UPPSALA
UNIVERSITET

UPTEC X 15 027
Examensarbete 30 hp
Augusti 2015

Identification of new regulatory mechanisms that determine coagulation FXI plasma concentration

Niklas Handin

UPPSALA
UNIVERSITET

Degree Project in Bioinformatics

Masters Programme in Molecular Biotechnology Engineering,

Uppsala University School of Engineering

UPTEC X 15 027	**Date of issue 2015-08-30**
Author **Niklas Handin**	
Title (English) **Identification of new regulatory mechanisms that determine coagulation FXI plasma concentration**	
Title (Swedish)	
Abstract FXI is a protein in the coagulation cascade in humans proven to be involved in the propagation and stabilization of developing thrombi in previous studies. The regulatory mechanisms for FXI are not well understood. Therefore an investigation to find regulatory mechanisms for FXI was performed. A meta-analysis was conducted, consisting of six Genome-wide association studies in which the trait was FXI levels in plasma. Three genome wide significant loci were found that also were significant in a replication study, in which three cohorts not included in the discovery set were used. Functional annotation, pathway analysis and eQTL analysis of these three loci yielded three genes believed potentially responsible for the regulation of FXI. Two of these three genes were chosen for a microRNA binding prediction search. Several microRNAs were found, and one was analyzed with a luciferase reporter assay.	
Keywords FXI, F11, blood plasma, coagulation, GWAS, Meta-analysis, pathway analysis, eQTL, miRNA, luciferase reporter assay	
Supervisors **Maria Sabater, co-sup Lars Maegdefessel** **Karolinska Institutet**	
Scientific reviewer **Bengt Persson** **Uppsala University & Karolinska Institutet**	
Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **74**

Biology Education Centre Biomedical Center Husargatan 3, Uppsala
Box 592, S-751 24 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 471 4687

Identification of new regulatory mechanisms that determine coagulation FXI plasma concentration

Niklas Handin

Populärvetenskaplig sammanfattning

Blodproppar är när en propp bildas på grund av att blod koagulerar inuti blodkärl. När en blodpropp lossnar och förs med i blodsystemet kan den sedan fastna när kärlen blir för smala. Detta kan medföra stroke och andra livshotande skador på kroppen. Faktor XI är ett protein som är delaktig i blodkoagulationen. Det har visat sig att Faktor XI har stor inverkan på utbredning och stabilisering av blodproppar. Förutom att höga nivåer av Faktor XI ökar risken för blodproppar så har det även visat sig att låga nivåer inte medför de stora problemen med inre blödningar som Faktor XIII och Faktor IX har.

Därför letade vi efter mekanismer som kan reglera Faktor XI. Vi hittade tre specifika platser i genomet som var signifikant associerade med Faktor XI-nivåer i blodplasma. Vi undersökte vad i dessa platser som har inverkan på nivåerna av Faktor XI. Slutligen gjordes en undersökning efter små RNA som kunde minska uttrycket av Faktor XI genom att binda till Faktor XI-genen. Från de flertal möjliga små RNA som hittades valdes en att bli validerad med ett försök *in vitro* (i levande celler). Försöket visade att det finns indikationer på att valt RNA kan förändra uttrycket av Faktor XI.

Examensarbete 30 hp
Civilingenjörsprogrammet i Molekylär bioteknik

Uppsala universitet, augusti 2015

Table of Contents

Abbreviations

aPTT	Activated Partial Thromboplastin Time
ATE	Arterial ThromboEmbolism
EAF	Effect Allele Frequency
eQTL	Expression Quantitative Trait Loci
GC	Genomic Control
GWAS	Genome Wide Association Study
LD	Linkage Disequilibrium
MAC	Minor Allele Count
MDS	MultiDimensional Scaling
N	Number of studies/individuals
QC	Quality Control
SE	Standard Error
SNP	Single-Nucleotide Polymorphism
TFBS	Transcription Factor Binding Site
VTE	Venous ThromboEmbolism

Chapter 1 Introduction

Arterial thromboembolism (ATE) and venous thromboembolism (VTE) are major public health problems today. It is estimated that 300,000 – 600,000 individuals are affected by VTE each year in the United States alone. It is also considered the leading preventable cause of death in hospitals in the United States (Beckman et al. 2010).

A thrombus is defined as a specific blood clot and thrombosis is the general condition of blood clotting somewhere in the blood circulatory system. Thromboembolism is a combination of thrombosis and embolism, in which a mobile thrombus is lodging and occluding the bloodstream in a downstream location.

FXI is involved in the propagation and stabilization of developing thrombi *in vivo* (Gailani and Broze 1991). FXI is a serine protease expressed in the liver and one of the central and initiating enzymes in the coagulation cascade (Fujikawa et al. 1986).

Heparin treatment is used as an anticoagulant to prevent and treat postoperative venous thromboembolism after total knee replacement. Lowering FXI levels in blood with an antisense oligonucleotide has been shown to be more effective than low molecular weight heparin (enoxaparin) treatment (Büller et al. 2015). FXI antisense treatment also had fewer patients experience clinically relevant bleeding complications than enoxaparin. Since high FXI levels are considered a risk factor for thrombosis, lowering FXI could be used for prevention of venous thrombosis (Büller et al. 2015). However, the genetic regulation of FXI is not well understood.

To find an association between common genetic variants and a diseases or disease-related traits that can be quantitative measured, the Genome-Wide Association Study (GWAS) approach has been applied successfully in previous investigations (Hindorff et al. 2009). But individual GWAS can be too small to give the necessary power to detect associations, when accounting for the number of multiple independent tests. Combining multiple GWAS to a single meta-analysis can be an effective approach to increase the prediction power.

Therefore we performed a meta-analysis of GWAS data with association between common genetic variants and FXI levels in plasma. Expression analyses in several tissues were also conducted. To further investigate regulation of FXI levels, microRNA (miRNA) binding prediction to genes associated with FXI levels was performed. The chosen miRNAs were validated using a luciferase reporter assay.

1.1 Aims

The aims of this thesis are to 1) find regulatory mechanisms of FXI levels in plasma using a meta-analysis of GWAS data, 2) functionally annotate the discovered regulatory mechanisms and 3) find miRNAs with predicted binding sites in mRNAs that regulate FXI expression levels.

1.2 Ethics

It is ethically motivated to restrict the distribution and sharing of genome-wide genotype and phenotype data to protect individual's privacy. In this thesis no patient data or individual genotyped data were handled. The data used was an association between a specific single nucleotide polymorphism (SNP) and FXI levels. The association was calculated by researchers that investigated the specific SNPs, with no possibility to trace the data back to a single individual. We therefore conclude that there is no breach in any individual's integrity.

Chapter 2 Materials and Methods

2.1 Data

The discovery set used consists of 16 809 samples from six European cohorts. The data consist of imputed SNPs and their association with FXI plasma levels in blood. The association was calculated using a linear regression model between every SNP and natural log-transformed levels, adjusted for age, sex, population stratification (and case-control status when necessary). The six different cohorts (FXI-1 to FXI-6) measured FXI either with an activity-based assay (Coagulometry: FXI-2, FXI-3 and FXI-6) or with an antigen method (ELISA like: FXI-1, FXI-4 and FXI-5).

2.2 Meta-data Quality Control

The six association studies were processed through a cleaning script (Supplemental Script 1) to ensure data quality for further downstream analysis and to create a common structure. For quality control the software EasyQC (Winkler et al. 2014) version: 9.0 for imputed 1000G data was used. The script performs several checks to validate the data. SNPs with missing alleles, p-value, beta, standard error (SE), effect allele frequency (EAF), individuals or imputation information were removed from the study. SNPs with values outside the range of the attribute (for example p-value not in the interval [0,1]) were removed from the study. SNPs were filtered by minor allele count (MAC, >5), and imputation quality (>0.3). Monomorphic SNPs were excluded. To reduce the file size, all data were converted to four significant digits.

2.3 Meta-analysis

GWAMA (Magi et al. 2010, Mägi and Morris 2010) version 2.1 was used to perform a meta-analysis of GWAS data after quality control (QC). Genomic Control (GC) was used to correct for inflation of the individual studies. We repeated the analysis with random effects to account for heterogeneity.

2.4 Selecting top SNPs

We selected a list of five relevant top SNPs for downstream analysis using the following criteria: SNP with lowest p-value in a region of 2 Mbp, p-value ($<10^{-6}$), not on sex chromosomes, present in all cohorts and showing the same effect direction in all cohorts. Regions containing the five top SNPs were plotted with regional plots. LocusZoom 1.3 (Pruim et al. 2010) was used to create regional plots. LocusZoom visualize the location of SNPs and genes in the chosen region. It also shows the linkage disequilibrium (LD) between the target SNP and all others. SNPs of interest were plotted with 500 kb distance both up and down-stream of the SNP's location. Three of the selected SNPs were genome-wide significant (*p-value of* $5*10^{-8}$ (Barsh et al. 2012)) and two were not genome-wide significant but showed suggestive association p-values (lower than 10^{-6}).

To produce regional plots and Manhattan plots, all autosome SNPs in all cohorts with a p-value of < 0.05 were plotted. The selection procedure was performed by scripting in AWK.

2.5 Functional annotation

Functional annotation was done for the 5 top SNPs with ANNOVAR version 2014-11-12 (Wang et al. 2010). ANNOVAR gives the option to collect data from several databases. The databases used to functional annotate our SNPs of interest were refGene, knownGene, ensGene, cytoBand, evofold, gwasCatalog, phastConsElements46way, targetScanS, tfbsConsSites, wgRna, wgEncodeBroadHmmHuvecHMM, and wgEncodeBroadHmmHepg2HMM. RegulomeDB (Boyle et al. 2012) was also used as an alternative to ANNOVAR to find evidence on SNPs that affect the binding regions for a regulatory element.

2.6 Adiponectin association

An association between adiponectin levels and our top SNP (rs710446) was done in PLINK. ADIPOQ that encodes for adiponectin has been associated with cardiovascular disease (Gable et al. 2006) and lies in proximity (100 kbp downstream) of our top SNP (rs710446). The covariates used were age, gender and population stratification. The data consisted of 3711 individuals, of which 3440 individuals (1674 males, 1792 females) passed QC. All participants in this database were of European decent and have at least three vascular risk factors. Adiponectin levels were measured with a double antibody radioimmunoassay (Millipore).

The natural logarithm levels adiponectin was used due to the untransformed levels of adiponectin showing non-normal distribution. Other transformations were also tested (Supplemental Figure 2).

2.7 Gene-based association study

Versatile Gene-based Association study (VEGAS) (Liu et al. 2010) version 0.8.27 was used to find genes that could be associated with FXI levels. VEGAS is unable to run on the hg19 build of the Homo sapiens and on hapmap data. Therefore it had to be extended with the 1000genome data and also extended to hg19. This was done with AWK scripting, TABIX and PLINK v1.90b2m (Purcell et al. 2007) on data from the 1000 genomes ftp site (*ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/*). After this,VEGAS was run successfully with the population set to 1000gEUR. All autosome SNPs with p-value <0.05 and in all cohorts were used as input.

To choose the top genes after VEGAS, a selection step was performed. Genes with rank 50 in gene p-value or best SNP-p-value were selected and concatenated. Both, gene p-value and SNP p-value, were required to be lower than 10^{-4} in order for the gene to be selected. SNPs with the lowest p-value in a gene with LD $\Delta^2 > 0.5$, and with SNPs in another gene with a lower p-value were removed.

2.8 Pathway analysis

Pathway analysis was performed on the raw output from VEGAS. All interactions between the genes with lower p-value than 10^{-3} were collected from the databases GeneMANIA, MENTHA, BioGrid, intact, MINT, UniProt and Reactome. The software to collect the data and to visualize the result was CytoScape Version 3.2.0 (Saito et al. 2012). PINBPA version 1.1.6 (Wang et al. 2015) was used to map the data.

Gene Relationships Among Implicated Loci (GRAIL) Beta from Broad institute (Raychaudhuri et al. 2009) was used to get a further understanding of the pathway of our genes. Since GRAIL does not work with 1000 genomes data, we converted our SNPs from 1000 genomes hg19 to hapmap CEU hg18. This was done (Supplemental Script 2) with the following parameters R^2 (0.2) and distance (100 kb). SNPs that had a p-value less than $1*10^{-8}$ from the VEGAS output was used as input to GRAIL.

Pathway Analysis by Randomization Incorporating Structure (Paris) version 1.1.3 (Yaspan et al. 2011) was also used to find pathways (Supplemental Script 3). All autosome SNPs present in all cohorts with p-value < 0.05 were used as input to the Paris analysis.

2.9 Expression Quantitative Trait Loci (eQTL)

eQTL analysis was performed on the SNPs with the lowest p-value for each genes selected after VEGAS, to see if SNPs may cause an expression change in *cis* genes at 250kb. eQTL was done in collaboration with another group at Karolinska institutet in Sweden. The samples used are from the Advanced Study of Aortic Pathology (ASAP) at the Karolinska University Hospital, Stockholm and include patients undergoing aortic valve surgery. The different tissues used for the eQTL are mammary artery intima-media (89 samples), liver (212 samples), aortic media (138 samples), aortic adventitia (133 samples), and heart (127 samples).

The RNA was isolated with Trizol (BRL-Life Technologies) and treated with RNasefree DNase set (Qiagen) following the manufacturer’s instruction. RNA quality was analyzed with the Agilent 2100 bioanalyzer (Agilent Technologies Inc, Palo Alto, Calif), and quantity was measured using NanoDrop (Thermo Scientific Waltham, Mass). Gene expression was generated using Affymetrix ST 1.0 Exon arrays (Affymetrix, Santa Clara, CA) and Affymetrix Meta prob set (Affymetrix). Whole gene variation with genotype and QC procedures have been reported (Folkersen et al. 2010).

The eQTL was performed using a linear regression model and corrected with FDR and Bonferroni separately.

2.10 **miRNA binding prediction**

To find miRNA believed to bind and change expression of genes associated with FXI levels, a scoring algorithm was produced. The scoring algorithm is described in the following lines. miRNAs get one point for every passed of the following steps: TargetScan (context+ score < 0.045), miRanda (Score > 145), SVR good hits, miRWalk 2 5' UTR algorithms (>60%). miRWalk 2 3' UTR algoritms (>50%), miRWalk 2 promoter (>75%). In MirWalk not all algorithms can predict binding in all regions, therefore the same cut-off cannot be used for all regions. Only miRNA that were expressed in the liver and had at least 4 points were selected.

To get the most updated data available, a local installation of TargetScan v 6.2 (Friedman et al. 2009, Garcia et al. 2011, Grimson et al. 2007, Lewis et al. 2005) and miRanda 3.3a (John et al. 2004) was performed to run on mirBase release 21 (Griffiths-Jones et al. 2008, 2006, Griffiths-Jones 2004, Kozomara and Griffiths-Jones 2014, 2011) so that new miRNA not featured in TargetScan's or miRanda's online algorithms could be detected. The regions (3' UTR, 5' UTR and promoter) to investigate for miRNA:mRNA binding were collected from TargetScan's own database.

SVR files from August 2010 Release of microRNA.org (Betel et al. 2010) were used. Good mirSVR scores (defined from microRNA.org) with both conserved and non-conserved sites were used, but not those with insufficient mirSVR scores.

MiRWalk 2 (Dweep et al. 2011) is a concatenation of different algorithms and databases. All available databases and algorithms (maximal of 12) of miRWalk 2 were chosen for 3' UTR, 5' UTR and promoter regions. All databases/algorithms cannot be utilized on all three regions (3' UTR, 5'UTR and promoter), because not all algorithms in miRWalk 2 were built to work on all regions. We were interested in the total number of algorithms that predict binding between miRNA and target gene.

MiRWalk 1 (Dweep et al. 2011) was also used to see if any difference exists in relation to to miRWalk 2. The same parameters as for miRWalk 2 were used.

Tarbase (Vlachos et al. 2015) was used to investigate if miRNAs found were already experimentally validated and to give the miRNAs a miTG score.

To find out which miRNAs were expressed in liver, three different databases were used: 1) SmirnaDB version 2009-05-08 from Swiss Institute of Bioinformatics. All miRNA that had more than one clone count in liver samples were assumed to be expressed in liver; 2) microRNA.org version 2010-11-01 (Landgraf et al. 2007). The data here is represented in the form of clone counts normalized to the total number of miRNAs that were cloned in each library. All miRNAs with a value above 0.00072 of the normalized clone counts were assumed to be expressed in liver; 3) data used in the article by Salloum-Adfar et al. (2014), which established the cut-off at 500 au were also utilized as well. Any miRNA that passed any of the selections was assumed to be expressed in the liver.

2.11 **Luciferase reporter assay**

To validate if the miRNA of interest changes the expression of FXI, a luciferase reporter assay was performed. In a luciferase reporter assay, cells are transfected with a vector containing a luciferase gene. Cells are then stimulated to induce expression of the luciferase. It is then possible to quantify the expression of the luciferase using the substrate, if any, to detect luciferase activity using a luminometer. By adding a 3'UTR section after the luciferase gene in the vector, and subsequent co-transfection with a miRNA, one can assess the effect a specific miRNA has to different vectors. In this case six different conditions were tested: 1. Empty vector and a scrambled miRNA, 2. Empty vector and hsa-miR-145-5p, 3. Empty vector and has-miR-181-5p, 4. FXI vector and a scrambled miRNA, 5. FXI vector and hsa-miR-145-5p, 6. FXI vector and hsa-miR-181-5p. The empty vector is the vector without the 3' UTR of FXI and the FXI vector contains the 3' UTR of FXI (Supplemental figure 1).

HEK293 (Human embryonic kidney cells, Sigma St. Louis, United States of America) were used due to their ease of growth and transfection. The HEK cells were cultured in Dulbecco's Modified Eagle Medium and 10% Fetal bovine serum (DMEM, FBS, Life technologies, Carlsbad, United States of America).

To perform the Luciferase reporter assay, cell transfection is necessary. Cells were seeded in a 24 well plate 24 hours before transfection. The cells were co-transfected (FuGENE® HD Transfection Reagent, Promega United States of America, Madison) when at 70% confluence with 10 nM of double-stranded miRNA (miR-145 or miR-181a or mirVana™ miRNA Mimic, Life technologies, Carlsbad, United States of America) and 100 ng of vector (Active Motif, LightSwitch™ GoClone™ Collection, 3' UTR FXI vector or EMPTY_3UTR, Carlsbad, United States) according to manufacturer's instructions. The amount of miRNA was later increased to 20 nM and the vector increased to 400 ng to improve the transfection efficiency.

The luciferase assay (LightSwitch™ Luciferase Assay Kit, Carlsbad, United States) was used as previously described (Maegdefessel et al. 2014). The luminescent was measured in a 96 well plate for 2 seconds (GLOMAX MULTI+ Detection system, Promega United States of America, Madison).

Chapter 3 Results

3.1 Meta-data Quality Control

Six different (FXI-1 to FXI-6) cohorts were QC tested. QC is of importance to validate the data and see that it is of good quality. A QQ-plot is a plot comparing two probability distribution here representing the expected vs. observed –$\log_{10}$(p-value). From the QQ plot (Figure 1) there is an elevation of the curve prematurely for FXI -3 and FXI-5. An elevated curve suggests false positives. The elevation of FXI-5 was enough to be excluded from all analysis downstream of EasyQC.

The Effect Allele Frequency (EAF) plot is used to find data management errors, analytical errors or miss-specified effect alleles, which no cohorts showed (Figure 2).

The Lambda-N shows problems with population stratification. FXI-4 was above the threshold of 1.1, but not considered enough to be removed from downstream analysis (Figure 3).

The P-Z plot indicates problems with beta estimates, p-values or/and SE values if there is deviation from the identity line. FXI-3 does not align with the identity line and deviates lower and above the identity line. However, it was not deemed to be enough to be removed from downstream analysis (Figure 4). See Supplemental Table 1-3 for more QC-data.

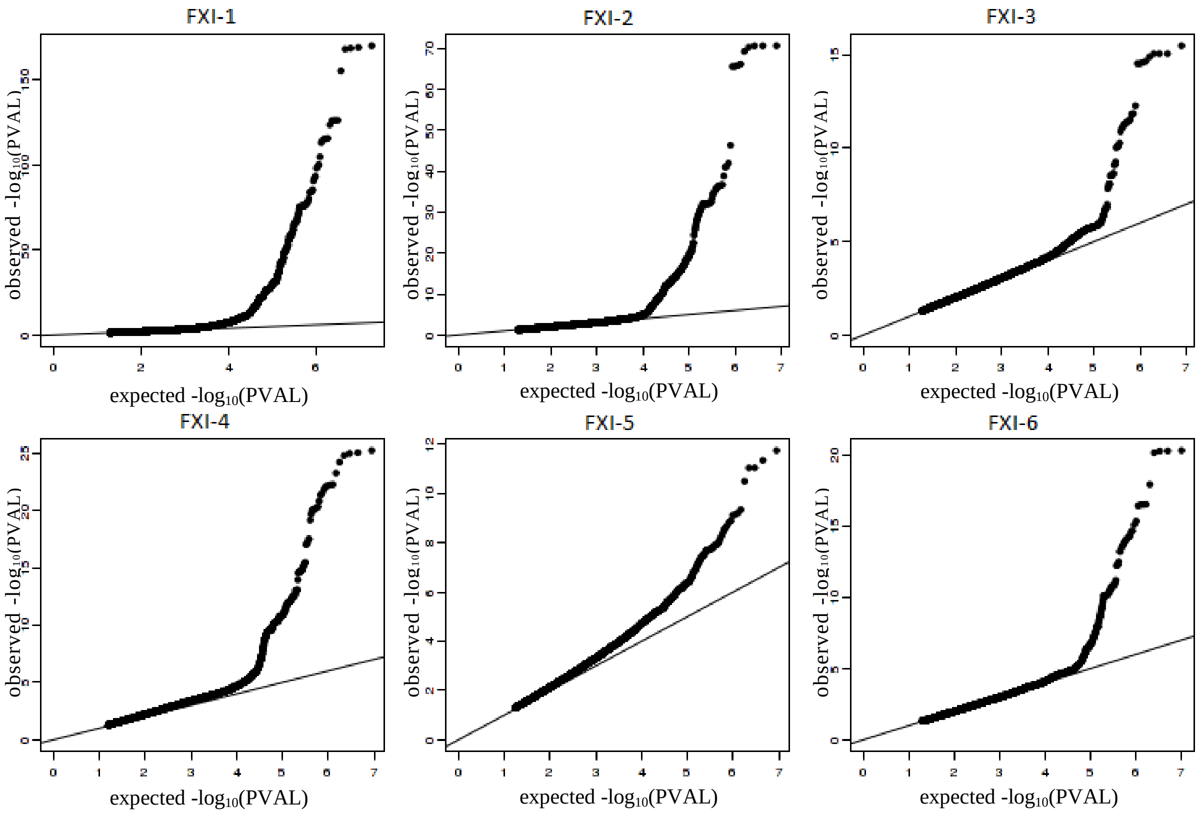


Figure 1. QQ-plot. The correlation between expected –log$_{10}$(p-value) and observed –log$_{10}$(p-value) for the different cohorts.

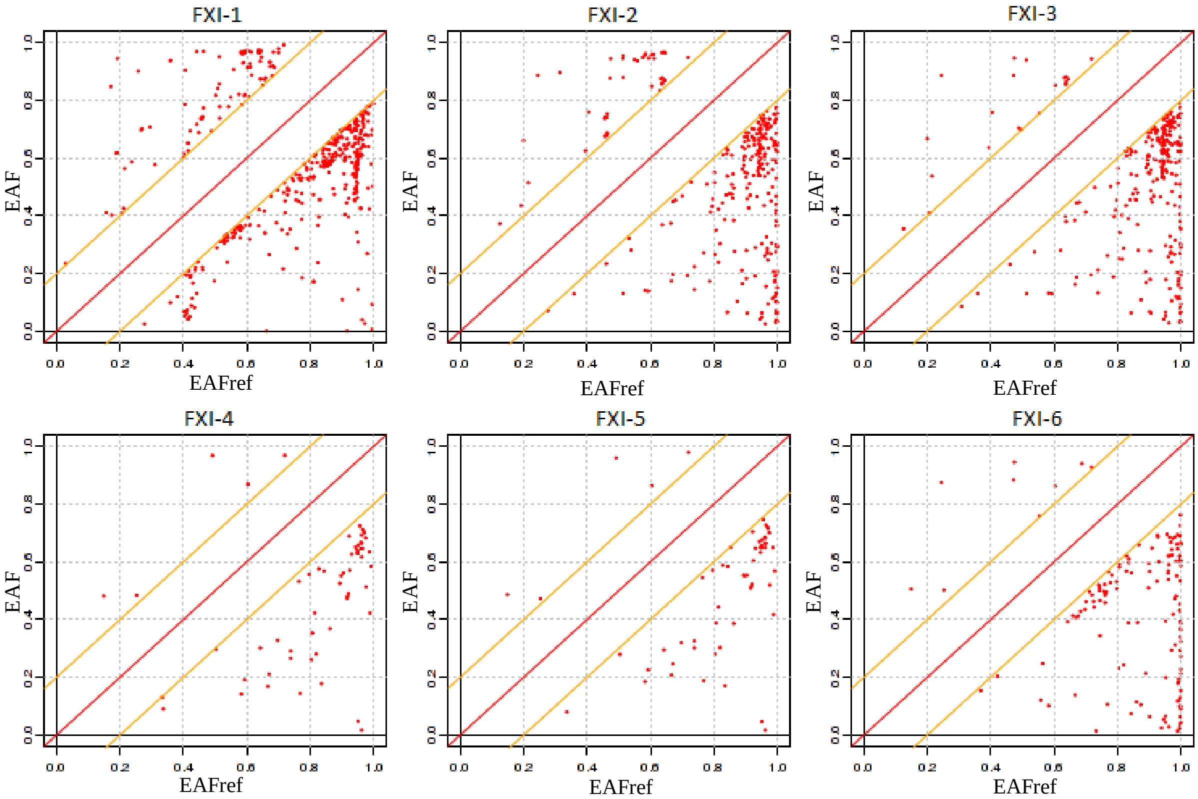


Figure 2. EAF of the reference vs EAF in the samples from the six cohorts. The outliers here are more likely to be samples of non-European ancestry than data management errors, analytical errors or miss-specified effect alleles. The region between the yellow lines indicates differences between EAF that is negligibly small. See Winkler et all 2014 for more information about the EAF plots.

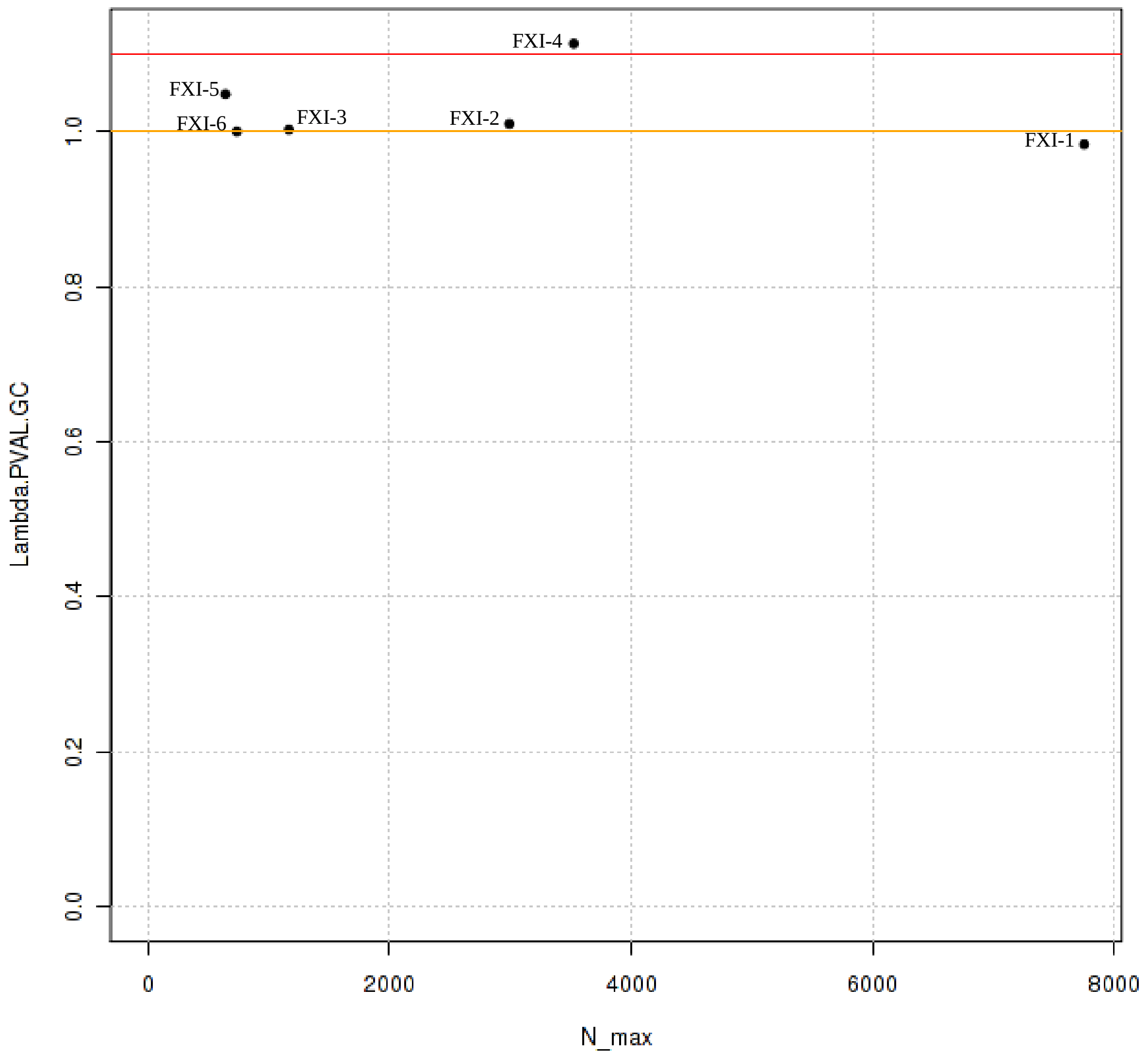


Figure 3. Lambda-N plot to evaluate if the different cohorts have issues with population stratification. From the left FXI-5, FXI-6, FXI-3, FXI-2, FXI-4 and FXI-1.Yellow line indicates no population stratification, values above the red line indicates problems with population stratification. See Winkler et al. 2014 for more information about Lambda-N plots.

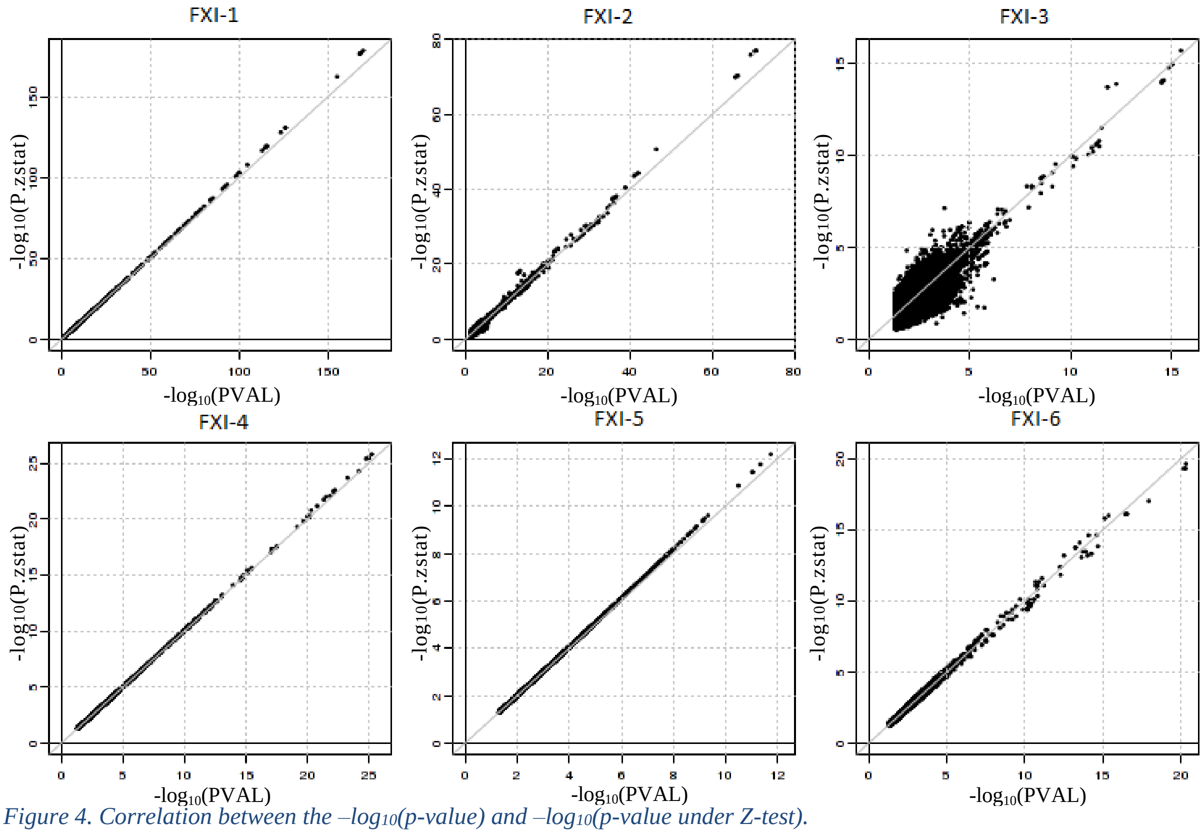


Figure 4. Correlation between the –log10(p-value) and –log10(p-value under Z-test).

3.2 Meta-analysis

3.2.1 Discovery and Manhattan-plot

From the meta-analysis of the GWAS data in the discovery set approximately 18 million SNPs were obtained. A Manhattan-plot was created (Figure 5 and 6) to illustrate the distribution of strong associations between SNPs and FXI levels (–log(p-values)) along the genomic coordinates from the meta-analysis. Results from Manhattan plot show three genome-wide significant loci and two suggestive regions (with p-values between 10^{-6} and $5*10^{-8}$). After selection for top SNPs (see methods) we have three genome-wide significant SNPs and two suggestive SNPs (Table 1).

3.2.2 Replication

To validate the findings from the meta-analysis, a replication study was done by other members of the same KI group. The replication was done in three other cohorts not present in our discovery set. The combined data from the three cohorts consists of 2058 individuals with European ancestry. The three genome-wide significant top SNPs (rs710446, rs4253417, rs780094) are significant in the replication set (Table 2). The two SNPs under the significant threshold for GWAS were not significant in the replication set.

3.2.3 Heterogeneity

The three significant SNPs have high heterogeneity in the discovery set. A search for the cause was done. We repeated the GWAMA run but left one of the cohorts out at the time. But there are not one cohort that alone were responsible for the high heterogeneity. Two different ways to measure FXI were used in the discovery set. To investigate if the high heterogeneity is caused by the different ways of measuring FXI, the cohorts with coagulometry and ELISA like methods were separated and then rerunning the analysis. The heterogeneity decreased but not enough to say that the heterogeneity are due to the different ways to measure FXI (Supplemental Table 4).

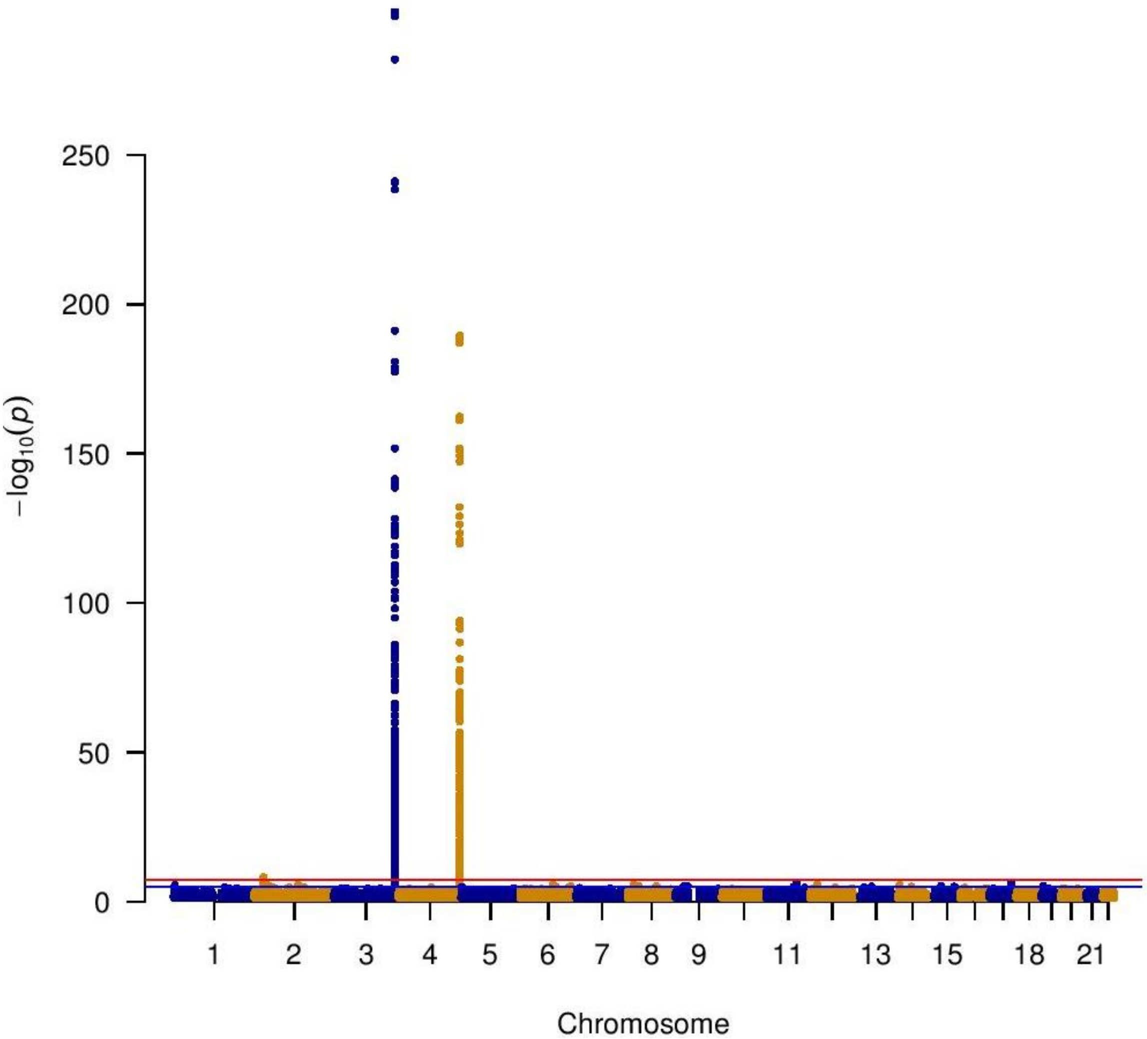


Figure 5. Manhattan plot of the results from GWAMA where the X-axis is genomic location increasingly and the y-axis is the –$\log_{10}$(p-value) where p-value represent the association with FXI levels from the Meta-analysis. The red line indicates genome-wide significant (-$\log_{10}(5*10^{-8})$). The blue line indicates a suggestive line (-$\log_{10}(1*10^{-5})$).

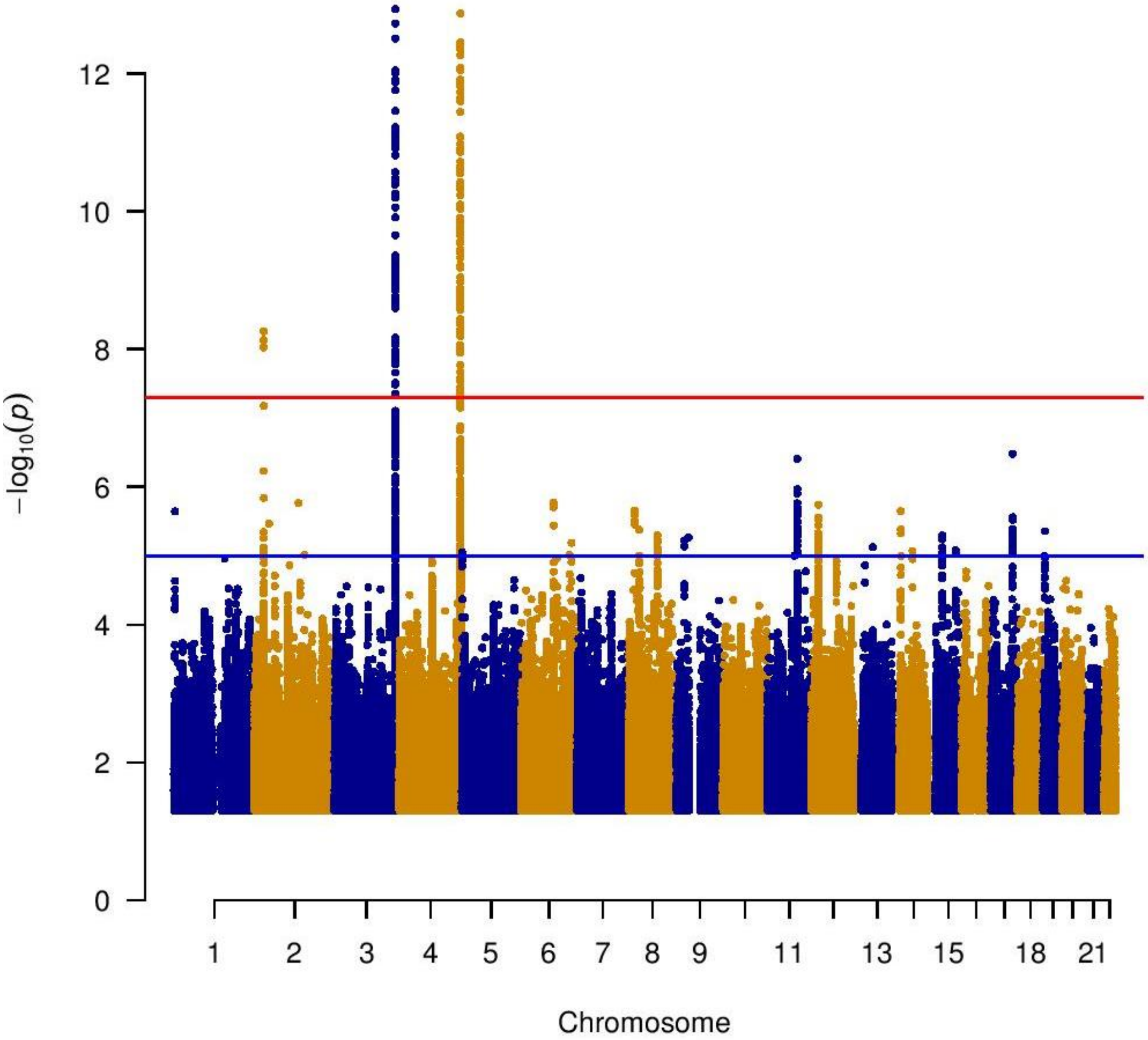


Figure 6. Manhattan plot of the results from GWAMA where the X-axis is genomic location increasingly and the y-axis is the -log$_{10}$*(p-value) where p-value represent the association with FXI levels from the Meta-analysis. The red line indicates genome-wide significant (-log*$_{10}$*(5*10*$^{-8}$*)). The blue line indicates a suggestive line (-log*$_{10}$*(1*10*$^{-5}$*)). This figure is a zoomed version of figure 5.*

Table 1. Top 5 SNPs after filtering step.

rs-number	chromosome	location	beta	SE	P-value	I^2	Effects*
rs710446	3	186459927	-0.0890	0.0024	6.95×10^{-299}	0.8510	-----
rs4253417	4	187199005	-0.0735	0.0025	3.04×10^{-190}	0.8882	-----
rs780094	2	27741237	0.0147	0.0025	5.50×10^{-09}	0.5520	+++++
rs78802760	17	66163686	-0.0196	0.0038	3.30×10^{-07}	0.0000	-----
rs199841773	11	92254694	0.0135	0.0027	3.88×10^{-07}	0.2335	+++++

**Sign of beta in the different cohorts*

Table 2. Replication of top 5 SNP.

rs-number	chromosome	beta	SE	P-value	FDR p-value*	Effects**
rs710446	3	-0.1246	0.0094	8.86×10^{-40}	7.97×10^{-39}	---
rs4253417	4	-0.0898	0.0098	6.51×10^{-20}	2.93×10^{-19}	---
rs780094	2	0.0238	0.0096	0.01372	0.0412	+++
rs78802760	17	-0.0082	0.0138	0.5534	1.24515	++-
rs199841773	11	0.0063	0.0098	0.5174	0.93132	+-+

** FDR corrected p-value calculated with Benjamini Hochberg procedure (#tests*P-value/Rank, #tests is nine)*

*** Sign of beta in the different cohorts*

3.2.4 Regional plots

Regional plots of the five top SNPs are displayed in figures 7-11. The SNP with rs-number rs199841773 did not have information to calculate the LD. Therefore SNP rs505383 was used to get information about the LD in that region.

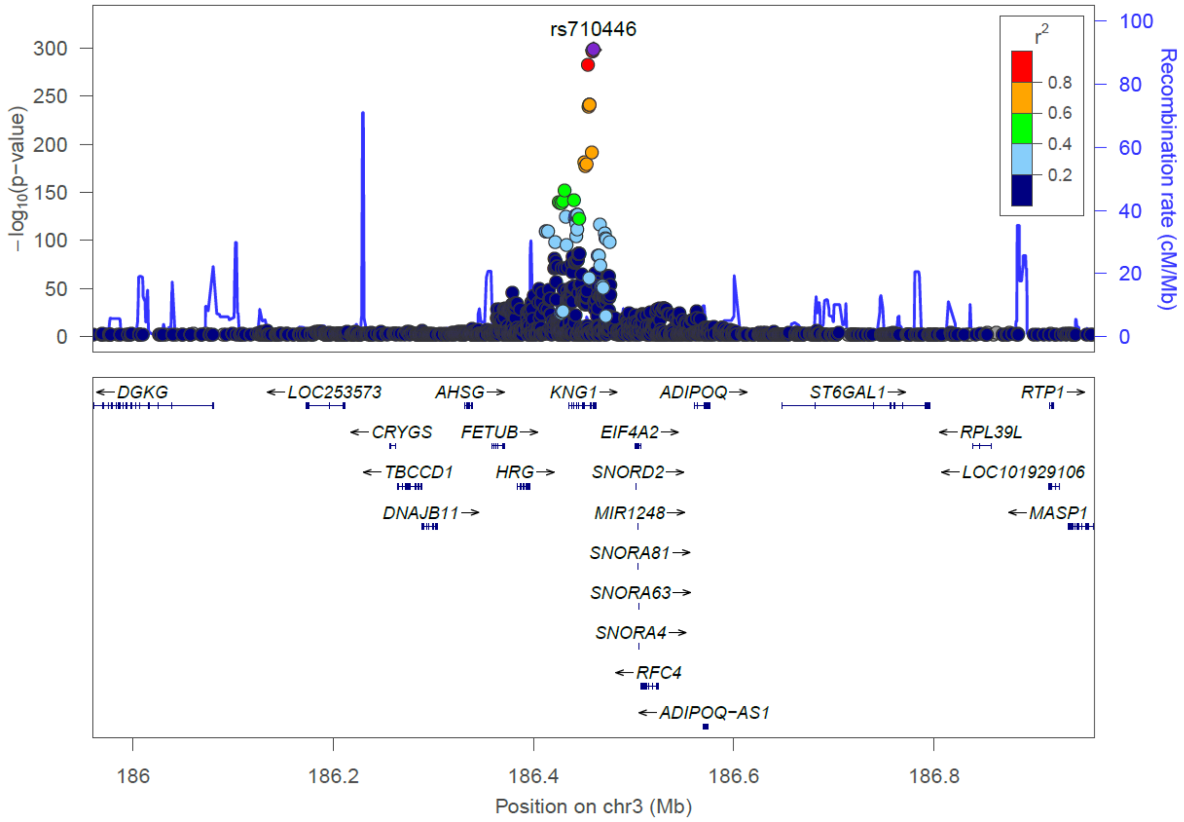


Figure 7. Regional plot of rs710446. On the x-axis genomic location and below genes in that loci. On the y-axis –log₁₀(p-value) from GWAMA and also the recombination rate (cM/Mb). The color of the SNPs represent the r^2 value where red is r^2 close to one and dark blue close to zero.

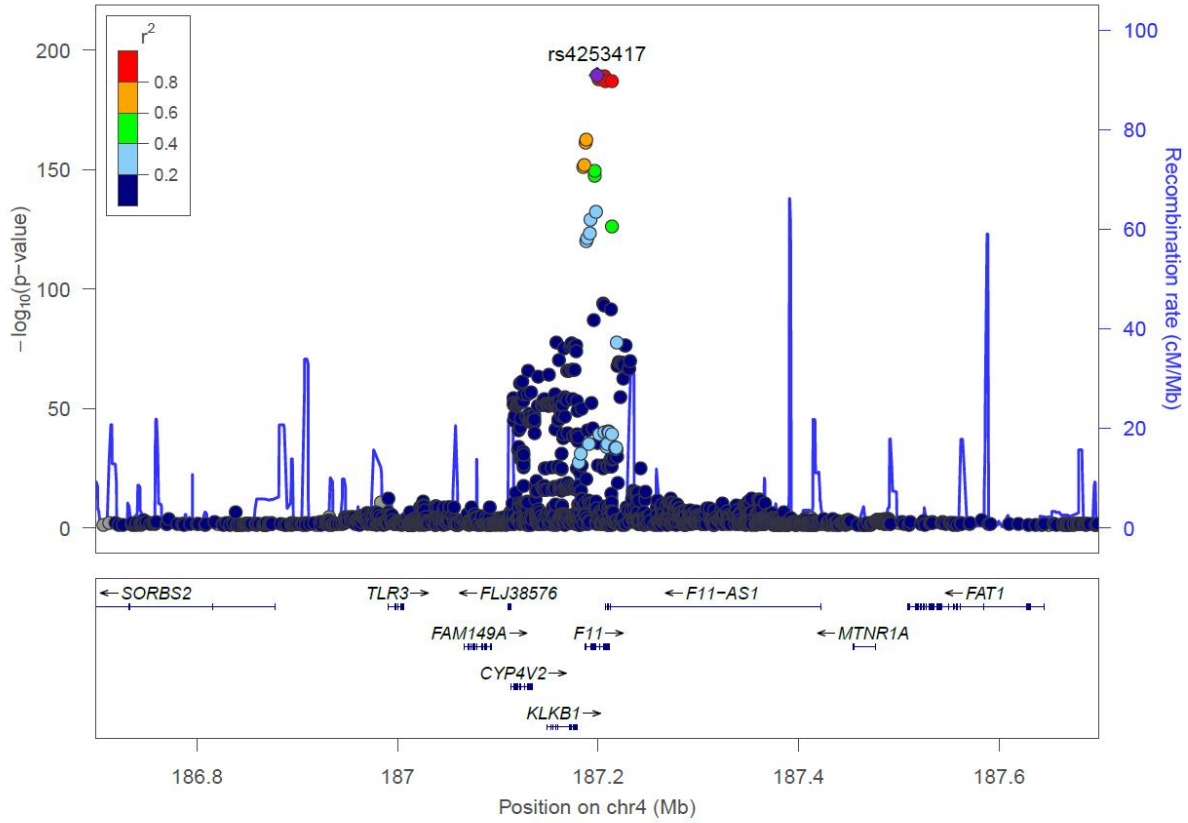


Figure 8. Regional plot of rs4253417. On the x-axis genomic location and below genes in that loci. On the y-axis –log$_{10}$(p-value) from GWAMA and also the Recombination rate (cM/Mb). The color of the SNPs represent the r^{2}value where red is r^2 close to one and dark blue close to zero.

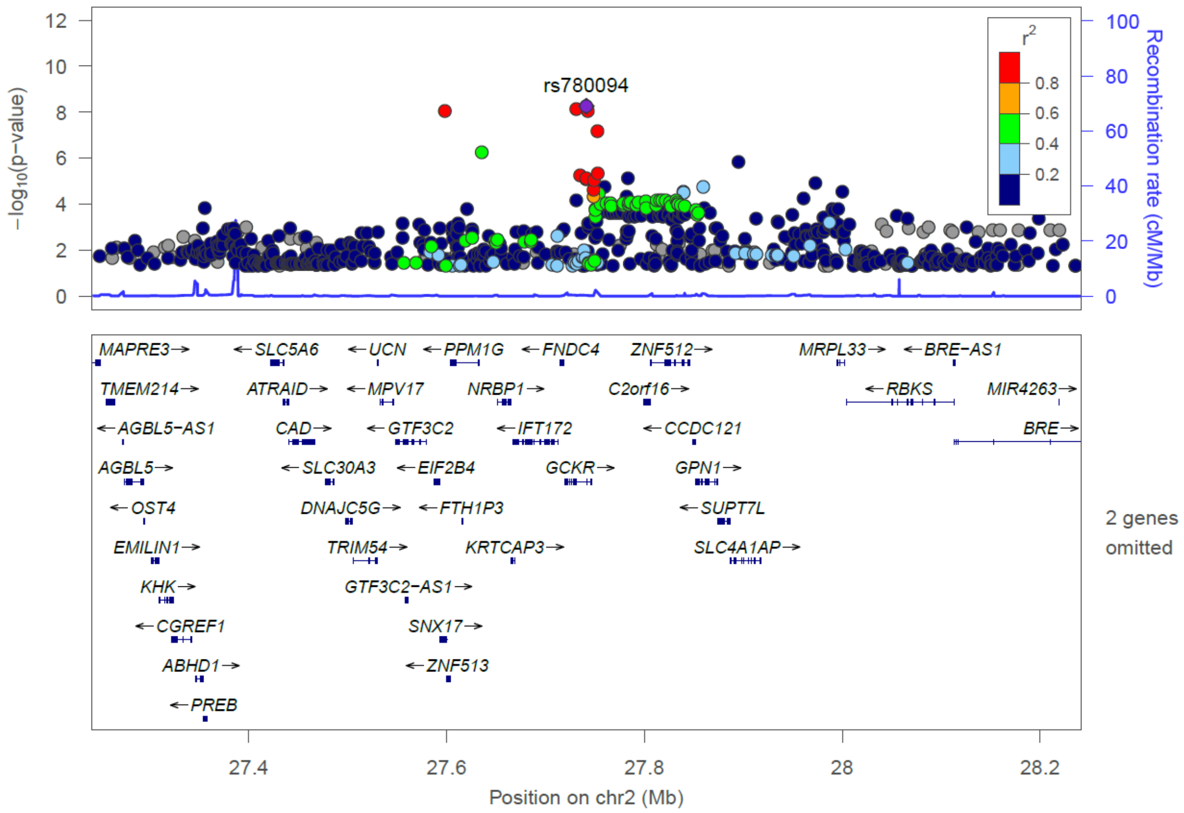


Figure 9. Regional plot of rs780094. On the x-axis genomic location and below genes in that loci. On the y-axis –log$_{10}$(p-value) from GWAMA and also the Recombination rate (cM/Mb). The color of the SNPs represent the r^{2}value where red is r^2 close to one and dark blue close to zero.

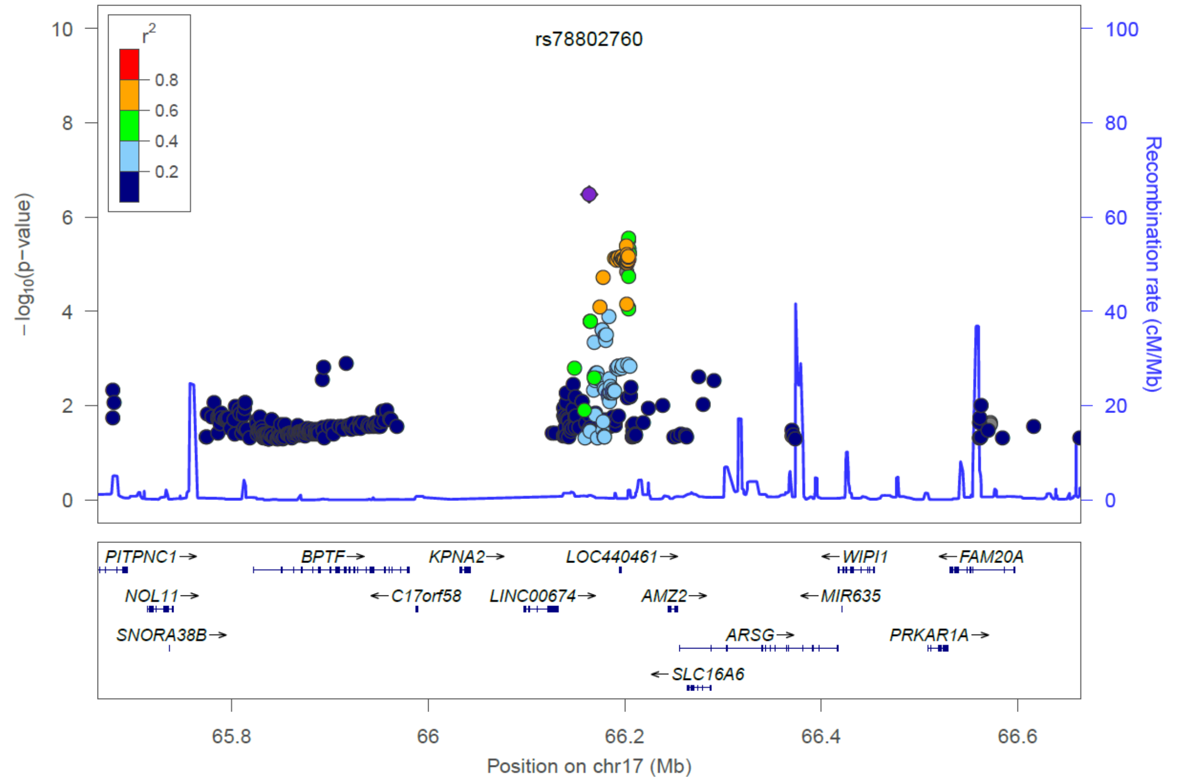


Figure 10. Regional plot of rs78802760. On the x-axis genomic location and below genes in that loci. On the y-axis –log$_{10}$(p-value) from GWAMA and also the Recombination rate (cM/Mb). The color of the SNPs represent the r^{2}value where red is r^2 close to one and dark blue close to zero.

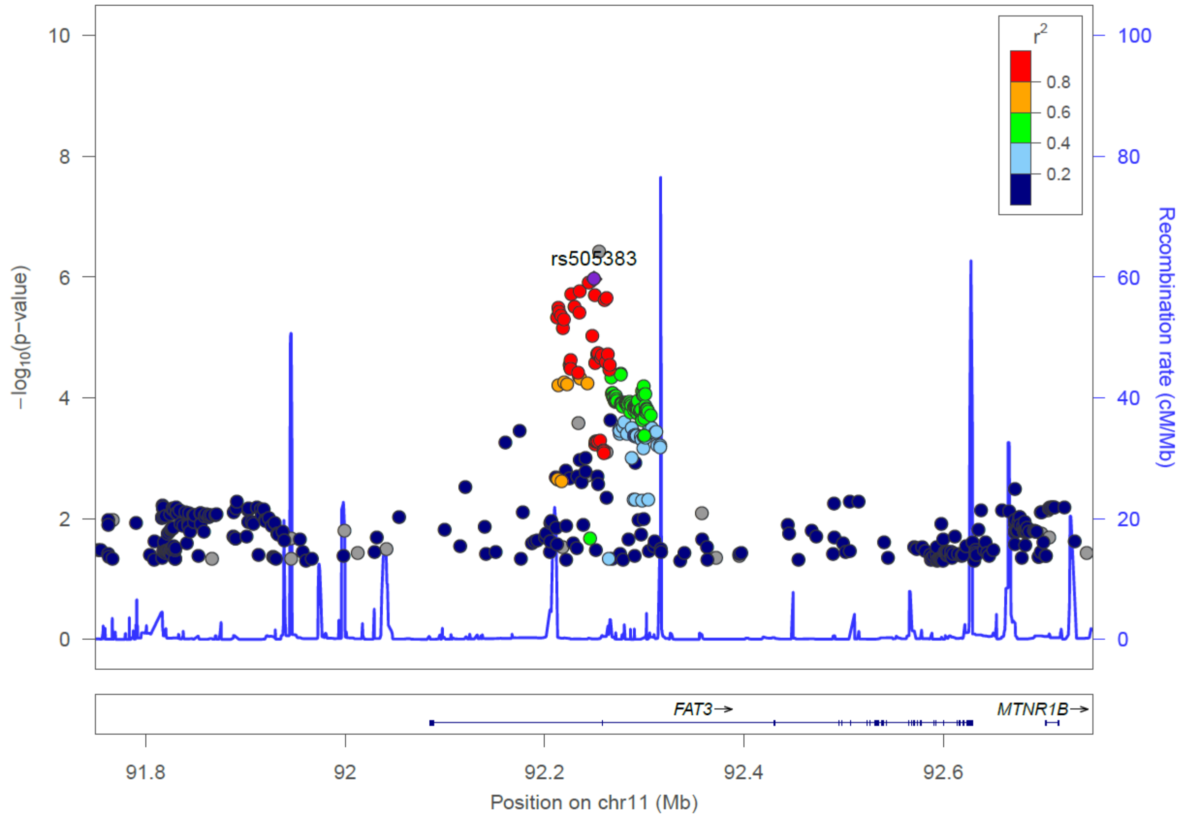


Figure 11. Regional plot of rs505353. On the x-axis genomic location and below genes in that loci. On the y-axis –log_{10}(p-value) from GWAMA and also the Recombination rate (cM/Mb). The color of the SNPs represent the r^2value where red is r^2 close to one and dark blue close to zero. The SNP rs505353 is just below rs199841773 so it was used to get an idea of LD in that region because rs199841773 does not give enough information to calculate a good LD score.

3.3 Functional annotation

Relevant information from Annovar for the top five SNPs are shown in table 3. Rs4253417 and rs780094 are intronic regions in the *FXI* and *GCKR* genes respective while rs710446 is in the exonic region of the *KNG1*gene. Rs710446 was found to associate (in previous GWAS) in activated partial thromboplastin time (aPTT). Rs780094 has been identified in several GWAS studies (see table 3). None of these SNPs was in any known transcription factor's binding site but from the chromHMM predictions we see that rs780094 got classified as a strong enhancer. chromHMM is a software utilizing the Hidden Markov Model (HMM) for characterizing chromatin states.

Table 3. Information from Annovar for the top 3 SNPs.

rs-number	Gene	Func	Exonic Function	GWAS Catalog	BroadHmm Hepg2HMM
rs710446	KNG1	exonic	Non synonymous SNV	Activated partial thromboplastin time	Weak Enhancer
rs4253417	FXI	intronic	NA	NA	Weack transcribed
rs780094	GCKR	intronic	NA	Uric acid levels, Metabolic syndrome, Urate levels, Calcium levels, Metabolic traits, Fasting insulin-related traits (interaction with BMI), Fasting glucose-related traits (interaction with BMI), Phospholipid levels (plasma), C-reactive protein, Triglycerides, LDL cholesterol, Fasting glucose-related traits, Fasting insulin-related traits	Strong Enhancer
rs78802760	None	intergenic	NA	NA	Heterochromatin, low signal
rs199841773	FAT3	intronic	NA	NA	Heterochromatin, low signal*

**BroadHmmHuvecHMM*

From RegulomeDB, rs780094 was predicted to likely affect binding. The supporting data for this location is Transcription factor (TF) binding, matched TF motif and DNase peak. Rs710446, rs4253417 and rs199841773 got no data supporting binding. Rs78802760 has TF binding or DNase peak, and represent minimal binding evidence in regulomeDB.

3.4 Adiponectin association

Top SNP (rs710446 in KNG1) is close (100 kbp upstream) to the gene ADIPOQ. An association test between rs710446 and adiponectin levels were conducted to see if there exists an association with adiponectin levels. The result was not significant with a p-value of 0.1996 and a BETA of -0.0297 (SE=0.0232).

3.5 Gene association

VEGAS generates a list of genes and its gene-based test statistics. This gives a list of genes that could be associated with our phenotype. Table 4 shows the top 20 genes based on SNP p-value. GCKR and SNX17 are the genes with the best association in the locus on chromosome 2, where rs780094 is the SNP with lowest p-value in GCKR. The SNX17 gene only has four SNPs inside and could be biased. In the locus on chromosome 3 (where rs710446 has the lowest p-value), there are several genes with a p-value of 0. But the gene KNG1 has the lowest SNP p-value, most SNPs and the highest sum of all the chi-squared tests in this locus. The locus on chromosome 4 (rs4253417 lowest p-value) has five genes with a p-value of 0 (FXI, FXI-AS1, KLKB1, CYP4V2, TLR3). After selecting the top genes from the VEGAS output, a total of 31 genes were selected. The output from VEGAS was also used to analyze pathways.

Table 4. Top 20 Genes (Based on SNP-pvalue) from VEGAS

Chr	Gene	nSNPs	nSims	Start	Stop	Test[*]	p-value	Best-SNP	SNP p-value
2	GCKR	26	10^6	27719705	27746550	294	1.80×10^{-05}	rs780094	5.50×10^{-9}
2	SNX17	4	10^6	27593362	27600400	50	0	rs4665972	9.46×10^{-9}
2	SLC4A1AP	10	10^6	27886337	27917847	80	3.60×10^{-05}	rs2178198	1.44×10^{-6}
3	KNG1	144	10^6	186435097	186462199	44478	0	rs710446	6.95×10^{-299}
3	RFC4	27	10^6	186507681	186524484	1601	0	rs266728	3.70×10^{-29}
3	FETUB	23	10^6	186358148	186370797	1255	0	rs6767451	7.96×10^{-29}
3	ADIPOQ	35	10^6	186560462	186576252	878	0	rs73185702	5.48×10^{-27}
3	HRG	17	10^6	186383746	186396023	833	0	rs1042445	4.92×10^{-26}
3	EIF4A2	8	10^6	186501360	186507685	332	0	rs266720	1.23×10^{-22}
3	ADIPOQ-AS1	9	10^6	186569675	186573912	184	0	rs2241766	8.09×10^{-10}
3	AHSG	15	10^6	186330849	186339107	197	1.00×10^{-06}	rs35799453	2.19×10^{-8}
4	FXI	58	10^6	187187117	187210835	13421	0	rs4253417	3.04×10^{-190}
4	FXI-AS1	492	10^6	187207251	187422212	18851	0	rs2289252	7.33×10^{-188}
4	KLKB1	111	10^6	187148671	187179625	16549	0	rs4253253	3.39×10^{-78}
4	CYP4V2	76	10^6	187112673	187134617	9584	0	rs2276918	2.16×10^{-66}
4	TLR3	16	10^6	186990308	187006252	240	0	rs75357674	4.20×10^{-13}
4	FAM149A	54	10^6	187065994	187093817	527	3.00×10^{-06}	rs114742882	1.98×10^{-9}
4	FLJ38576	11	10^6	187110185	187112644	133	9.90×10^{-05}	rs35641294	3.54×10^{-7}
6	MCHR2	75	10^6	100367785	100442114	758	1.30×10^{-05}	rs11155193	1.69×10^{-6}
11	FAT3	221	10^6	92085261	92629635	2671	7.00×10^{-06}	rs505383	1.08×10^{-6}

[*] *Is the sum of all the chi-squared one df statistics of that gene.*

3.6 Pathway analysis

HRG and AHSG interacts with KNG1 through a secondary mechanism (Figure 12 and 13). We can therfore conclude that there exists a connection in form of a protein interaction between the locus on chromosomes 3 (rs710446 lowest p-value) and 4 (rs4253417 lowest p-value).

The Interaction map indicates that KNG1, FXI and KLKB1 interact with each other (top left corner of Figure 14). Apart from that, there is no cluster of top genes

Pathway analysis suggests that our data is a good representation of *complement and coagulation cascades*. This provides evidence that the analysis is a good representation of our FXI phenotype. Pathways with p-values below 0.0001 are not given and the top 3 pathways (Table 5) all have scores below 0.0001. Therefore they are sorted by Simple Feature Count (A single SNP in an area of linkage equilibrium) (Yaspan et al. 2011).

The GRAIL output shows a list of 20 keywords describing our set of SNPs associated with FXI levels. Some keywords were plasma, coagulation and blood. For the full keywords set and other output from GRAIL see Supplemental Table 5-6.

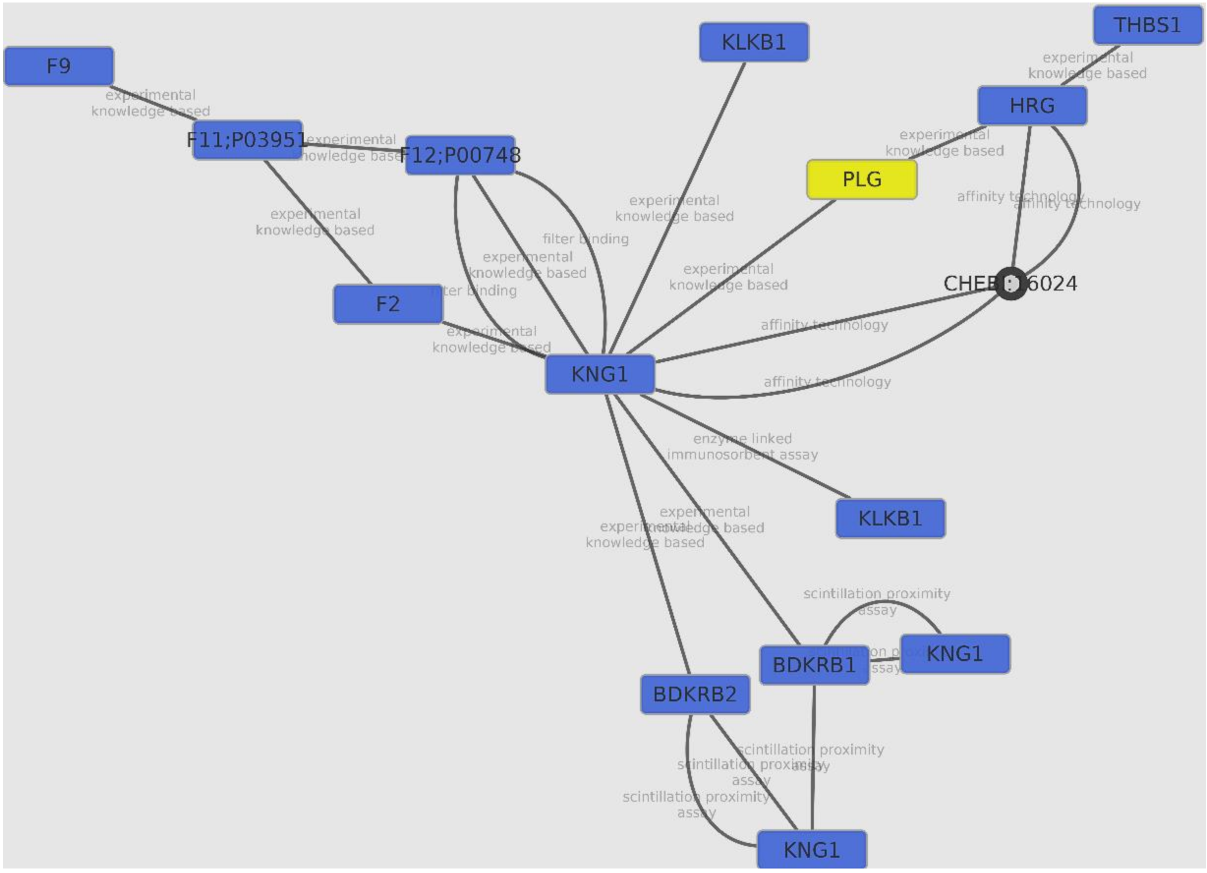

KLKB1
THBS1
F9
experimental knowledge based
HRG
experimental knowledge based
F11;P03951
experimental knowledge based
F12;P00748
PLG
experimental knowledge based
affinity technology
affinity technology
experimental knowledge based
experimental knowledge based
filter binding
experimental knowledge based
filter binding
experimental knowledge based
CHEBI:16024
F2
experimental knowledge based
affinity technology
KNG1
affinity technology
enzyme linked immunosorbent assay
experimental knowledge based
experimental knowledge based
KLKB1
scintillation proximity assay
BDKRB1
KNG1
BDKRB2
scintillation proximity assay
scintillation proximity assay
scintillation proximity assay
KNG1

Figure 12 Interaction map were lines indicate interactions. The text on the lines says how this interaction is proven. CHEBI:16924 is D-mannose

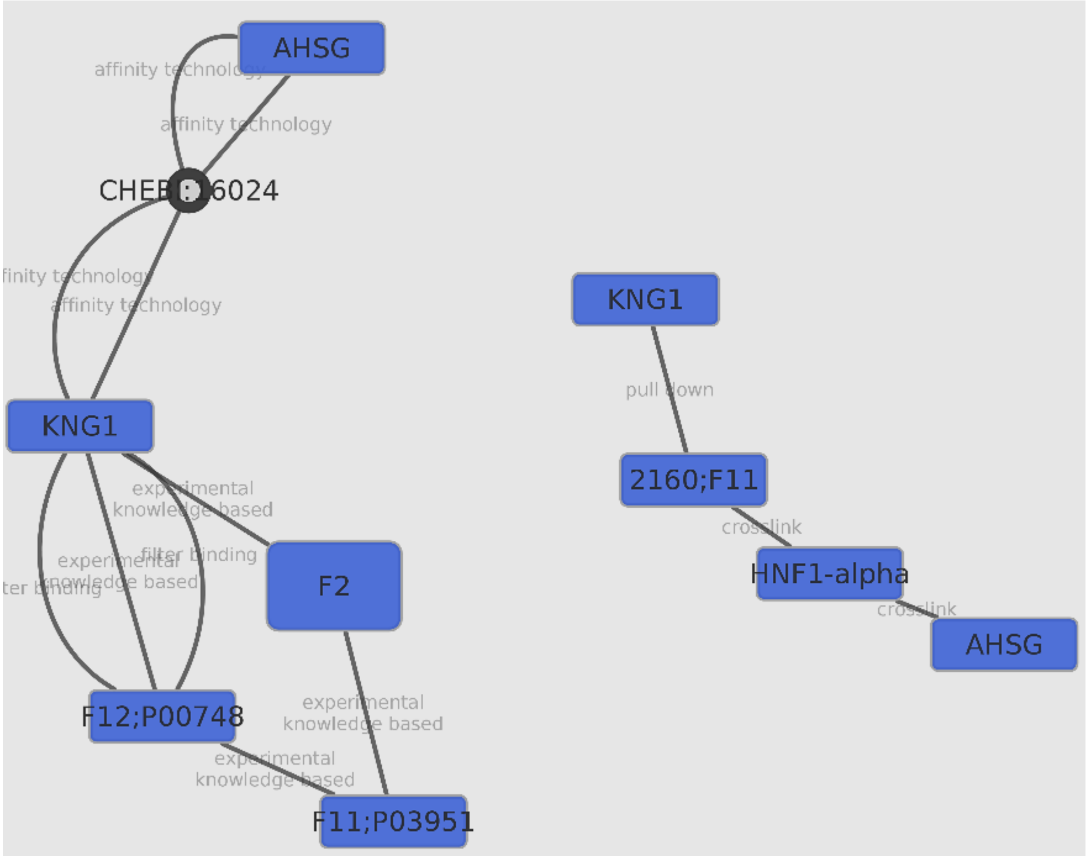


Figure 13. Interaction map were lines indicate interactions. The text on the lines says how this interaction is proven. CHEBI:16924 is D-mannose

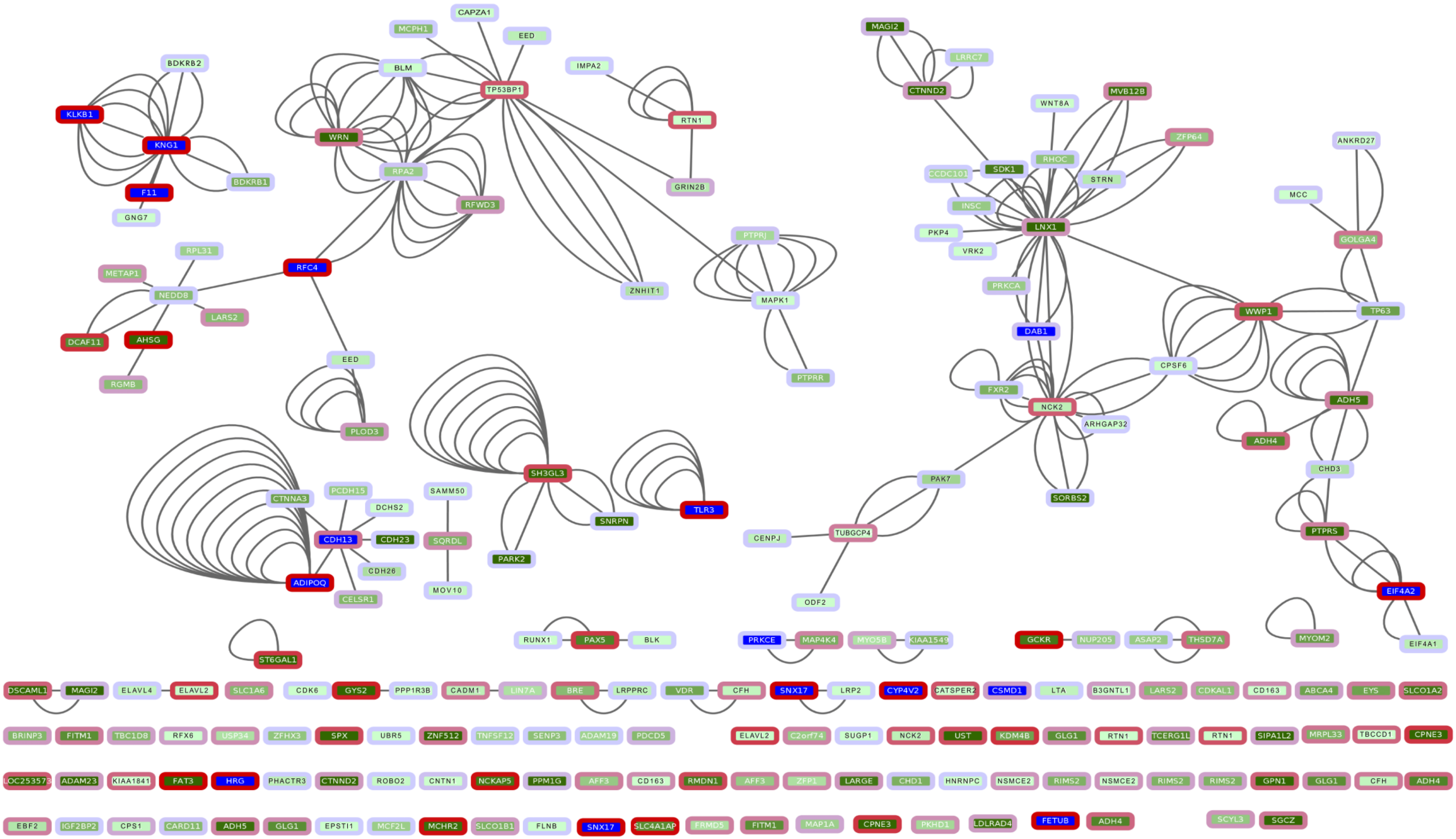


Figure 14. Interaction map of genes from VEGAS with p-value less then 10^{-3}. Lines indicate interactions. Several databases was used that is why there could be more than one line on each interaction. The border color represent the SNP with the lowest p-value in that gene were the redder the color the lower it is. The inside color represent the p-value for the gene given by VEGAS were the bluer the color the lower.

Table 5. Top 3 Pathways from PARIS

KB Name	Pathway ID	Pathway Name	Description	Simple Feature Count**	Gene Count*
KEGG	197339	has04610	Complement and coagulation cascades - Homo sapiens (human)	328	3
Gene Ontonology	51644	GO:0005576	The space external to the outermost structure of a cell. For cells without external protective or external encapsulating structures this refers to space outside of the plasma membrane. This term covers the host cell environment outside an intracellular parasite. [GOC:go_curators]	285	5
NetPath	119	NetPath_9	netpath/NetPath_9_GeneReg.tsv	274	2

** Genes that contain one or more feature with a p-value at or below 1e-08.*

*** P-values of 0.005 or less.*

3.7 eQTL

The SNPs with lowest p-value inside the genes from selection after VEGAS were used as input to the eQTL. The three top SNPs that are significant from GWAMA (rs710446, rs4253417 and rs780094) do not have any significant hits. Liver is the most relevant tissue because FXI is highly expressed there. Rs710446 seems to effect KNG1 expression levels to some extent in liver (Figure 15). Rs4253417 have nothing close to significant in liver (Figure 16). Rs780094 have to some extent association to the expression of GCKR in liver (Figure 17). This is also in line with the results from Annovar and RegulomeDB that both suggest that this SNP could be in an important region for expression.

Three other SNPs (*rs62323564 on LNX1 in liver p-value 1.27*10^{-4}, rs2508175 on ALG8 in aorta adventitia p-value 5.86*10^{-5} and rs2508175 on KCTD21 in liver p-value 5.91*10^{-5}*) was found to be significant with FDR (Figure 18-19). Two of the three hits are significant with Bonferroni correction (p-value of 6.85*10^{-05}).

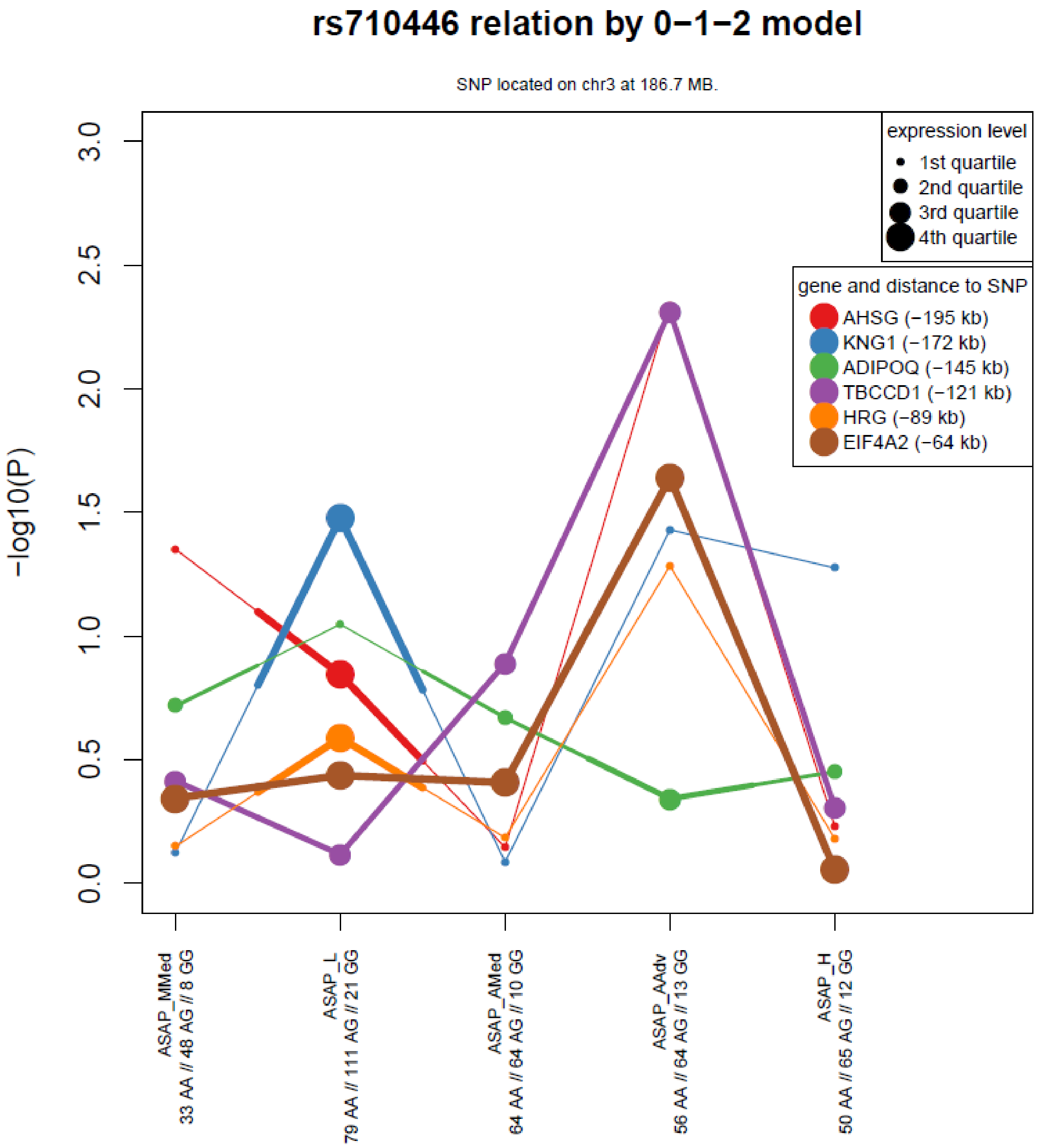


Figure 15. Plot of the different tissues and the corresponding p-value for cis genes of the SNP. The expression levels is indicated with the size of the dot for each tissue. ASAP is Advanced Study of Aortic Pathology Patients. MMed = mammary artery intima-media (89 samples), L= liver (212 samples), AMed = aorta media (138 samples), AAdv= aorta adventitia (133 samples) and H = heart (127 samples).

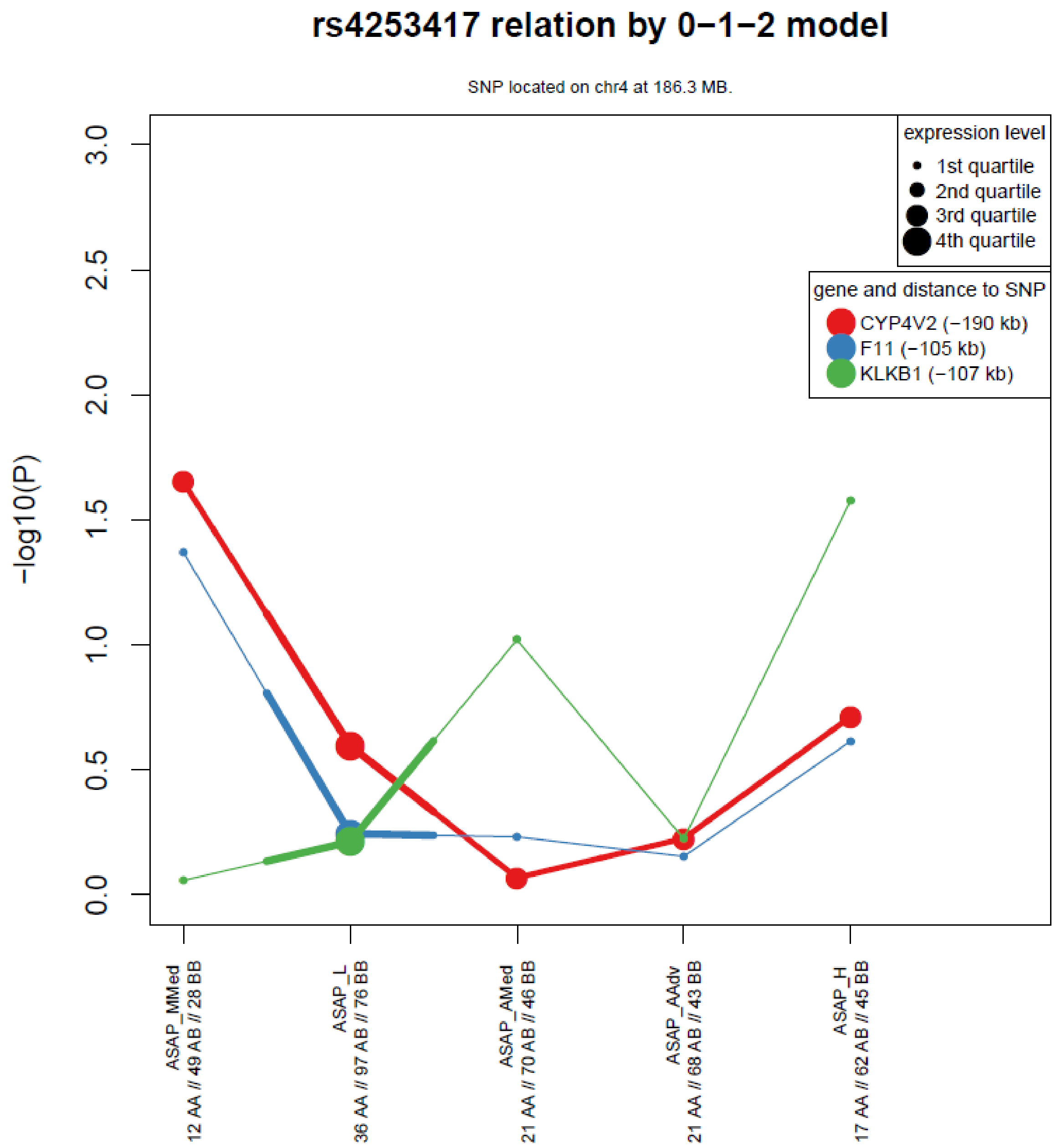


Figure 16. Plot of the different tissues and the corresponding p-value for cis genes of the SNP. The expression levels is indicated with the size of the dot for each tissue. ASAP is Advanced Study of Aortic Pathology Patients. MMed = mammary artery intima-media (89 samples), L= liver (212 samples), AMed = aorta media (138 samples), AAdv= aorta adventitia (133 samples) and H = heart (127 samples).

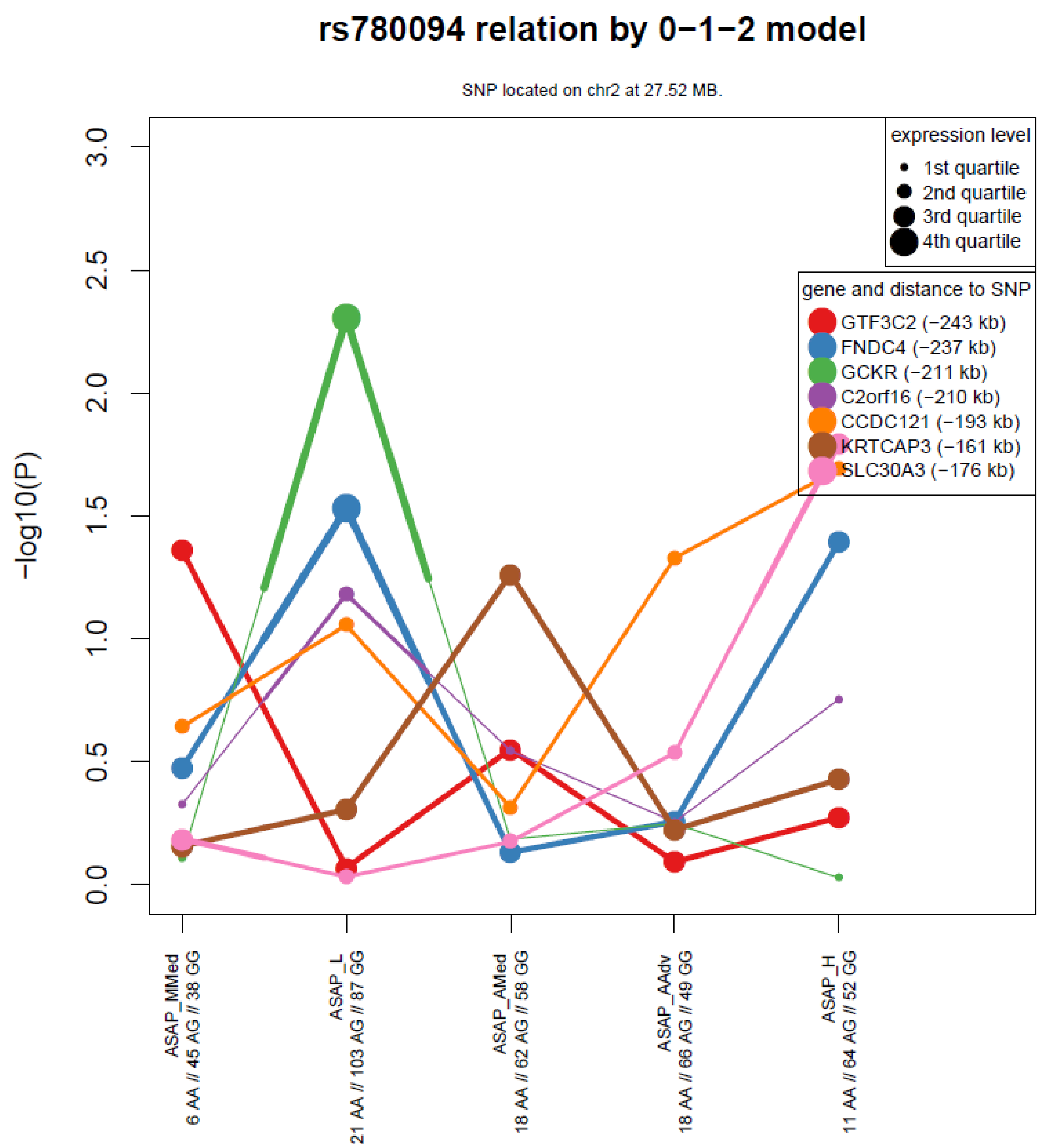


Figure 17. Plot of the different tissues and the corresponding p-value for cis genes of the SNP. The expression levels is indicated with the size of the dot for each tissue. ASAP is Advanced Study of Aortic Pathology Patients. MMed = mammary artery intima-media (89 samples), L= liver (212 samples), AMed = aorta media (138 samples), AAdv= aorta adventitia (133 samples) and H = heart (127 samples).

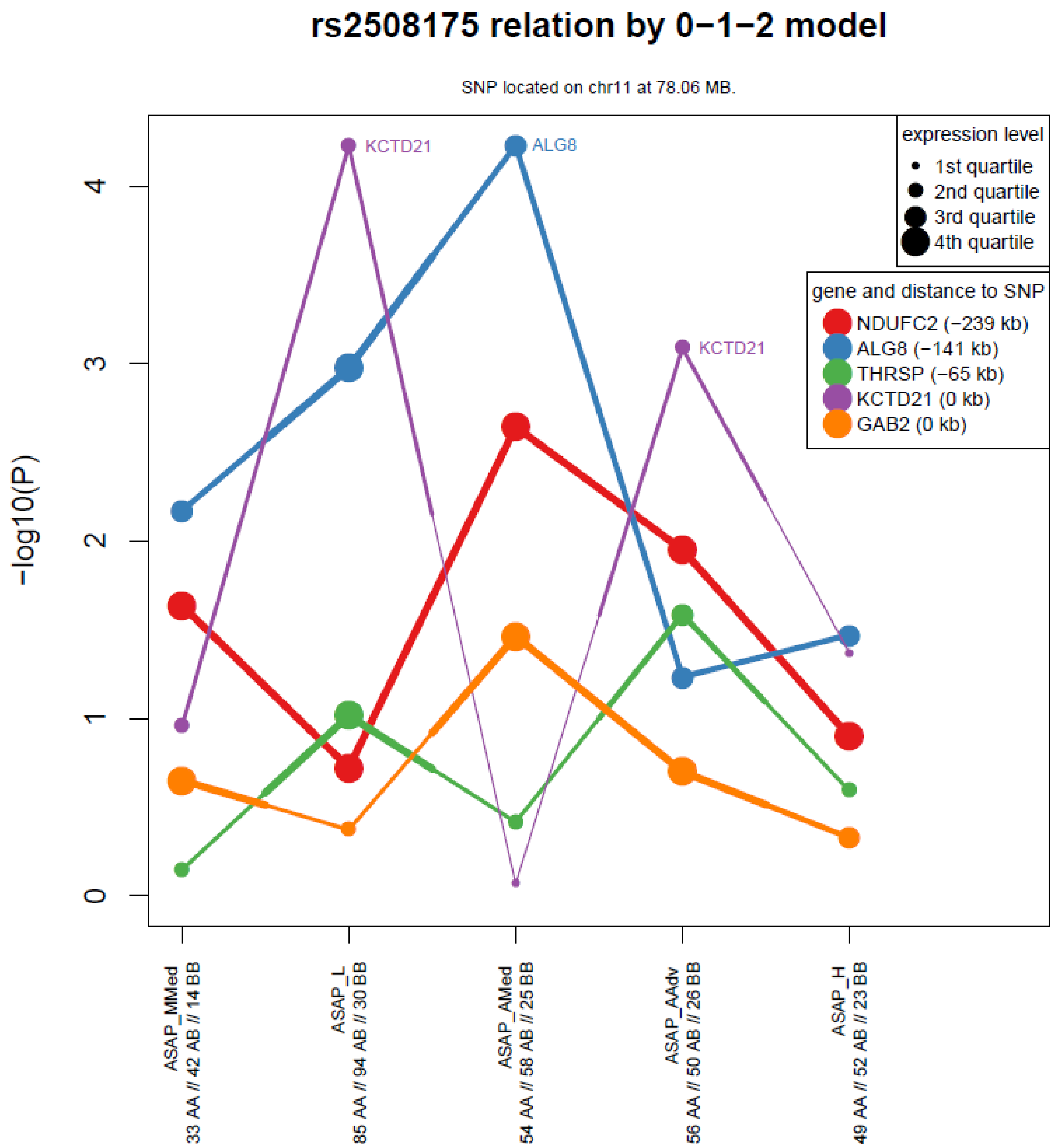


Figure 18. Plot of the different tissues and the corresponding p-value for cis genes of the SNP. The expression levels is indicated with the size of the dot for each tissue. ASAP is Advanced Study of Aortic Pathology Patients. MMed = mammary artery intima-media (89 samples), L= liver (212 samples), AMed = aorta media (138 samples), AAdv= aorta adventitia (133 samples) and H = heart (127 samples).

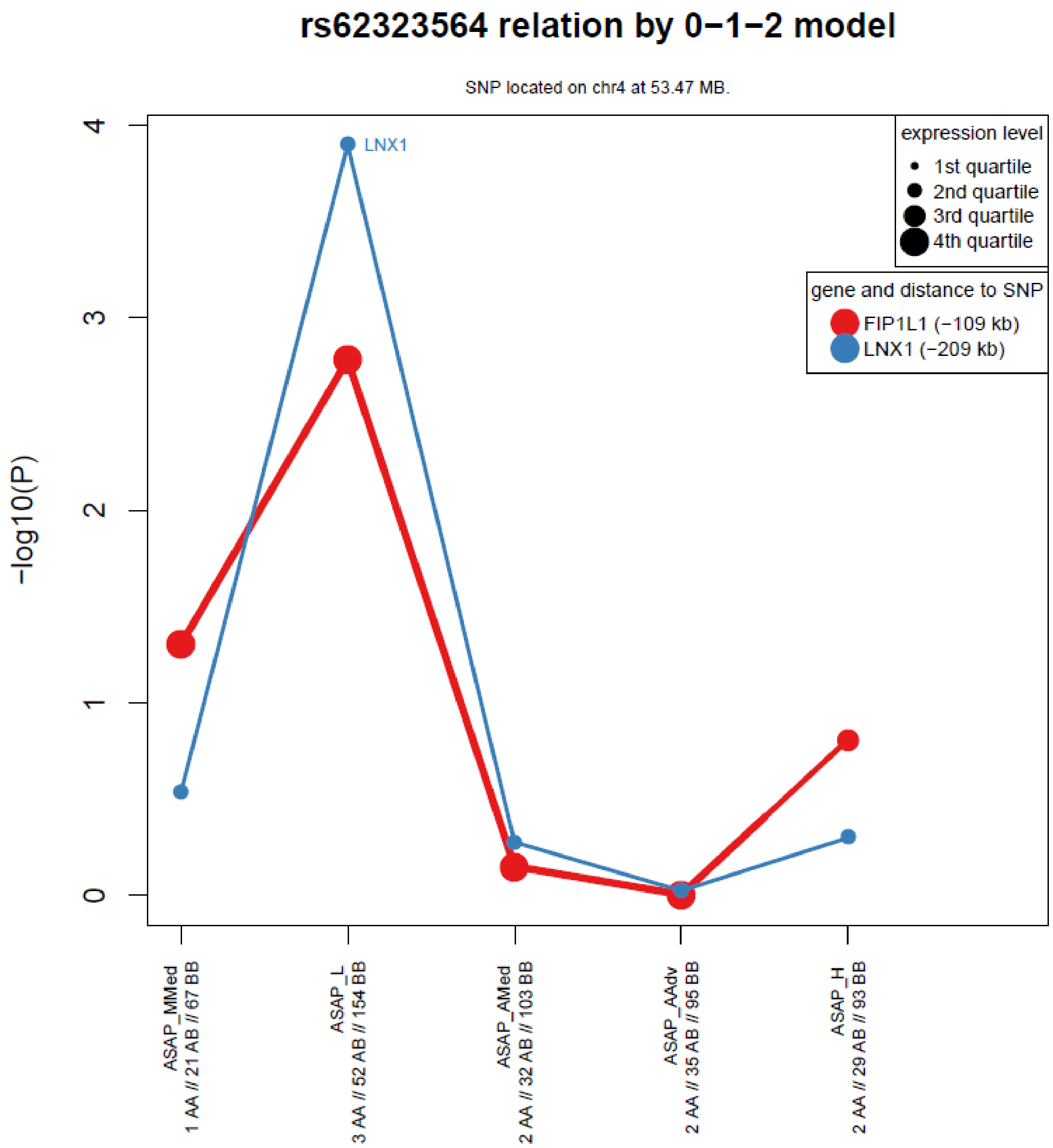


Figure 19. Plot of the different tissues and the corresponding p-value for cis genes of the SNP. The expression levels is indicated with the size of the dot for each tissue. ASAP is Advanced Study of Aortic Pathology Patients. MMed = mammary artery intima-media (89 samples), L= liver (212 samples), AMed = aorta media (138 samples), AAdv= aorta adventitia (133 samples) and H = heart (127 samples).

3.8 miRNA binding prediction

The top two genes (KNG1 and FXI) were analyzed regarding their targeting by miRNAs through using a miRNA prediction algorithm. Since rs780094 is substantially weaker associated with FXI, it was excluded from the analysis. The gene KNG1 did not have any miRNA that passed through the prediction algorithm. FXI have several miRNA with binding prediction (Table 6). Some of them were already validated (Salloum-Asfar et al. 2014) and were therefore left out, except for miR-181a-5p as it worked as a positive control.

Table 6. miRNA passing prediction algorithm. Predicted to bind to accordingly with MirWalk outcome in methods. #Algorithms: The number of algorithms predicted to bind from MirWalk 2 ordered as stated in Binding column. TargetScan context+ score (Friedman et al. 2009). miRanda total score (Enright et al. 2003). SVR: miRSVR score (Betel et al. 2010). MirWalk number of algorithms predicting binding to 5' region. Tarbase miTG score (Vlachos et al. 2015). Green indicates best value of the 5 SNPs for that prediction tool. Blue is already tested miRNA (Salloum-Asfar et al. 2014).

miRNA	Binding	#Algorithms	TargetScan	miRanda	SVR	MirWalk1	Tarbase
hsa-miR-15a-5p	3' and promoter	6,3	-0.114	145	-0.1365	5	0.46
hsa-miR-15b-5p	3' and promoter	5,3	-0.114	149	-0.1365	5	0.464
hsa-miR-145-5p	3' and promoter	7,2	-0.287	155	-1.1535	6	0.751
hsa-miR-150-5p	3' and promoter	6,4	-0.111	140	-0.2513	5	0.637
hsa-miR-424-5p	3' and promoter	7,2	-0.133	157	-0.1338	5	0.555
hsa-miR-181a-5p	3', 5' and promoter	6,3,2	-0.047	145	-0.3645	6	0.515
hsa-miR-16-5p	3' and promoter	6,3	-0.114	146	-0.129	5	0.464
hsa-miR-23a-3p	3' and promoter	5, 2	-0.073	140	-0.6707	5	0.664

3.9 Luciferase reporter assay

The miRNA miR-145, miR-181a from the miRNA prediction step and a scrambled miRNA were tested for expression effects on a vector with the FXI 3’ UTR region with a Luciferase reporter assay. The concentration of miRNA was 10 nM and the amount of vector was 100 ng when transfecting the cells. The samples from these experiments had low luminescence for the FXI vector. We believe that the transfection was at fault. Therefore an increase of miRNA and FXI vector was preformed to get a better transfection. The samples with the low luminescence were excluded. The luminescence from these experiments can be seen in Supplemental Table 8.

When increasing the miRNA concentration to 20 nM and the amount of vector to 400 ng a stronger luminescence signal was detectable. Four replications were conducted for each of the six combinations. From Figure 20 we can see that the miR-145 and miR-181 have a lower luminescence mean than the scrambled when the vector carrying the luminescence is the vector with the FXI 3’ UTR region. This is not the case when the vector carrying the luminescence is the vector without the FXI 3’ UTR region. It should be noticed that the standard error of the mean is a lot bigger for the empty vector compared to the FXI vector.

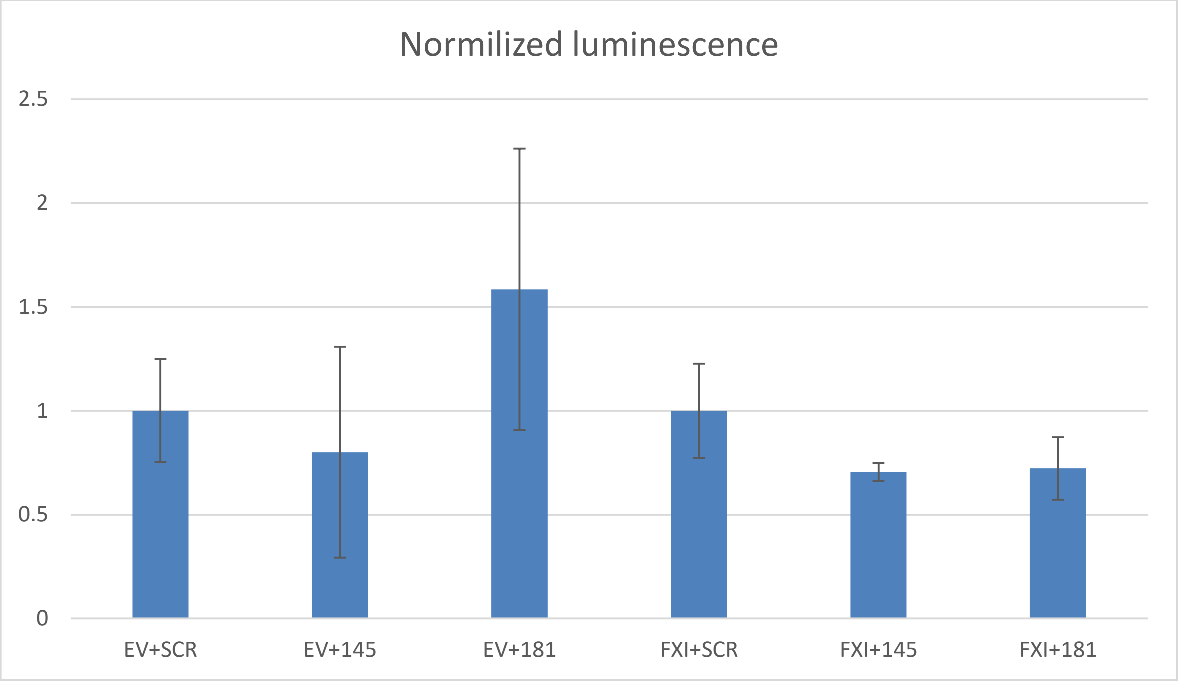


Figure 20. Column plot of the normalized mean value of the luminescence from the luciferase reporter assay. A total of four samples of each combinations were produced and used. Values are normalized with the mean value of scrambled (SCR) miRNA for respective vector. Error bars indicate the standard error of the mean value. Only values from samples were miRNA concentration are 20 nM and the amount of vector are 400 ng are included. For raw luminescence values se Supplement Table 7. EV is the empty vector, FXI is the vector with the FXI 3' UTR region insert, 145 is the miR-145 mimic, 181 is the miR-181 mimic and SCR is the scrambled miRNA.

Chapter 4 Discussion

Elevated levels of FXI levels in plasma have been associated with venous thrombosis (Meijers et al. 2000) and ischemic stroke (Yang et al. 2006), which are both severe diseases that could potentially be fatal. Recent studies have identified common SNPs, both within the FXI gene (Bezemer et al. 2008, Li et al. 2009), and within the KNG1 gene (Sabater-Lleal et al. 2012) that are associated with FXI levels. But the regulatory mechanisms are not fully understood and the studies need to be extended with a meta-analysis to increase power. We therefore present a meta-analysis of GWAS data conducted in European-ancestry individuals, aimed to find new regulatory mechanisms for FXI levels and functionally annotate the findings.

Our findings show three loci (located in chromosome two, three and four) with SNPs passing the genome wide significant threshold. The SNPs with lowest p-value in each of the loci are rs710446 (chromosome three), rs4253417 (chromosome four) and rs780094 (chromosome two). The combination of eQTL, pathway analyses, gene association and functional annotations indicate that the most plausible candidate genes for these regions are KNG1 (rs710446), FXI (rs4253417) and GCKR (rs780094).

The KNG1 gene encodes the high-molecular-weight kininogen (HK) that is part of the contact pathway with FXI, FXII and plasma prekallikrein (Colman and Schmaier 1997). FXI circulating in blood is almost always in a complex with HK, and HK deficiency has been associated with low FXI levels (Maas et al. 2011). This was also indicated by the interaction map in our pathway analysis.

The SNP with lowest p-value in the KNG1 locus (Rs710446) is located in an exon, causing an amino acid exchange. Therefore the amino acid could change the activity that this protein has and in turn change the FXI levels in blood. But from the eQTL we can also see some evidence (not significant) that the SNP changes the expression of KNG1. Therefore we have evidence for both theories. The one that is the underlying cause of the association to FXI levels cannot be concluded. It is plausible that both theories can be present at the same time and that is causing the association. From the gene association we can see that KNG1 is probably the gene of interest in this loci. This SNP is close to ADIPOQ that earlier has been associated with cardiovascular diseases (see methods). Association analysis between rs710446 and adiponectin levels showed that there was no significant association between the two. Rs710446 has been associated with activated partial thromboplastin time (aPTT) in other studies (Houlihan et al. 2010). This supports our results because aPPT reflects the function of the intrinsic coagulation pathway, which FXI is a part of.

The SNP with the lowest p-value in chromosome 4 is rs4253417 and it is located in an intron of the FXI gene, which encodes for the coagulation factor XI. Rs4253417 has no indication of changing the expression from the functional annotation or the eQTL. The reason why this SNP is associated with FXI level could be that it affects some post-transcriptional regulation factors. This will not be detected on the eQTL because it only measures the mRNA levels (transcriptional levels). This was not further investigated in our analyses. An expansion of the eQTL looking for *trans*-elements could also be done to answer if it actually effects *trans*-elements. From the gene association analysis, we found several genes (FXI, FXI-AS1 and/or KLKB1) that were associated in this region (chromosome four). The FXI-AS1 is the anti-sense to FXI and KLKB1 encodes Kallikrein B, Plasma (Fletcher Factor) 1, which interacts directly with KNG1 that then interacts with FXI according to the pathway analysis.

Finally, the SNP with the lowest p-value in the GCKR locus (Rs780094) is located in an intronic region of GCKR. The GCKR gene encodes for a glucokinase (hexokinase 4) regulator. The GCKR loci has been associated with FVII levels (Smith et al. 2010) and many more, but this is the first time that the GCKR loci has been significantly associated with FXI levels. From the functional annotation we got evidence that this SNP (Rs780094) changes the expression (strong enhancer) of nearby genes. The eQTL shows that this SNP (Rs780094) changes the expression of GCKR but the p-value was not significant when corrected with FDR. Even if the eQTL is not significant enough to be significant after FDR correction there is evidence from two independent sources. The gene association rank GCKR high in this loci, but also SNX17. When looking at the regional plot we see that there is a SNP (rs4665972) with low p-value in this gene. That could be why SNX17 is ranked so highly. Rs4665972 and Rs780094 are highly linked and because rs780094 has approximately a two-fold lower p-value, we believe that the other SNP only ‘hitchhiked’ to popularity and is of minor importance.

The three significant SNPs rs710446 (KNG1 gene), rs4253417 (FXI gene) and rs780094 (GCKR) have high heterogeneity. We investigated if there was one cohort that was responsible and if it was because of non-functional versus functional measurements of the phenotype. But none of this theory holds, and the underlying cause of this is more complicated and remains unknown. The three top SNPs were significant in the replication set and therefore further investigations of the heterogeneity was deemed unnecessary.

We see that our gene/SNP contains information that describes cardiovascular disease from the pathway analysis. We found that it describes the complement and coagulation cascades for Homo sapiens (PARIS) and that some of the keywords describing our set would be plasma, coagulation and blood. With this we can conclude that our study have relevant data in it.

The miRNA prediction gave five miRNA with prediction to bind FXI that have not been previously described in relation to FXI regulation, and three already investigated miRNAs (Salloum-Asfar et al. 2014). No miRNA was predicted to bind to KNG1 following the scoring algorithm created (see methods). The miRNA with best score from most algorithms (5 of 6, Table 6) used was has-miR-145-5p.

Has-miR-145-5p was then validated with a luciferase reporter assay. The mean luminescence did increase to desired amount (Figure 20). The results shows that when miR-145-5p and miR-181-5p (positive control) are present with the FXI vector a lower mean luminescence can be found compared to a scrambled (negative control). These results indicate that miR-145-5p interact with the 3’ UTR region of FXI and lower its expression. But one should take into consideration that only four replications were done.

In summary, we report a meta-analysis of five GWAS studies of FXI levels in blood. The Meta-analysis reveals three loci (KNG1 gene, FXI gene and GCKR gene) that regulate FXI levels. The KNG1 and FXI loci were described before in a previous GWAS study, but the GCKR locus is a novel finding. For the SNP with lowest p-value in the GCKR gene locus (rs780094) there is evidence that it changes the expression of GCKR. A miRNA prediction for miRNAs that can bind and regulate the FXI gene was done finding has-miR-145-5p as a candidate. Has-miR-145-5p was validation through a luciferase reporter assay for the FXI gene finding that it possibly binds and regulates the FXI gene.

With this we have contributed to the understanding of FXI regulation, which may help determine risk of VTE and inspire the development of a promising medicine.

4.1 Limitations

The number of samples used in the eQTL may not be enough to produce necessary power. In the eQTL we look at 35 SNPs and that could make the multiple testing problem too big for the number of samples used to produce p-values significant after correction to multiple testing.

The high heterogeneity in the discovery set for some SNPs is high and may hide significant SNPs or produce false positives in our set.

All look-ups from databases are dependent on what is currently available. There is a possibility that relevant data is not there or that the data is not completely accurate.

The luminescence reporter assay only contains four replications. A bigger sample size would increase the trustworthiness of the experiment.

Chapter 5 Acknowledgements

First of all, I would like to thank Maria Sabater for giving me the opportunity to be a part of this educational and interesting project. She has given me a great insight in this part of bioinformatics as well as selflessly taught me about it, which I will take with me for the rest of my career. I want to thank Lars Maegdefessel for the help arranging the miRNA part of this project as well as guiding me. He has also given me a valuable perspective of how science is performed. I want to thank Bengt Persson for both being my subject reader and giving important feedback throughout the project. I also want to thank Jesper Gadin for the help preforming the eQTL and giving me the opportunity to learn more about it. An important part of the project was done in the laboratory and there I would like to say thank you to Yuhuang Li and Ekaterina Chernogubova for helping me out and teaching me. A special thank you to Bengt Sennblad for all the help with IT related things and statistical software’s as well as helpful discussions. I want to thank Alexandra Bäcklund and Rona Strawbridge for helping me out when lost, both in the project and at CMM, and all the laughs we had. I want to say thanks to Greg Korzunowicz for preforming the last lab work on the luciferase reporter assay experiment. I also want to thank the rest of the research groups at CMM taking part of this interesting project. Last I want to express my gratitude to my family for supporting me, and my lovely girlfriend for showing understanding and patience when I was occupied by the project as well as being my training audience when I needed to practice for presentations.

Chapter 6 References

Barsh GS, Copenhaver GP, Gibson G, Williams SM. 2012. Guidelines for Genome-Wide Association Studies. PLoS Genet 8: e1002812. doi:10.1371/journal.pgen.1002812

Beckman MG, Hooper WC, Critchley SE, Ortel TL. 2010. Venous thromboembolism: a public health concern. Am. J. Prev. Med. 38: S495–501. doi:10.1016/j.amepre.2009.12.017

Betel D, Koppal A, Agius P, Sander C, Leslie C. 2010. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. 11: R90. doi:10.1186/gb-2010-11-8-r90

Bezemer ID, Bare LA, Doggen CJM, Arellano AR, Tong C, Rowland CM, Catanese J, Young BA, Reitsma PH, Devlin JJ, Rosendaal FR. 2008. Gene variants associated with deep vein thrombosis. JAMA 299: 1306–1314. doi:10.1001/jama.299.11.1306

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22: 1790–1797. doi:10.1101/gr.137323.112

Büller HR, Bethune C, Bhanot S, Gailani D, Monia BP, Raskob GE, Segers A, Verhamme P, Weitz JI, FXI-ASO TKA Investigators. 2015. Factor XI antisense oligonucleotide for prevention of venous thrombosis. N. Engl. J. Med. 372: 232–240. doi:10.1056/NEJMoa1405760

Colman RW, Schmaier AH. 1997. Contact system: a vascular biology modulator with anticoagulant, profibrinolytic, antiadhesive, and proinflammatory attributes. Blood 90: 3819–3843

Dweep H, Sticht C, Pandey P, Gretz N. 2011. miRWalk – Database: Prediction of possible miRNA binding sites by “walking” the genes of three genomes. J. Biomed. Inform. 44: 839–847. doi:10.1016/j.jbi.2011.05.002

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. 2003. MicroRNA targets in Drosophila. Genome Biol. 5: R1. doi:10.1186/gb-2003-5-1-r1

Folkersen L, Hooft F van’t, Chernogubova E, Agardh HE, Hansson GK, Hedin U, Liska J, Syvänen A-C, Paulsson-Berne G, Franco-Cereceda A, Hamsten A, Gabrielsen A, Eriksson P. 2010. Association of Genetic Risk Variants With Expression of Proximal

Genes Identifies Novel Susceptibility Genes for Cardiovascular Disease. Circ. Cardiovasc. Genet. 3: 365–373. doi:10.1161/CIRCGENETICS.110.948935

Friedman RC, Farh KK-H, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 19: 92–105. doi:10.1101/gr.082701.108

Fujikawa K, Chung DW, Hendrickson LE, Davie EW. 1986. Amino acid sequence of human factor XI, a blood coagulation factor with four tandem repeats that are highly homologous with plasma prekallikrein. Biochemistry (Mosc.) 25: 2417–2424

Gable DR, Hurel SJ, Humphries SE. 2006. Adiponectin and its gene variants as risk factors for insulin resistance, the metabolic syndrome and cardiovascular disease. Atherosclerosis 188: 231–244. doi:10.1016/j.atherosclerosis.2006.02.010

Gailani D, Broze GJ. 1991. Factor XI activation in a revised model of blood coagulation. Science 253: 909–912

Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. 2011. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat. Struct. Mol. Biol. 18: 1139–1146. doi:10.1038/nsmb.2115

Griffiths-Jones S. 2004. The microRNA Registry. Nucleic Acids Res. 32: D109–D111. doi:10.1093/nar/gkh023

Griffiths-Jones S, Grocock RJ, Dongen S van, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34: D140–D144. doi:10.1093/nar/gkj112

Griffiths-Jones S, Saini HK, Dongen S van, Enright AJ. 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res. 36: D154–D158. doi:10.1093/nar/gkm952

Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell 27: 91–105. doi:10.1016/j.molcel.2007.06.017

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. 106: 9362–9367. doi:10.1073/pnas.0903103106

Houlihan LM, Davies G, Tenesa A, Harris SE, Luciano M, Gow AJ, McGhee KA, Liewald DC, Porteous DJ, Starr JM, Lowe GD, Visscher PM, Deary IJ. 2010. Common Variants of Large Effect in F12, KNG1, and HRG Are Associated with Activated

Partial Thromboplastin Time. Am. J. Hum. Genet. 86: 626–631. doi:10.1016/j.ajhg.2010.02.016

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. 2004. Human MicroRNA targets. PLoS Biol. 2: e363. doi:10.1371/journal.pbio.0020363

Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 39: D152–D157. doi:10.1093/nar/gkq1027

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 42: D68–D73. doi:10.1093/nar/gkt1181

Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foà R, Schliwka J, Fuchs U, Novosel A, Müller R-U, Schermer B, Bissels U, Inman J, Phan Q, Chien M, Weir DB, Choksi R, De Vita G, Frezzetti D, Trompeter H-I, Hornung V, Teng G, Hartmann G, Palkovits M, Di Lauro R, Wernet P, Macino G, Rogler CE, Nagle JW, Ju J, Papavasiliou FN, Benzing T, Lichter P, Tam W, Brownstein MJ, Bosio A, Borkhardt A, Russo JJ, Sander C, Zavolan M, Tuschl T. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 129: 1401–1414. doi:10.1016/j.cell.2007.04.040

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15–20. doi:10.1016/j.cell.2004.12.035

Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG, Macgregor S. 2010. A Versatile Gene-Based Test for Genome-wide Association Studies. Am. J. Hum. Genet. 87: 139–145. doi:10.1016/j.ajhg.2010.06.009

Li Y, Bezemer ID, Rowland CM, Tong CH, Arellano AR, Catanese JJ, Devlin JJ, Reitsma PH, Bare LA, Rosendaal FR. 2009. Genetic variants associated with deep vein thrombosis: the F11 locus. J. Thromb. Haemost. JTH 7: 1802–1808. doi:10.1111/j.1538-7836.2009.03544.x

Maas C, Oschatz C, Renné T. 2011. The plasma contact system 2.0. Semin. Thromb. Hemost. 37: 375–381. doi:10.1055/s-0031-1276586

Maegdefessel L, Spin JM, Raaz U, Eken SM, Toh R, Azuma J, Adam M, Nagakami F, Heymann HM, Chernugobova E, Jin H, Roy J, Hultgren R, Caidahl K, Schrepfer S, Hamsten A, Eriksson P, McConnell MV, Dalman RL, Tsao PS. 2014. miR-24 limits aortic vascular inflammation and murine abdominal aneurysm development. Nat. Commun. 5: 5214. doi:10.1038/ncomms6214

Magi R, Lindgren CM, Morris AP. 2010. Meta-analysis of sex-specific genome-wide association studies. Genet. Epidemiol. 34: 846–853. doi:10.1002/gepi.20540

Mägi R, Morris AP. 2010. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11: 288. doi:10.1186/1471-2105-11-288

Meijers JC, Tekelenburg WL, Bouma BN, Bertina RM, Rosendaal FR. 2000. High levels of coagulation factor XI as a risk factor for venous thrombosis. N. Engl. J. Med. 342: 696–701. doi:10.1056/NEJM200003093421004

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. 2010. LocusZoom: regional visualization of genome-wide association scan results. Bioinforma. Oxf. Engl. 26: 2336–2337. doi:10.1093/bioinformatics/btq419

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81: 559–575

Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, International Schizophrenia Consortium, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ. 2009. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 5: e1000534. doi:10.1371/journal.pgen.1000534

Sabater-Lleal M, Martinez-Perez A, Buil A, Folkersen L, Souto JC, Bruzelius M, Borrell M, Odeberg J, Silveira A, Eriksson P, Almasy L, Hamsten A, Soria JM. 2012. A Genome-Wide Association Study Identifies KNG1 as a Genetic Determinant of Plasma Factor XI Level and Activated Partial Thromboplastin Time. Arterioscler. Thromb. Vasc. Biol. 32: 2008–2016. doi:10.1161/ATVBAHA.112.248492

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang P-L, Lotia S, Pico AR, Bader GD, Ideker T. 2012. A travel guide to Cytoscape plugins. Nat. Methods 9: 1069–1076. doi:10.1038/nmeth.2212

Salloum-Asfar S, Teruel-Montoya R, Arroyo AB, García-Barberá N, Chaudhry A, Schuetz E, Luengo-Gil G, Vicente V, González-Conejero R, Martínez C. 2014. Regulation of

Coagulation Factor XI Expression by MicroRNAs in the Human Liver. PLoS ONE 9. doi:10.1371/journal.pone.0111713

Smith NL, Chen M-H, Dehghan A, Strachan DP, Basu S, Soranzo N, Hayward C, Rudan I, Sabater-Lleal M, Bis JC, de Maat MPM, Rumley A, Kong X, Yang Q, Williams FMK, Vitart V, Campbell H, Mälarstig A, Wiggins KL, Van Duijn CM, McArdle WL, Pankow JS, Johnson AD, Silveira A, McKnight B, Uitterlinden AG, Wellcome Trust Case Control Consortium;, Aleksic N, Meigs JB, Peters A, Koenig W, Cushman M, Kathiresan S, Rotter JI, Bovill EG, Hofman A, Boerwinkle E, Tofler GH, Peden JF, Psaty BM, Leebeek F, Folsom AR, Larson MG, Spector TD, Wright AF, Wilson JF, Hamsten A, Lumley T, Witteman JCM, Tang W, O'Donnell CJ. 2010. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. Circulation 121: 1382–1392. doi:10.1161/CIRCULATIONAHA.109.869156

Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, Anastasopoulos I-L, Maniou S, Karathanou K, Kalfakakou D, Fevgas A, Dalamagas T, Hatzigeorgiou AG. 2015. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res. 43: D153–159. doi:10.1093/nar/gku1215

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38: e164–e164. doi:10.1093/nar/gkq603

Wang L, Matsushita T, Madireddy L, Mousavi P, Baranzini SE. 2015. PINBPA: cytoscape app for network analysis of GWAS data. Bioinforma. Oxf. Engl. 31: 262–264. doi:10.1093/bioinformatics/btu644

Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, Ferreira T, Fall T, Graff M, Justice AE, Luan J, Gustafsson S, Randall JC, Vedantam S, Workalemahu T, Kilpeläinen TO, Scherag A, Esko T, Kutalik Z, Heid IM, Loos RJF, The Genetic Investigation of Anthropometric Traits (GIANT) Consortium. 2014. Quality control and conduct of genome-wide association meta-analyses. Nat. Protoc. 9: 1192–1212. doi:10.1038/nprot.2014.071

Yang DT, Flanders MM, Kim H, Rodgers GM. 2006. Elevated factor XI activity levels are associated with an increased odds ratio for cerebrovascular events. Am. J. Clin. Pathol. 126: 411–415. doi:10.1309/QC259F09UNMKVP0R

Yaspan BL, Bush WS, Torstenson ES, Ma D, Pericak-Vance MA, Ritchie MD, Sutcliffe JS, Haines JL. 2011. Genetic analysis of biological pathway data through genomic randomization. Hum. Genet. 129: 563–571. doi:10.1007/s00439-011-0956-2

Appendix

Supplemental Script 1. EasyQC script

```
##############################################################################################
###################

##### EasyQC-script to perform study-level and Meta-level QC on imputed 1000G data

##### EasyQC version: 9.0

##### Programmer: Thomas Winkler, 2014-09-22

##### Contact: thomas.winkler@klinik.uni-regensburg.de

##### Edited by Niklas Handin 2015-01-26

##############################################################################################
###################

### Please DEFINE here format and input columns of the following EASYIN files

DEFINE  --pathOut /media/disken/NikHan/easyQC/Results/

               --strMissing NA

               --strSeparator ,

               --acolIn 
Markername;Chr;pos;NonEffect_allele;Effect_allele;Freq;Beta;Se;Pval;Ntotal;Imp_info;Imputation
_type

               --acolInClasses 
character;character;integer;character;character;numeric;numeric;numeric;numeric;integer;numeri
c;numeric

               --acolNewName 
SNP;CHR;POS;OTHER_ALLELE;EFFECT_ALLELE;EAF;BETA;SE;PVAL;N;IMPUTATION;IMPUTATION_TYPE



## Please DO NOT CHANGE --acolNewName values because these reflect the column names used 
throughout the script

## If the study used different column names, please amend the respective value at --acolIn, 
the column will then

## be automatically renamed to the respective --acolNewName value

### Please DEFINE here all input files:

EASYIN  --fileIn /media/disken/NikHan/Orginal/FXI-1

### FXI-2 FXI-3 have differences for some params

EASYIN  --fileIn /media/disken/NikHan/Orginal/FXI-2

        --acolIn Markername;Chromosome;Position;Non.Effect Allele;Effect 
Allele;Freq;Beta;SE;p.value;N_total;Imp_info;Imputation
```

```
#Markername,Chromosome,Position,Effect Allele,Non-Effect
Allele,Freq,Beta,SE,N_total,Imputation,Imp_info,p-value



EASYIN  --fileIn /media/disken/NikHan/Orginal/FXI-3

        --acolIn
Markername;Chromosome;Position;NonEffectAllele;EffectAllele;Freq;Beta;SE;pval;N_total;Imp_info
;Imputation

#Markername,Chromosome,Position,EffectAllele,NonEffectAllele,Freq,Beta,SE,N_total,Imputation,I
mp_info,pval



### FXI-4 adhers to the default format

EASYIN  --fileIn /media/disken/NikHan/Orginal/FXI-4

               --acolIn
Markername;Chromosome;Position;NonEffectAlleleA2;EffectAlleleA1;FreqA1;Beta;SE;Pvalue;Ntotal;I
mp_info;Imputation

#Strand,Markername,Chromosome,Position,EffectAlleleA1,NonEffectAlleleA2,FreqA1,Beta,SE,Pvalue,
Ntotal,Imputation,Imp_info



### FXI-5 has differences for some params

EASYIN  --fileIn /media/disken/NikHan/Orginal/FXI-5

               --acolIn
Markername;Chromosome;Position;NonEffectAlleleA2;EffectAlleleA1;FreqA1;Beta;SE;Pvalue;Ntotal;I
mp_info;Imputation

#Strand,Markername,Chromosome,Position,EffectAlleleA1,NonEffectAlleleA2,FreqA1,Beta,SE,Pvalue,
Ntotal,Imputation,Imp_info



### FXI-6 has differences for some params

EASYIN  --fileIn /media/disken/NikHan/Orginal/FXI-6

        --strSeparator TAB

               --acolIn
Markername;Chromosome;Position;NonEffectAllele;EffectAllele;FREQ;Beta;SE;P.value;N_total;Imp_i
nfo_r2;Imputation

#Markername    Chromosome       Position EffectAllele    NonEffectAllele FREQ    Beta    SE
P-value  N_total Imputation      Imp_info_r2

##############################################################################################
###################

## EASYQC Scripting interface:

START EASYQC

####################

## 1. Sanity checks:
```

```
CLEAN --rcdClean is.na(EFFECT_ALLELE)&is.na(OTHER_ALLELE) --strCleanName
numDrop_Missing_Alleles

CLEAN --rcdClean is.na(PVAL) --strCleanName numDrop_Missing_P

CLEAN --rcdClean is.na(BETA) --strCleanName numDrop_Missing_BETA

CLEAN --rcdClean is.na(SE) --strCleanName numDrop_Missing_SE

CLEAN --rcdClean is.na(EAF) --strCleanName numDrop_Missing_EAF

CLEAN --rcdClean is.na(N) --strCleanName numDrop_Missing_N

CLEAN --rcdClean is.na(IMPUTATION) --strCleanName numDrop_Missing_Imputation

CLEAN --rcdClean PVAL<0|PVAL>1 --strCleanName numDrop_invalid_PVAL

CLEAN --rcdClean SE<=0|SE==Inf|SE>=10 --strCleanName numDrop_invalid_SE

CLEAN --rcdClean abs(BETA)>=10 --strCleanName numDrop_invalid_BETA

CLEAN --rcdClean EAF<0|EAF>1 --strCleanName numDrop_invalid_EAF

CLEAN --rcdClean IMPUTATION<0 --strCleanName numDrop_invalid_IMPUTATION

## This is important for data reduction, because some studies report an unnecessary large
number of significant digits

EDITCOL --rcdEditCol signif(EAF,4) --colEdit EAF

EDITCOL --rcdEditCol signif(BETA,4) --colEdit BETA

EDITCOL --rcdEditCol signif(SE,4) --colEdit SE

EDITCOL --rcdEditCol signif(PVAL,4) --colEdit PVAL



####################

## 2. Prepare files for filtering and apply minimum thresholds:

## Exclude monomorphic SNPs:

CLEAN --rcdClean (EAF==0)|(EAF==1) --strCleanName numDrop_Monomorph

## Create column with minor allele count:

ADDCOL --rcdAddCol signif(2*pmin(EAF,1-EAF)*N,4) --colOut MAC

## If you do not want to apply filters at this stage, please comment out the following rows or
amend the

## filter thresholds according to your needs.

## Change

CLEAN --rcdClean N<30 --strCleanName numDrop_Nlt30

CLEAN --rcdClean MAC<=6 --strCleanName numDrop_MAClet6

CLEAN --rcdClean (!is.na(IMPUTATION))&IMPUTATION<0.3 --strCleanName numDrop_lowImpQual

####################

#### 3. Harmonization of allele coding (I/D)
```

```
## The aim of this step is to compile uniform allele codes A/C/G/T or I/D from different
versions f given alleles

HARMONIZEALLELES       --colInA1 EFFECT_ALLELE

                                    --colInA2 OTHER_ALLELE

####################

## 4. Harmonization of marker names (compile 'cptid')



CREATECPTID --fileMap
/media/disken/NikHan/easyQC/rsmid_map.1000G_ALL_p1v3.merged_mach_impute.v1.txt.gz

                      --colMapMarker rsmid

                      --colMapChr chr

                      --colMapPos pos

                      --colInMarker SNP

                      --colInA1 EFFECT_ALLELE

                      --colInA2 OTHER_ALLELE

                      --colInChr CHR

                      --colInPos POS

## TO DO:      Define the path to the reference file
'rsmid_map.1000G_ALL_p1v3.merged_mach_impute.v1.txt.gz' at --fileMap.

##                    The mapping file can be found on our website www.genepi-
regensburg.de/easyqc.

##                    In case CHR or POS are not given in the input files, please remove "--
colInChr CHR" and "--colInPos POS" from the

##                    command and remove "CHR;POS;" from --acolIn and --acolNewName as well
as the respective "character;integer;"

##                    from --acolInClasses

####################

## 5.Filter duplicate SNPs

## This will count duplicates and throw out the SNP with the lower sample size:

CLEANDUPLICATES --colInMarker cptid

                              --strMode samplesize

                              --colN N

## The duplicates are written to the output in a separate file "*duplicates.txt"

####################

## 6. AF Checks

### TO DO:     Define the path to the reference file
'allelefreq.1000G_[ANCESTRY]_p1v3.impute_legends.noMono.noDup.noX.v2.gz' at --fileRef:
```

```
###             Please use the reference file ancestry that matches the ancestry of the study

MERGE   --colInMarker cptid

               --fileRef
/media/disken/NikHan/easyQC/allelefreq.1000G_EUR_p1v3.impute_legends.noMono.noDup.noX.v2.gz

                       --acolIn cptid;a0;a1;eaf

                       --acolInClasses character;character;character;numeric

               --strRefSuffix .ref

               --colRefMarker cptid

               --blnWriteNotInRef 1


ADJUSTALLELES  --colInA1 EFFECT_ALLELE

                              --colInA2 OTHER_ALLELE

                              --colInFreq EAF

                              --colInBeta BETA

                              --colRefA1 a0.ref

                              --colRefA2 a1.ref

                              --blnMetalUseStrand 1

                              --blnRemoveMismatch 1

                              --blnRemoveInvalid 1


## All mismatches will be removed (e.g. A/T in input, A/C in reference)

AFCHECK --colInFreq EAF

               --colRefFreq eaf.ref

               --numLimOutlier 0.2

               --blnPlotAll 0


## blnPlotAll 0 causes that only outlying SNPs with |Freq-Freq.ref|>0.2 will be plotted (way
less computational time)

####################

## 7. Rearrange columns and Write CLEANED output

GETCOLS --acolOut
cptid;SNP;EFFECT_ALLELE;OTHER_ALLELE;EAF;IMPUTATION;IMPUTATION_TYPE;BETA;SE;PVAL;N;MAC

WRITE   --strPrefix CLEANED.

               --strMissing .

               --strMode gz

####################
```

```
## 8.  Plot Z versus P
PZPLOT  --colBeta BETA
               --colSe SE
               --colPval PVAL
####################
## 9.  QQ plot
QQPLOT  --acolQQPlot PVAL
               --numPvalOffset 0.05
               --strMode subplot
####################
## 10. Summary Stats post-QC
CALCULATE --rcdCalc max(N,na.rm=T) --strCalcName N_max
GC      --colPval PVAL --blnSuppressCorrection 1
RPLOT   --rcdRPlotX N_max
               --rcdRPlotY Lambda.PVAL.GC
               --arcdAdd2Plot abline(h=1,col='orange');abline(h=1.1,col='red')
               --strAxes lim(0,NULL,0,NULL)
               --strPlotName GC-PLOT
####################
## 11. SE-N Plot - Trait transformation
CALCULATE --rcdCalc median(SE,na.rm=T) --strCalcName SE_median
CALCULATE --rcdCalc median(1/sqrt(2*EAF*(1-EAF)), na.rm=T) --strCalcName c_trait_transf
RPLOT   --rcdRPlotX sqrt(N_max)
               --rcdRPlotY c_trait_transf/SE_median
               --arcdAdd2Plot abline(0,1,col='orange')
               --strAxes zeroequal
               --strPlotName SEN-PLOT
STOP EASYQC
##############################################################################################
###################
```

Supplemental Script 2. Script to find a substitution SNP with the highest LD

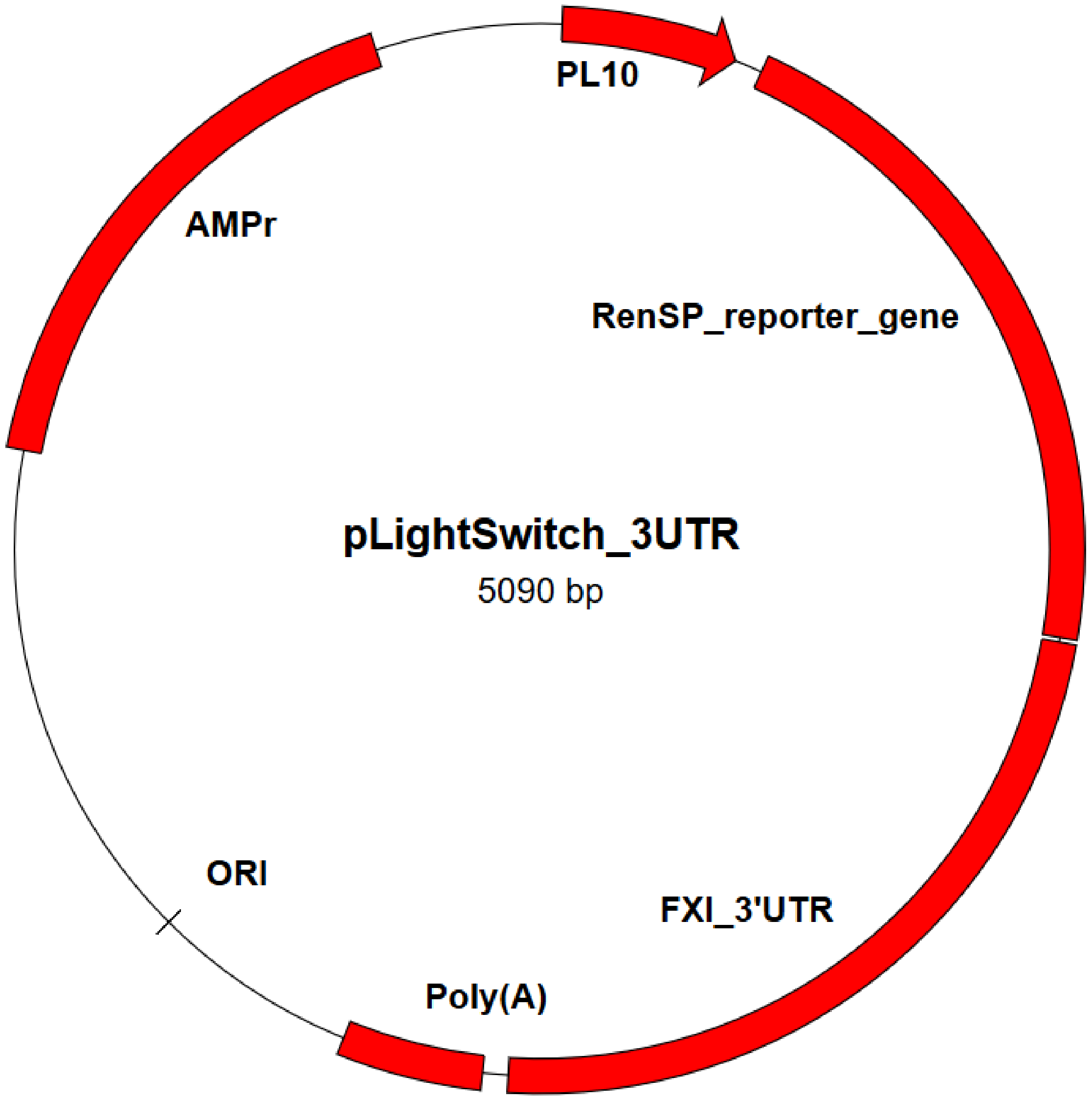
```
#!/bin/sh

# A simple script to map a rs number to a reference database.
# if it doesn't exist in the database it will look for the best approximation to that rs
(Accordingly to LD)
# example call:
#  ./Prox.sh -r /media/disken/MariaS/refpanel_1000G -b 100 -i small-snp-list.txt -m map2.txt -
o myout.txt -c 0.4
# -r the reference pathway separate files for every chromosome (should look like *.chr[1-22].*
were '*' is any number of characters)
# -b How many kb to use
# -i input file with the SNPs and the chromosomes
# -m the database with desired rs numbers
# -o output file
# -c cut-off for R2. SNP with lower value then this will not be examined. (The SNP with
highest R2 will be chosen)

### Functions
Print_temp_chrSNPfiles ()
{
tr -d '\r' < $1 | awk '
      BEGIN {
       FS="\t";
       OFS="\t";
   }
   NR==1 {
     for (f=1; f<=NF; f++) {
               if (tolower($f) == "chr") {
                      chr=f;
               }
          if (tolower($f) == "snp") {
                      SNP=f;
          }
          }
          if (length(chr) == 0) {
          print "Could not find column chr";
          exit;
          }
          if (length(SNP) == 0) {
          print "could not find column SNP";
          exit;
          }
     next
   }
   {
   if ($chr ~ /^[0-9]{1,2}/ && $SNP ~/^rs.*/) {
   print $SNP >> "temp/chr"$chr".temp"
   }
   else {
   print "Syntax error in input file", "chr=" $chr, "SNP=" $SNP;
   }
   }
'
}
while [ "$1" != "" ]; do
    case $1 in
        -r | --reference )                   shift
                                reference=$1
                                ;;
        -i | --inputfile )                   shift
                                                            inputfile=$1
                                ;;
        -m | --mapfile )                     shift
                                                            mapfile=$1
                                ;;
               -o | --outfile )                     shift
                                                            outFile=$1
                                ;;
               -c | --cutoff )                              shift
                                                            cutoff=$1
                                ;;
        -b | --kb )                                  shift
                                                                    kb=$1
                                ;;
    esac
    shift
done
DIRECTORY=.;
rm -rf temp
mkdir -p temp
file "$mapfile" | grep 'Zip archive data' &> /dev/null
if [ $? == 0 ]; then
        {
        echo "Map file is ziped";
        unzip "$mapfile" -d temp/ > /dev/null;

        if [ $(ls -1 temp | wc -l) -gt 1 ]; then
                {
                echo "Mapfile is multiple files. script will exit";
                exit;
                }
        elif [ $(ls -1 temp | wc -l) == 1 ]; then
                {
                mapfile="temp/$(ls temp)"
                }
        else
                {
                echo "internal error";
                exit;
                }
        fi
```

```
}
fi

#Do alot of cheeks!!!
#if reference is file take the path to the file

Print_temp_chrSNPfiles "$inputfile";
echo "Created temp SNP files";
count=0;
for i in $DIRECTORY/temp/chr*.temp
do


        filename=$(basename "$i"):
        filename_no_ext="${filename%.*}";
        PathtoSpecificCHR="${reference}/*${filename_no_ext}.*";
        finalRefPath=$(ls $PathtoSpecificCHR);
        echo "$finalRefPath";
        while read SNP; do
        count=$((count + 1));
        (plink --vcf "$finalRefPath" --r2 --ld-snp "$SNP" --ld-window-kb "$kb" --ld-window
99999 --out temp/temp"$filename_no_ext""$SNP" | grep 'No valid variants specified by' &>
/dev/null;
        if [ $? == 0 ]; then
        echo "$SNP failed in plink" >> "$outFile.error"
        fi
        )&#>/dev/null
        if [ $count == 5 ]; then
        {
                        wait;
                        count=0;
        }
        fi
        done <$i
done
        wait;
        echo "Done with all ld files";
        echo "old_rs-no        new_rs-no      R2" >> "$outFile.out";
        echo "old_rs-no">> "$outFile.notMaped";
for i in $DIRECTORY/temp/*.ld
        do
        SortedFile=$(sort -k7 -r -n "$i");
        found=0;
        SortedFile=$(echo "$SortedFile" | awk '{if ($6 != "") print $0}');
        SortedFile=$(echo "$SortedFile" | awk '{if ($6 != ".") print $0}');
        SortedFile=$(echo "$SortedFile" | awk -v cutoff="${cutoff}" '{if ($7 > cutoff) print
$0}');
        while read line
                do
                echo "$line";
                filename=$(basename "$i");
                filename_no_ext="${filename%.*}";
                oldRS=$(echo "$filename_no_ext"| cut -d's' -f 2);
                oldRS="rs${oldRS}"
                newRS=$(echo "$line" | awk '{print $6}');
                newR2=$(echo "$line" | awk '{print $7}');
                if zgrep -q "${newRS}\s" "$mapfile"; then
                {
                        echo "$oldRS   $newRS  $newR2" >> "$outFile.out";
                        found=1;
                        break;
                }
                fi
        done <<< "$SortedFile";
        if [ $found == "0" ]; then
                echo "$oldRS did not have any approximation rs...";
                echo "$oldRS">> "$outFile.notMaped";
        fi
done
rm -r temp
```

Supplemental Script 3. Configuration script for PARIS

```
# Variations data
VARIATION_FILENAME     variations.bn

# BioFilter data
SETTINGS_DB     bio-settings.cn

# List the various knowledge base (by KB ID) separated by spaces
INCLUDE_KNOWLEDGE      1 2 3 4

# filename containing snp,pvalue
DATA_SOURCE     ../Input/DataSource2

# Set the population ID to match the population your data is drawn
# from so that LD patterns can be used to expand the gene boundaries.
POPULATION      EUR

# Prefix used for all reports
#REPORT_PREFIX

# Single word to describe data followed by optional long description
# which can contain spaces (no returns, though). These will be used in some of the reports.
REPORT_NAME     Test1

# Loads all aliases and generates a text report containing their associations
LOAD_ALL_ALIASES       YES

# Write reports in html format (not all reports support HTML formatting
HTML_REPORTS    NO
```

```
# Target number of features inside each bin. Paris will define bins to
# be as close as possible to this number, but seldom will the count be exact.
BIN_SIZE       10000

# Number of permutations to be performed on each pathway.
P_COUNT 10000

# Threshold for determining the significance of a pathway (based on permutations)
PATHWAY_SIG_THRESH    0.005

# Threshold for determining if a SNP is significant.
RESULTS_SIG_THRESH    0.00000001

# How many base pair locations up and down stream do we expand gene boundaries
GENE_BOUNDARY_EXTENSION       50000

# Set the random seed used in permutations
RANDOM_SEED    1371

# ON/OFF to ignore pvalues of zero. If they aren't ignored, they will be
# counted as insignificant
IGNORE_PVALUES_OF_ZERO ON

# ON/OFF to allow features common to multiple genes in the same pathway to
# be counted multipe times
ALLOW_REDUNDANT_FEATURES      OFF

# Columnar location used for chromosome (1-22XY
COL_CHROMOSOME 1

# Columnar location of the RS (can have rs prefix (caps or not) or just be a
# numerical value
COL_RSID       2

# Columnar location of the pvalue to be used
COL_PVALUE     3

# The lower bound for borderline pvalues (set this to equal
# REFINEMENT_THRESHOLD_MAX to not perform refinement)
REFINEMENT_THRESHOLD_MIN      0.0001

# The upper bound for borderline pvalues (set this to equal
# REFINEMENT_THRESHOLD_MIN to not perform refinement)
REFINEMENT_THRESHOLD_MAX      0.001

# The number of repeteated ptests performed when a pvalue is determined to be
# borderline
REFINEMENT_REP_COUNT   1000

# When writing pathway investigation reports, do we show all pathways or only
# the signficant ones?
SHOW_ALL_ASSOCIATED_PATHWAYS  OFF

# User defined group file
#USER_PATHWAY_FILE
```



Supplemental Figure 1.Luciferase reporter assay vector FXI. PL10 is a constitutive promoter. The RenSP_reporter_gene is the gene producing the luciferase protein (synthetic renilla luciferase). FXI_3'UTR is the 3' UTR region of FXI. The FXI_3'UTR is absent in the empty vector.

Supplemental Table 1. numSNPsIn Number of SNPs in the input. numSNPsOut: the number of SNPs in the cleaned output file after all Quality checks. Missing X is if the SNP had a missing value in field X

Name	numSNPsIn	numSNPsOut	Missing_Alleles	Missing_P	Missing_BETA	Missing_SE	Missing_EAF	Missing_N	Missing_Imputation
FXI-1	22896100	18181775	0	0	0	0	0	0	0
FXI-2	7856324	7312395	0	0	0	0	0	0	0
FXI-3	7858383	7304354	0	0	0	0	0	0	0
FXI-4	9119903	8653402	0	0	0	0	0	0	0
FXI-5	9118482	8652206	0	0	0	0	0	0	0
FXI-6	17864829	11834987	0	0	0	0	0	0	0

Supplemental Table 2. Invalid_X is when the value interred in X is not possible (see methods). Monomorphic the number of Monomorphic SNPs. Nlt30: the number off SNPs that is in only 30 individuals. MAClet6: the number of SNPs that have MAC lesser or equal to six. lowImpQual: when the imputation quality is lower than 0.3.

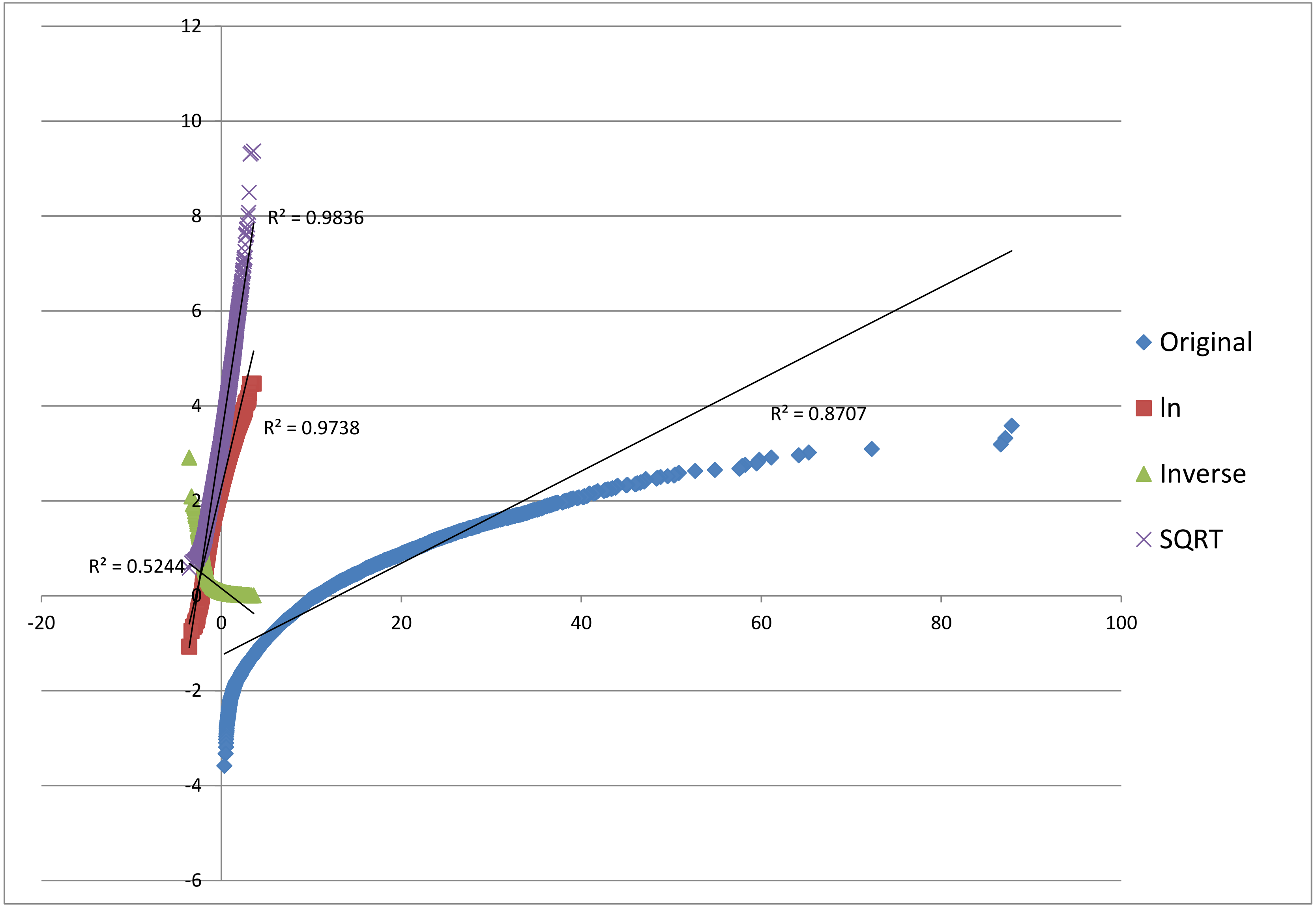
Name	invalid_PVAL	invalid_SE	invalid_BETA	invalid_EAF	Invalid_IMPUTATION	Monomorphic	Nlt30	MAClet6	lowImpQual
FXI-1	0	0	0	0	0	0	0	4031446	682850
FXI-2	0	0	0	0	0	0	0	0	543716
FXI-3	0	0	0	0	0	0	0	0	553816
FXI-4	0	0	0	0	0	0	0	0	466501
FXI-5	0	0	0	0	0	0	0	0	466276
FXI-6	0	208616	37778	0	0	0	0	5783062	0

Supplemental Table 3. BothAllelesMissing: When both NonEffect allele and Effect_allele is missing. #_Recoded_DEL

Name	BothAlleles Missing	#_Recoded_DEL	Recoded_MACH_R	Recoded_SEQ	Invalid_Alleles	cor_eaf.ref_EAF	numOutlier	N_max	Lambda PVAL.GC	SE_median	C trait transf
FXI-1	0	2370	0	1161393	0	0.9986	402	7753	0.9782	0.0245	7.9120
FXI-2	0	0	923956	0	0	0.9977	252	2992	1.0103	0.0065	1.9125
FXI-3	0	0	925256	0	0	0.9974	238	1165	1.0023	0.0112	1.9125
FXI-4	0	0	829421	0	0	0.9979	151	3525	1.1131	0.0132	2.1744
FXI-5	0	0	824896	0	0	0.9965	162	640	1.0518	0.0345	2.1799
FXI-6	0	0	1543964	0	0	0.9876	18380	734	1.0014	0.0250	2.6291

Supplemental Table 4 In column one stand the SNPs to investigate the I2 values. They are with the lowest p-value first. Then there is there cptID. All other values is there represented I2 values for that SNP. The color indicates if the value is better (Green) or worse (Red) then when all cohorts is used. Columns 4-8 is when one cohort is left out. Indicated with Not x when x is not present. Column 9-10 is when the functional/non-functional is group together.

rs-number	cptID	All cohorts	Not FXI-1	Not FXI-2	Not FXI-3	Not FXI-6	Not FXI-4	Non-Functional (FXI-1,4)	Functional (FXI-2,3,6)
rs710446	3:186459927	0.85103	0.68312	0.846085	0.831084	0.873126	0.881878	0.796892	0.788736
rs4253417	4:187199005	0.888161	0.85746	0.813931	0.893963	0.904609	0.905612	0	0.815705
s780094	2:27741237	0.552008	0.44773	0	0.664005	0.663237	0.572868	0	0
rs4253421	4:187204937	0.860664	0.84329	0.771123	0.875255	0.894768	0.843852	0.717135	0
rs76438938	3:186461524	0.797651	0.29667	0.825684	0.695803	0.845591	0.847659	0	0.342115
rs505383	11:92249613	0.220546	0.41478	0.363102	0.399218	0.352605	0	0.728373	0
rs2045869	12:21707920	0	0	0	0	0	0	0	0
rs78802760	17:66163686	0	0	0	0	0	0	0	0
rs266728	3:186523301	0.237957	0	0.332609	0.416213	0	0.394544	0.066389	0
rs4253253	4:187158433	0.382195	0.52128	0.488142	0.496356	0.244486	0.046796	0.672989	0.363967



Supplemental Figure 2. The three different transformations investigated (ln, inverse and square root) and the original plotted against the data if it was normally distributed. The R^2 value is a good value to score the different transformations.

Supplemental Table 5. Keywords from GRAIL when our SNPs associated with FXI levels was used as input.

Keywords
kallikrein
histidine
plasma
bradykinin
glycoprotein
rich
weight
initiation
matter
heparin
coagulation
white
plasminogen
chain
translation
factor
heavy
blood
eukaryotic
molecular

Supplemental Table 6. Genes of interests and similar genes from GRAIL when all associated FXI SNPs was used as input.

GENE	GRAIL p-value	SELECTED SIMILAR GENES (Rank in parantheses)
KNG1	0.000247	KLKB1(1), FXI(13), EIF4A2(136), NRBP1(702), XAB1(758), HRG(807), PPM1G(924), CAD(1241), ZNF512(1437), ADIPOQ(1510)
KLKB1	0.000349	KNG1(2), FXI(17), EIF4A2(237), ADIPOQ(288), ZNF512(480), PPM1G(727), GCKR(844), XAB1(1097), NRBP1(1211), CAD(1483)
HRG	0.002322	KNG1(35), FETUB(150), KLKB1(233), PPM1G(613), EIF4A2(627), NRBP1(646), XAB1(1134), FXI(1153)
FXI	0.002946	KLKB1(2), KNG1(12), EIF4A2(187), C2orf28(197), CYP4V2(438), PPM1G(712), CAD(733), ZNF512(847), XAB1(1064), NRBP1(1275), ZNF513(1758)
EIF2B4	0.020076	EIF4A2(13), PPM1G(253), XAB1(584), RFC4(657), NRBP1(1373), CAD(1478)

Supplemental Table 7. Luminecense values for samples with 400 ng of vector and 20 nM of miRNA. EV is the empty vector, FXI is the vector with the FXI 3' UTR region insert, 145 is the miR-145 mimic, 181 is the miR-181 mimic and SCR is the scrambled miRNA.

EV+SCR	EV+145	EV+181	FXI+SCR	FXI+145	FXI+181
65 707,60	24 825,60	51 918,10	57 763,70	28 462,30	39 587,60
289 346,00	55 056,10	458 399,00	30 208,20	22 961,30	21 963,10
242 660,00	127 475,00	189 252,00	19 135,10	30 551,20	15 225,60
321 153,00	527 904,00	755 968,00	50 927,70	29 543,80	37 381,00

Supplemental Table 8. Luminecense values for samples with 100 ng of vector and 10 nM of miRNA. EV is the empty vector, FXI is the vector with the FXI 3' UTR region insert, 145 is the miR-145 mimic, 181 is the miR-181 mimic and SCR is the scrambled miRNA.

EV+SCR	EV+145	EV+181	FXI+SCR	FXI+145	FXI+181
128 932,00	94 949,80	150 312,00	5 932,19	12 101,00	5 758,09
133 035,00	98 241,10	95 315,60	7 042,88	8 508,97	6 720,17
152 230,00	86 128,10	89 223,00	9 322,67	16 110,40	8 299,30
131 072,00	453 366,00	188 907,00	7 084,41	11 175,00	17 490,20
154 869,00	574 066,00	270 990,00	15 650,80	16 155,00	26 649,50
293 633,00	802 536,00	303 455,00	17 324,90	15 361,30	18 201,40