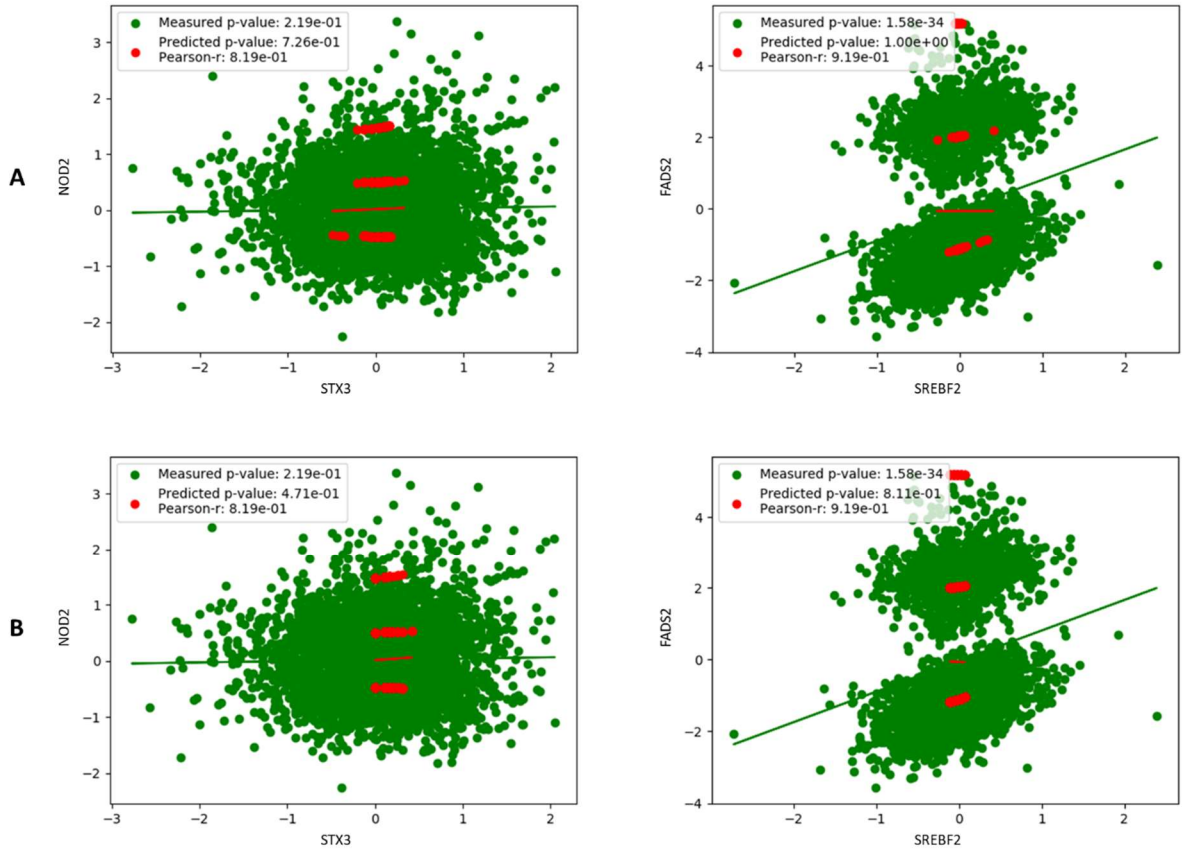


Figure 13: The comparison plots visualizing the predictions of the interaction models for NOD2 and FADS2 initialized using the measured values of NOD2 and FADS2. On the x-axis is the expression of STX3 or SREBF2, on the y-axis is the expression of NOD2 or FADS2. Green dots indicate measured values of expression while red dots indicated predicted values of expression, the regression lines are drawn using the dots of the same colors and the p-values are of those regression lines. In the legend are p-values of the regression lines for both the measured and predicted values as well the Pearson-r value[33] between the predicted and measured values of the samples[34]. **A,** Results when the models of STX3 and SREBF2 use SNPs identified using the eQTLgen pipeline. **B,** Results when the models of STX3 and SREBF2 use the SNPs in the toRegressOut.log.gz.

For the first set of 10 genes, five of the top interactions (FAM118A, CUTALP, ERAP2, DDX11, CHURC1) are smaller than the threshold *(figure S16)*. In addition, the top interacting gene has changed for ERAP2, PEX6, PSPHP1, ACCS, RPS26 and CHURC1.

For the second set of 10 genes *(figure S16),* the number of top interactions above the threshold increased to six (NKX3-1, NFXL1, ATOH8, ZNF589, ZNF132, MYBL2), the interaction p-values in general also became lower. In addition, the top interacting gene has changed for ZNF266, NKX3-1, ZNF83, PAX8, NR2F6, ZNF132. The peak graphs for the top interactions only changed for NKX3-1, ZNF83, PAX8, NR2F6, ZNF132. No graphs were produced for NKX3-1, PAX8 and ZNF132. In case of NR2F6 the graphs have peaks which are closer to the SNP than was the case for the initialization using the predicted values *(Supplementary data 5)*. ZNF83 has some graphs where the peaks are closer to the SNP, the other looked more or less the same.

There were χ^2-contingency tests performed for more of the genes when initialization was performed using the measured expression. Only for PAX8 and NR2F6 were no interactions considered to have been predicted when the interaction p-value threshold is set to 10^{-6}. Here ZNF266 had the best test outcome *(figure 14)*, with four replicates having a significant difference. Like before, there were genes *(figure S17-S23)* where a significant difference was found more often or with a greater significance but in all those cases the experiment from which the data originated had replicates missing TF binding information about one or more chromosomes. In all cases where a significant difference was found, all predicted

interactions had no peaks near them according to CHiP-seq data. Thus, the interactions identified still have a lower probability to occur in vivo as well.

A

	filetype	interaction peak	interaction no peak	no interaction peak	no interaction no peak	chi^2-statistic	p-value	significant with a probability of 95.0%
ENCSR466VYP_ENCFF532RAU	bigBed	0	1	666	14922	5.113212002	0.023744322	True
ENCSR466VYP_ENCFF480LQT	bigBed	0	1	474	15114	7.480366157	0.006237539	True
ENCSR466VYP_ENCFF483FIW	bigBed	0	1	992	14596	3.195874236	0.073824289	False
ENCSR466VYP_ENCFF950GUZ	Bed	0	1	462	15126	7.693734951	0.005541284	True
ENCSR466VYP_ENCFF064AIE	Bed	0	1	661	14927	5.15739156	0.023147637	True
ENCSR466VYP_ENCFF715VBR	Bed	0	1	970	14618	3.284582206	0.069933425	False

B

	filetype	interaction peak	interaction no peak	no interaction peak	no interaction no peak	chi^2-statistic	p-value	significant with a probability of 95.0%
ENCSR466VYP_ENCFF138RFM	bigBed	0	1	960	14628	3.326254558	0.068181977	False
ENCSR466VYP_ENCFF380MRC	Bed	0	1	940	14648	3.412271579	0.064713307	False

Figure 14: The results of the χ2 contingency tests for ZNF266 from the output of interaction identification with initialization using measured values. The index is composed as {experiment}_{replicate}, the columns are as following: *filetype* indicates in what type of file the data was stored, a .Bed or .bigBed file; *interaction peak* indicates the number of interaction models for which an interaction was predicted and had a CHiP-seq binding peak near them; *interaction no peak* indicates the number of interaction models for which an interaction was predicted but had no CHiP-seq binding peak near them; *no interaction peak* indicates the number interaction model for which no interactions were predicted but had a CHiP-seq binding peak near them; *no interaction no peak* indicates the number interaction model for which no interactions were predicted and had no CHiP-seq binding peak near them; *chi^2-statistic* gives the χ2-statistic of the test; *p-value* gives the p-value of the test; *significant with a probability of 95.0%* indicates whether or not there is a significant difference between the models of with predicted interactions and those without, in regard to how often they have a CHiP-seq binding peak near their modulating SNP. **A,** the results for the data of the IDR thresholded Peaks. **B,** the results for the data of the Conservative IDR thresholded peaks.

Conclusions and discussion

The aim was to develop a method to identify interactions between genes in order to enable reconstructions of GRNs. This was done by modeling the interactions between two genes in the form of an equation. The assumption was that the likelihood of the interaction occurring in vivo would correlate with the p-value of the interaction component of the model.

Methods to identify Cis-eQTLs were developed with the intent of making it easier to correct for LD between Cis-eQTLs. These methods are based on using the residuals of previously identified Cis-eQTLs to identify new Cis-eQTLs. This is similar to the approach utilized by the eQTLgen pipeline[14,15].

The predictions made using Cis-eQTL SNPs identified by the Cis-eQTL identification algorithm display a smaller range of values than the measurements *(figures 6, S4, S5)*. The predictions show high p-values indicating a high probability of any association between the variable being due to chance*(figures 6, S4, S5)*. The Pearson rank correlation between predictions and the measurements is low*(figures 6, S4, S5)*, indicating a weak correlation between the predictions and the measurements and therefore a low accuracy for the predictions. Taken together these results are an indication that the Cis-eQTL identification algorithm does not work as it is supposed to. Possible causes for this are given below

along with potential solutions. Because the Cis-eQTL identification algorithm does not work as intended, it is recommended to use other programs for any future Cis-eQTL identification attempts. It has to be noted however that the predictions made using eQTLgen pipeline identified Cis-eQTL SNPs *(figure 6, S6)* do not have a higher accuracy than the other predictions. Thus, using other eQTL mapping programs may not necessarily identify Cis-eQTL SNPs which yield more accurate predictions when used to initialize a model.

The interaction identification algorithm did not work as intended considering that the χ^2-contingency tests *(figures 12, 14, S12-S15 and S17-S23)* show a decreased probability of occurring in vivo for identified potential interactions. In addition, the SNPs of the top interactions never show any overlap with binding peaks of CHiP-seq experiments *(figure 11, Supplementary data 5)*, and sometimes there are no binding peaks near the SNP.
Initializing the interaction models using measured values of the TF gene, yields smaller p-values for the model *(figures 10, S11 and S16)*, thus making it less likely that the predictions made by the model are the result of random noise. This did not increase the probability of identified potential interactions to occur in vivo.

The cause as to why both algorithms did not work as intended is unknown. It may because cell types can differ strongly in expression of genes, even between different blood cell types[38]. Different cell type ratios may cause strong differences in gene expression between samples as the expression is given per sample. This can result in expression differences being attributed to an eQTL SNP while they actually result from differences in cell type ratios. The reverse also holds true; eQTL effects can be masked because the cell type ratios of several samples are almost the same. This can also lead to problems for the interaction models. Therefore, the cell type ratios of the different samples should be identified and corrected for.
Another possibility is that there are strong epigenetic differences between the samples themselves[39]. These epigenetic differences may cause the chromatin to locally condense in some samples, leaving one SNP accessible in one sample but inaccessible in the other. This could result in certain eQTL SNPs being falsely dismissed as not significant because the SNP is simply inaccessible in most of the samples. This can occur in the interaction models, and some identified interactions might not have been replicated in vivo simple because of differences in accessibility of the interacting SNP. To prevent this from occurring, ATAC-seq[40] should be performed to see which SNPs are accessible in a certain individual and which are not.

Some additional explanations are possible for the unsuccessfulness of the interaction identification algorithm. It may be possible that the levels of RNA expression are difficult to translate into protein levels, because of post transcriptional modifications such as splicing[41]. As the proteins are the transcription factors, this means that high RNA expression may not be a good indicator for high amount of a corresponding protein. This may result in the false identification of interactions when e.g., the RNA expression of gene is high but the concentration of the corresponding protein is actually low. For the samples used in the project protein levels were unavailable. It may be interesting for a future research project to attempt the interaction identification algorithm using protein levels instead of RNA expression levels. Doing so, may result in the identification of potential interactions which do have a higher probability of occurring in vivo. This may even explain some of the inabilities of the Cis-eQTL identification algorithm as some Cis-eQTLs affect the splicing of genes[42] instead of the expression.
One other aspect which could explain the problems with interaction identification is time. The response of the interacting gene to the TF is unlikely to be instantaneous, as transcription and translation do take some time. Thus it is likely there is some time lag between the expression of the TF gene and the expression change of the interacting gene[8,43]. This would also mean that the response of an interacting

gene in not always immediately detectable and may be missed during measurements. Which may result in some interactions not being predicted. Time series data can be used to prevent this from occurring.

All above explanations presume some fault in the acquiring or preprocessing of the data, but there may also be a fault in the interaction identification algorithm itself. As shown in *figures 12, 14, S12-S15* and *S17-S23* having a TF gene model that can accurately predict expression, is no indicator for a small interaction p-value. Thus, the method performed here to find genes best capable of being predicted by Cis-eQTLs alone may not be that accurate. Especially considering FADS2 is number 14 best predictable *(Supplementary data 3)*, while it is known that FADS2 has a transcription factor in SREBF2[9]. A useful avenue of research in the future is thus to develop a better algorithm for identifying genes whose prediction depends as little as possible on the expression of other genes. Preferably finding genes who are only subject to self-regulation[44], such genes would be loose ends in a gene regulatory network and would be a good place to start the process of reconstructing these networks. If this cannot be done, it would instead be wise to begin with a gene which codes for a protein which does not have any gene targets and work backwards from there. Start by identifying the TF genes influencing that gene. Another possible cause may be that the algorithm itself is too simple. Interactions were modeled as a single Cis-eQTL SNP interacting with a single TF gene, while in reality multiple TF genes can interact with multiple SNPs. Thus, the model may simply be too simple to be accurate. In addition, only a single Cis-eQTL SNP was considered per gene. This was done due to time constraints as running this for all Cis-eQTL SNPs would take a very long time. Even so attempting this in the future may be wise as the potential interactions which could be found in vivo may simply have been missed because they were not considered.

Of all these possible explanations as to why the algorithms did not work as intended, the mismatch between RNA expression and protein levels seems like the most likely explanation to this author's opinion. It has to be noted however that none of the explanations are mutually exclusive however thus multiple explanations may be correct. A general advise for future research is: to initialize interaction models using the measured values of the TF genes; to attempt more advanced filtering methods than were attempted than used in this project; to correct for cell type, epigenetic and time differences between the samples and to use protein levels instead of only RNA expression level if possible. In addition it may also be wise to use other programs for network reconstruction (such as ANASE[45]) and then model the reconstructed network using a modified version of the interaction identification algorithm described here.

Acknowledgements
I would like to thank prof. dr. L.H. (Lude) Franke for the opportunity to work as part of his research group; Dr. P. (Patrick) Deelen for the daily guidance and supervision during the project; H.H. (Henry) Wiersma, MSc for aid in developing the code for the algorithms and Dr. H.J. (Harm-Jan) Westra. for the generation of the combined *.vcf* files and the aid on how the use the eQTLgen pipeline.

Supplementary Information

Supplementary Figures
Supplementary figures S1-S23

Supplementary Data 1
Tables containing the models for STX3, SREBF2, NOD2 and FADS2

Supplementary Data 2
.zip file containing the settings files used to run the eQTLgen pipeline.

Supplementary Data 3

Table ranking gene by their ability of having their expression predicted by their Cis-eQTLs alone. Per gene the Pearson rank correlation (r-value) and R^2 value between the predictions and the measurements are given, as is the p-value of the model making the predictions.

Supplementary Data 4

Table displaying all CHiP-seq experiment from the ENCODE database[17] that were used given per TF gene. Given is per experiment the cell type in which the experiment as wells as any genetic alterations to the cell line (if applicable) and the experiments that were used as control. References are given for all experiments and genetic modifications.

Supplementary Data 5

.zip file containing the peak plots discussed but not shown in the main text.

References

1. Corcos, A. F. & Monaghan, F. V. Mendel's Work and Its Rediscovery: A New Perspective. *CRC. Crit. Rev. Plant Sci.* **9**, 197–212 (1990).
2. Reid, J. & Ross, J. Mendel's genes: Toward a full molecular characterization. *Genetics* **189**, 3–10 (2011).
3. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* vol. 169 1177–1186 (2017).
4. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
5. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 368 (2013).
6. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
7. Davidson, E. & Levine, M. Gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* vol. 102 4935 (2005).
8. Van Der Wijst, M. G. P., De Vries, D. H., Brugge, H., Westra, H. J. & Franke, L. An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome Medicine* vol. 10 1–15 (2018).
9. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
10. Vandereyken, K., Van Leene, J., De Coninck, B. & Cammue, B. P. A. Hub protein controversy: Taking a closer look at plant stress response hubs. *Frontiers in Plant Science* vol. 9 694 (2018).
11. Cavalli, M. *et al.* Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals. *Genomics* **107**, 248–254 (2016).
12. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
13. Slatkin, M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* vol. 9 477–485 (2008).
14. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* vol. 18 10 (2018).
15. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease

associations. *Nat. Genet.* **45**, 1238–1243 (2013).

16. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-.).* **316**, 1497–1502 (2007).

17. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* vol. 22 1813–1831 (2012).

18. van Greevenbroek, M. M. J. *et al.* The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *Eur. J. Clin. Invest.* **41**, 372–379 (2011).

19. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: Study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).

20. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44**, 1172–1180 (2015).

21. Schoenmaker, M. *et al.* Evidence of genetic enrichment for exceptional survival using a family approach: The Leiden Longevity Study. *Eur. J. Hum. Genet.* **14**, 79–84 (2006).

22. Hofman, A. *et al.* The rotterdam study: 2014 objectives and design update. *Eur. J. Epidemiol.* **28**, 889–926 (2013).

23. Ligthart, L. *et al.* The Netherlands Twin Register: Longitudinal Research Based on Twin and Twin-Family Designs. *Twin Res. Hum. Genet.* **22**, 623–636 (2019).

24. Huisman, M. H. B. *et al.* Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *J. Neurol. Neurosurg. Psychiatry* **82**, 1165–1170 (2011).

25. VCFv & BCFv. The Variant Call Format Specification. 1–36 (2020).

26. Pedersen, B. S. & Quinlan, A. R. Cyvcf2: Fast, flexible variant analysis with Python. *Bioinformatics* **33**, 1867–1869 (2017).

27. McKinney, W. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python High Perform. Sci. Comput.* 1–9 (2011).

28. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).

29. Van Rossum, G. The Python Library Reference, release 3.8.2. (2020).

30. Gerald, B. A Brief Review of Independent, Dependent and One Sample t-test. *Int. J. Appl. Math. Theor. Phys.* **4**, 50 (2018).

31. UCSC Genome Browser. Genome Browser FAQ. (2000).

32. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

33. Yeager, K. LibGuides: SPSS Tutorials: Pearson Correlation.

34. Bishara, A. J. & Hittner, J. B. Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychol. Methods* **17**, 399–417 (2012).

35. Thisted, R. A. *What is a P-value? galton.uchicago.edu* http://galton.uchicago.edu/~thisted/Distribute/pvalue.pdf (1998).

36. Pahl, H. L. Activators and target genes of Rel/NF-κB transcription factors. *Oncogene* vol. 18 6853–6866 (1999).

37. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* **24**, 1202–1205 (2016).

38. Palmer, C., Diehn, M., Alizadeh, A. A. & Brown, P. O. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* **7**, 1–15 (2006).

39. David Sweatt, J. & Tamminga, C. A. An epigenomics approach to individual differences and its translation to neuropsychiatric conditions. *Dialogues Clin. Neurosci.* **18**, 289–298 (2016).

40. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015**, 21.29.1-21.29.9 (2015).

41. Franks, A., Airoldi, E. & Slavov, N. Post-transcriptional regulation across human tissues. *PLoS Comput. Biol.* **13**, e1005535 (2017).

42. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**, 1–16 (2021).

43. Heigwer, F. *et al.* Time-resolved mapping of genetic interactions to model rewiring of signaling pathways. *Elife* **7**, (2018).

44. Fournier, T. *et al.* Steady-state expression of self-regulated genes. *Bioinformatics* **23**, 3185–3192 (2007).

45. Xu, Q., Georgiou, G., Veenstra, G. J., Zhou, H. & van Heeringen, S. ANANSE: An enhancer network-based computational approach for predicting key transcription factors in cell fate determination. *Biorxiv* 2020.06.05.135798 (2020) doi:10.1101/2020.06.05.135798.